pedocs
Open Access Erziehungswissenschaften

DIPF

Meinhardt, Claudia

Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern

Berlin : Logos Verlag 2018, XIV, 420 S. - (Studien zum Physik- und Chemielernen; 256) - (Dissertation, Martin-Luther-Universität Halle-Wittenberg, 2018)

Quellenangabe/ Reference:
Meinhardt, Claudia: Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern. Berlin : Logos Verlag 2018, XIV, 420 S. - (Studien zum Physik- und Chemielernen; 256) - (Dissertation, Martin-Luther-Universität Halle-Wittenberg, 2018) - URN: urn:nbn:de:0111-pedocs-333848 - DOI: 10.25656/01:33384; 10.30819/4712

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-333848
https://doi.org/10.25656/01:33384

in Kooperation mit / in cooperation with:

https://www.logos-verlag.de

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: http://creativecommons.org/licenses/by-nc-nd/4.0/deed.de - Sie dürfen das Werk bzw. den Inhalt unter folgenden Bedingungen vervielfältigen, verbreiten und öffentlich zugänglich machen: Sie müssen den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen. Dieses Werk bzw. dieser Inhalt darf nicht für kommerzielle Zwecke verwendet werden und es darf nicht bearbeitet, abgewandelt oder in anderer Weise verändert werden.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use

This document is published under following Creative Commons-License: http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en - You may copy, distribute and transmit, adapt or exhibit the work in the public as long as you attribute the work in the manner specified by the author or licensor. You are not allowed to make commercial use of the work or its contents. You are not allowed to alter, transform, or change this work in any other way.

By using this particular document, you accept the above-stated conditions of use.

Kontakt / Contact:

peDOCS
DIPF | Leibniz-Institut für Bildungsforschung und Bildungsinformation
Informationszentrum (IZ) Bildung
E-Mail: pedocs@dipf.de
Internet: www.pedocs.de

Mitglied der Leibniz-Gemeinschaft

Studien zum Physik- und Chemielernen

H. Niedderer, H. Fischler, E. Sumfleth [Hrsg.]

256

Claudia Meinhardt

Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern

λογος

Studien zum Physik- und Chemielernen

Herausgegeben von Hans Niedderer, Helmut Fischler und Elke Sumfleth

Diese Reihe im Logos-Verlag bietet ein Forum zur Veröffentlichung von wissenschaftlichen Studien zum Physik- und Chemielernen. In ihr werden Ergebnisse empirischer Untersuchungen zum Physik- und Chemielernen dargestellt, z. B. über Schülervorstellungen, Lehr-/Lernprozesse in Schule und Hochschule oder Evaluationsstudien. Von Bedeutung sind auch Arbeiten über Motivation und Einstellungen sowie Interessensgebiete im Physik- und Chemieunterricht. Die Reihe fühlt sich damit der Tradition der empirisch orientierten Forschung in den Fachdidaktiken verpflichtet. Die Herausgeber hoffen, durch die Herausgabe von Studien hoher Qualität einen Beitrag zur weiteren Stabilisierung der physik- und chemiedidaktischen Forschung und zur Förderung eines an den Ergebnissen fachdidaktischer Forschung orientierten Unterrichts in den beiden Fächern zu leisten.

Hans Niedderer Helmut Fischler Elke Sumfleth

Studien zum Physik- und Chemielernen

Band 256

Claudia Meinhardt

Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern

Logos Verlag Berlin

Studien zum Physik- und Chemielernen
Hans Niedderer, Helmut Fischler, Elke Sumfleth [Hrsg.]

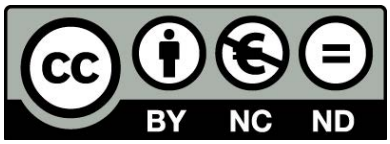
Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

© Copyright Logos Verlag Berlin GmbH 2018

Alle Rechte vorbehalten.

ISBN 978-3-8325-4712-7



Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
D-10243 Berlin
Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
http://www.logos-verlag.de

Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern

Claudia Meinhardt

[1]

gefördert durch die

DFG Deutsche Forschungsgemeinschaft

[1] Mikael Damkier/Shutterstock.com

Dissertation an der Naturwissenschaftlichen Fakultät II der Martin-Luther-Universität Halle-Wittenberg

Veröffentlichungen im Rahmen des Dissertationsprojekts

Teile dieser Dissertation sind bereits in den folgenden Werken/Veröffentlichungen erschienen:

Meinhardt, C. (2011). *Pilotierung eines Instruments zur domänenspezifischen Selbstwirksamkeitserwartung von Physik-Lehramtsstudierenden, unveröffentlichte Masterarbeit*, Universität Potsdam.

Rabe, T., Meinhardt, C. & Krey, O. (2012). Entwicklung eines Instruments zur Erhebung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 293–315.

Rabe, T., Krey, O. & Meinhardt, C. (2013a). Physikdidaktische Selbstwirksamkeitserwartungen zukünftiger Physiklehrkräfte I. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012* (S. 635–637). Münster: LIT.

Rabe, T., Krey, O. & Meinhardt, C. (2013b). Physikdidaktische Selbstwirksamkeitserwartungen (zukünftiger) Physiklehrkräfte II – eine Projektskizze. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012* (S. 638–640). Münster: LIT.

Meinhardt, C., Rabe, T. & Krey, O. (2014a). Pre-service teachers' self-efficacy beliefs for teaching physics. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research For Evidence-based Teaching and Coherence in Learning. Part 13 (co-eds. L. Avraamidou & M. Michelini)* (S. 244–250). Nicosia, Cyprus: European Science Education Research Association.

Meinhardt, C., Rabe, T. & Krey, O. (2014b). Qualitative Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen (SWE) in physikdidaktischen Handlungsfeldern. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in München 2013* (S. 558–560). Kiel: IPN.

Meinhardt, C., Rabe, T. & Krey, O. (2015). Quantitative Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern – Erste Ergebnisse. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 283–285). Kiel: IPN.

Meinhardt, C., Krey, O. & Rabe, T. (2016a). Measurement Invariance as a Validity Argument of a Physics Teaching Efficacy Questionnaire. In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto & K. Hahl (Hrsg.), *Electronic Proceedings of the ESERA 2015 Conference. Science Education Research: Engaging Learners for a Sustainable Future, Part 13 (co-eds. M. Evagorou & M. Michelini)* (S. 2135–2140). Helsinki, Finland: University of Helsinki.

Meinhardt, C., Rabe, T. & Krey, O. (2016b). Selbstwirksamkeitserwartungen (angehender) Physiklehrkräfte: Ausgewählte Ergebnisse einer Validierungsstudie. In C. Maurer (Hrsg.), *Authentizität und Lernen - das Fach in der Fachdidaktik. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Berlin 2015* (S. 337–339). Universität Regensburg.

Meinhardt, C., Rabe, T. & Krey, O. (2016c). *Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern. Skalendokumentation.* Didaktik der Physik, Martin-Luther-Universität Halle-Wittenberg. Verfügbar unter: http://www.pedocs.de/frontdoor.php?source_opus=11818, abgerufen am 26.02.2016.

Dank

Ein bisschen misstrauisch nehme ich zur Kenntnis, dass dieser aufregende aber auch anstrengende Lebensabschnitt des Promovierens nun tatsächlich ein Ende gefunden hat. Ohne zahlreiche Unterstützerinnen und Unterstützer hätte ich ihn nicht (so gut) gemeistert.

Mein größter Dank gilt Thorid und Olaf, denn aufgrund eurer Projektidee ergab sich für mich überhaupt erst die Chance zu promovieren. Ihr habt an mich geglaubt und mir großes Vertrauen entgegengebracht. Ich danke euch sehr für diese einmalige Lerngelegenheit.

Thorid – dir möchte ich besonders danken, denn du hast mich auf meinem Weg immer unterstützt. Du warst offen für Ideen und du hast dir sehr viel Zeit für Gespräche und Diskussionen genommen. Vor allem hast du es ermöglicht, dass diese Arbeit trotz des vorskizzierten Projekts zu meiner eigenen Arbeit werden konnte. Du hast mich eigene Wege einschlagen und Ideen umsetzen lassen und dabei auch so manchen Umweg in Kauf genommen. Du hast es ausgehalten, wenn alles mal wieder etwas länger gedauert hat, weil ich es perfekt machen wollte. Danke auch für die vielen schönen Momente – ob beim Konzert, im Garten oder bei einer Tasse Tee.

Olaf – du bist mein größter Kritiker und gleichzeitig mein wohlwollendster Begleiter. Du scheinst unerschütterlich davon überzeugt zu sein, dass ich alles schaffen kann. Du hast geduldig mit mir Probleme diskutiert, Fehler gesucht und mich bei Rückschlägen motiviert. Deine Haltung, alles zu hinterfragen und genau verstehen zu wollen, hat mich manchmal zwar einige Nerven gekostet (☺), aber prinzipiell dazu angestiftet, mich nicht mit weniger als möglich zufrieden zu geben. Ich danke dir von ganzem Herzen!

Ich danke allen Hiwis, die im Laufe der Zeit an meinem Projekt mitgearbeitet haben. Ich weiß, einige Aufgaben waren öde hoch drei! Danke Stefan, Jens, Christoph und Lisa – ohne euch hätte alles noch sehr viel länger gedauert… Außerdem danke ich allen Studienteilnehmerinnen und -teilnehmern, die die Fragebögen ausgefüllt haben und allen Studienseminarleiterinnen und -leitern sowie Dozentinnen und Dozenten, die meine Befragung unterstützt bzw. in ihren Veranstaltungen durchgeführt haben und ohne deren Engagement die vorliegende Arbeit nicht in dieser Form möglich gewesen wäre.

Außerdem danke ich den beiden externen Gutachtern dieser Arbeit, Prof. Dr. Claudia von Aufschnaiter und Prof. Dr. Knut Neumann, für ihre Geduld, ihre investierte Zeit in ausführliche Gutachten und ihr Engagement auch der Verteidigung beizuwohnen und diese mit kritischen Fragen zu bereichern.

Danke Familie – ihr habt mich unterstützt, auch wenn ich mal wieder länger gearbeitet habe. Care-Pakete sind super! Danke Rici, dass du immer da bist; danke fürs Zuhören, Aufmuntern und kluge-Ratschläge-Verteilen. Ein großer Dank gilt Monika und Martin, die beim Korrekturlesen keine Mühen gescheut haben.

Auf ins nächste Abenteuer!

Inhaltsverzeichnis

Veröffentlichungen im Rahmen des Dissertationsprojekts i

Dank iii

1. Vorsatz 1

I Theorie

2. Entwicklung und „Validierung“ eines Testinstruments 10
2.1. Validität . . . 11
2.1.1. Definition und grundlegende Eigenschaften . . . 11
2.1.2. Klassische Arten der Validität . . . 14
2.1.3. Konstruktvalidität als übergeordnetes Konzept . . . 18
2.1.4. Argumentbasierter Ansatz der Validierung . . . 22
2.1.5. Zum Verhältnis von Validität zu anderen Testgütekriterien . . . 25
2.1.6. Zusammenfassung . . . 26
2.2. Entwicklung eines Testinstrumentes . . . 27
2.2.1. Konstruktirrelevante Varianz . . . 28
2.2.2. Klassische Beurteilungsfehler . . . 30
2.2.3. Aspekte der Testplanung . . . 31
2.2.4. Aspekte der Testkonstruktion . . . 33
2.2.5. Itemerprobung, Review und Überarbeitung . . . 37
2.2.6. Besonderheiten der Fragebogenmethode . . . 37
2.2.7. Zusammenfassung . . . 40

3. Selbstwirksamkeitserwartungen 41
3.1. Mehrperspektivische theoretische Annäherung an das Konstrukt . . . 42
3.1.1. Die sozial-kognitive Theorie Banduras . . . 43
3.1.2. Exkurs: Philosophische Randnotizen . . . 45
3.1.3. Selbstwirksamkeitserwartungen als Kontroll-Konstrukt . . . 46

3.2. Analyse der Definition: Eigenschaften und Struktur von Selbstwirksamkeitserwartungen . . . 49
3.2.1. Die Definition des Konstruktes . . . 50
3.2.2. Weitere Eigenschaften des Konstruktes, die sich aus der Definition und der theoretischen Einbettung ergeben . . . 54
3.2.3. Weitere strukturelle Eigenschaften des Konstruktes . . . 56
3.3. Quellen von Selbstwirksamkeitserwartungen . . . 57
3.4. Abgrenzung und Beziehung zu anderen Konstrukten . . . 60
3.4.1. Persönlichkeit . . . 61
3.4.2. Einstellungen ("attitudes") . . . 62
3.4.3. Selbstkonzept . . . 63
3.4.4. Selbstwertgefühl . . . 69
3.4.5. Effektanzmotivation ("effectance motivation") . . . 70
3.4.6. Kontrollüberzeugungen ("locus of control") . . . 70
3.4.7. Handlungsergebniserwartungen („outcome expectancies") . . . 72
3.5. Die Bedeutung von Selbstwirksamkeitserwartungen . . . 73
3.6. Zur Erfassung des Konstruktes . . . 75
3.6.1. Spezifitätsniveau . . . 75
3.6.2. Anleitung zur Konstruktion und Validierung von Selbstwirksamkeitserwartungsskalen . . . 77
3.7. Exkurs: Kollektive Selbstwirksamkeitserwartungen . . . 80
3.8. Zusammenfassung . . . 81

4. Von „teacher efficacy" zu „Lehrer-Selbstwirksamkeitserwartungen" 83
4.1. Definitionsversuche . . . 85
4.1.1. Definition nach Rotter . . . 86
4.1.2. Definition nach Bandura . . . 87
4.1.3. Theoriekonfusion: Wie Gibson und Dembo (1984) versuchen, „teacher effficacy" und Banduras sozial-kognitive Theorie zusammenzubringen . . . 88
4.1.4. Das Modell von Tschannen-Moran, Woolfolk Hoy und Hoy (1998): Ausweg aus der Krise? . . . 90
4.1.5. Das Ende der begrifflichen Verwirrung: Dellinger, Bobbett, Olivier und Ellett (2008) bringen Licht ins Dunkel . . . 93
4.1.6. Vertiefte Auseinandersetzung mit dem Ansatz von Dellinger, Bobbett, Olivier und Ellett (2008): Auf dem Weg zu einem besseren Verständnis des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen . . . 95
4.1.7. Das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen im deutschen Sprachraum . . . 100

4.1.8. Zusammenfassung . . . 101
4.2. Abgrenzung und Beziehung zu verwandten Konstrukten . . . 102
4.2.1. Lehrer-Selbstkonzept/“teacher self-concept“ . . . 102
4.2.2. Kontrollüberzeugungen von Lehrpersonen/“teacher locus of control“ . . . 102
4.3. Zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen . . . 104
4.3.1. Personal Teaching Efficacy Vignette Scale . . . 106
4.3.2. Teacher Efficacy Scale – TES . . . 111
4.3.3. Instrument zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen nach Schmitz und Schwarzer (2000) . . . 114
4.3.4. Teacher Self-Efficacy Scale – TSES . . . 116
4.3.5. Ohio State Teacher Efficacy Scale – OSTES . . . 117
4.3.6. Teacher Confidence Scale – TCS . . . 120
4.3.7. Norwegian Teachers Self-Efficacy Scale – NTSES . . . 120
4.3.8. Mehrdimensionale Skala der Lehrer-Selbstwirksamkeitserwartung – MSLS . . . 121
4.3.9. Teachers’ Efficacy Beliefs System-Self – TEBS-Self . . . 122
4.3.10. Self-Efficacy Measure for Student Teachers in Competence-Based Education . . . 123
4.3.11. Zusammenfassung . . . 124
4.3.12. Ableitungen für die Instrumentenentwicklungen . . . 127
4.4. Empirische Basis zur Relevanz des Konstruktes . . . 128
4.4.1. Lehrervariablen . . . 132
4.4.2. Schülervariablen . . . 135
4.4.3. Zusammenfassung . . . 136
4.5. Quellen von Lehrer-Selbstwirksamkeitserwartungen . . . 137
4.5.1. Einflussfaktoren auf Lehrer-Selbstwirksamkeitserwartungen . . . 138
4.6. Zur Entwicklung von Lehrer-Selbstwirksamkeitserwartungen . . . 140
4.6.1. Ergebnisse querschnittlich angelegter Studien . . . 140
4.6.2. Ergebnisse längsschnittlich angelegter Studien . . . 141
4.6.3. Zusammenfassung und Kommentar zur Problematik der Interpretation der Studienergebnisse aufgrund der Erfassung des Konstruktes mittels defizitärer Instrumente . . . 143
4.7. Zur Ausprägung von Lehrer-Selbstwirksamkeitserwartungen in unterschiedlichen Gruppen . . . 144
4.8. Zusammenfassung . . . 145

5. Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern (“science teaching self-efficacy beliefs“) 149
5.1. Zur Definition des Konstruktes . . . 150

5.2. Zur Erfassung des Konstruktes 151
5.2.1. Science Teaching Efficacy Beliefs Instrument – STEBI 154
5.2.2. Self-Efficacy Teaching and Knowledge Instrument for Science Teachers – SETAKIST 156
5.2.3. Self-Efficacy Beliefs about Equitable Science Teaching and Learning – SEBEST 156
5.2.4. Selbstwirksamkeitserwartung bezogen auf das Physik-Lehren . 157
5.2.5. Zusammenfassung 158
5.3. Abschließende Bemerkungen 158

6. Von der „guten“ Lehrperson – Selbstwirksamkeitserwartungen im Forschungsfeld der Lehrerprofessionalisierung 160
6.1. Kompetenzorientierung in der Lehrerbildung 162
6.1.1. Kompetenz – Annäherung an eine Definition 162
6.1.2. Modellierung der Kompetenzen von (angehenden) Lehrpersonen 164
6.1.3. Kompetenzentwicklung von Lehrpersonen 169
6.1.4. Messung von Lehrerkompetenzen 170
6.2. Kompetenzmodellierungen von (angehenden) Physiklehrkräften 172
6.2.1. Projekte mit dem Schwerpunkt der Kompetenzmodellierung von (angehenden) Physiklehrkräften 173
6.2.2. Physikdidaktisches Wissen 176
6.3. Zusammenfassung 179

II Instrumentenentwicklung, Forschungsfragen und Studiendesign

7. Theoriebasierte Instrumentenentwicklung – Erste Schritte 183
7.1. Theoretische Ableitungen 184
7.1.1. Arbeitsdefinition 184
7.1.2. Konstruktionsregeln bez. des Itemstammes 185
7.1.3. Itemkonstruktion unter Berücksichtigung des Bezugs zu anderen Konstrukten 188
7.1.4. Wahl des Antwortformates 188
7.1.5. Reflexion des formalen Status zugrunde liegender Theorien . . 188
7.2. Struktur des Konstruktes 189
7.2.1. Festlegung des Spezifitätsniveaus 189
7.2.2. Inhalte/Handlungsfelder 190
7.2.3. Dimensionen 192
7.2.4. Hierarchie 192
7.2.5. Zusammenfassung 193

7.3. Vorgehen bei der Itemkonstruktion . . . 194
7.4. Testplanung . . . 194
7.5. Festlegung der zu validierenden Testwertinterpretation . . . 196
7.5.1. Intendierte Testwertinterpretation . . . 196
7.5.2. Zugrunde liegende Annahmen der intendierten Testwertinterpretation . . . 197
7.5.3. Gewählte Verfahren zur Prüfung der Gültigkeit getroffener Annahmen . . . 199
7.5.4. Konkurrierende Testwertinterpretationen . . . 201
7.5.5. Abschätzung potentieller Konsequenzen . . . 201
7.6. Zusammenfassung . . . 202

8. Forschungsfragen, Hypothesen, Studiendesign 204
8.1. Forschungsfragen und Hypothesen . . . 204
8.1.1. Übergeordnete Forschungsfrage . . . 204
8.1.2. Teilforschungsfragen und zugehörige Hypothesen . . . 204
8.2. Projektdesign . . . 209
8.2.1. Übersicht des Projektablaufs . . . 209
8.2.2. Vorstellung einzelner Projektteile . . . 210

III Empirie

9. Exkurs: Pilotstudie I 215
9.1. Stichprobe . . . 215
9.2. Fragebogen . . . 216
9.3. Methodische Anmerkungen . . . 216
9.4. Analyseverfahren und benutzte Analysesoftware . . . 216
9.5. Ergebnisse . . . 216
9.5.1. Reliabilitätsanalyse . . . 216
9.5.2. Konfirmatorische Faktorenanalyse . . . 217
9.5.3. Korrelationsanalyse . . . 217
9.5.4. Varianzanalyse . . . 219
9.6. Interpretation der Ergebnisse und Validitätsargumente . . . 219
9.6.1. Reliabilitätsanalyse . . . 219
9.6.2. Konfirmatorische Faktorenanalyse . . . 220
9.6.3. Korrelationsanalyse . . . 220
9.6.4. Varianzanalyse . . . 221
9.7. Itemrevision . . . 221
9.8. Zusammenfassung . . . 221

10. Pilotstudien II und III (qualitativer Natur) 223
10.1. Interviewstudie . . . 223
10.1.1. Verfahren der Gewinnung von Interviewpartnern . . . 223
10.1.2. Stichprobe . . . 224
10.1.3. Methodische Anmerkungen . . . 224
10.1.4. Datenerhebung – Interviewdurchführung . . . 226
10.1.5. Datenaufbereitung und Auswerteverfahren . . . 227
10.2. Expertenbefragung . . . 228
10.2.1. Verfahren der Gewinnung von Experten . . . 228
10.2.2. Methodische Anmerkungen – Aufbau des Fragebogens . . . 228
10.2.3. Stichprobe . . . 229
10.2.4. Datenaufbereitung und Auswerteverfahren . . . 229
10.3. Exkurs: Akzeptanzbefragung bez. des Antwortformates . . . 230
10.4. Ergebnisse der Pilotstudien II und III . . . 230
10.4.1. Umsetzung der Arbeitsdefinition . . . 231
10.4.2. Itemverständnis . . . 231
10.4.3. Aktivierung kognitiver Prozesse bei der Itembeantwortung . . . 232
10.4.4. Authentizität der Items . . . 232
10.4.5. Exkurs: Itemschwierigkeiten . . . 233
10.4.6. Eignung des Antwortformats . . . 234
10.4.7. Angemessenheit der Abbildung der Handlungsfelder . . . 237
10.4.8. Trennschärfe der Konstrukte . . . 237
10.4.9. Exkurs: Relevanz der Handlungsfelder . . . 237
10.5. Itemrevision . . . 239
10.5.1. Vorgehen und Ableitungen bez. der Itemüberarbeitung . . . 239
10.5.2. Exemplarische Illustration der Itemüberarbeitung . . . 241
10.6. Interpretation der Ergebnisse und Validitätsargumente . . . 242
10.6.1. Aktivierung kognitiver Prozesse und Umsetzung der Arbeitsdefinition . . . 242
10.6.2. Itemverständnis . . . 243
10.6.3. Relevanz und Authentizität . . . 243
10.6.4. Eignung des Antwortformates . . . 244
10.6.5. Trennschärfe der Konstrukte . . . 244
10.7. Zusammenfassung . . . 244

11. Weiterentwicklung des Instruments 245
11.1. Überarbeitung der Itemkonstruktionsregeln . . . 245
11.2. Überarbeitung des Antwortformates . . . 246
11.3. Zusammensetzung des Fragebogens . . . 246
11.3.1. Demografie . . . 246

11.3.2. Items . . . 246
11.3.3. Zusätzliche Konstrukte . . . 248
11.4. Instruktionen . . . 250
11.5. Anordnung der Skalen . . . 251
11.6. Überlegungen bez. der Tendenz zu sozialer Erwünschtheit . . . 252
11.7. Papier-und-Bleistift versus Online-Befragung . . . 252

12. Hauptstudie **253**
12.1. Datenerhebung . . . 254
12.1.1. Rekrutierung von Teilnehmerinnen und Teilnehmern . . . 254
12.1.2. Durchführung der Datenerhebung . . . 255
12.2. Datenaufbereitung . . . 255
12.2.1. Datenselektion . . . 255
12.3. Datenqualität . . . 256
12.3.1. Fehlende Werte . . . 257
12.3.2. Verteilung der Item- und Skalenwerte . . . 257
12.4. Stichprobe und deskriptive Analysen . . . 258
12.4.1. Detaillierte Beschreibung der befragten Kohorten . . . 259
12.4.2. Beschreibung relevanter Subgruppen . . . 260
12.5. Analyseverfahren und benutzte Analysesoftware . . . 261
12.6. Methodische Anmerkungen . . . 261
12.6.1. Anmerkungen zur klassischen Item- & Skalenanalyse . . . 262
12.6.2. Konfirmatorische Faktorenanalyse . . . 262
12.6.3. Raschanalyse . . . 264
12.6.4. Korrelationsanalysen . . . 267
12.6.5. Mittelwertvergleiche . . . 267
12.7. Ergebnisse . . . 268
12.7.1. Eindimensionalität der Skalen . . . 268
12.7.2. Exkurs: Raschanalyse . . . 270
12.7.3. Zuverlässigkeit der Skalen . . . 270
12.7.4. Eignung des Antwortformates . . . 273
12.7.5. Messinvarianz . . . 274
12.7.6. Dimensionsanalyse . . . 277
12.7.7. Abgrenzbarkeit zu anderen Konstrukten . . . 279
12.7.8. Abbildung theoretischer Erwartungen . . . 281
12.8. Interpretation der Ergebnisse und Validitätsargumente . . . 285
12.8.1. Eindimensionalität der Skalen . . . 285
12.8.2. Zuverlässigkeit der Skalen . . . 285
12.8.3. Eignung des Antwortformates . . . 287
12.8.4. Messinvarianz . . . 287

12.8.5. Dimensionsanalyse 288
12.8.6. Abgrenzbarkeit zu anderen Konstrukten 289
12.8.7. Abbildung theoretischer Erwartungen 289
12.9. Externe Validierung 290
12.10. Zusammenfassung 290

IV Schluss

13. Zusammenfassung, Diskussion und Ausblick 293
13.1. Zusammenfassung der Ergebnisse 293
13.1.1. Ergebnisse auf theoretischer Ebene 293
13.1.2. Ergebnisse auf empirischer Ebene in Bezug auf die Forschungsfragen 296
13.1.3. Das wichtigste Ergebnis: Das Instrument 302
13.2. Diskussion 303
13.2.1. Zur Reflexion der Itementwicklung und -überarbeitung 303
13.2.2. Zum übergeordneten Validitätsargument 303
13.2.3. Zur Subjektivität und einem mittleren Spezifitätsniveau 305
13.3. Ausblick 307

V Anhang

A. Auswahl an Instrumenten zur Erfassung des Selbstkonzeptes 311
A.1. Instrumente zur Erfassung des Selbstkonzeptes von Schülerinnen und Schülern. 311
A.2. Instrumente zur Erfassung des Selbstkonzeptes von (angehenden) Lehrerinnen und Lehrern. 314

B. Itemsammlung selten verwendeter Instrumente zur Erfassung von „Lehrer-Selbstwirksamkeitserwartungen“ 315

C. Exemplarische Darstellung der Bedeutung des Konstruktes ohne Berücksichtigung der Erhebungsinstrumente 317

D. Itemsammlung selten verwendeter Instrumente zur Erfassung von „science teaching self-efficacy beliefs“ 319

E. Vergleich der „Teacher Efficacy Scale“ und des „Science Teaching Efficacy Beliefs Instrument“ 320

F. Physikdidaktische Themenfelder **321**

G. Skalenversionen **324**
G.1. Skalenversion 1 (V1): Pilotstudie I 324
G.1.1. Skala zum Experimentieren/Planung 324
G.1.2. Skala zum Experimentieren/Durchführung 324
G.1.3. Skala zum Elementarisieren/Planung 325
G.1.4. Skala zum Elementarisieren/Durchführung 325
G.1.5. Skala zum Umgang mit Schülervorstellungen/Planung 325
G.1.6. Skala zum Umgang mit Schülervorstellungen/Durchführung . . 326
G.2. Skalenversion 2 (V2): Pilotstudie II und III 327
G.2.1. Skala zum Experimentieren/Planung 327
G.2.2. Skala zum Experimentieren/Durchführung 327
G.2.3. Skala zum Elementarisieren/Planung 327
G.2.4. Skala zum Elementarisieren/Durchführung 328
G.2.5. Skala zum Umgang mit Schülervorstellungen/Planung 328
G.2.6. Skala zum Umgang mit Schülervorstellungen/Durchführung . . 329
G.2.7. Skala zum Umgang mit Aufgaben/Planung 329
G.2.8. Skala zum Umgang mit Aufgaben/Durchführung 330
G.3. Skalenversion 3 (V3): Hauptstudie 331
G.3.1. Skala zum Experimentieren/Planung 331
G.3.2. Skala zum Experimentieren/Durchführung 332
G.3.3. Skala zum Elementarisieren/Planung 333
G.3.4. Skala zum Elementarisieren/Durchführung 334
G.3.5. Skala zum Umgang mit Schülervorstellungen/Planung 335
G.3.6. Skala zum Umgang mit Schülervorstellungen/Durchführung . . 336
G.3.7. Skala zum Umgang mit Aufgaben/Planung 337
G.3.8. Skala zum Umgang mit Aufgaben/Durchführung 338
G.4. Skalenversion 4 (V4): Zur Verwendung empfohlen 339
G.4.1. Skala zum Experimentieren/Planung 339
G.4.2. Skala zum Experimentieren/Durchführung 340
G.4.3. Skala zum Elementarisieren/Planung 341
G.4.4. Skala zum Elementarisieren/Durchführung 342
G.4.5. Skala zum Umgang mit Schülervorstellungen/Planung 343
G.4.6. Skala zum Umgang mit Schülervorstellungen/Durchführung . . 344
G.4.7. Skala zum Umgang mit Aufgaben/Planung 345
G.4.8. Skala zum Umgang mit Aufgaben/Durchführung 346

H. Zum Interview **347**
H.1. Skalen – Interviewversion 347

H.2. Interviewleitfaden 348
H.2.1. Phase I – Anmerkungen zur Einführung 348
H.2.2. Phase II – Mögliche Impulse bez. des Vierschritts 348
H.2.3. Phase III – Anmerkungen zur Gesamtskala 349
H.3. Beispiel Interviewnotizen 351
H.4. Beispielübersicht Itemauswertung (Interviews) 352

I. Zur Expertenbefragung 355
I.1. Expertenfragebogen (Physikdidaktik) 355
I.2. Expertenfragebogen (Psychologie) 360
I.3. Beispielübersicht Itemauswertung (Expertenbefragung) 365
I.4. Beispielübersicht Skalenauswertung (Expertenbefragung) 366

J. Fragebogen zur Akzeptanzbefragung bez. des Antwortformates 367

K. Beispiel einer zusammenfassenden Übersicht zur Itemüberarbeitung nach Pilotstudien II & III 369

L. Fragebogen – Hauptstudie 370

M. Anreizsysteme und Aufrufe 376
M.1. Flyer 376
M.2. Foreneinträge 377

N. Dokumentation Datenselektion 379
N.1. Dokumentation für die Stichprobe der Lehrkräfte 379
N.2. Dokumentation für die Stichprobe der Referendare 380
N.3. Dokumentation für die Stichprobe der Studierenden 381

O. Messinvarianz – zusätzliche Skalen 383
O.1. Mehrgruppenvergleich: Personengruppe 383
O.2. Mehrgruppenvergleich: Geschlecht 386
O.3. Mehrgruppenvergleich: Studiengang und Schulform 387
O.4. Mehrgruppenvergleich: Praxiserfahrung 388

P. Mittelwertvergleiche 389
P.1. Zusätzliche Skalen 389
P.2. Neukonstruierte Skalen 391

Literatur 392

Skalendokumentation 420

1. Vorsatz

„We echo the need for new or revised measure(s) that would reliably assess science teaching efficacy and its components.“

Cakiroglu, Capa-Aydin & Woolfolk Hoy, 2012, S. 458

Der PISA-Schock im Jahre 2000 löste getragen vom politischen Willen nicht nur im schulischen Bereich viele Neuerungen aus, sondern auch eine Reihe von Forschungsanstrengungen, die sich auf die Lehrerprofessionalisierung konzentrierten. In den Fokus der empirischen Bildungsforschung rückten insbesondere die Kompetenzen der Lehrkräfte (vgl. z. B. Baumert & Kunter, 2006). Die Wahrnehmung, dass der Kompetenzbegriff dabei häufig einseitig auf Fähigkeiten oder Wissen(skomponenten) reduziert und somit bezüglich „professioneller“ Lehrkräfte eine ganz spezifische, ggf. eingeschränkte Perspektive eingenommen wurde, kann als Ausgangspunkt für die hier vorliegende Arbeit angesehen werden.

Selbstwirksamkeitserwartungen sind eines der Konstrukte, die im „Kompetenzforschungsgerangel“ zunächst weniger Beachtung erfuhren. Dabei spielen sie nicht nur im Kompetenzkonstrukt selbst (je nach Auslegung als determinierende oder konstituierende Facette) eine wichtige Rolle (vgl. Zlatkin-Troitschanskaia & Seidel, 2011). Auch wird ihnen unabhängig davon im Rahmen der sozial-kognitiven Theorie Banduras (1977b, 1986, 1997) eine hohe Handlungsrelevanz zugesprochen, was mit Blick auf Lehrpersonen als handelnde, Unterricht gestaltende Akteure von Bedeutung ist. Im Allgemeinen rekurrieren (Lehrer-)Selbstwirksamkeitserwartungen auf die persönliche Gewissheit, bestimmte (unterrichtsbezogene) Tätigkeiten ausführen zu können, auch wenn sich dabei Widerstände in den Weg stellen. Man geht davon aus, dass Personen, die überzeugt sind, solch schwierige Aktivitäten bewältigen zu können, nicht nur eher eine Handlungsinitiative ergreifen, sondern auch ein höheres Maß an Anstrengungsbereitschaft sowie Zeit und Beharrlichkeit in die Bewältigung der Aufgabe investieren (vgl. Bandura, 1997). Es ist einsichtig, dass die genannten Konsequenzen für die Planung und Durchführung von Unterricht in höchstem Maße relevant sind.

Insgesamt konnte zu Beginn der Promotionszeit ein Desiderat dahingehend formuliert werden, dass kein deutschsprachiges Instrument existierte, welches Selbstwirksamkeitserwartungen von Lehrpersonen auf einem unterrichtsfachspezifischen Niveau erfasste und somit auch für fachdidaktische Forschungsfragen zur Lehrerprofessionalisierung anschlussfähig gewesen wäre. Als Indikator für das tatsächliche Fehlen eines solchen

Instruments erwies sich das Interesse mehrerer Gruppen der Forschungsgemeinschaft in Form regelmäßiger Anfragen auch und sogar in den sehr frühen Phasen der Instrumentenentwicklung zum Einsatz der Skalen in anderen Projekten und Studien.[1]

Wie bereits angedeutet, sind Lehrer-Selbstwirksamkeitserwartungen auch unabhängig vom Blickwinkel der Lehrerprofessionalisierung seit Längerem Gegenstand von Forschungsbemühungen, sodass das formulierte Desiderat ggf. zweifelhaft erscheint. Im Verlauf des Dissertationsprojekts konnten durch eine intensive Recherche sowie eine kriteriengeleitete Evaluation existierender Erhebungsinstrumente jedoch nicht nur das Bestehen des Desiderats als solches bestätigt, sondern auch Gründe für eben dieses herausgearbeitet werden. Diese liegen insbesondere in der Schwierigkeit der theoriekonformen Anpassung des Spezifitätsniveaus für domänenspezifische Operationalisierungen des Konstruktes der Selbstwirksamkeitserwartungen, wobei dies konkret für die Domäne des Lehrerberufs (d. h. für Operationalisierungen von Lehrer-Selbstwirksamkeitserwartungen) aufgezeigt werden konnte (vgl. Abschnitte 4.3 und 5.2). In Operationalisierungen wird häufig ein Konstrukt namens „teacher efficacy“ unter dem Deckmantel der sozial-kognitiven Theorie Banduras (1997) erhoben, wobei in der Regel unklar bleibt, wie dieses Konstrukt tatsächlich definiert ist und welche Bezüge zur genannten Theorie bestehen. In vielen Fällen ergaben sich hinsichtlich des „teacher efficacy“-Konstrukts Konfundierungen mit anderen Konstrukten wie „outcome expectancies“ (Handlungsergebniserwartungen) oder „locus of control“ (Kontrollüberzeugungen). Als Konsequenz werden in dieser Arbeit die Konstrukte „teacher efficacy“ und „teacher self-efficacy beliefs“ unterschieden, wobei letztere Lehrer-Selbstwirksamkeitserwartungen repräsentieren und auf die theoretischen Arbeiten Banduras, das heißt auf die sozial-kognitive Theorie rekurrieren.[2]

Eine Folge des Fehlens geeigneter Instrumente (bzw. der Schwierigkeit der Interpretation der Ergebnisse von Studien, die mit Hilfe bereits existierender, ggf. unzureichender Instrumente gewonnen wurden) besteht darin, dass für eine Vielzahl von postulierten Zusammenhängen des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen im Bereich des Lehrens und Lernens zu anderen Konstrukten wie beispielsweise zu Schülerleistungen oder Schülermotivation die empirische Basis als äußerst dürftig anzusehen ist und damit die Bedeutsamkeit des Konstrukts im Wesentlichen auf Plausibilitätsargumenten und theoretischen Ableitungen beruht (vgl. Abschnitt 4.4). Insofern kann davon ausgegangen werden, dass ein neues, theoriekonformes Erhebungsinstrument für zukünftige Forschung zum Schließen dieser Lücken von Relevanz ist.

Das nunmehr detailliert beschriebene Desiderat und seine Folgen aufgreifend verfolgt

[1] Mit Verweis auf den ggf. unzureichenden Entwicklungsstand wurden die Items in der Regel zur Verfügung gestellt (vgl. auch Auflistung in Abschnitt 12.9).

[2] Beachtet werden sollte, dass die Verwendung der Begrifflichkeit „teacher self-efficacy (beliefs)“ durch andere Forschende nicht automatisch bedeutet, dass tatsächlich Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras erfasst werden (vgl. Abschnitt 4.3).

die vorliegende Arbeit das Ziel, ein Instrument zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen hinsichtlich der Planung und Durchführung von Physikuntersicht zu entwickeln und einem Validierungsprozess zu unterziehen. Sie lässt sich damit in das Paradigma der (quantitativen) empirischen Bildungsforschung einordnen. Diese Ausrichtung kann – oder muss – auch kritisch hinterfragt werden, denn dass man den Selbstwirksamkeitserwartungen von (angehenden) Physiklehrpersonen mittels Fragebogenerhebung „auf die Schliche“ kommt, kann ggf. ganz grundsätzlich in Frage gestellt werden. In dieser Arbeit wurde der Versuch unternommen, im Rahmen des quantitativen Ansatzes ein hohes Maß an Transparenz in Bezug auf die Itemüberarbeitung und Methodik sowie die Ergebnisdokumentation zu gewähren. Inwiefern der hier präsentierte Ansatz zur Erfassung von Selbstwirksamkeitserwartungen (angehender) Physiklehrpersonen in physikdidaktischen Handlungsfeldern geeignet und/oder gelungen sowie das formulierte Validitätsargument überzeugend sind, müssen die hoffentlich kritischen Leser dieser Arbeit beurteilen. In jedem Falle handelt es sich bei allen Anstrengungen hinsichtlich der Validierung um einen unabgeschlossenen Prozess, der in der Zukunft zahlreiche Ergänzungen erfahren sollte.

Neben der theoriebasierten Entwicklung eines Instrumentes zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern wird der Validierung des Instruments in dieser Arbeit große Aufmerksamkeit geschenkt. Validierung bedeutet letztlich, dass das Instrument einer umfangreichen Qualitätsprüfung unterzogen wird. Die Qualität des Instruments bemisst sich derweil daran, ob die mit diesem Instrument erhobenen Daten mit einer gewissen Sicherheit im beabsichtigten Sinne interpretiert werden können, also tatsächlich etwas über die Ausprägung der Lehrer-Selbstwirksamkeitserwartungen der befragten Kohorte bezüglich der spezifizierten Domäne aussagen. Entgegen der landläufigen sprachlichen Verwendung ist im Falle eines positiven Bescheids der Qualitätsprüfung folglich nicht das Instrument an sich als valide zu bezeichnen – vielmehr sind es die Aussagen, die aus den erhobenen Daten abgeleitet werden (vgl. AERA, APA, NCME, 2014, S. 11 ff.). Insgesamt wird hier dem Procedere des argumentbasierten Ansatzes der Validierung gefolgt (vgl. u. a. Kane, 1992; Messick, 1995; AERA, APA, NCME, 2014), in dessen Rahmen aufgrund einer Vielzahl von Indizien, die in einzelnen Studien gesammelt werden, ein vorläufiges (!) Validitätsargument formuliert wird, das im besten Falle die festgelegte Testwertinterpretation stützt. Um dieses Validitätsargument überprüfbar zu machen, wird in dieser Arbeit detailliert dargestellt, welche Annahmen der zu validierenden Testwertinterpretation zugrunde liegen und welche Maßnahmen ergriffen werden, um die Gültigkeit dieser Annahmen zu prüfen. Dieses (ggf. übertrieben kleinschrittig erscheinende) Vorgehen wird gewählt, um die theoretischen Vorarbeiten Kanes und Messicks tatsächlich umzusetzen.

Der Validierungsprozess ist methodisch vielfältig angelegt, sodass einzelne Annahmen mehrperspektivisch geprüft werden können. Auch wenn in einem Großteil der Studien

ein „quantitativer Blickwinkel“ eingenommen wird, so können gerade die Ergebnisse der qualitativ angelegten Validierungsstudien (Expertenbefragung, Interviews) für die Itementwicklung und -überarbeitung genutzt werden. Für die quantitativ ausgerichtete Hauptstudie werden systematisch sowohl Methoden der klassischen wie der probabilistischen Testtheorie angewandt, um die Forschungsfragen zu beantworten. Diese „methodische Kreuzvalidierung“ hat im Wesentlichen zum Ergebnis, dass bezüglich beider Paradigmen ähnliche Antworten auf betreffende Forschungsfragen formuliert, jedoch je nach methodischem Fokus einzelne Forschungsfragen unterschiedlich detailliert beantwortet werden können.

Im Rückblick könnte die Itementwicklung und Überarbeitung fast als exzessiv bezeichnet werden, wenn man den zeitlichen wie methodischen Aufwand betrachtet. Insgesamt werden die Items drei Überarbeitungszyklen unterzogen. Zwischenschritte in der Itemüberarbeitung werden dabei so weit es geht transparent gemacht. Alle Ergebnisse, insbesondere der Hauptstudie sind in einer umfangreichen Skalendokumentation publiziert worden (vgl. Meinhardt, Rabe & Krey, 2016c), die integraler Bestandteil dieser Arbeit ist und den quantitativen Ergebnisteil erheblich ergänzt. Unweigerlich stellt sich die Aufwand-Nutzen-Frage. Die grundsätzliche Herangehensweise, d. h. die Durchführung aller getätigten Maßnahmen (und zwar in derselben Tiefgründigkeit) erscheint jedoch nicht nur plausibel, sondern auch notwendig zu sein. Insbesondere nach den Erkenntnissen, die sich aus der Analyse bereits existierender Instrumente ergeben, wird versucht, das Konzept „Validierung“ ernst zu nehmen.

Aufbau und Struktur der Arbeit

Die Arbeit unterteilt sich grob in die fünf Blöcke Theorie (I), Instrumententwicklung und Forschungsfragen (II), Empirie (III), Schluss (IV) sowie Anhang (V). Darüber hinaus wird der Arbeit zu Begutachtungszwecken die Dokumentation der Skalenentwicklung beigefügt, da diese einen integralen Bestandteil der Arbeit bildet (vgl. Abbildung 1.1).

Teil I: Theorie Im ersten Teil dieser Arbeit werden zunächst die Grundlagen hinsichtlich der Entwicklung eines Testinstrumentes sowie der Validierung eines solchen dargelegt (vgl. Kapitel 2). Schwerpunktmäßig wird die Validierung betrachtet, wobei ältere Ansätze zur Validierung (Unterteilung in Inhalts, Kriteriums- und Konstruktvalidität) neueren, argumentbasierten Konzepten gegenübergestellt und deren Vorzüge beschrieben werden. In drei weiteren Kapiteln schließt sich eine genaue Untersuchung des im Mittelpunkt dieser Arbeit stehenden Konstruktes der Selbstwirksamkeitserwartungen an, wobei mit jedem Kapitel ein spezifischeres Niveau des Konstruktes genauer betrachtet wird (Kapitel 3: *allgemeine* Selbstwirksamkeitserwartungen, Kapitel 4: *Lehrer*-Selbstwirksamkeitserwartungen, Kapitel 5: *Lehrer*-Selbstwirksamkeitserwartungen in *physikdidaktischen Handlungsfeldern*). In Kapitel 3 wird zunächst auf Grundzü-

Abbildung 1.1.: Überblick über den Aufbau der Arbeit.

ge der sozial-kognitiven Theorie nach Bandura (1997) fokussiert. Außerdem werden Selbstwirksamkeitserwartungen als Kontrollkonstrukt charakterisiert und eine theoriebasierte Definition des Konstrukts vorgenommen. Zudem werden dessen Merkmale und strukturelle Eigenschaften herausgearbeitet. Insgesamt wird durch dieses Vorgehen ein Kriterienkatalog erarbeitet, der im Anschluss dazu dient, Selbstwirksamkeitserwartungen von anderen Konstrukten abzugrenzen. Um die theoretische Basis zu vervollständigen, werden unter anderem Quellen und Entwicklungsmechanismen für Selbstwirksamkeitserwartungen und die Bedeutsamkeit des Konstruktes anhand theoretischer Erwägungen plausibel gemacht. Aus den Darstellungen ergeben sich Ableitungen hinsichtlich der Konstruktion und Validierung eines neuen Instrumentes beispielsweise bezüglich eines geeigneten Spezifitätsniveaus.

In einem nächsten Schritt werden in Kapitel 4 Lehrer-Selbstwirksamkeitserwartungen – also Selbstwirksamkeitserwartungen auf einem spezifischeren Niveau – analysiert. Dabei steht das „teacher efficacy“-Konstrukt im Fokus, da dieses häufig mit der Theorie Banduras in Zusammenhang gebracht wird. Es wird herausgearbeitet, dass in der historischen Entwicklung des Konstruktes die theoretische Fundierung zumeist unzureichend blieb und mit „teacher-efficacy“ in der Regel *keine* Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras („teacher self-efficacy beliefs“) abgebildet werden, auch wenn dies behauptet wird. Ein Modell von Dellinger, Bobbett, Olivier und Ellett (2008), welches einige der begrifflichen bzw. definitorischen Unzulänglichkeiten benennt, wird zuerst erläutert, um dann präzisiert und beispielsweise um die Entwicklungsmechanismen von Lehrer-Selbstwirksamkeitserwartungen erweitert zu werden (vgl. Unterabschnitt 4.1.6). Damit wird der Versuch unternommen, einen Beitrag zur Diskussion der Theorie des Konstruktes zu leisten. Ein weiterer Schwerpunkt des vierten Kapitels bildet die Recherche von Erhebungsinstrumenten, die beanspruchen, „teacher efficacy“ oder „teacher self-efficacy beliefs“ zu erheben (vgl. Abschnitt 4.3). Anhand des herausgearbeiteten Kriterienkatalogs wird beurteilt, ob und inwiefern mittels der Operationalisierungen der Konstrukte Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras erhoben werden, um daraus für die eigene Instrumententwicklung zu lernen. Die umfangreiche Literaturrecherche wird im Anschluss dazu verwendet, die empirische Basis hinsichtlich einer Reihe von Vermutungen zum Beispiel bezüglich der Relevanz des Konstruktes im Schulkontext zu eruieren.

In Kapitel 5 werden dann Instrumente recherchiert, mit deren Hilfe Lehrer-Selbstwirksamkeitserwartungen auf einem noch spezifischeren Niveau erfasst werden sollen, nämlich in physik- bzw. naturwissenschaftlichen Handlungsfeldern. Es können ähnliche Probleme dokumentiert werden, wie sie bereits für Skalen zu „teacher (self) efficacy“ dargelegt wurden, da die Instrumentarien häufig auf diesen Skalen beruhen und lediglich leichte Änderungen vorgenommen wurden, um sie auf die Domäne anzupassen.

Den Abschluss des Theorieteils bilden Überlegungen zur Einbettung dieser Arbeit

in einen größeren Forschungszusammenhang im Rahmen der Kompetenzorientierung in der Lehrerbildung (vgl. Kapitel 6). Dazu wird unter anderem der Kompetenzbegriff beleuchtet, zum Beispiel indem seine Parallelen bzw. Schnittmengen mit dem Konstrukt der Selbstwirksamkeitserwartungen herausgestellt werden. Zusätzlich werden einzelne Modellierungsansätze von Lehrerkompetenz und schwerpunktmäßig Konzeptualisierungsansätze der Kompetenzfacette „Physikdidaktisches Wissen“ beschrieben. In der Folge können beispielsweise in Bezug auf die Relevanz physikdidaktischer Handlungsfelder Erkenntnisse hinsichtlich der eigenen Instrumententwicklung abgeleitet werden. Ebenfalls andiskutiert werden potentielle sich anschließende Forschungsfragen.

Teil II: Instrumententwicklung und Forschungsfragen Das siebente Kapitel spielt für die Verzahnung von Theorie und Praxis eine wichtige Rolle, da anhand der herausgearbeiteten Grundlagen zur Instrumententwicklung und -validierung sowie zur Operationalisierung von Selbstwirksamkeitserwartungen auf spezifischen Niveaus konkrete Entscheidungen getroffen und begründet werden. Zunächst wird die Arbeitsdefinition vorgestellt und erläutert, inwiefern alle notwendigen Merkmale und strukturellen Eigenschaften des Konstrukts auf die gewählte Domäne übertragen werden. Zusätzlich werden theoriebasierte Itemkonstruktionsregeln expliziert sowie inhaltliche Entscheidungen wie die Wahl der Handlungsfelder gerechtfertigt. Außerdem wird in Anlehnung an die Standards (AERA, APA, NCME, 2014) und das umrissene Validierungskonzept die intendierte und damit zu validierende Testwertinterpretation festgelegt. Aus dieser werden zum einen eine Vielzahl zugrunde liegender Annahmen abgeleitet und zum anderen Maßnahmen, die zur Prüfung der Gültigkeit dieser Annahmen geeignet erscheinen. Aus den dargelegten Annahmen und Maßnahmen lassen sich sodann die Forschungsfragen ableiten, die in dieser Arbeit beantwortet werden sollen. Die Darstellung von Forschungsergebnissen in Kapitel 4 ermöglicht es darüber hinaus, theoriebasierte Hypothesen in Bezug auf die Forschungsfragen aufzustellen. Außerdem ergeben sich aus den dargestellten Maßnahmen Gestaltungsaspekte in Hinblick auf das Design der zu planenden Validierungsstudien, die abschließend in Kapitel 8 vorgestellt werden.

Teil III: Empirie Im dritten Teil der Arbeit werden die Umsetzung der zuvor beschriebenen Validierungsstudien beschrieben sowie deren Ergebnisse berichtet, die dann für die Beantwortung der Forschungsfragen bzw. hinsichtlich des zu formulierenden Validitätsarguments interpretiert werden. Beachtet werden muss, dass das Dissertationsprojekt durch zusätzliche Studien flankiert wird, die sich unter anderem aus der zwischenzeitlichen Finanzierung durch die DFG ergaben. Immer dann, wenn Studienergebnisse herangezogen werden, die nicht Teil dieser Arbeit sind, wird explizit oder durch den Hinweis auf einen „Exkurs“ darauf hingewiesen. Insbesondere bei Pilotstudie I (vgl. Kapitel 9), die eine separate Qualifizierungsarbeit darstellt (Masterabschluss), und

bei der Akzeptanzbefragung (vgl. Exkurs-Abschnitt 10.3) handelt es sich um solche Studien.

Wie bereits erwähnt, wurde ein Großteil der Ergebnisse der Hauptstudie in Form einer Skalendokumentation publiziert (vgl. Meinhardt u. a., 2016c). Häufig wird in Kapitel 12 deshalb auf diese verwiesen. Die Skalendokumentation enthält darüber hinaus methodische Anmerkungen, die die Generierung der Ergebnisse erläutern. Sie ist damit ein essentieller Bestandteil dieses Dissertationsprojektes und Teile der Argumentation können ohne Sichtung selbiger ggf. nicht nachvollzogen werden.

Teil IV: Schluss Alle Ergebnisse werden im vierten Teil der Arbeit zusammenfassend dargestellt und mit Bezug auf die intendierte Testwertinterpretation analysiert. Ein abschließendes Validitätsargument wird formuliert. Die konkurrierenden Testwertinterpretationsmöglichkeiten werden gegeneinander abgewogen. Ein Ausblick mit Ideen für potentielle Anschlussfragestellungen bildet den Abschluss dieser Arbeit.

Teil V: Anhang Im letzten und fünften Teil der Arbeit können unter anderem alle Skalenversionen eingesehen und miteinander verglichen werden.

Teil I.

Theorie

2. Entwicklung und „Validierung“ eines Testinstruments

> „Validation was once a priestly mystery, a ritual performed behind the scenes, with the professional elite as witness and judge. Today it is a public spectacle combining the attractions of chess and mud wrestling.“
>
> *Cronbach, 1988, S. 3*

Cronbach bemüht im obigen Zitat natürlich ein sehr starkes Bild, welches jedoch in Zeiten großer Testprogramme wie z. B. den Schulleistungsstudien PISA oder TIMSS eine gewisse Berechtigung erhält. Es sei an dieser Stelle beispielsweise an die höchst emotional geführte PISA-Debatte erinnert. Das Abschneiden in Tests hat für den Einzelnen und die Gesellschaft immer größere Relevanz, wenn dieses über Zukunft und Teilhabe entscheidet. Die Frage nach der Vertrauenswürdigkeit solcher Tests erweist sich damit als zentral. Ob und unter welchen Voraussetzungen Entscheidungen auf Grundlage bestimmter Testergebnisse als gerechtfertigt angesehen werden, wird in der Regel in umfänglichen sogenannten Validierungsstudien ermittelt. Da es sich bei der vorliegenden Arbeit dem Anliegen nach um eine solche – wenn auch im Maßstab zu oben genannten Studien deutliche kleinere Validierungsstudie handelt – soll im nachfolgenden Abschnitt das Konzept der Validität einführend dargestellt werden. An dieser Stelle soll darauf hingewiesen werden, dass das Konzept der Validität ganz allgemein im Rahmen empirischer Forschung von Bedeutung ist. Nachfolgende Ausführungen fokussieren allerdings eher auf die *quantitative* empirische Forschung, weil die vorliegende Arbeit stärker an diesem Forschungsparadigma orientiert ist.

Im Laufe des Kapitels wird sich zeigen, dass der Begriff der Validität in den letzten 60 Jahren eine dynamische Entwicklung erfahren hat. Trotzdem halten sich mittlerweile überholte Vorstellungen, z. B. Validität als dreigliedriges Konzept oder als Eigenschaft eines Tests aufzufassen, hartnäckig. Die regelmäßig erscheinenden „Standards for Educational and Psychological Testing“ (vgl. AERA, APA, NCME, 2014) eignen sich, um ein aktuelles Bild des Konstruktes der Validität zu zeichnen. Dieses fußt im Wesentlichen auf drei konzeptionellen Grundannahmen, deren Erläuterung Gegenstand des nachfolgenden Kapitels sind. Demnach bezieht sich Validität auf die Interpretation von Testergebnissen, wird mittels eines argumentbasierten Ansatzes beurteilt und muss als ganzheitliches Konzept begriffen werden.

Darüber hinaus wird sich zeigen, dass die Testentwicklung und -validierung untrennbar miteinander verbunden sind, da der Entwicklung gerade das Ziel zugrunde liegt, mittels des neuen Instrumentes möglichst verlässliche Aussagen über die Ausprägung des zu messenden Merkmals bei einer bestimmten Gruppe von Personen treffen zu können. Prinzipien der Testkonstruktion sowie Besonderheiten bei der Erfassung sogenannter Persönlichkeitsvariablen (wie Selbstwirksamkeitserwartungen) werden daher im zweiten Teil dieses Kapitels thematisiert. Insbesondere werden Möglichkeiten betrachtet, potentielle Quellen von Invalidität bereits bei der Testkonstruktion zu minimieren, die die Ursache *konstruktirrelevanter Varianz* darstellen, das heißt einen Varianzanteil bilden, der nicht auf das eigentlich interessierende Konstrukt zurückzuführen ist.

2.1. Validität

„Validity is [...] the most fundamental consideration in developing tests and evaluating tests.“

AERA, APA, NCME, 2014, S. 11

Validität gehört neben Reliabilität, Objektivität und Fairness zu den klassischen Gütekriterien, mit deren Hilfe Messinstrumente bzw. diagnostische Verfahren hinsichtlich ihrer Qualität beurteilt werden können (vgl. z. B. Moosbrugger & Kelava, 2012). Wenn Testwerte mit einer hohen Wahrscheinlichkeit tatsächlich Auskunft über ein interessierendes Merkmal geben, so gelten diese Testwertinterpretationen als valide. Der Validität kommt dabei als übergeordnetem Kriterium eine besondere Rolle zu, denn wenn Testwerte nicht in beabsichtigter Weise interpretiert werden dürfen, dann bleibt das zuverlässige, objektive oder faire Verfahren der Erhebung dieser Werte ohne Bedeutung (vgl. Hartig, Frey & Jude, 2012, S. 144). Im nächsten Abschnitt wird dieser erste Zugang zum Konzept der Validität vertiefend behandelt. Anschließend wird die klassische Einteilung in Inhalts-, Kriteriums- und Konstruktvalidität beschrieben, sowie die Ablösung dieser Konzeptionalisierung durch eine diese Arten integrierende, zusammenführende Sichtweise auf das Konstrukt dargestellt. Der Prozess des Evaluierens im Rahmen einer Validierungsstudie wird dann detaillierter betrachtet und darüber hinaus die Wechselseitigkeit des Gütekriteriums der Validität zu den anderen genannten – insbesondere zur Reliabilität – thematisiert.

2.1.1. Definition und grundlegende Eigenschaften

Traditionell, aber vereinfacht (!) spricht man von einem „validen Test“, wenn man annimmt, dass dieser das misst oder abbildet, was er messen soll (vgl. u. a. Rost, 2004, S. 33 f.; Bortz & Döring, 2006, S. 200; Bühner, 2011, S. 61; Moosbrugger & Kelava,

2012, S. 13). In dieser Sprechweise wird dem Testinstrument selbst die Eigenschaft der Validität zugesprochen. Präziser formulieren AERA, APA, NCME (2014, S. 11), dass Validierung ein Evaluationsprozess ist, in dem darüber befunden wird, in welchem Maße theoretische Argumente oder empirische Belege angeführt werden können, die eine bestimmte Interpretation der Testwerte hinsichtlich einer festgelegten Testnutzung rechtfertigen. In dieser heute zumindest im theoretischen Sinne paradigmatischen Auffassung von Validität[1] geht es demnach darum, ob valide Aussagen (Testwertinterpretationen) aus einem Test gewonnen werden können oder nicht. Die Validierung eines Testinstrumentes hat damit zum Ziel, möglichst viele Argumente zu sammeln, die eine bestimmte Interpretation der Testwerte stützen oder diese verwerfen.

Man ist versucht anzunehmen, dass es sich bei der vermeintlichen „Ausschärfung“ um ein Sprachspiel handeln könnte, um eine Pingeligkeit. Jedoch lenkt diese Ausschärfung die Aufmerksamkeit auf eine fundamentale Implikation des Validitätskonstruktes: „No test permits interpretations that are valid for all purposes or in all situations“ (AERA, APA, NCME, 2014, S. 23). Aber auch die spezifische Interpretation der Testwerte eingeschränkt auf einen bestimmten Zweck, auf eine bestimmte Population und eine genau definierte Situation ist nicht unumstößlich oder dem Test nach einer Validierungsphase für immer zugeschrieben. Mit Berücksichtigung des Verlaufs wissenschaftlicher Entwicklungen sollten theoretische Annahmen grundsätzlich und fortwährend in Frage gestellt, sowie Anwendungskontexte auf ihre Angemessenheit bei jeder erneuten Testanwendung geprüft werden. Testanwendern kommt damit die Verantwortung und Aufgabe zu, ihre jeweiligen Interpretationen gewissenhaft zu begründen und sich nicht ausschließlich auf frühere Argumentationen zurückzuziehen. Die Verwendung der Phrase „Validität eines Tests“ ist damit aus dieser Perspektive als unangemessen einzustufen (vgl. AERA, APA, NCME, 2014, S. 11), weil sie Tatsachen suggeriert (Ein Test ist valide.), die unter Umständen einer eher „unwissenschaftlichen“ Praxis der Testanwendung (blinde Testanwendung) Vorschub leisten.[2]

Zusammenfassend ergeben sich aus der Definition und deren Deutung folgende Eigenschaften des Validierungsprozesses:

- Die Validierung eines Tests bezieht sich immer auf eine (oder mehrere) spezifische intendierte Testwertinterpretation(en), die deshalb zu Beginn des Validierungspro-

[1] Diese Konzeptionalisierung des Konstruktes der Validität ist dabei insofern *theoretisch* paradigmatisch, als dass sie als Konsens einer fortdauernden wissenschaftlichen Auseinandersetzung formuliert ist (vgl. AERA, APA, NCME, 2014), ihre praktische Umsetzung jedoch in vielfältigen anderen Wissenschaftsdisziplinen Anwendung findet und somit ggf. vom theoretisch formulierten Ideal abweicht.

[2] An dieser Stelle klaffen das Verständnis der Autorin des Begriffes der „Wissenschaftlichkeit“ und die gängige Praxis ggf. weit auseinander und man könnte anmerken, dass Wissenschaft sich über Tätigkeiten definiert, die in der Wissenschaft/Praxis betrieben werden, wozu eben teilweise auch eine eher blinde Testanwendung gehört.

zesses transparent gemacht werden muss (müssen) (vgl. u. a. AERA, APA, NCME, 2014, Standard 1.0; Kane, 1992).

- Die Testvalidierung ist ein evaluativer Prozess, an dessen vorläufigem Ende nach gründlicher Abwägung möglichst vieler Hinweise, Indizien, empirischer Belege, theoretischer Argumente etc. ein übergeordnetes Validitätsargument formuliert wird, welches eine bestimmte Testwertinterpretation stützt oder nicht (vgl. u. a. Cronbach, 1988, S. 4).
- Grenzen des Urteils über die Befürwortung der jeweiligen Testwertinterpretation speisen sich aus den Einschränkungen und Grenzen der theoretischen und empirischen Belege. Dabei können nie alle möglichen für- oder widersprechenden Argumente geprüft werden (vgl. u. a. Cronbach, 1988, S. 3).
- Im Validierungsprozess werden immer sowohl für- als auch widersprechende Argumente bezogen auf eine stimmte Testwertinterpretation zu Tage treten. Das Ziel eines Validierungsprozesses muss daher die möglichst unvoreingenommene Prüfung einer bestimmten Testwertinterpretation sein (vgl. u. a. Cronbach, 1988, S. 3), auch und gerade wenn die begründete Hoffnung der Verteidigung der intendierten Interpretation unter Umständen den Ausgangspunkt des Prozesses markiert.
- Die Diskussion möglicher Einschränkungen und Grenzen stellt eine Stärkung des späteren Validitätsargumentes dar, weil diese bei der Testanwendung und Testrevision berücksichtigt werden können (vgl. u. a. Cronbach, 1988, S. 4). Den zweifelhaftesten, problematischsten Annahmen einer intendierten Testwertinterpretation, ebenso wie der Interpretation potentiell entgegengesetzter, d. h. rivalisierender Hypothesen sowie der Antizipation möglicher negativer Konsequenzen einer Testwertinterpretation oder Testnutzung sollte im Validierungsprozess besondere Aufmerksamkeit zukommen, weil aus deren Entkräftung eine Stärkung des abschließenden, übergeordneten Validitätsargumentes resultiert (vgl. u. a. Kane, 2001; AERA, APA, NCME, 2014, S. 12 f.).
- Sowohl Argumentationsketten der Testentwickler als auch Ableitungen, die potentielle Testanwender treffen, müssen immer wieder kritisch vor dem Hintergrund der jeweils neuen Anwendungssituation als auch des wissenschaftlichen Fortschritts hinterfragt werden. Der Validierungsprozess ist damit als niemals abgeschlossen gekennzeichnet. Ein übergeordnetes Validitätsargument basiert daher zwangsläufig immer auf unvollständigen und vorläufigen Informationen (vgl. u. a. Cronbach, 1988, S. 4 f.).
- Wann die Validitätsfrage zunächst erschöpfend geklärt ist, muss im konkreten Fall begründet werden. Welche Prüfmethoden Anwendung finden, muss jeweils aus der festgelegten, zu validierenden Testwertinterpretation abgeleitet werden. Aufgrund der Vielzahl möglicher Testwertinterpretationen sowie potentieller Möglichkeiten

der Pro- oder Kontraargumentation kann kein allgemeingültiges Routineverfahren oder Abbruchkriterium benannt werden (vgl. u. a. Hartig u. a., 2012, S. 145).

2.1.2. Klassische Arten der Validität

Bereits vor rund 30 Jahren befand Cronbach (1988, S. 4): „The 30-year-old idea of three types of validity [...] is an idea whose time has gone.” Gleichwohl hat die klassische Dreiteilung in Inhalts-, Kriteriums- und Konstruktvalidität das Verständnis dieses Konstruktes über lange Zeit maßgeblich bestimmt und dient auch heute noch als Ausgangspunkt, das Konstrukt der Validität zu erläutern (vgl. z. B. Bortz & Döring, 2006, S. 200-203; Bühner, 2011, S. 61-64; Moosbrugger & Kelava, 2012; Schmiemann & Lücken, 2014). Viele Elemente dieser Troika sind in neueren Ansätzen der Beschreibung des Validitätskonstruktes integriert (siehe folgende Abschnitte), sodass eine kurze Erläuterung dieser drei benannten Validitätsarten gerechtfertigt erscheint. Zusätzlich sollen jeweils mögliche Verfahren beschrieben werden, mit deren Hilfe die jeweilige „Validität“ u. a. untersucht werden kann.

Inhaltsvalidität

Die zentrale Frage der Inhaltsvalidität ist die nach dem Verhältnis von Testinhalten zu dem zu messenden latenten Konstrukt. Inhaltsvalidität ist gegeben, wenn Testinhalte das Konstrukt möglichst gut abbilden. Dabei ist es wichtig, die Testinhalte als Einheit von Items/Aufgaben und zugehörigem Antwortformat zu betrachten, denn das Antwortformat beeinflusst maßgeblich die Art der erhobenen Testwerte, die interpretiert werden müssen. Fragen, die im Rahmen der Inhaltsvalidierung gestellt werden können sind u. a.: Welche Inhaltsbereiche oder Facetten des Konstruktes werden durch die Testinhalte abgedeckt? Ist das Konstrukt durch die Itemwahl eher eng oder eher breit abgebildet? Stellen die Items einen angemessenen Ausschnitt des zum Konstrukt gehörenden Itemuniversums dar und ist das Konstrukt vollständig durch die Testinhalte repräsentiert?

Die Basis, um solchen oder ähnlichen Fragen im Rahmen von Validierungsstudien auf den Grund zu gehen, ist eine fundierte theoretische Einbettung sowie eine detaillierte Definition und Beschreibung des Konstruktes. Dazu zählen in Abhängigkeit von der jeweiligen intendierten Testwertinterpretation zum Beispiel die begründete Festlegung oder Eingrenzung von Inhaltsbereichen des Konstruktes, systematisches und theorieorientiertes Ableiten von Iteminhalten, die begründete Festlegung eines Antwortformates, die begründete Darstellung von Itemkonstruktionsregeln oder die begründete Auflistung kognitiver Prozesse, die als relevant für die Beantwortung der Testaufgaben gelten.

Diese erarbeiteten Materialien können in weiteren Schritten als Reflexionsfolien einer Selbst- oder Fremdkontrolle beispielsweise durch Expertenurteile dienen. Kriteriengelei-

tet können so die Repräsentativität und Vollständigkeit der Abbildung des Konstruktes mittels der gewählten Items, die Passung des Antwortformates oder die korrekte Anwendung der Konstruktionsregeln beurteilt werden. Dabei kann jedes einzelne Item auf seine Konstruktpassung untersucht werden, aber auch das Zusammenwirken aller Items mit Blick auf eine bestmögliche Repräsentation des Konstruktes. Auch die Itemerprobung rekurriert auf die theoretische Fundierung, da in dieser mit Hilfe von Personen der potentiellen Probandengruppe überprüft werden kann, ob die Itemkonstruktion in dem Sinne gelungen ist, dass durch die Items erwartete kognitive Prozesse angeregt werden, oder diese im intendierten Sinne verstanden werden (vgl. Schmiemann & Lücken, 2014, S. 109-112; Hartig u. a., 2012, S. 148-152).

Kriteriumsvalidität

Kriteriumsvalidität, auch externe Validität, liegt dann vor, wenn von einem Testergebnis auf ein inhaltlich relevantes Kriterium (zum Beispiel das Verhalten der Person in einer anderen Testsituation) mit hoher Sicherheit geschlossen werden kann, also eine hohe Korrelation zwischen dem zu messenden Konstrukt und dem Kriterium vorliegt. Voraussetzung für die Einschätzung ist in jedem Fall ein plausibles und etabliertes „Außenkriterium“. Dies bedeutet, dass die Testwerte des „Kriteriums“ valide Interpretationen ermöglichen und eine theoretisch begründete Hypothese über die erwarteten Zusammenhänge vorliegt.

Im Rahmen der Kriteriumsvalidierung werden demnach Hypothesen getestet, die sich ebenfalls aus der theoretischen Fundierung des Konstruktes ergeben. Wenn die Testwerte in der beabsichtigten Weise interpretiert werden können, müssten sich die theoretisch erwartbaren starken Zusammenhänge (plausible Beziehungen) zu anderen Konstrukten empirisch nachweisen lassen.

Es werden drei Arten der externen Validität unterschieden, wobei sich der Unterschied jeweils durch den Zeitpunkt der Erhebung des Außenkriteriums manifestiert: Bei der Übereinstimmungsvalidität (konkurrente Validität), wird das externe Kriterium zeitgleich oder zeitnah zum eigentlich zu erfassenden Konstrukt erhoben. Im Rahmen der Prüfung der Vorhersagevalidität (prognostische bzw. prädiktive Validität) wird die Ausprägung der Kriteriumsvariable zu einem späteren Zeitpunkt erhoben als die zu validierenden Testwerte, während hinsichtlich der Prüfung auf retrospektive Validität die Annahmen über Zusammenhänge mit bereits erhobenen Variablen ausgewertet werden.

Diese Art der Validität spielt eine große Rolle, wenn auf Grundlage der Testwertinterpretation z. B. weitreichende Entscheidungen getroffen werden sollen. Empirische Belege der erwarteten Zusammenhänge sind dann sowohl ein Indiz für die Validität der Testwertinterpretation als auch für die Validität der abgeleiteten Entscheidungen (vgl. Schmiemann & Lücken, 2014, S. 112 ff.; Hartig u. a., 2012, S. 162-166).

Konstruktvalidität

Im Mittelpunkt der Konstruktvalidität steht die Interpretation der Testwerte auf Basis der konstruktkonstituierenden Theorie. Zum einen geht es um die Prüfung der theoretisch postulierten inneren Struktur eines Konstruktes, zum anderen um die Analyse der vermuteten Beziehungen zu anderen Konstrukten. Das Ziel ist eine schrittweise Überprüfung möglichst aller postulierten Zusammenhangsstrukturen. Da hohe Korrelationen zu anderen Konstrukten (Kriterien) im Rahmen der Kriteriumsvalidität geprüft werden, kann diese als Teil der Konstruktvalidität aufgefasst werden. Wenn die theoretisch angenommenen und empirisch ermittelten Strukturen übereinstimmen, so wird die These der Konstruktvalidität gestärkt. Dass der theoretischen Fundierung eine tragende Rolle zukommt, wird deutlich, wenn die Ursprünge des Konzeptes betrachtet werden.

Das Konzept der Konstruktvalidität geht auf Cronbach und Meehl (1955) zurück, deren Ausgangspunkt ein streng hypothetisch-deduktives Forschungsparadigma bildet. Dieses Paradigma, später „starkes Programm“ genannt, setzt das Vorhandensein einer formalen Theorie, d. h. eines nomologischen Netzes aus theoretischen Axiomen, beobachtbaren Variablen, sowie zwischen Axiomatik und Empirie vermittelnden Schluss- bzw. Korrespondenzregeln voraus (vgl. Abbildung 2.1). In der Praxis sind zugrunde liegende (psychologische) Theorien dagegen häufig unvollständig, wenig strukturiert, nicht streng axiomatisch, oder sie enthalten nicht alle interessierenden Konstrukte – das nomologische Netz muss folglich häufig als unzureichend bezeichnet werden (vgl. Hartig u. a., 2012, S. 153 f.). Theorien werden tatsächlich in den seltensten Fällen aus einem Guss formuliert, um dann empirischen Prüfungen unterzogen zu werden. Dieser Idealvorstellung ist sich auch Cronbach (1988, S. 14) bewusst, wenn er eingesteht: „[T]here is no hope of developing in the short run the „nomological networks“ we once envisioned.“ Das „Format“ der Konstruktvalidierung wird damit jedoch nicht aufgegeben, sondern zugunsten eines „schwachen Programms“, d. h. einer Validierung ohne strikte formale theoretische Basis, weiter verfolgt. Im Zuge der Validierung muss daher der epistemische Status, also z. B. der Grad der Formalisierung der zugrunde liegenden Theorie reflektiert werden. Das Fehlen genügend formaler Theorien rückt häufig operational definierte Konstrukte in den Fokus des Validierungsprozesses (vgl. Hartig u. a., 2012, S. 147 f.).

Die Grundlage für die Untersuchung auf Konstruktvalidität bildet daher wie angedeutet eine möglichst fundierte theoretische Einbettung des Konstruktes (im Sinne des schwachen Ansatzes). Der begründeten Darlegung der Beziehungen zu anderen Konstrukten sowie zu erwartender interner Strukturen kommt dabei eine zentrale Rolle zu, da aus dieser empirisch prüfbare Hypothesen abgeleitet werden können. Theoretische Annahmen hinsichtlich der inneren Struktur des Konstruktes können sich beispielsweise auf Itemschwierigkeiten, Korrelationen der Items untereinander oder die Dimensionalität des Konstruktes beziehen.

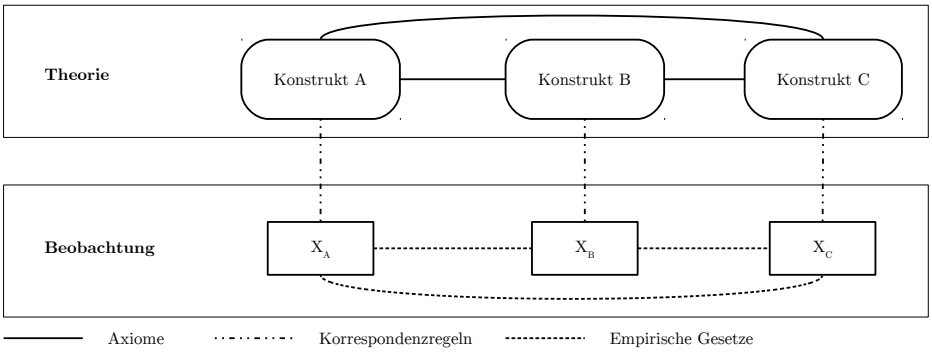


Abbildung 2.1.: Nomologisches Netz nach Cronbach und Meehl (1955). Quelle: nach Hartig u. a. (2012, S. 154).

Zur Prüfung der Hypothesen können verschiedene Verfahren angewandt werden. Klassischerweise werden Korrelationsanalysen durchgeführt. Wichtig ist das Abschätzen der Größe des erwarteten Zusammenhangs, sodass Grenzwerte zur Beurteilung abgeleitet werden können. Wird ein sehr hoher Zusammenhang zu anderen Konstrukten erwartet und empirisch bestätigt, so spricht man von konvergenter Validität. Wird kein oder nur ein schwacher Zusammenhang erwartet und empirisch gefunden, so handelt es sich um die sogenannte diskriminante Validität. Klassische Methoden zur Prüfung der postulierten inneren Struktur des Konstruktes (im Rahmen quantitativer empirischer Bildungsforschung) sind beispielsweise Faktorenanalysen, Methoden der Item Response Theory, aber auch Korrelationsanalysen, Expertenurteile oder Interviewstudien (vgl. Schmiemann & Lücken, 2014, S. 115 ff.; Hartig u. a., 2012, S. 152-162).

Abschließende Bemerkungen

Dass es keine einfachen, rezeptartig abzuarbeitenden Routinen der Validitätsprüfung spezifischer Testwertinterpretationen geben kann, ist bereits angeklungen – in den obigen Darstellungen jedoch ggf. nicht immer explizit enthalten. Deshalb soll dies im Folgenden an einigen Beispielen konkretisiert werden.

Zunächst ist bereits die Methodenauswahl keine Fingerübung. Die Ableitung adäquater Prüfverfahren ergibt sich jeweils aus der intendierten und zunächst priorisierten Testwertinterpretation. Die Sinnhaftigkeit einer Validierungsmaßnahme muss dabei jeweils im Kontext aller relevanten Fragestellungen und Konsequenzen bezüglich der Testnutzung kritisch betrachtet werden. Im Rahmen einer Leistungstestvalidierung skizziert Cronbach (1988, S. 7) beispielsweise die Problematik curricularer Validitätsargumente: „Matching tests to the curriculum produces a spurious validity argument, wherever the curricular aim is no higher than to have students reproduce authority's responses to a catechism."

Auch für die Einordnung und Wertung möglicher Resultate einer Prüfmethode

gibt es keine einfachen Lösungen. So wie positive Ergebnisse – positiv im Sinne der intendierten Testwertinterpretation – niemals die Validität endgültig und abschließend festschreiben, so führen negative Ergebnisse nicht zwangsläufig zu einer Falsifikation der Testwertinterpretationshypothese. Die geprüften Annahmen können falsch sein, oder die Testwertinterpretationen involvierter Konstrukte unzuverlässig. So wie es keine letzte Verifikation geben kann, gibt es auch keine endgültige, letzte Falsifikation.

Darüber hinaus resultieren weitere Fragen hinsichtlich der Interpretation der Ergebnisse einzelner Validierungsmaßnahmen. Was bedeutet beispielsweise im jeweiligen Fall positiv, was negativ im Sinne der Testwertinterpretation? Korrelationen können Grenzwerte über- oder unterschreiten, weil deren Höhe durch die Reliabilität aller verwendeten Instrumente beeinflusst ist. Expertenurteile können heterogen ausfallen, Indizien der konvergenten Validität die Erhebung des interessierenden Konstruktes in Frage stellen. Wann führen Korrelationsinterpretationen zu Zirkelschlüssen?

Der Wert eines übergeordneten Validitätsargumentes hängt damit wesentlich von der Überzeugungskraft der argumentativen Kette ab, die sich aus vielen einzelnen Interpretationsschritten zusammensetzt. Ergebnisse, die aus einem blinden Empirismus heraus resultieren, tragen nicht zur Plausibilität dieser Kette bei. Das Prinzip der Validität ad absurdum führt beispielsweise eine diskriminante Validierung „in alle Richtungen“. Dabei geht es nicht darum, das interessierende Konstrukt gegen irgendwelche anderen Konstrukte ohne theoretische Bezüge zum zu validierenden Konstrukt abzugrenzen. „Vielmehr gilt es nachzuweisen, dass kein oder nur ein geringer Zusammenhang zu einem – möglicherweise ähnlichen – Konstrukt besteht, das aber eben nicht erfasst werden soll“ (Schmiemann & Lücken, 2014, S. 116).

2.1.3. Konstruktvalidität als übergeordnetes Konzept

Aus den Darstellungen zu den drei „traditionellen“ Validitätsarten lässt sich bereits erahnen, dass diese unter Umständen Überlappungsbereiche aufweisen. Aspekte der Inhaltsvalidität können z. B. als Voraussetzungen weiterer Prüfungen angesehen werden und Korrelationen zu anderen Konstrukten spielen sowohl im Rahmen der Kriteriums- als auch der Konstruktvalidierung eine Rolle. Validität als ein einheitliches Konstrukt aufzufassen, welches nicht in verschiedene Arten unterteilt werden kann, gilt heute als state-of-the-art. In den Standards findet sich die Unterteilung jedenfalls längst nicht mehr, sie wird sogar explizit aufgegeben und als „historical nomenclature“ bezeichnet (vgl. AERA, APA, NCME, 2014, S. 14). Der Konstruktvalidität kommt dabei die Rolle zu, die anderen beiden Arten gewissermaßen in sich zu integrieren und als übergeordnetes Konstrukt zu fungieren. Aus historischer Perspektive formulieren Wainer und Braun (1988, S. 1): „It is particularly interesting to watch the evolution of construct validity from a junior member of a troika including criterion validity and content validity, to a

coequal status and, finally, to representing 'the whole of validity from a scientific point of view'."

Als ein vehementer Verfechter des skizzierten integrativen Ansatzes gilt Samuel Messick. Eine tragende Säule seiner Validitätstheorie nimmt der argumentative Ansatz (geprägt durch Cronbach (1988, S. 4)) der Rechtfertigung einer intendierten Testwertinterpretation unter Einbeziehung möglichst vieler Evidenzen ein, der zwangsläufig die Bedeutung verschiedener Validitätsarten negiert. „Hence, validity becomes a unified concept, and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely construct validity" (Messick, 1995, S. 744). Messick geht allerdings noch einen Schritt weiter, indem er im Rahmen der Validierung konsequent die Betrachtung oder Antizipation der jeweiligen Testnutzung sowie der sich aus der Testnutzung potentiell ergebenen positiven wie negativen Konsequenzen einfordert. Erst der Einbezug *aller* relevanten Belege, auch solcher, die die Testnutzung und deren Implikationen adressieren, „is what is meant by validiy as a unified concept" (Messick, 1995, S. 747). Insofern stellt die Validierung ein Zusammenspiel der Beurteilung der Bedeutung der in Frage stehenden Testwertinterpretation sowie deren Angemessenheit (im Sinne der Testnutzung) und der Nützlichkeit (im Sinne potentieller Konsequenzen) dar.

Das Zusammenführen einer Vielzahl unterschiedlicher Evidenzen zu einem übergeordneten Validitätsargument ist nach Messick ein komplexes Unterfangen. Aus diesem Grunde unterscheidet er sechs Validitätsfacetten, nämlich inhaltsbezogene („content"), substantielle („substantive"), strukturelle („structural"), generalisierende ("generalizability"), externe („external") sowie konsequenzenbezogene („consequential") Validitätsaspekte, die dann der Nuancierung, Differenzierung und präziseren Argumentation dienen sollen (vgl. Messick, 1995, S. 744 ff.). Inhaltsbezogene Validitätsaspekte beziehen sich auf die Vollständigkeit und Angemessenheit der Abbildung des jeweiligen Konstruktes, während sich substantielle Aspekte auf die Vollständigkeit und Angemessenheit der Abbildung kognitiver Prozesse, die mit dem Konstrukt assoziiert werden, fokussieren. Beide Aspekte ergänzen sich durch das Element der Repräsentativität: „The issue of domain coverage refers not just to the content representativeness of the construct measure but also to the process representation of the construct and to the degree to which these processes are reflected in construct measurement" (Messick, 1995, S. 745). Unter strukturellen Aspekten fasst Messick Facetten der inneren Struktur des Konstruktes, aber auch Fragen der Passung von Scoringmethoden oder Antwortformaten bezüglich des Konstruktes zusammen. Fragen nach der Verallgemeinerung für verschiedene Gruppen, Items oder Settings werden unter dem generalisierenden Validitätsaspekt summiert, während klassische Analysen der konvergenten oder diskriminanten Validierung zu den externen Aspekten zu zählen sind. Schließlich umfasst der „consequential aspect" alle Fragen nach Implikationen, die sich aus einer bestimmten Testwertinterpretation

inkl. der Nutzung der Testwerte ergeben. Im Sinne einer einheitlichen Validitätstheorie fordert Messick (1995, S. 747): „Evidence pertinent to *all* of these aspects needs to be integrated into an overall judgement“ (Hervorhebungen C. M.).

Abschließende Bemerkungen

Messicks Ideen sind in den Standards (AERA, APA, NCME, 2014) insoweit präsent, als die Validitätsdefinition die Testwertinterpretation in Relation zur Testnutzung direkt adressiert (AERA, APA, NCME, 2014, S. 11, Hervorhebungen C. M.): „Validity refers to the degree to which evidence and theory support the interpretations of test scores *for proposed uses of tests*.“ Dieses Verständnis des Validitätskonzeptes wird darüber hinaus als zentrales Qualitätskriterium in Form eines übergeordneten Standards (Standard 1.0) festgeschrieben, der die klare Artikulation der intendierten Testwertinterpretation bezüglich einer spezifischen Nutzung sowie die Angabe geeigneter Evidenzen für die Angemessenheit selbiger für jede Art von Test und von jedem Testanwender einfordert (AERA, APA, NCME, 2014, S. 23).

Messicks Forderung, das Bewerten potentieller Konsequenzen als integralen Bestandteil des Validierungsprozesses zu betrachten, hat in der Forschungslandschaft zu einer großen Debatte geführt (vgl. u. a. Moss, 1995; Crocker, 1997). Die Relevanz der Facette ist zumeist unbestritten, jedoch wird vereinzelt die Praktikabilität hinterfragt. So können in der Testentwicklung in den seltensten Fällen Konsequenzen beobachtet oder evaluiert werden. Konsequenzen können höchstens antizipiert werden, wobei diese Antizipation sehr wahrscheinlich unvollständig bleibt. Darüber hinaus impliziert das Konzept der „consequential validity“ einen Kausalzusammenhang zwischen der Testnutzung und spezifischen Beobachtungen, der in der Praxis schwerlich nachzuweisen sein dürfte (vgl. z. B. Reckase, 1998). Andere Autoren plädieren für eine engere Definition des Validitätsbegriffes (vgl. u. a. Popham, 1997; Mehrens, 1997; Borsboom, Mellenbergh & van Heerden, 2004; Lissitz & Samuelsen, 2007), wobei ihr Hauptargument darauf zielt, dass die Beurteilung der Auswirkungen bestimmter Testnutzungen nicht mehr der eigentliche Kern einer Validierung sei, sondern eher als „political issue“ betrachtet werden müsse (vgl. u. a. Mehrens, 1997, S. 18). Grundsätzlich argumentieren die Befürworter des umfassenden Ansatzes, dass das Abschätzen möglicher Konsequenzen dem Konzept der Testvalidierung quasi inhärent ist (vgl. Shepard, 1997), gerade weil die Grundfrage jeweils lautet, ob Testwerte tatsächlich wie intendiert, d. h. hinsichtlich ihrer intendierten Nutzung interpretiert werden können – ob die jeweilige Testwertinterpretation also ihrem Zweck oder Ziel gerecht wird. Messick (1995, S. 744) betont, dass Konsequenzen der Testnutzung nach Cronbach und Meehl (1955) Teil des nomologischen Netzes sind und damit per se Teil der Validierung. Gerade nicht bedachte Konsequenzen können den intendierten Zweck einer Untersuchung nachhaltig beeinflussen. Auch wenn im Validierungsprozess das Abschätzen von Implikationen und Konsequenzen ggf. nicht

immer im Mittelpunkt steht, so besteht der Verdienst Messicks darin, die Wichtigkeit der „consequential validity“ stärker ins Bewusstsein der Testentwickler und -anwender gerufen zu haben.

Die Standards positionieren sich klar bezüglich der Streitfrage, ob Betrachtungen zu Konsequenzen der Testnutzung essentieller Bestandteil einer jeden Testnutzung sein sollten, indem sie diese als eine *mögliche* Quelle für Validitätsargumente benennen (vgl. AERA, APA, NCME, 2014, S. 19 ff.). Dies impliziert, dass Validierungsszenarien denkbar sind, die nicht unbedingt auf den „consequential aspect“ eingehen. Shepard (1997, S. 6) weist in dem Zusammenhang darauf hin, dass in den seltensten Fällen zweckungebundene Validierungen durchgeführt werden und sobald eine spezifische Testnutzung adressiert wird, die Betrachtung verschiedenster Implikationen zum integralen Bestandteil der Validierung wird. Daher befürworten die Standards eine klare Positionierung hinsichtlich der Kosten-Nutzen-Rechnung (Nützlichkeitsanalyse), der Antizipation möglicher negativer Konsequenzen sowie der Gewichtung zu erwartender positiver und negativer Auswirkungen (AERA, APA, NCME, 2014, S. 11). Zusätzlich weisen sie auf Folgendes hin: „Validation is the joint responsibility of the test developer and the test user“ (AERA, APA, NCME, 2014, S. 13). Damit wird die Relevanz potentieller Konsequenzen hervorgehoben, aber eine potentielle Überforderung z. B. auf Seiten der Testentwickler (Validierung der Testwertinterpretation bezogen auf jegliche Anwendungskontexte) vermieden. Shepard (1997, S. 8) bringt es auf den Punkt, indem sie sagt: „When users appropriate tests for purposes not sanctioned and studied by the test developers, users become responsible for conducting the needed validity investigation.“ Ob damit wiederum eine Überforderung der Testanwender einhergeht, müsste näher beleuchtet werden – tatsächlich scheint die „zusätzliche“ Verantwortung hinsichtlich der spezifischen „Zweckvalidierung“ den meisten Anwendern unter Umständen nicht bewusst zu sein.

Kane (2001, S. 337) präsentiert eine Lösung des scheinbaren Dilemmas. Für ihn besteht das übergeordnete Validitätsargument aus zwei Teilen – dem deskriptiven Teil und dem präskriptiven Teil. Der deskriptive Teil bezieht sich auf die intendierte Testwertinterpretation und letzterer auf das Treffen bestimmter Entscheidungen, die auf der jeweiligen Testwertinterpretation beruhen und somit Konsequenzen der Testnutzung und -interpretation betreffen. Darüber hinaus schlägt Kane (2001, S. 337) vor: „Given the differences in the descriptive and prescriptive parts of the argument, it might be useful in many cases to evaluate the two parts of the interpretive argument separately.“ Damit verneint er nicht die Zusammengehörigkeit beider Teile im Sinne einer „unified validity“, aber berücksichtigt eine praktikable Aufteilung der Validierungsaufgaben in der Praxis, wobei er gleichzeitig die Verantwortung einer gerechtfertigten Nutzung eines Tests, so wie es auch die Standards nahe legen, auf Testentwickler und -anwender aufteilt (Kane, 2001, S. 337 f.).

2.1.4. Argumentbasierter Ansatz der Validierung

Die Auffassung, Validität als evaluativen Prozess zu begreifen, hat sich als Standard durchgesetzt (AERA, APA, NCME, 2014). Da Validitätsargumente auf vorläufigen und unvollständigen Informationen beruhen, werden an die Argumentation hohe Ansprüche gestellt. Dezidiert nimmt Kane (1992) den Argumentationsprozess in den Blick, weil er diesen als Kern des Validitätskonzeptes versteht. Wie bereits erwähnt, beginnt der Prozess mit der detaillierten Darlegung der zu untersuchenden Testwertinterpretation. Aus dieser wiederum können die *relevanten* Validierungsschritte systematisch abgeleitet werden. „The kinds of evidence needed for validation of a test-score interpretation can be identified systematically by focusing attention on the details of the interpretation“ (Kane, 1992, S. 527). Damit unterscheidet sich seine Herangehensweise erkennbar von der Ansicht Messicks, im Rahmen einer Validierung Evidenzen bezüglich *aller* Validitätsfacetten anzuführen (vgl. Ausführungen auf S. 19) .

Die Validierungsschritte können Kane zufolge deshalb direkt aus intendierten Testwertinterpretationen abgeleitet werden, weil diesen bestimmte Schlussfolgerungen immanent sind, die im Rahmen des Aufbaus einer Argumentationskette zur Stützung einer intendierten Testwertinterpretation eine große Rolle spielen und folglich im Validierungsprozess besonderer Aufmerksamkeit bedürfen, da deren Gültigkeit zu belegen ist (Kane, 2001, S. 329 f.). Die genaue Kenntnis und Ausdifferenzierung der Schussfolgerungen ist dabei der Aspekt, der im Validierungsprozess die Auswahl bestimmter Validierungsschritte vereinfacht, weil diese gerade darin bestehen, Argumente für die Gültigkeit der Schlussfolgerungen zur Verfügung zu stellen („supporting evidence“, Kane, 2001, S. 330). Kane (2001, S. 330) unterscheidet standardmäßig bewertende („evaluation“), generalisierende/verallgemeinernde („generalization“), extrapolierende („extrapolation“), erklärende („explanation“) sowie entscheidungsrelevante („decision making“) Schlüsse.[3]

Sowohl bewertende als auch generalisierende Schlüsse müssen in der Regel für jede Testwertinterpretation gerechtfertigt werden. Bewertende Schlüsse liegen dann vor, wenn im Zuge der Testwertinterpretation von der Sinnhaftigkeit der Zuweisung eines Wertes zu einer Bearbeitungsperformanz ausgegangen wird. Daher ist es im Zuge der Validierung u. a. notwendig, die Wertzuweisung bzw. das Scoringverfahren plausibel zu machen. Generalisierende Schlüsse werden gezogen, wenn der Testwert unabhängig vom spezifisch gewählten Itemset sowie unabhängig von der Erhebungssituation etc. als Merkmalsausprägung interpretiert werden kann (vgl. Kane, 2013, S. 10).

> „[W]e are not so much interested in making claims about how well the test taker

[3] Kane ändert zum Teil die Wahl der Begrifflichkeiten. So beschreibt er evaluierende Schlüsse in früheren Beiträgen unter dem Stichwort „observation“ (vgl. Kane, 1992, S. 529), während er in aktuellen Beiträgen diese Art der Schlussfolgerung als „scoring inference“ bezeichnet (vgl. Kane, 2013, S. 10).

> did on a particular set of tasks on a particular day in a particular year; that is a matter of fact. Typically, we want to make claims about how well the test taker can perform in some larger domain of tasks over some range of occasions and conditions of observation [...]“ (Kane, 2013, S. 10).

Wird der Testwert dazu verwendet, Performanz in anderen Kontexten oder die Ausprägung anderer Konstrukte vorherzusagen, so liegt ein extrapolierender Schluss vor (vgl. Kane, 1992, S. 529; Kane, 2013, S. 11). Von einem erklärenden Schluss spricht man dann, wenn der vorliegende Testwert durch Anwendung einer hinreichend formalen Theorie kausal mit der Ausprägung eines anderen Merkmals oder dem Wirken eines bestimmten Mechanismus plausibel gemacht werden kann.[4] Entscheidungsbasierte Schlüsse spielen dann eine Rolle, wenn der Testwert als Begründung dafür genutzt wird, bestimmte Maßnahmen zu ergreifen.

Sowohl für Testvalidierende als auch Testanwendende ist es bedeutsam, dass nicht jeder intendierten Testwertinterpretation alle benannten Schlüsse inhärent bzw. für das Validitätsargument relevant sind. Wird beispielsweise bezüglich der Testwertinterpretation keine spezifische Testnutzung adressiert, so entfällt in der Regel die Prüfung der Gültigkeit extrapolierender und entscheidungsbasierter Schlüsse. Ist das interessierende Konstrukt zwar in einem nomologischen Netz verankert, aber nicht stringent in kausale Strukturen eingebettet, so können per se keine erklärenden Schlüsse gerechtfertigt werden. Unterschiede in den Testwertausprägungen können dann lediglich zur Kenntnis genommen werden. Zusätzlich sind die vorgestellten Schlüsse nicht isoliert, sondern als Teil einer umfangreichen Argumentationskette zu verstehen. Beispielsweise kann die Gültigkeit eines extrapolierenden Schlusses im Zuge einer Validierung nur dann gerechtfertigt werden, wenn sowohl die Gültigkeit bewertender als auch verallgemeinernder Schlüsse als Argumentationsgrundlage dienen.

Die mit der spezifischen Testwertinterpretation und deren innewohnenden Schlüssen verbundenen sowie zumeist aus der zugrunde liegenden Theorie resultierenden Annahmen müssen in einem nächsten Schritt zur Ableitung jeweils gezielter Überprüfungsmaßnahmen offen gelegt werden. Besonderes Augenmerk legt Kane (1992, 2001) auf das Aufdecken möglicher konkurrierender Interpretationen, die im Zuge der Validierung entkräftet werden können. Insgesamt ergibt sich folgender standardisierter Ablauf (vgl. u. a. Kane, 2001, S. 330):

1. Festlegung der Testwertinterpretation,
2. Spezifizierung aller Annahmen, die der intendierten Testwertinterpretation zugrunde liegen,
3. Identifikation potentiell konkurrierender Interpretationen, sowie der strittigsten Annahmen,

[4] Ein Beispiel kann u. a. bei Hartig u. a. (2012, S. 151) nachgelesen werden.

4. Evaluation der Annahmen mit Fokus auf den problematischsten Annahmen bzw. den alternativen Interpretationen,
5. Revision der Testwertinterpretation oder zugrunde liegender Annahmen und erneute Evaluation.

Neben der ausführlichen Beschreibung des Argumentationsablaufes besteht Kanes Verdienst in der detaillierten, grundsätzlichen Charakterisierung der Argumentationsführung. Es kennzeichnet diese als „praktische Argumentation“ und grenzt sie damit gegenüber streng logischen, mathematischen Argumenten ab. Allein die Tatsache, dass nie alle Annahmen geprüft werden können, fokussiert auf die Eigenschaft der Plausibilität und Überzeugungskraft eines Validitätsargumentes. Nie können im mathematischen Sinne Beweise angeführt werden, da Annahmen in der Regel nicht als Axiome gelten können. Im Rahmen der „praktischen Argumentation“ tragen eine Vielzahl unabhängiger Evidenzen zur Rechtfertigung der Gültigkeit ein und derselben Annahme (Triangulation) sowie das Entkräften antizipierter Kontraargumente zur Stärkung der Überzeugungskraft des Validitätsargumentes bei.

Um überzeugend zu sein, sollte die Argumentation den folgenden drei Kriterien genügen (vgl. Kane, 1992, S. 528):

1. Klarheit der Argumentation (präzise, klare, detaillierte, umfassende Formulierung aller Annahmen und Schlüsse),
2. Kohärenz der Argumentation (nachvollziehbare, schlüssige, theoriegestützte Darstellung der Argumentationskette),
3. Plausibilität der Argumentation (Anführung ausreichender theoretischer wie empirischer Evidenzen).

Zusammenfassend stellen sich die Grundzüge des argumentbasierten Ansatzes der Validität wie folgt dar: „First, state the claims that are being made in a proposed interpretation or use [...], and second, evaluate these claims (the validity argument)“ (Kane, 2013, S. 9).

Abschließende Bemerkungen

Die Standards (AERA, APA, NCME, 2014) beschreiben eine Art Synthese aus den Ansichten Kanes und Messicks. Als gemeinsamer Nenner und tragendes Fundament kann dabei die Definition betrachtet werden, in der die Testwertinterpretation auf den spezifischen Testnutzen bezogen und eine Argumentation für oder gegen die intendierte Testwertinterpretation eingefordert wird. Die Standards verzichten zwar auf die Begrifflichkeiten Messicks, orientieren sich aber inhaltlich an den Validitätsaspekten, die er beschreibt, wenn es um die Quellen für mögliche Evidenzen im Rahmen einer Validierung geht (vgl. AERA, APA, NCME, 2014, S. 14-20):

- Evidenz basierend auf den Testinhalten,
- Evidenz basierend auf dem Antwortprozess,
- Evidenz basierend auf der inneren Struktur des Konstruktes,
- Evidenz basierend auf Relationen zu anderen Variablen (konvergente, diskriminante, kriteriale, generalisierende),
- Evidenz basierend auf Betrachtungen (möglicher) Konsequenzen der Testnutzung.

Gleichzeitig befürworten die Standards im Sinne Kanes die Auswahl geeigneter Methoden in Abhängigkeit der zu prüfenden Testwertinterpretation: „[E]ach type of evidence [...] is not required in all settings. Rather support is needed for each proposition that underlies a proposed test interpretation for a specified use“ (vgl. AERA, APA, NCME, 2014, S. 14). Die Betrachtung von Konsequenzen der Testnutzung ist damit nicht zwangsläufig Teil der Validierung oder der Testentwicklung.

2.1.5. Zum Verhältnis von Validität zu anderen Testgütekriterien

Die Reliabilität ist ein weiteres Gütekriterium zur Beurteilung (psychologischer) Messinstrumente oder Tests und ein Maß für die Messgenauigkeit der Erhebung des latenten Konstruktes (vgl. Schermelleh-Engel & Werner, 2012, S. 120). Intuitiv einsichtig erscheint es, dass Messgenauigkeit eine wesentliche Voraussetzung für Validität darstellt, denn „if test scores vary substantially over replications of the testing procedure, it is hard to give the scores any consistant interpretation“ (Kane, 2013, S. 29).

Moss (1994) weist darauf hin, dass Reliabilität in der quantitativen empirischen Forschung oft mit statistischen Kennzahlen wie Cronbachs Alpha gleichgesetzt wird, was die ursprüngliche Bedeutung des Reliabilitätskonzeptes (Messgenauigkeit) auf Generalisierungsaspekte einschränkt. Im psychometrischen Sinne liegt damit jeder Reliabilitätsaussage eine Generalisierung zugrunde. Zudem beruhen zu validierende Testwertinterpretationen in der Regel auf generalisierenden Schlüssen, sodass nicht nur intuitiv, sondern ganz praktisch Reliabilität als notwendige, wenn auch nicht hinreichende Bedingung für Validität anzusehen ist (vgl. Kane, 1992, S. 529). Ob dies auch in anderen Settings wie zum Beispiel in hermeneutischen Ansätzen ebenso der Fall ist, muss gesondert diskutiert werden und ist ggf. nicht trivial. Die Debatte dazu kann u. a. bei Moss (1994, 2004), Mislevy (2004) oder Li (2003) nachverfolgt werden.

Wenn jedoch generalisierende Testwertinterpretationen, wie in den meisten psychometrischen Anwendungen, zur Debatte stehen, so spielt die Standardisierung und damit Objektivierung von Testsituationen, sowie Testdurchführungen und Testauswertungen eine zentrale Rolle. In diesem Falle ist also das Testgütekriterium der Objektivität Voraussetzung für Reliabilität und damit für Validität (vgl. u.a. Rost, 2004, S. 33; Schermelleh-Engel & Werner, 2012, S. 120).

Zusätzlich ist Reliabilität im psychometrischen Sinne nicht nur Voraussetzung für Validität, sie beschränkt diese zugleich. Dies wird anschaulich, wenn Korrelationskoeffizienten als Validitätsmaße interpretiert werden. Diese sind im Sinne der klassischen Testtheorie durch die Reliabilitätsmaße der im Koeffizienten enthaltenen Konstrukte beschränkt (Rost, 2004, S. 392; Bortz & Döring, 2006, S. 202; Moosbrugger & Kelava, 2012, S. 159). Dieser Beschränkung wohnt ein Dilemma inne. Denn der Versuch, die Reliabilitäten zu erhöhen, führt oft zu einer Homogenisierung der Items. Sehr homogene Items bilden jedoch in den seltensten Fällen das Konstrukt in seiner Vielschichtigkeit und seinem Facettenreichtum ab, was jedoch ein Ziel hinsichtlich der Validität wäre. Man spricht dann von einer „zu Tode homogenisierten Skala“ (Bühner, 2011, S. 249). Diese Problematik ist auch unter dem Validitäts-Reliabilitäts-Dilemma bekannt (vgl. u. a. Rost, 2004, S. 392 ff.; Rost, 2013, S. 364).

2.1.6. Zusammenfassung

„In the very earliest discussions of test validity, some writers said that a test is valid if it measures ‚what it purports to measure‘. That raised, in primitive form, a question about truth. Other early writers, saying that a test is valid if it serves the purpose for which it is used, raised a question about worth“ (Cronbach, 1988, S. 5). Cronbach greift in diesem Zitat zwei der eher veralteten, wenn auch heute in Lehrbüchern weitgehend noch präsenten Auffassungen von Validität auf. In beiden Auffassungen wird dem Testinstrument die Eigenschaft der Validität zugeschrieben. Zusätzlich geht mit diesem Validitätsverständnis in der Regel eine Gliederung des Konstruktes in verschiedene Arten einher. Man fragt sich beispielsweise, ob der Test inhaltsvalide ist.

Mit der Fokussierung des Konstruktes auf die Einschätzung der Glaubwürdigkeit bestimmter Testwertinterpretationen setzt sich eine fundamentale Änderung im Verständnis des Konstruktes der Validität durch. Zur Einschätzung der Angemessenheit der intendierten Interpretation müssen zahlreiche Belege angeführt werden, die zwar verschiedene Aspekte der Validität beleuchten können, aber letztlich durch Zusammenführung zu einem übergeordnetem Urteil (Validitätsargument) ein ganzheitliches Konstrukt der Validität repräsentieren. In die Argumentation können auch Betrachtungen über Konsequenzen der Testnutzung einfließen. Linn (1997, S. 15) fasst den begrifflichen Wandel wie folgt zusammen:

> „The early primitive form of the question about truth – that the test measures what it purports to measure – has evolved into an elaborate theory of construct validity. In a similar manner, the notion that validity is concerned with the degree to which a test achieves certain aims is more fully and coherently articulated in the discussion of the need to evaluate intended positive and plausible unintended negative effects of test use.“

Im Rahmen dieses argumentbasierten Ansatzes werden an die Argumentation verschiedene Anforderungen gestellt, um die Überzeugungskraft eines Validitätsargumentes zu stärken. Diese sollte klar, kohärent und plausibel sein sowie parallele Argumentationsstränge enthalten und potentielle Kontraargumente berücksichtigen. Die Abschätzung potentieller Implikationen wird in der Regel als essentiell für den Validierungsprozess angesehen, auch wenn Entwicklungsszenarien denkbar sind, die keinen Schwerpunkt auf bestimmte Testnutzungen legen. Zumeist werden dann pragmatische Herangehensweisen präferiert, wie eine Aufteilung der Validierungsverantwortung auf Testentwickler und Testnutzer.

Insgesamt sollen die beschriebenen Anforderungen an den argumentbasierten Validierungsprozess für dieser Arbeit handlungsleitend sein. Neben der detaillierten theoretischen Fundierung in den nachfolgenden Kapiteln werden dazu beispielsweise in Kapitel 7, Abschnitt 7.5 die intendierte und damit zu validierende Testwertinterpretation sowie konkurrierende Interpretationen aufgezeigt und konkrete Maßnahmen zur Überprüfung selbiger abgeleitet. Aus den Ergebnissen dieser Maßnahmen im Zuge mehrerer Validierungsstudien (vgl. Kapitel 9 bis 12) werden jeweils Argumente hinsichtlich der Validität der angestrebten Testwertinterpretation abgeleitet (vgl. z. B. Abschnitte 9.6, 10.6 und 12.8) und abschließend zu einem übergeordneten Validitätsargument zusammengeführt und diskutiert (vgl. Abschnitt 13.1).

2.2. Entwicklung eines Testinstrumentes

Wie bereits deutlich geworden sein sollte, beginnt die Validierung einer Testwertinterpretation nicht erst nach Fertigstellung eines neuen Instrumentes – vielmehr müssen Validitätsaspekte während des gesamten Entwicklungsprozesses beachtet und reflektiert werden. Entwicklungs- und Valdierungsarbeit gehen damit Hand in Hand. Wie in den vorangegangenen Abschnitten dargestellt, beginnt die Validierung mit der sorgfältigen theoretischen Fundierung und der (begründeten) Festlegung wichtiger Parameter (Zielgruppe, Antwortformat, etc.). Offensichtlich werden aus diesen Betrachtungen konkrete Hinweise für die Aufgaben- oder Itemkonstruktion und damit die Instrumententwicklung abgeleitet. Zudem ist der Entwicklungsprozess wesentlich durch das Bestreben beeinflusst, konstruktirrelevante Varianz bei der Erhebung des interessierenden Konstruktes zu vermeiden, da diese als Hauptursache für Invalidität angesehen wird (vgl. AERA, APA, NCME, 2014, S. 12 f.). Letztlich spiegelt sich das Wechselverhältnis von Validierung und Entwicklung darin, dass Resultate der Validierungsstudien immer Möglichkeiten eröffnen, das Instrument zu optimieren oder weiterzuentwickeln.

Im Folgenden soll zunächst die Problematik konstruktirrelevanter Varianz erläutert werden, sowie häufiges, in der Regel unbewusstes Verhalten der Probanden im Rahmen der Itembeantwortung aufgegriffen werden, welches zur Verfälschung der Messergebnisse

beiträgt. Um konstruktirrelevante Varianz durch die beschriebene Art der Verzerrung zu minimieren bzw. zu vermeiden, muss der Prozess des Beantwortens zunächst ausführlich modelliert werden. Aus diesem Modell ergeben sich einige allgemeine Prinzipien wie Verständlichkeit, Sachlichkeit, Eindeutigkeit, etc., die als Leitgedanken die Testkonstruktion bestimmen.

2.2.1. Konstruktirrelevante Varianz

„Construct-irrelevance refers to the degree to which test scores are affected by processes that are extraneous to the test's intended purpose“ (AERA, APA, NCME, 2014, S. 12). Prozesse, die solche konstruktirrelevante Varianz erzeugen, sind im Wesentlichen in systematischen Fehlerquellen bei der Itembeantwortung zu suchen, die dann zu einer „Kontamination“ des zu erhebenden Konstruktes führen. Zu systematischen Fehlbeantwortungen kann es zum Beispiel kommen, wenn einige Subgruppen, die den Test beantworten, jeweils in irgendeiner Art und Weise benachteiligt oder bevorteilt werden z. B. aufgrund bestimmter Formulierungen (Gender-Stereotype, Fachbegriffe, ...) oder aufgrund der Praxis der Testdurchführung. Dieses Beispiel hebt dabei den Punkt der Testfairness zur Vermeidung systematischer Fehler hervor (AERA, APA, NCME, 2014, Kapitel 3). Weitere typische systematische Fehler, die im Rahmen des Antwortprozesses eine Rolle spielen, sind unter den Schlagwörtern „soziale Erwünschtheit“ und „Akquieszenz“ bekannt. Bevor diese Begrifflichkeiten näher erläutert und Möglichkeiten der Vermeidung konstruktirrelevanter Varianz aufgezeigt werden, soll zunächst der Prozess der Itembeantwortung detaillierter betrachtet werden, da diesem in Bezug auf das Zustandekommen systematischer Fehler eine besondere Bedeutung zukommt.

Modell der Itembearbeitung

In der Regel kann die Analyse der kognitiven Prozesse, die durch die Darbietung des Items ausgelöst werden und die auf die Beantwortung des Items abzielen, weiterhelfen, wenn nach konstruktirrelevanter Varianz gesucht wird. All jene Prozesse, die nicht auf das eigentlich interessierende Konstrukt fokussieren, sind daher als Fehlerquelle bei der Itembeantwortung zu betrachten. Insgesamt kann es zu Fehlern in jedem Stadium der Itembeantwortung kommen. Es werden die Stadien Verständnis des Items, Abruf relevanter Informationen aus dem Langzeitgedächtnis, Urteilsbildung unter Bezugnahme auf als relevant erachteter Informationen, Wahl der Antwortkategorie sowie Abgabe der Antwort beispielsweise durch Ankreuzen unterschieden (vgl. Jonkisz, Moosbrugger & Brandt, 2012, S. 58; nach Podsakoff, MacKenzie, Lee & Podsakoff, 2003). Jedes der fünf Stadien unterliegt dabei unterschiedlichen Fehleranfälligkeiten. Unter Umständen versteht der Befragte das Item in nicht intendierter Weise oder es werden aufgrund bestimmter Reize vornehmlich negative oder positive Assoziationen zum Item abgerufen.

Das Urteil kann beeinflusst werden, wenn der Proband aus der Zusammenstellung der Items eine Art Globalurteil fällt und dann nicht jedes Item für sich bewertet, sondern das Globalurteil heranzieht. In den meisten Fällen müssen die Urteile aktiv konstruiert werden; die Urteilsbildung ist aus diesem Grunde von den zur Verfügung stehenden Informationen abhängig. Informationen, die das Urteil verfälschen, können unter anderem aus der Itemreihenfolge, vorangestellten Instruktionen, dem Erhebungskontext oder missverständlichen Formulierungen stammen. Die Wahl einer Antwortkategorie und das tatsächliche Antwortgeben kann von einer Reihe von „Fehlern“ begleitet sein, beispielsweise von einer Tendenz zur Mitte, oder der Korrektur des Urteils hinsichtlich eines möglicherweise „erwünschteren“ Urteils (vgl. folgende Abschnitte).

In dem dargestellten Modell der Itembeantwortung wird davon ausgegangen, dass das Durchlaufen aller fünf Schritte die Wahrscheinlichkeit erhöht, eine „optimale“ Antwort zu geben, die also eher Rückschlüsse auf die tatsächliche Ausprägung des Konstruktes zulässt. Werden dagegen kognitive Prozesse nur oberflächlich tangiert oder gar übersprungen – beispielsweise wenn nach dem Lesen ohne Abwägung eine Antwortkategorie ausgewählt und angekreuzt wird – so sinkt höchstwahrscheinlich die Qualität der Itembeantwortung. Deshalb ist es wichtig, Faktoren zu identifizieren, die den Ablauf des Antwortprozesses möglicherweise beeinflussen, um in einem nächsten Schritt über die gezielte Manipulation dieser Faktoren den Antwortprozess zu optimieren (Bühner, 2011, S. 109).

Jonkisz u. a. (2012, S. 57 f.) nennen den Faktor der Motivlage der Probanden. Handelt ein Proband beispielsweise aus altruistischen Motiven heraus und ist damit dem Test gegenüber eher wohlwollend eingestellt, dann ist es sehr wahrscheinlich, dass sich der Proband viel Mühe gibt und folglich unbewusst alle Stadien der Itembeantwortung hinreichend tiefgehend vollzieht, was die Wahrscheinlichkeit für eine Konsistenz zwischen tatsächlicher Konstruktausprägung und Antwortverhalten erhöht. Denkbar ist aber auch, dass Befragte aus Zufall oder aufgrund von Verpflichtungen am Test teilnehmen. Dann bearbeiten sie den Test ggf. ebenfalls wohlwollend, jedoch nur oberflächlich, oder sie beantworten den Test nach äußeren bzw. formalen Kriterien, sodass Antworten unter Umständen in gar keinem Zusammenhang zur tatsächlichen Konstruktausprägung stehen (vgl. Jonkisz u. a., 2012, S. 58).

Aus dem Modell der Itembeantwortung und insbesondere aus den potentiellen Fehlerquellen bezüglich der einzelnen Stadien ergeben sich Ableitungen bzw. grundsätzliche Prinzipien in Bezug auf die Testkonstruktion. So sollten Items möglichst verständlich, sowie intersubjektivierbar („eindeutig“) formuliert sein. Auch Sachlichkeit scheint von größter Wichtigkeit zu sein, um den Abrufprozess nicht in bestimmte Richtungen zu lenken und damit die Urteilsbildung zu verzerren. Auch die Anordnung der Items kann den Urteilsprozess beeinflussen und muss daher bei der Testzusammenstellung berücksichtigt werden. Bezüglich der Wahl und Abgabe des Urteils ergeben sich Konsequenzen sowohl

für die Testkonstruktion als auch für die Testdurchführung (vgl. nachfolgender Abschnitt 2.2.2). Da sich auch aus den Motivlagen der Probanden potentielle Verzerrungen bei der Itembeantwortung ergeben können, sollten diese bei der Akquise der Probanden sowie die Testdurchführung berücksichtigt werden. Für die Testkonstruktion ergeben sich Möglichkeiten der Formulierung geeigneter Instruktionen zur Beeinflussung der Motivation der Testteilnehmerinnen und Testteilnehmer.

2.2.2. Klassische Beurteilungsfehler

Nachfolgend werden in Kürze klassische Fehler dargestellt, die sich auf die letzten beiden Stadien der Itembeantwortung beziehen. Es werden Konsequenzen für die Testkonstruktion und -durchführung abgeleitet, um konstruktirrelevante Varianz zu vermeiden. Im Rahmen der Latent-State-Trait-Theorie können Beurteilungsfehler wie die Tendenz zur sozialen Erwünschtheit zum Beispiel als Teil der Interaktion von Situation und Person explizit im Modell berücksichtigt werden (vgl. Ziegler & Bühner, 2009).

Soziale Erwünschtheit

Das Verhalten Befragter, in einem Test nicht die individuell am besten passendste Antwortkategorie zu wählen, sondern diejenige, welche vom Befragten als möglicherweise von ihm erwünscht oder erwartet eingeschätzt wird, verfälscht in hohem Maße die Messung des latenten Konstruktes (Effekt der „sozialen Erwünschtheit“). Im Falle eines Antwortverhaltens entsprechend der sozialen Erwünschtheit täuscht der Proband nicht nur den Tester, sondern auch sich selbst. Die Tendenz sich selbst in einem möglichst guten Licht darzustellen, ist dabei eine auch im Alltag häufig anzutreffende, sehr menschliche Eigenschaft. Die Tendenz wird im Befragungskontext gefördert, wenn in der Testsituation Personen anwesend sind, zu denen Befragte sonst in einem Abhängigkeitsverhältnis stehen, weil sie beispielsweise durch diese beurteilt werden. Auch wenn Testergebnisse weitreichende Konsequenzen für die Befragten nach sich ziehen (Einstellung, Auswahl, ...) ist die Wahrscheinlichkeit, sich besser darzustellen, ggf. größer. Beim Einsatz unterschiedlicher Skalen oder ganzer Testbatterien, ist dann die Verzerrung auf solchen Skalen unter Umständen umso größer, je mehr die Skalen von den Befragten als relevant für die zu treffenden Entscheidungen (Einstellung etc.) erachtet werden. Zusammenfassend sagen Ziegler und Bühner (2009, S. 550): “Consequently, faking is not just simply a response distortion. Rather, it depends on contextual variables.“

Um ein Antwortverhalten möglichst unabhängig von vermeintlichen, sozialen Erwartungen zu erhalten, können Maßnahmen ergriffen werden wie zum Beispiel die Zusicherung absoluter Anonymität, „die Betonung, dass mit der Befragung ein neuer Fragebogen oder eine Theorie auf dem Prüfstand stehen und nicht etwa die befragte

Person [...] [oder] dass es keine richtigen oder falschen Antworten gibt und es nur auf die persönliche Einschätzung ankommt“(Mummendey & Grau, 2014, S. 49 f.). Weitere Maßnahmen wären die Durchführung des Tests durch unabhängige/unbekannte Tester, der Einsatz sogenannter Kontrollskalen[5] zur Quantifizierung des Ausmaßes sozial erwünschten Antwortens oder das Modellieren der Verzerrung als systematischen Fehler im Rahmen der Strukturgleichungsmethodik (vgl. Jonkisz u. a., 2012, S. 59 f.; Bühner, 2011, S. 126). Weitere Methoden der Kontrolle sozial erwünschten Antwortens berichten z. B. Mummendey und Grau (2014, S. 174-190).

Antworttendenzen

Die Neigung jeweils eine bestimmte Kategorie zu wählen, d. h. eine Tendenz zur Zustimmung („Akquieszenz“), Ablehnung oder zur Mitte zu zeigen, verringert die Item- und damit die Skalenvarianz. Um diesem Phänomen entgegen zu wirken, sollten beispielsweise keine „Mittelkategorien“ zur Wahl gestellt und keine extremen Bezeichnungen der Skalenpole verwendet werden. Um Zustimmungs- oder Ablehnungstendenzen aufzudecken, werden häufig Items in ihr Gegenteil umformuliert und das Antwortverhalten auf Konsistenz geprüft. Diese sogenannte Umpolung von Items kann jedoch weitere Schwierigkeiten nach sich ziehen und muss wohl überlegt sein. Die Tendenz zuzustimmen, tritt womöglich dann gehäuft auf, wenn Probanden mit den Items überfordert sind, sodass die Zielgruppenadäquatheit solchen Antworttendenzen unter Umständen entgegen wirkt (vgl. Jonkisz u. a., 2012, S. 60 ff.; Mummendey & Grau, 2014, S. 160).

2.2.3. Aspekte der Testplanung

An eine fundierte theoretische Einbettung des zu erhebenden latenten Konstruktes, die „eine intensive Literaturrecherche sowie die Aufarbeitung von vorhandenen Theorien, empirischen Befunden und allenfalls bereits existierenden Tests“ (Jonkisz u. a., 2012, S. 28) umfasst, schließen sich Betrachtungen an, die bei der Testkonstruktion berücksichtigt werden sollten (vgl. Bühner, 2011, S. 84; Jonkisz u. a., 2012). Es werden Aspekte der Testplanung (Art der Items, Zielgruppe, Fragen der Administration etc.) und Aspekte bezüglich der eigentlichen Testkonstruktion (Konstruktionsstrategien, Itemformate, Konstruktionsregeln, etc.) unterschieden, wobei die beiden Facetten in einem Wechselverhältnis zueinander stehen. Die Konstruktion eines Testes kann offensichtlicherweise von Aspekten der Testplanung wie der Zielgruppe beeinflusst sein, während die Konstruktion in umgekehrter Richtung Auswirkungen auf die weitere Planung, Zusammenstellung oder Administration des Testes haben kann. Folglich geht es nicht um die schrittweise Abarbeitung aller Aspekte, sondern um einen ineinandergreifenden

[5] Die Aussagekraft solcher Skalen wird zumeist sehr kritisch gesehen, da diese Skalen einerseits nicht immun sind gegen das Phänomen, welches sie eigentlich erfassen sollen und andererseits höchstwahrscheinlich konfundiert sind mit anderen Persönlichkeitsmerkmalen (Bühner, 2011, S. 128).

Prozess, der viele der genannten Aspekte integriert. Alle Maßnahmen der Testplanung und Konstruktion haben zum Ziel, eine Messung zu ermöglichen, die möglichst wenig durch konstruktirrelevante Varianz verfälscht ist. Das bedeutet, dass jede Maßnahme oder Entscheidung daraufhin beurteilt werden muss, inwieweit und in welche Richtung der Prozess der Itembeantwortung (Verständnis, Interpretation, Urteilsbildung oder Urteilabgabe) potentiell beeinflusst wird. Die Umsetzung der Leitgedanken wie Sachlichkeit und Verständlichkeit, die sich aus dem Modell der Itembeantwortung ergeben (vgl. Abschnitt 2.2.1), schlagen sich in konkreten Regeln der Itemkonstruktion und Itemzusammenstellung nieder.

Betrachtungen bez. der Merkmalsart Aus den Vorbetrachtungen sollte sich beispielsweise ergeben, ob es sich eher um ein quantitatives oder qualitatives Merkmal handelt, welches erfasst werden soll. Ein quantitatives Merkmal ist dabei ein solches, welches abgestufte Ausprägungskategorien aufweisen kann. Auch über die Struktur des Merkmals sollte Klarheit herrschen, also darüber, ob es sich um ein ein- oder mehrdimensionales Konstrukt handelt. Weiterhin ist von Interesse, ob das interessierende Konstrukt eher als zeitlich veränderlich konzeptionalisiert ist („state“) oder eben nicht („trait“) und somit ggf. Situationsspezifika eine besondere Rolle spielen, die u. U. zu Verzerrungen bei der Itembeantwortung führen können (vgl. z. B. Jonkisz u. a., 2012, S. 28 f.; Kelava & Schermelleh-Engel, 2012).

Betrachtungen bez. der Testart Zusätzlich müssen die Testart (Leistungs- oder Persönlichkeitstest) festgelegt und deren jeweilige Eigenarten berücksichtigt werden. Insbesondere Persönlichkeitstests, die in der Regel Selbstauskünfte der Probanden erfragen, unterliegen einer gewissen Anfälligkeit für Verfälschungen (vgl. Abschnitt 2.2.2). Dieser „Anfälligkeit“ kann durch einer Reihe von Maßnahmen begegnet werden (Standardisierung der Erhebungssituation, Entgegenwirkung klassischer Beurteilungsfehler, Intersubjektivierung der Iteminhalte, Verringerung der Augenscheinvalidität, ...), sodass im besten Fall von einem „objektiven“ Persönlichkeitstest ausgegangen werden kann (vgl. Jonkisz u. a., 2012, S. 29-32).

Betrachtungen bez. des Geltungsbereiches und der Zielgruppe Darüber hinaus müssen Entscheidungen z. B. über den anvisierten Geltungsbereich, d. h. über mögliche Anwendungskontexte des zu entwickelnden Instrumentes getroffen werden. Über den Geltungsbereich wird maßgeblich festgelegt, welche Inhalte durch die Items repräsentiert werden müssen. Der Geltungsbereich wird zumeist durch die intendierte Testwertinterpretation, die zu Beginn einer Testentwicklung/Testvalidierung transparent gemacht wird, beschrieben. Die Festlegung der Zielgruppe oder mehrerer Zielgruppen zieht konkrete Maßnahmen bei der Itemkonstruktion nach sich, da Items dann eindeutig

formuliert werden müssen und zwar in dem Sinne, dass jede Gruppe dasselbe unter einem Item versteht. Merkmale der Zielgruppe (z. B. Alter, Sprache, ...), die für die Testkonstruktion relevant sein können, müssen daher sorgfältig beschrieben werden. Eine Vernachlässigung relevanter Merkmale kann zu einer Benachteiligung/Bevorteilung einzelner (Sub)Gruppen und damit zu einer verzerrten Messung führen, da Itemschwierigkeiten dann für diese Gruppen variieren (vgl. Jonkisz u. a., 2012, S. 32 ff.; Bühner, 2011, S. 87-91).

Betrachtungen bez. der Testzeit und der Testlänge Die Testlänge, d. h. die Anzahl der Items zur Erfassung eines Konstruktes, hängt im Wesentlichen vom Geltungsbereich und der zugrunde liegenden Definition des Konstruktes ab. Je heterogener das Konstrukt aufzufassen ist, desto mehr Items erscheinen für eine valide Erfassung notwendig. Es ist zu beachten, dass eine präzisere Messung mit einer Erhöhung der Itemzahl einher geht. Gedeckelt ist die Testlänge jedoch durch die Testzeit, die in der Realität erheblichen Einfluss auf die Motivation der Befragten und die Praktikabilität des Tests hat. Insofern muss ein Kompromiss zwischen Testlänge und Testzeit gefunden werden (vgl. Jonkisz u. a., 2012, S. 34 f.).

Betrachtungen bez. der Testadministration Es müssen begründete Entscheidungen über das Format und die Form der Testung getroffen werden. Bezüglich des Formats kommen beispielsweise Papier-&-Bleistift-Tests oder Online-Erhebungen in Frage. Zur Form gehören Fragen der Gruppen- oder Einzeltestung sowie der Standardisierung des Testverfahrens (vgl. Jonkisz u. a., 2012, S. 35 f.).

2.2.4. Aspekte der Testkonstruktion

Nachfolgend werden Facetten der eigentlichen Testkonstruktion beschrieben, also Aspekte der Itemkonstruktion und der Zusammenstellung der entwickelten Items zu einem Test unter der Perspektive der Vermeidung konstruktirrelevanter Varianz beleuchtet.

Festlegung/Reflexion der Konstruktionsstrategie In Abhängigkeit des Vorliegens einer mehr oder weniger elaborierten (formalen) Theorie können verschiedene Strategien angewandt werden, um Items zu konstruieren. Ist die theoretische Basis eher schwach, so ist die Vorgehensweise meist intuitiv und erfahrungsbasiert. Bei Vorhandensein einer fundierten theoretischen Grundlage kann die Konstruktion rational erfolgen, d. h. Items können z. B. nach Regeln konzipiert werden, die sich stringent aus der Theorie ableiten lassen. Voraussetzung für die Itemkonstruktion ist naheliegenderweise die theoriegeleitete Definition des Konstruktes und dessen gründliche theoretische Einbettung. In der Praxis werden meist beide Vorgehensweisen miteinander kombiniert. Wichtig ist, die

Vorgehensweise offen zu legen und zu reflektieren. Weitere mögliche Konstruktionsstrategien sind die kriteriumsorientierte, prototypische oder faktorenanalytische (auch: induktive) Konstruktion (vgl. Jonkisz u. a., 2012, S. 36 ff.; Bühner, 2011, S. 92-96; Mummendey & Grau, 2014, S. 63 ff.).

Festlegung der Art des Itemstammes Jedes zu konstruierende Item eines Persönlichkeitstests besteht aus einem Itemstamm, der eine Aussage oder ein Statement umfasst und einem zugehörigen Antwortformat, welches festlegt, wie Probanden auf den Itemstamm reagieren können. Das Statement oder die Aussage kann dabei nach verschiedenen Kriterien formuliert werden, wobei begründet werden sollte, welche Kriterien bei der Itemkonstruktion berücksichtigt werden.

Beispielsweise kann eine Aussage mit dem Ziel konstruiert werden, das interessierende Merkmal möglichst direkt zu erfassen (Beispiel: „Sind Sie fröhlich?“). Bei dieser Vorgehensweise ist zu beachten, dass direkte Aussagen stark unterschiedlich interpretiert werden können und damit eine Intersubjektivierbarkeit nicht immer gegeben ist. Beim indirekten Erfassen würde über das Beschreiben verschiedener Indikatoren auf das interessierende Merkmal geschlossen werden. Weiterhin können Items so formuliert werden, dass sie hypothetische oder biografiebezogene Aussagen beschreiben, wobei hypothetische Items auf Situationen abheben, die Probanden womöglich noch nie erfahren haben, also fiktive Szenen darstellen, während Aussagen letzteren Typs auf Situationen anspielen, die höchstwahrscheinlich von allen Befragungsteilnehmern schon einmal erfahren wurden. Es liegt nahe, dass hypothetisch formulierte Items mit einer höheren Wahrscheinlichkeit Antworten hervorbringen, die auf Fehlurteilen beruhen. Ggf. ist jedoch gerade die Diskrepanz von Selbsteinschätzung und „tatsächlichem“ Verhalten von Bedeutung, was jeweils transparent gemacht werden muss. Darüber hinaus können Items eher konkret oder eher abstrakt formuliert sein, wobei konkretere Items situationsbezogene Aspekte berücksichtigen und daher unter Umständen weniger Interpretationsspielraum zulassen, sodass wiederum von einer höheren Intersubjektivierbarkeit auszugehen ist. Auch ist der Grad der Personalisierung eines Items bei der Konstruktion zu berücksichtigen. Zu persönlich wirkende Items können Fehlantworten provozieren, während Items, die eher unpersönlich formuliert sind, möglicherweise lediglich globale, wenig aussagekräftige Einschätzungen der Befragten erfassen. Auch zu beachten ist die potentielle Stimulusqualität eines formulierten Items, also die Anlage des Items, Emotionen hervorzurufen. Sachlich formulierte Items verringern dabei die Wahrscheinlichkeit einer verzerrten Antwort (vgl. Jonkisz u. a., 2012, S. 62 ff.).

Festlegung des Antwortformates Der Typ eines Items wird neben der Art und Weise der Formulierung des Itemstammes durch das Antwortformat bestimmt. Klassischerweise werden freie, gebundene oder atypische Antwortformate unterschieden. Im Bereich

der Erfassung von Persönlichkeitsmerkmalen spielen gebundene Antwortformate und insbesondere solche eine wichtige Rolle, die eine Beurteilung in Bezug auf das Ausmaß der Zustimmung oder Ablehnung der Aussage des Itemstammes einfordern. In den meisten Fällen werden solche Beurteilungen durch den Einsatz diskreter Ratingskalen realisiert. Der Befragte kann demnach zwischen verschiedenen Antwortkategorien wählen, die bestimmte Abstufungen der Zustimmung/Ablehnung darstellen, weshalb solche Formate auch Stufenantwortaufgaben genannt werden (vgl. Jonkisz u. a., 2012, S. 50). Hinsichtlich des Abstufungsformates müssen die Differenziertheit, Polarität und Bezeichnung berücksichtigt werden (vgl. Bühner, 2011, S. 110-116).

Bezüglich der Differenziertheit ist davon auszugehen, dass mehr als sieben Stufen den Informationsgehalt, der aus Antworten gewonnen werden kann, nicht weiter erhöhen (vgl. z. B. Jonkisz u. a., 2012, S. 51; Bühner, 2011, S. 111). Allerdings führt eine höhere Stufigkeit des Antwortformates (ab 5 Stufen) unter Umständen eher zu einer Ablehnung des Modells im Rahmen konfirmatorischer Faktorenanalysen oder Raschanalysen (vgl. Bühner, 2011, S. 116). Sehr wenige Stufen verhindern dagegen eher das Fällen hinreichend differenzierter Urteile.

Auch ob eine gerade oder ungerade Anzahl von Antwortkategorien präsentiert werden, kann Einfluss auf die Erfassung des Konstruktes haben. Aufgrund der Tendenz zur Mitte bei der Beantwortung (vgl. Abschnitt 2.2.2) wird von der Verwendung neutraler „Mittekategorien“ eher abgeraten, d. h. eine gerade Anzahl von Stufen wird präferiert. Eine zusätzliche „Weiß nicht“-Antwortmöglichkeit ist ggf. im Validierungsprozess sinnvoll, wenn über die inhaltliche Passung zur gewählten Probandengruppe Unklarheit herrscht (vgl. z. B. Jonkisz u. a., 2012, S. 50-54; Bühner, 2011, S. 110 f.). Zusammenfassend erscheinen damit vier oder sechs Stufen ggf. kombiniert mit einer zusätzlichen „Weiß-nicht“-Kategorie eine naheliegende Wahl zu sein. Darüber hinaus kann die Skala bi- oder unipolar gewählt werden. Bei einer unipolaren Skala kann der Grad der Zustimmung oder der Ablehnung nur in eine Richtung anwachsen, während eine bipolare Skala sowohl über ein Maximum der Ablehnung und ein Maximum der Zustimmung als auch dazwischen über einen Nullpunkt verfügt (vgl. z. B. Jonkisz u. a., 2012, S. 51 ff.; Bühner, 2011, S. 111 f.).

Die Bezeichnung der Skalenstufen kann verbal oder numerisch erfolgen. Die Nummerierung der Antwortalternativen legt für die Befragten eine zugrunde liegende Intervallskala nahe, garantiert eine solche Lesart jedoch nicht. Die Nummernzuordnung ist nicht trivial, Auswirkungen beschreiben z. B. Boone, Townsend und Staver (2011). Die Ausformulierung der Bedeutung jeder einzelnen Kategorie führt dazu, dass Probanden weniger Interpretationsarbeit zum Beispiel beim Enkodieren der Kategoriennummern leisten müssen. Die kognitive Entlastung kann zu einer positiveren Einstellung gegenüber der Testbearbeitung führen. Um die erhobenen Daten als intervallskaliert betrachten zu können, sollten verbale Bezeichnungen so gewählt sein, dass sie gleiche Abstände

zwischen den Antwortkategorien nahe legen. In Studien zeigte sich, dass die Bezeichnungen stimmt nicht - stimmt wenig - stimmt mittelmäßig - stimmt ziemlich - stimmt sehr diese Forderung im Rahmen der Zustimmungsabfrage am ehesten abdeckt (vgl. Mummendey & Grau, 2014, S. 79 f.; nach Rohrmann, 1978). In der Praxis wird meist eine Mischform beider Formen der Bezeichnung der Antwortkategorien verwendet, um die Vorteile beider Formate zu nutzen. Um missverständliche Kategorieninterpretationen zu vermeiden, kann gänzlich auf Ratingskalen verzichtet werden und statt dessen die Verwendung dichotomer Items Anwendung finden. Bei diesem Vorgehen kann dann jedoch weniger genau zwischen Gruppen oder Personen differenziert werden. Auch grafische oder symbolhafte Formate finden Anwendung (vgl. z. B. Jonkisz u. a., 2012, S. 52 f.; Bühner, 2011, S. 112-116).

Die Wahl eines Antwortformates muss in jedem Fall unter Berücksichtigung der Passung von Itemstamm und Antwortformat und in Bezug auf die „bestmögliche“ Erfassung des latenten Konstruktes begründet werden. Entscheidungen bezüglich des Itemstammes und des Antwortformates bestimmen nicht nur das „wie“ der Erfassung eines Konstruktes, sondern maßgeblich „was“ erfasst wird und sind daher als konstruktkonstituierend aufzufassen (vgl. Mummendey & Grau, 2014, S. 41 f.).

Beachtung von Itemkonstruktionsregeln Aus den vorangegangenen Überlegungen, der Modellierung der Itembeantwortung und der Prämisse der Vermeidung konstruktirrelevanter Varianz können allgemeine Regeln für die Itemkonstruktion abgeleitet werden und darüber hinaus Regeln, die die Verständlichkeit, die Eindeutigkeit und die Itemschwierigkeit[6] betreffen (vgl. Tabelle 2.1).

Betrachtungen bez. des Testaufbaus Bezüglich des strukturellen Testaufbaus muss die Anordnung und Zusammenstellung der Testinhalte (Subtests) festgelegt und begründet werden. Der Testaufbau hängt in der Regel davon ab, ob ein- oder mehrdimensionale Konstrukte erfasst werden sollen.
Hinsichtlich der Itemreihenfolge und Zusammensetzung des Tests sollten nachfolgende Punkte erwogen werden (vgl. Jonkisz u. a., 2012, S. 68; Mummendey & Grau, 2014, S. 41):

- aufsteigende Schwierigkeit, Eisbrecherfunktion weniger schwieriger Items nutzen,
- Aktualisierungseffekte berücksichtigen (wechselseitige Erschwerung oder Erleichterung der Items),
- Konsistenzeffekte berücksichtigen, Randomisierungen erwägen,

[6] Items müssen prinzipiell Varianz im Antwortverhalten der Probanden erzeugen, d. h. Personen unterschiedlicher Merkmalsausprägung sollten unterschiedliche Antwortkategorien wählen.

- Kontexteffekte berücksichtigen (durch Itemreihenfolgen oder vorangestellte Instruktionen), offene Mitteilung der Testabsicht abwägen.

Auch eher formale Aspekte der Testgestaltung, des Layouts und der Instruktion können, wie bereits geschildert, Einfluss auf die Itembeantwortung nehmen und sollten daher im Rahmen der Testkonstruktion berücksichtigt werden (vgl. Jonkisz u. a., 2012, S. 69):

- Einfachheit, Übersichtlichkeit, Lesbarkeit gewähren,
- zur Mitarbeit anregen,
- Erläuterungen zur Beantwortung geben,
- zur spontanen, wahrheitsgetreuen, vollständigen Beantwortung auffordern,
- Anonymität zusichern und gewähren,
- Kontaktpersonen/Verantwortliche nennen,
- nur notwendige demographische Angaben abfragen,
- ggf. Forschungsinteresse transparent machen.

2.2.5. Itemerprobung, Review und Überarbeitung

Zur Entwicklung eines Instrumentariums zur Erfassung eines interessierenden Konstruktes gehört die Pilotierung bzw. erstmalige Erprobung, deren Ziel im Wesentlichen in der Überprüfung und Verbesserung der Items besteht. Dazu werden beispielsweise Experten zu den Items befragt, statistische Itemanalysen vorgenommen oder Probanden (retrospektiv) bezüglich der Testdurchführung und ihres Antwortverhaltens (Debriefing, lautes Denken, Interview) befragt (vgl. Jonkisz u. a., 2012, S. 70 f.). Unter Itemanalysen werden im Rahmen der klassischen Testtheorie zum Beispiel deskriptivstatistische Analysen der Schwierigkeiten, Varianzen und Trennschärfen der Items verstanden sowie eine Reliabilitätsanalyse[7] (vgl. Kelava & Moosbrugger, 2012; Mummendey & Grau, 2014, S. 97-100). Studien dieser Art sind als Validierungsteilstudien zu bezeichnen (vgl. Abschnitt 2.1).

2.2.6. Besonderheiten der Fragebogenmethode

Zwei Kritikpunkte werden immer wieder angebracht, wenn die Fragebogenmethode zur Erhebung von Persönlichkeitseigenschaften Anwendung findet. Zum einen sind Persönlichkeitsmerkmale durch ihre relative Variabilität, d. h. im Umkehrschluss durch ihre relative Stabilität gekennzeichnet. Zum anderen ergibt sich der scheinbare Widerspruch

[7] Eine Analyse der Trennschärfen wird auch als Teil der Reliabilitätsanalyse verstanden (vgl. Bühner, 2011, Kap. 5)

Tabelle 2.1.: Übersicht über Itemkonstruktionsregeln.

allgemeine Regeln (vgl. z. B. Jonkisz u. a., 2012, S. 67):
• auf Aktualität der Items achten
• Wertungen vermeiden
• emotionale Assoziationen vermeiden
• Suggestionen vermeiden
• Überprüfung der Passung von Itemstamm und Antwortformat
Verständlichkeit (vgl. z. B. Jonkisz u. a., 2012, S. 64 f.; Bühner 2011, S. 134 f.):
• Items positiv formulieren, Negationen vermeiden
• einfache Sätze formulieren, komplizierte Satzgefüge vermeiden
• Abkürzungen und Fachbegriffe vermeiden
• Sprache an Zielgruppe anpassen
• doppelte Verneinungen vermeiden
Eindeutigkeit (vgl. z. B. Jonkisz u. a., 2012, S. 65 f.; Bühner 2011, S. 133 f.):
• Universalausdrücke, Verallgemeinerungen („immer“, ...) vermeiden
• falls notwendig, Definitionen angeben
• Mehrdeutigkeiten vermeiden
• pro Item nur eine Aussage formulieren, sprachl. Verknüpfungen vermeiden
• Items unabhängig vom Vorwissen halten
• Zeitspannen genau definieren
• negative Itempolungen eher vermeiden
• irrelevante oder redundante Formulierungen vermeiden
Itemschwierigkeiten (vgl. z. B. Jonkisz u. a., 2012, S. 66 f.):
• sehr leichte und sehr schwere Items eher vermeiden
• Items unterschiedlicher Schwierigkeit konstruieren

der objektiven Messung sehr subjektiver Konstrukte, die unter Umständen besonders leicht verfälschbar sind. Beide Dilemmata lassen sich nicht gänzlich auflösen, aber doch hinreichend kontrollieren.

Die Subjektivität der Messung[8] muss als erstes als unverrückbare Tatsache *geschätzt*

[8] Es liegt eine Subjektivität in zweifacher Hinsicht vor. Das Individuum gibt Auskunft über etwas, das anderen Personen offensichtlicherweise nicht oder nur in Ausschnitten zugänglich ist. Darüber hinaus kann das Individuum bei der Beurteilung einen subjektiven Maßstab anlegen (vgl. Mummendey & Grau, 2014, S. 165).

werden. Gerade solche Informationen, die aus einer ehrlichen, möglichst unverfälschten Selbstdarstellung resultieren, begründen in den meisten Fällen das Forschungsinteresse. Forschende müssen dabei jeweils einschätzen, inwiefern die erhobenen Daten vertrauenswürdig, also unverfälscht sind. Darüber hinaus obliegt es ihnen, das Nötigste dazu beizutragen, den Messprozess so zu gestalten, dass mit einer hohen Wahrscheinlichkeit „ehrliche“ Antworten zu erwarten sind. Wie die vorangegangenen Abschnitte zeigen, können zum Zweck dieser Objektivierung einige Optionen genutzt werden (z. B. Instruktionen, Standardisierung der Testsituation, Messwiederholung, ...). „Genaue und ehrliche Urteile sind zu erwarten, wenn die befragte Person zu gründlichem Nachdenken motiviert ist, also die Bedeutsamkeit der Untersuchung anerkennt, wenn sie nicht unter Zeitdruck steht oder abgelenkt ist, wenn die Befragung anonym ist bzw. keine individuellen Folgen für die befragte Person zu erwarten sind“ (Mummendey & Grau, 2014, S. 39).

Bei allen Bemühungen um eine Erfassung möglichst zutreffender Selbsteinschätzungen hat die Erhebungssituation jedoch unabhängig vom Grad der Objektivierung *immer* einen Einfluss auf die Messung. „[M]an [geht] heute davon aus, dass in jede psychodiagnostische Messung sowohl konsistente Merkmale der Person als auch inkonsistente Merkmale der Situation einfließen“ (Kelava & Moosbrugger, 2012, S. 364). Diese Problematik zielt auf die relative Stabilität der interessierenden Konstrukte und es stellt sich die Frage, welche Schlüsse man aus einer Messung ziehen kann, die aufgrund der situativen Einflüsse unter Umständen nie replizierbar ist. Doch auch bezüglich dieser Überlegung können Kontrollstrategien aufgezeigt werden, sodass eine Messung mehr Varianz der stabileren, eher überdauernden Anteile eines Konstruktes aufklärt als der situativen.[9] Die Situation setzt sich immer aus äußeren Komponenten (Rahmenbedingungen, Stimmungen, Zweck der Untersuchung, ...) zusammen und aus solchen, die direkt durch die Interaktion mit den Testinhalten kreiert werden. Zunächst müssen sich Testkonstrukteure darüber im Klaren sein, dass auf Interaktionseffekte Einfluss genommen werden kann, da die Formulierung der Items (Abstraktionsgrad, Direktheit, ...) über die Stabilität des zu messenden Merkmals mitbestimmt (vgl. Mummendey & Grau, 2014, S. 35). Darüber hinaus sind bestimmte (standardisierte) Befragungssituationen eher dafür prädestiniert, das Antwortverhalten zu beeinflussen. Dazu zählen beispielsweise Testungen zu einem Zweck, der für die Befragten von besonderer Relevanz ist (Test im Rahmen eines Auswahlverfahrens). Dann kann es zu Interaktionseffekten kommen, beispielsweise zur Verfälschung einzelner Items, die von den Befragten als besonders wichtig erachtet werden, die in der konkreten Testsituation schwer zu kontrollieren sind. Es besteht dann jedoch die Möglichkeit, diese potentiellen Verzerrungen in das Messmodell zu integrieren und damit zu quantifizieren (vgl. Ziegler & Bühner,

[9] Die Umkehrung trifft ebenfalls zu: Unter Umständen ist man gerade an der situativen Ausprägung und nicht am überdauernden Anteil des Konstruktes interessiert.

2009; Kelava & Schermelleh-Engel, 2012).

Zusammenfassend lässt sich sagen, dass „Frageb[ö]gen [...] relativ „objektive“, d. h. weitgehend unabhängig vom Untersucher anwendbare Selbstbeschreibungsverfahren dar[stellen], die Verfahren subjektiver Selbstdarstellung sind. Mit ihnen wird es möglich, die Selbstrepräsentation von Personen und ihre Veränderung unter bestimmten [...] Bedingungen zu erforschen“ (Mummendey & Grau, 2014, S. 19). Mummendey und Grau (2014, S. 18) empfehlen trotzdem aufgrund der Problematik der Sensibilität in Bezug auf die Situativität und Subjektivität, Fragebogenergebnisse nicht zur Individualdiagnostik zu verwenden. Es sei jedoch daran erinnert, dass auch scheinbar „objektive“ Tests, wie Leistungsfeststellungen, ähnlichen Problemen ausgesetzt sind. Erhebungssituationen und damit auch Interaktionseffekte können einen Einfluss auf die Messung ausüben, und inwiefern zu messende Konstrukte tatsächlich stabil sind, muss unter diesem Hintergrund ebenfalls kritisch beleuchtet werden (vgl. Debatte um Kompetenz-/Performanzerfassung in Abschnitt 6.1.4).

2.2.7. Zusammenfassung

Festzuhalten ist: „Die Gestaltung eines Fragebogens beeinflusst [...] die Antworten“ (Mummendey & Grau, 2014, S. 45). Daher ist jede einzelne Entscheidung angefangen von den Itemformulierungen, über das Antwortformat, die Itemzusammenstellung, die Itemreihenfolge, die Erhebungssituation, etc. im Rahmen der Validierung dahingehend zu hinterfragen, welchen Einfluss sie auf die Erhebung des interessierenden Konstruktes und damit die Interpretation der intendierten Testwertinterpretation ausübt.

Ausgehend von einer umfassenden theoretischen Fundierung des interessierenden Konstruktes in den nachfolgenden Kapiteln 3, 4 und 5) und dessen Einordnung in die Forschungslandschaft in Kapitel 6 werden in Kapitel 7 Entscheidungen transparent gemacht, diskutiert und begründet, die eine Vielzahl der zuvor beschriebenen Facetten der Instrumententwicklung betreffen. Um ein übergeordnetes Validitätsargument zu formulieren, wird auf diese Entscheidungen abschließend erneut Bezug genommen (vgl. Kapitel 13).

3. Selbstwirksamkeitserwartungen

„[C]ontrol is central in human lives."

Bandura, 1997, S. 2

In der frühen Menschheitsgeschichte kann das Kontrollieren einer Situation als überlebenswichtig eingeschätzt werden. Aber auch heutzutage kann es im täglichen Leben entscheidend sein, Kontrolle über Gegebenheiten zu haben, beispielsweise beim Autofahren. Umgangssprachlich drückt sich die Bedeutung in Redewendungen und Haltungen aus: So sollte man stets „Herr der Lage sein" und Kontrollverlust wird zumeist als sozial unerwünschtes Verhalten angesehen. Für viele Menschen gilt ganz grundsätzlich: „Uncertainty in important matters is highly unsettling" (Bandura, 1997, S. 2). Dass ein „Gefühl von Kontrolle" (a sense of control) für die geistige Gesundheit bedeutend ist, gilt heute als wissenschaftlich gut abgesicherte Tatsache (vgl. zusammenfassend Skinner, 1996, S. 549).

Mit dem Verändern oder Kontrollieren einer Situation gehen in der Regel das Erreichen gewünschter Ziele (in früheren Zeiten z. B. das Überleben!) und damit Erfahrungen der eigenen Wirksamkeit einher. Das Handeln ist in dem beschriebenen Sinne intentional (Ziele, Absichten, gewünschte Resultate) und das Bewusstmachen der eigenen Wirksamkeit ein Akt von Selbstreflexion. Zusammengenommen führen Handlung und Reflexion zu einer Einschätzung der eigenen Wirksamkeit und damit auch der eigenen Fähigkeiten; den Selbstwirksamkeitserwartungen: „Perceived self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, S. 3).

Für Bandura ist das Hervorbringen intentionaler oder zielgerichteter Handlungen ein Kernaspekt seiner Theorie menschlichen Handelns – der sozial-kognitiven Theorie (vgl. Bandura, 1986, 1997). In dieser nehmen Selbstwirksamkeitserwartungen eine weitere Schlüsselposition ein, da sie neben anderen Faktoren, die handlungsleitend wirken können, als eine Art Trigger fungieren und damit Handlungen überhaupt auslösen oder gänzlich verhindern können. „If people believe they have no power to produce results, they will not attempt to make things happen (Bandura, 1997, S. 3). Selbstwirksamkeitserwartungen haben damit im konzeptionellen Rahmen Banduras einen großen Einfluss darauf, ob und inwiefern der Versuch einer Kontrollausübung unternommen wird.

Diese zunächst grobe Einführung in das Konstrukt der Selbstwirksamkeitserwartungen soll an dieser Stelle für einen ersten Zugang und eine thematische Orientierung genügen.

Ziel des nachfolgenden Kapitels ist es, schrittweise den hier skizzierte ersten Eindruck zu ergänzen, auszuschärfen und zu vervollständigen. In einem ersten Schritt wird dazu eine theoretische Annäherung an das Konstrukt der Selbstwirksamkeitserwartungen versucht, indem die Grundzüge der einbettenden sozial-kognitiven Theorie erläutert werden. In einem zweiten Schritt werden Selbstwirksamkeitserwartungen aus der bereits angedeuteten, übergeordneten Perspektive heraus betrachtet und als „Kontroll-Konstrukt“ aufgefasst. In diesem Zusammenhang werden ein Klassifizierungsschema und ein Begriffsnetz eingeführt, welche letztlich eine bessere Charakterisierung des Konstruktes erlauben. Genutzt werden diese, wenn im Anschluss daran das Konstrukt möglichst präzise definitorisch gefasst werden soll.

Mit Blick auf die in dieser Arbeit angestrebte Instrumententwicklung erscheint es besonders relevant, nicht nur das Konstrukt und seine Eigenschaften gut zu kennen, sondern auch Gemeinsamkeiten und Unterschiede zu anderen (verwandten) Konstrukten klar zu benennen, denn das eigentlich interessierende Konstrukt der Selbstwirksamkeitserwartungen gewinnt durch dieses Vorgehen ebenfalls an Kontur. Deshalb wird in diesem Kapitel auch der Versuch unternommen, Selbstwirksamkeitserwartungen von anderen Konstrukten wie zum Beispiel Handlungsergebniserwartungen, dem Selbstkonzept oder Kontrollüberzeugungen abzugrenzen. Auch konkurrierende Testwertinterpretationen können durch dieses Vorgehen leichter antizipiert und von der eigentlich interessierenden Testwertinterpretation unterschieden werden. Schließlich ergeben sich aus den Darstellungen des Kapitels die theoretische wie praktische Bedeutung des Konstruktes sowie konkrete Hinweise dazu, wie das Konstrukt operationalisiert und gemessen werden kann.

3.1. Mehrperspektivische theoretische Annäherung an das Konstrukt

> „People's beliefs in their efficacy affect almost everthing they do: how they think, motivate themselves, feel, and behave.“
>
> *Bandura, 1997, S. 19*

Nachfolgend werden zuerst die Kernpunkte der sozial-kognitiven Theorie Banduras dargestellt, aus denen sich unter anderem auch die besondere theoretische Bedeutung des Konstruktes der Selbstwirksamkeitserwartungen ableiten lässt (vgl. Abschnitt 3.5). Indem Selbstwirksamkeitserwartungen spezifische Einschätzungen eigener Fähigkeiten adressieren, die im Rahmen der sozial-kognitiven Theorie als zentral bezüglich der Handlungsinitiierung angesehen werden, erlauben sie entsprechend der theoretischen Annahmen ebenfalls eine Abschätzung darüber, ob und inwiefern ein Individuum auf eine Situation potentiell kontrollierend einwirken wird oder nicht. In diesem Sinne

würden Personen mit hoch ausgeprägten Selbstwirksamkeitserwartungen mit größerer Wahrscheinlichkeit solche (kontrollierenden) Handlungen aufnehmen, auf die sich die Selbstwirksamkeitserwartungen beziehen als Personen mit diesbezüglich niedrig ausgeprägten Selbstwirksamkeitserwartungen. Demnach können Selbstwirksamkeitserwartungen als „Kontroll-Konstrukt“ aufgefasst werden, was die Verwendung von Begrifflichkeiten und Konzepten zur Charakterisierung des Konstruktes erlaubt, die nicht der sozial-kognitiven Theorie entstammen, sondern auf den Systematisierungsansatz von Kontroll-Konstrukten Skinners u.a. (1988, 1996) zurückgehen. Insgesamt führt dieser Ansatz zu einem besseren strukturellen Verständnis des Konstruktes der Selbstwirksamkeitserwartungen und wird deshalb nachfolgend eingeführt und erläutert.

3.1.1. Die sozial-kognitive Theorie Banduras

> „People are agentic operators in their live course not just onlooking hosts of brain mechanisms orchestrated by environmental events.“
>
> *Bandura, 1997, S. 5*

Die sozial-kognitive Theorie wird ausführlich bei Bandura (1986, 1989, 1997) beschrieben. An dieser Stelle können jedoch lediglich überblicksartig die Eckpunkte der Theorie erläutert werden. Im Allgemeinen beschreibt die sozial-kognitive Theorie das Zustandekommen menschlichen Handelns sowie menschlicher Kognitionen, Emotionen und Motivationen (vgl. Maddux, 2002, S. 279). Eine Annahme besteht darin, dass Menschen grundsätzlich, wenn auch unter Berücksichtigung zusätzlicher Faktoren, Einfluss auf ihre Handlungen nehmen können: „People can exercise influence over what they do. Most human behavior, of course, is determined by many interacting factors, and so people are contributers to, rather than the sole determiners of, what happens to them“ (Bandura, 1997, S. 3).

Indem Individuen auf ihr Handeln gezielt Einfluss nehmen können, ist das menschliche Handeln als intentionaler Akt gekennzeichnet. Dies bedeutet jedoch nicht, dass anvisierte Ziele tatsächlich durch Handlungen erreicht werden. „Whether the exercise of that agency has beneficial or detrimental effects or produces unintended consequences is another matter“ (Bandura, 1997, S. 3). Ein Individuum ist jedoch aufgrund seiner Selbstreflexivität fähig, die Qualität der ausgeführten Handlung sowie die aus der Handlung resultierenden Konsequenzen zu erkennen und zu bewerten. Es erfährt damit etwas über seine eigene Wirksamkeit. Die hier dargestellte getrennte Betrachtungsweise (Handlungen ausführen vs. Handlungskonsequenzen beurteilen) ist ein zentraler Aspekt der Theorie Banduras, der für die Operationalisierung des Konstruktes konkrete Auswirkungen hat (vgl. Abschnitte 3.2.1 bzw. 4.3).

Die zentrale Annahme der sozial-kognitiven Theorie besteht darin, dass für das

Aufnehmen einer Handlung die Überzeugung in die eigene Wirksamkeit, genauer die Überzeugung in die eigene Fähigkeit bestimmte Handlungen auszuführen (und damit ggf. bestimmte Ziele zu erreichen) von besonderer Bedeutung ist: „Efficacy belief [...] is a major basis of action“ (Bandura, 1997, S. 3). Aufgrund dieser Bedeutungszuschreibung können Selbstwirksamkeitserwartungen als das zentrale Element der sozial-kognitiven Theorie Banduras angesehen werden.

Menschliches Handeln als triadische, reziproke Verursachung Nach Bandura entsteht menschliches Handeln aus einer triadischen, reziproken Interaktion von drei Determinanten. Zu diesen zählen die internalen Faktoren eines Individuums (kognitive, affektive, biologische Ereignisse), das Verhalten des Individuums und die Umwelt (vgl. Abbildung 3.1). Die jeweils bidirektionalen Beziehungen der drei interagierenden Faktoren sind dabei situationsabhängig unterschiedlich stark ausgeprägt (vgl. Bandura, 1997, S. 6). „Determinism does not imply the fatalistic view that people are only pawns of external forces. Reciprocal causation provides people with opportunities to exercise some control over their destinies as well as setting limits on self-direction“ (Bandura, 1997, S. 8).

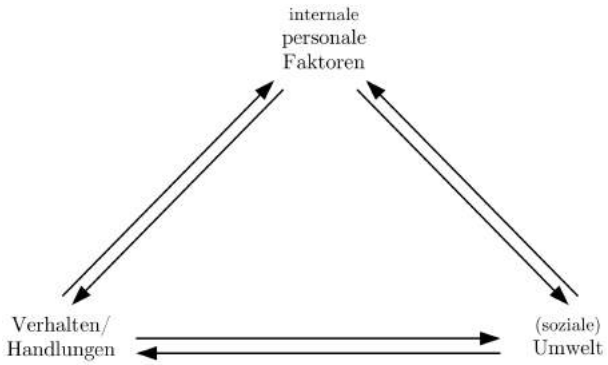


Abbildung 3.1.: Prinzip der triadischen, reziproken Verursachung (nach Bandura, 1997, S. 6).

Entscheidend in diesem Zusammenhang ist die Charakterisierung menschlichen Handelns als proaktiv und generierend, weshalb einzelne Handlungen auch nicht auf eine dualistische Perspektive (Selbst vs. Umwelt) zurückgeführt werden können: „Human behavior cannot be fully understood solely in terms of either social structural factors or psychological factors. A full understanding requires an integrated causal perspective in which social influences operate through self-processes that produce the actions“ (Bandura, 1997, S. 6).

Zusammenfassung Banduras sozial-kognitive Theorie orientiert sich an den folgenden zentralen Grundannahmen:

1. Der Mensch kann sein Handeln beeinflussen.
2. Jedes Handeln wird durch (soziale) Rahmenbedingungen und individuelle Kognitionen beeinflusst.
3. Der Mensch handelt proaktiv und kreativ.
4. Das Handeln erfolgt i. d. R. intentional.
5. Das Handeln muss getrennt von (beabsichtigten) Konsequenzen analysiert werden.
6. Selbstwirksamkeitserwartungen sind für die Handlungsinitiative zentral.

3.1.2. Exkurs: Philosophische Randnotizen

Die Annahmen Banduras sozial-kognitiver Theorie tangieren unweigerlich philosophische Fragestellungen, die hier schlaglichtartig angedeutet werden.

Reduktionismusdebatte Nach Bandura sind Selbstwirksamkeitserwartungen Kognitionen/mentale Zustände und beruhen damit unweigerlich auf spezifischer Gehirnaktivität. Selbstwirksamkeitserwartungen sind jedoch ontologisch nicht mit Gehirnaktivität gleichzusetzen bzw. auf Gehirnaktivität reduzierbar. Sie sind mehr als ein neurobiologisches Aktivitätsmuster – sie können als mentale Zustände, d. h. als emergente Eigenschaft spezifischer Gehirnaktivitäten aufgefasst werden (vgl. Bandura, 1997, S. 4). Darüber hinaus beschreibt Bandura das Verhältnis psychologischer und neurobiologischer Theorien im Rahmen der Reduktionismusdiskussion, indem er dasselbe Argument der „Nichtreduzierbarkeit“ anwendet: „The fact that cognition is a cerebral occurence does not mean that the laws expressing functional relations in psychological theory are reducible to those in neurophysiological theory. One must distinguish between how cerebral systems function and the personal and social means by which they can be orchestrated to produce courses of action that serve different purposes“ (Bandura, 1997, S. 4). Eine rein physikalistische Sichtweise hat darüber hinaus damit Probleme, den determinierenden Einflusscharakter bestimmter mentaler Zustände sowie das Entstehen ganz neuer Gedanken zu erklären.

Determinismus Die Analyse von Handlungen im Rahmen der sozial-kognitiven Theorie berührt die fundamentale Frage nach der Determiniertheit oder Freiheit des menschlichen Handelns (vgl. Bandura, 1997, S. 7). Für Bandura schließen sich Determinismus und Freiheit nicht aus, sondern befruchten sich in friedlicher Koexistenz. In dem Sinne verwendet Bandura den Determinismus-Begriff nicht in einer strengen Form. „The term *determinism* is used here to signify the production of effects by events rather than in the doctrinal sense [...]“ (Bandura, 1997, S. 7). Bestimmte Restriktionen führen erst dazu, Handlungsoptionen abwägen zu können. Selbstregulative Fähigkeiten spielen damit eine

große Rolle, denn der Mensch muss situationale Gegebenheiten auf sich beziehen und Folgerungen ziehen können. Bandura geht davon aus, dass mit stärker ausgeprägten selbstregulativen Fähigkeiten eher „freie“ Handlungsentscheidungen getroffen werden können. Zu diesen Fähigkeiten zählen neben dem Abwägen von Alternativen und der Selbstmotivation das Vorhersehen und Gewichten möglicher Konsequenzen sowie die Einschätzung eigener Fähigkeiten (vgl. Bandura, 1997, S. 7).

Insofern folgt in Banduras Sinne der objektiven Einschränkung durch bestimmte (soziale) Rahmenbedingungen nicht unbedingt eine Einschränkung der persönlichen Freiheit. Schließlich kann eine Situation noch so zwingend sein, Menschen können trotzdem Handlungen unterlassen und sich beispielsweise verweigern. Bei solchen Abwägungen spielen jedoch auch Prinzipien wie selbst gesetzte Standards und damit einhergehend das Selbstwertgefühl eine große Rolle. Einige Handlungen werden also wahrscheinlicher, andere unwahrscheinlicher. Trotzdem kann eine Handlung nie mit absoluter Gewissheit vorhergesagt werden (vgl. Bandura, 1997, S. 6). Der Grund dafür liegt laut Bandura in der Fähigkeit des Menschen begründet, proaktiv und kreativ zu sein, also Neues hervorbringen zu können (vgl. Bandura, 1997, S. 4 f.). Der Mensch gestaltet und wird gestaltet und das wahrscheinlich sogar gleichzeitig. „In agentic transactions, people are both producers and products of social systems“ (Bandura, 1997, S. 6).

3.1.3. Selbstwirksamkeitserwartungen als Kontroll-Konstrukt

Nachfolgend soll eine weitere Perspektive genutzt werden, sich auf einer theoretischen Ebene dem Konstrukt der Selbstwirksamkeitserwartungen anzunähern. Dazu ist es nötig, Selbstwirksamkeitserwartungen oder „self-efficacy beliefs“ als „Kontroll-Konstrukt“ zu konzeptionalisieren (vgl. Skinner, 1996). Dies ist deshalb möglich, weil im Rahmen der sozial-kognitiven Theorie Selbstwirksamkeitserwartungen als entscheidend in Bezug auf die Handlungsinitiative und damit die Ausübung von Kontrolle angesehen werden. Bandura selbst charakterisiert sein Konstrukt in diesem Sinne, was zum Beispiel an dem gewählten Titel „Self-Efficacy: The Exercise of Control“ seines Hauptwerkes sichtbar wird (vgl. Bandura, 1997).

Auch Konstrukte, die durch die sozial-kognitive Theorie eng mit Selbstwirksamkeitserwartungen verknüpft sind (z. B. Handlungsergebniserwartungen/„outcome-expectancy“), werden den Kontroll-Konstrukten zugeordnet (vgl. Skinner, 1996). Eine sehr ergiebige Möglichkeit, Kontroll-Konstrukte voneinander zu unterscheiden, stellt Skinner (1996, S. 549) vor.[1] Vor dem Hintergrund der angestrebten präzisen definitorischen Fassung des Konstruktes der Selbstwirksamkeitserwartungen erscheint es relevant, diese Unterscheidungsmöglichkeit inkl. der verwendeten Begrifflichkeiten nachfolgend detaillierter vorzustellen.

[1] Für eine Kritik des Ansatzes vgl. Bandura (1997, S. 27 ff.).

Das Charakterisierungs- und Unterscheidungstool von Skinner (1996) kann unter anderem dann angewandt werden, wenn domänenspezifische Adaptionen auf ihre Passung zur sozial-kognitiven Theorie untersucht werden müssen. Beispielsweise existieren mittlerweile im Bereich der Lehrerbildung unter dem Stichwort der „Lehrer-Selbstwirksamkeitserwartungen“ eine Vielzahl von Adaptionsversuchen. Ob damit jedoch jeweils Selbstwirksamkeitserwartungen im Bandura'schen Sinne erfasst werden, kann mit Hilfe des nachfolgend beschriebenen Ansatzes nach Skinner (1996), wie gesagt, zielführender untersucht werden (vgl. Abschnitt 4).

Skinners Klassifizierungssystem für Kontroll-Konstrukte

Die folgenden Ausführungen beziehen sich, soweit nicht anders gekennzeichnet, auf den Artikel „A Guide to Constructs of Control“ von Skinner (1996). Insgesamt arbeitet die Autorin vier Unterscheidungsansätze heraus:

1. Unterscheidung danach, ob im objektiven („objective control“) oder subjektiven Sinne („subjective control“) Kontrolle vorliegt/ausgeübt werden kann oder ob Kontrollerfahrungen („experiences of control“) gemacht werden.
2. Unterscheidung von Akteur („agent“), Mittel/Methode („means“) sowie Ziel/ Zweck („ends“) der Kontrollausübung.
3. Unterscheidung von retro- und prospektiver Kontrolle.
4. Unterscheidung von spezifischer und genereller Kontrolle.

Die beiden letzteren Unterscheidungsmöglichkeiten sind für die Klassifizierung jedoch eher ungeeignet, denn „any kind of belief can appear as either retrospective or prospective and at any level of generality“ (Skinner, 1996, S. 555), weshalb nachfolgend die ersten beiden Ansätze erläutert werden.

Die Unterscheidung in tatsächlich, d. h. im objektiven Sinne vorliegende Kontrolle und selbst wahrgenommene Kontrolle ist deshalb wichtig, da das psychische Wohlbefinden sowie Handlungsinitiativen und/oder Erregungszustände durch subjektiv wahrgenommenes Kontrollempfinden beeinflusst werden (vgl. Skinner, 1996, S. 551). Von besonderer Bedeutung sind für Personen Kontrollerfahrungen („experiences of control“), das heißt solche Situationen, in denen eine Interaktion mit der Umwelt zu einem gewünschten Erfolg führt und das eigene Handeln als wirksam interpretiert wird. Darüber hinaus beeinflussen Kontrollerfahrungen nachweislich die subjektive Wahrnehmung des eigenen Kontrollvermögens („subjective control“) positiv (vgl. Skinner, 1996, S. 551).

Der zweite Unterscheidungsansatz beruht darauf, dass in der Regel Kontroll-Konstrukte derart operationalisiert werden, dass die jeweiligen Akteure oder Individuen, die Kontrolle ausüben („agents“), die (erwünschten oder unerwünschten) Ziele oder Konsequenzen, die durch Kontrollausübung erreicht werden („ends“), sowie die Mittel

und Wege, die zur Kontrollausübung genutzt werden („means“), in Relation zueinander gesetzt werden.[2] Das prototypische, intuitive Kontroll-Konstrukt rekurriert auf das Individuum als „agent“, dessen Handlungen als „means“ und Veränderungen in der (sozialen) Umwelt aufgrund der individuellen Handlungen als „ends“ (vgl. Skinner, 1996, S. 558). Grundsätzlich sind jedoch auch andere Konstellationen denkbar. Typische „means“ können neben Handlungen (Verhalten, Reaktionen, etc.) zum Beispiel Kognitionen oder Attribute (Fähigkeit, Persönlichkeit, Gene, etc.), aber auch externale Faktoren wie Glück, Zufall oder Schicksal sein. Typische „ends“ können als Resultate, Konsequenzen, Reaktionen, Prozesse, Performanz oder Emotionen konzeptualisiert sein (vgl. tabellarische Übersicht bei Skinner, 1996, S. 554).

In der Regel finden sich in den Operationalisierungen der Kontroll-Konstrukte drei mögliche Kombinationen der drei Faktoren „agent“, „means“ und „ends“ wieder und zwar in Form von „agent-ends“-, „agent-means“- und „means-ends“-Relationen (vgl. Abbildung 3.2). Das oben bereits eingeführte Konstrukt des subjektiven Kontrollempfindens („subjective control“) stellt dabei klassischerweise eine „agent-ends“-Relation dar: „Connections between people and outcomes prescribe the prototypical definitions of control. In general, control refers to the extent to which an agent can intentionally produce desired outcomes and prevent undesired ones [...]. When individuals believe they can do this, they are said to have personal control, perceived control, or a sense of control“ (Skinner, 1996, S. 554). Die beiden verbleibenden Relationen werden von Skinner (1996) wie folgt definiert:

- „Means-ends relations, whether perceived or objective, refer to the connection between particular classes of potential causes and desired and undesired outcomes“ (Skinner, 1996, S. 552).
- „Perceived or objective agent-means relations refer to the extent to which a potential means is available to a particular agent“ (Skinner, 1996, S. 553).

Beachtet werden muss, dass die Unterscheidung von „agent-ends“- und „means-ends“-Relationen teilweise schwierig ist, weil in Konstrukten, die als „means-ends“-Relation konzipiert sind, immer auch ein Kontrolle ausübender Akteur eine Rolle spielt. Umgekehrt kann in klassischen „agent-ends“-Konstrukten zumindest implizit auch ein „means“ enthalten sein. Es muss daher für jedes Konstrukt separat beurteilt werden, ob in Bezug auf das Erreichen eines Ziels/Resultates („ends“) der Fokus eher auf dem Individuum („agent-ends“) oder dem Handeln des Individuums („means-ends“) liegt. Unabhängig davon ist es für die vorliegende Arbeit und bezüglich des anvisierten Ziels der möglichst genauen Charakterisierung des Konstruktes der Selbstwirksamkeitserwartungen von

[2] „Ends refer to the desired and undesired outcomes over which control is exerted, agents refer to the individuals or groups who exert control, and means refer to the pathways through which control is exerted“ (vgl. Skinner, 1996, S. 552).

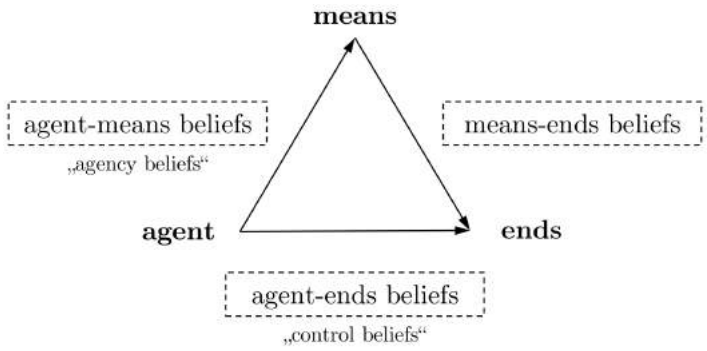


Abbildung 3.2.: Schema der möglichen Relationen zwischen „agent“, „means“ und „ends“ (nach Skinner, Chapman & Baltes, 1988, S. 118).

besonderer Relevanz, dass Selbstwirksamkeitserwartungen oder „self-efficacy beliefs“ mit Hilfe der eingeführten Terminologie und entsprechend der ersten Unterscheidungsmöglichkeit zu den subjektiven Kontroll-Konstrukten gehören und entsprechend der zweiten Unterscheidungsmöglichkeit als *„agent-means“-Relation* aufgefasst werden müssen (vgl. Abschnitt 3.2 und Skinner, 1996, S. 553).

Zusammenfassend folgt aus Skinners Ansatz, dass für das Vorliegen subjektiver oder objektiver Kontrolle zwei Bedingungen erfüllt sein müssen: Erstens muss das Individuum über die nötigen Mittel („means“) verfügen (kompetent sein) und zweitens muss entsprechend der individuellen Einschätzung mindestens eines der Mittel („means“) zu einem bestimmten Ziel („ends“) führen (Passung Mittel/Ziel). Diese beiden Bedingungen werden in der Literatur häufig auch als Kompetenz- („competence“) und Kontingenzbedingung („contingency“) verstanden, wobei erstere eine „agent-means“- und letztere eine „means-ends“-Relation darstellt (vgl. Skinner, 1996, S. 559).

3.2. Analyse der Definition: Eigenschaften und Struktur von Selbstwirksamkeitserwartungen

Nachfolgend wird die eingangs bereits dargestellte Definition des Konstruktes (vgl. Einleitung zu Abschnitt 3 auf Seite 41) eingehender betrachtet. Es wird sich zeigen, dass das Konstrukt allein durch seine Definition nicht in vollem Umfang verständlich wird, ja sogar fehlinterpretiert werden kann. Im Besonderen wird durch die bloße Betrachtung der Definition eine Skinner'sche „means-ends“-Relation nahegelegt, die mit den theoretischen Prämissen Banduras jedoch nicht kompatibel ist. Daher müssen ergänzende theoretische Überlegungen Banduras angeführt werden, um Selbstwirksamkeitserwartungen entsprechend den Maßgaben der sozial-kognitiven Theorie tatsächlich als „agent-means“-Relation zu charakterisieren.

In einem zweiten Schritt können sodann weitere konstruktkonstituierende Merkmale abgeleitet werden. Zudem ergeben sich aus den Darstellungen einige strukturelle Eigenschaften des Konstruktes, die am Ende dieses Abschnitts aufgegriffen werden.

3.2.1. Die Definition des Konstruktes

Zunächst wird die wohl am weitesten verbreitete bzw. typischerweise zitierte Version der Definition von Selbstwirksamkeitserwartungen angegeben:

> „Perceived self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments" (Bandura, 1997, S. 3).

Diese kann wie folgt übersetzt werden: Selbstwirksamkeitserwartung (wahrgenommene Selbstwirksamkeit) bezieht sich auf den Glauben (die Überzeugung, die Meinung, die Erwartung, das Zutrauen) in die eigenen Fähigkeiten, bestimmte Handlungen planen oder durchführen zu können, um gesetzte Ziele zu erreichen (Leistungen zu erbringen). Analysiert man nun die Definition nach „agent", „means" und „ends" im Sinne Skinners (1996), so ergeben sich entsprechend der Darstellungen in Unterabschnitt 3.1.3 folgende Zuordnungen: Als Akteur wird das Individuum/die urteilende Person („agent": „beliefs in one's") benannt. Die konkreten Handlungen der Person können anhand der Definition als Kontrollmöglichkeit („means": „courses of action") und das Zielerreichen oder die Leistungserbringung als Konsequenzen der Kontrollhandlung („ends": „produce given attainment") interpretiert werden.

Da das Augenmerk eher auf den vom Individuum einzuschätzenden Fähigkeiten zu handeln („means") und damit nicht allein auf dem Individuum („agent") liegt, müsste das Konstrukt *aufgrund der Definition* folgerichtig als „means-ends"-Relation klassifiziert werden. Wie bereits in Unterabschnitt 3.1.3 erwähnt, handelt es sich bei dem Konstrukt der Selbstwirksamkeitserwartungen jedoch um eine „agent-means"-Relation. Zu dieser Einschätzung kann man allerdings nur gelangen, wenn man alternative, wenn auch weniger häufig zitierte Definitionen Banduras und/oder einige zusätzliche Informationen aus der Theorie Banduras heranzieht.

Aus der folgenden Definition des Konstruktes wird die „agent-means"-Relation zum Beispiel viel deutlicher: „Perceived self-efficacy is a judgement of one's ability to organize and execute given types of performances" (Bandura, 1997, S. 21) – das Individuum („agent") schätzt sich bezüglich seiner spezifischen Fähigkeit, Handlungen zu organisieren oder durchzuführen („means") selbst ein. Auch die folgende Definition betont die durch das Konstrukt dargestellte „agent-means"-Beziehung: „Perceived self-efficacy is defined as people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" (Bandura, 1986, S. 391).

Dass Selbstwirksamkeitserwartungen *keine* „means-ends“-Relation darstellen, folgt zudem explizit aus Banduras Prämisse der Unterscheidung von Selbstwirksamkeits- und Handlungsergebniserwartungen („outcome expectancies“), wobei letztere wie folgt definiert sind: „An outcome expectancy is defined as a person's estimate that a given behavior will lead to certain outcomes“ (Bandura, 1977a, S. 193). Damit ist das Konstrukt der Handlungsergebniserwartung als „means-ends“-Relation zu charakterisieren, denn ein bestimmtes Verhalten/bestimmte Handlungen („means“) werden mit bestimmten Ergebnissen oder Resultaten („ends“) in Verbindung gebracht. Selbstwirksamkeitserwartungen beziehen sich dagegen explizit und ausschließlich auf das eigene Handeln und nicht auf die Konsequenzen der eigenen Handlungen (vgl. Bandura, 1977a, S. 193). Sie stellen damit eine „agent-means“-Beziehung dar. Der Unterschied beider Konstrukte wird in Abbildung 3.3 dargestellt.

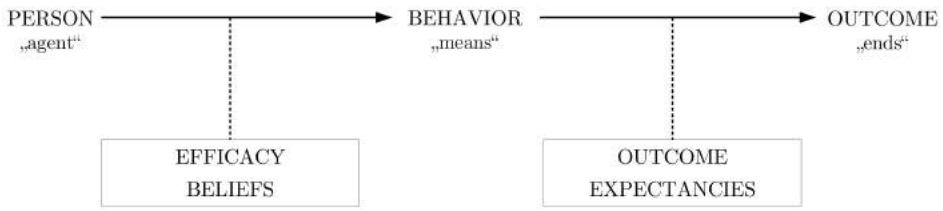


Abbildung 3.3.: Unterschied zwischen Selbstwirksamkeitserwartungen („efficacy beliefs“) und Handlungsergebniserwartungen („outcome expectancies“) nach Bandura (1977a, S. 193) unter Berücksichtigung der Klassifikation nach Skinner (1996).

Werden beide Einschätzungen, also die Selbstwirksamkeitserwartungen und die Handlungsergebniserwartungen eines Individuums zusammen betrachtet, so lässt sich nach Bandura (1997, S. 20) das Verhalten dieses Individuums und damit die konkrete Kontrollausübung am besten voraussagen. Im Sinne Skinners (1996) sind durch das Wechselspiel der beiden Konstrukte die Kompetenz- und Kontingenzbedingung zur Erklärung von „Kontrolle“ erfüllt, denn Selbstwirksamkeitserwartungen adressieren die Einschätzung des Individuums, über die nötigen Mittel der Kontrollausübung zu verfügen, während Handlungsergebniserwartungen die Einschätzung des Individuums bezüglich der Richtung der Kontrollausübung beinhalten (vgl. Unterabschnitt 3.1.3).

Im Rahmen der sozial-kognitiven Theorie werden die Beziehung von Selbstwirksamkeits- und Handlungsergebniserwartungen detailliert beschrieben und deren unterschiedlichen Rollen erläutert (vgl. Bandura, 1997, S. 19-24). Zunächst wird, wie in Abbildung 3.3 dargestellt, eine kausale Beziehung beider Konstrukte postuliert, wobei Selbstwirksamkeitserwartungen den Handlungsergebniserwartungen vorausgehen. Bandura (1997, S. 21) nimmt also an, dass die Einschätzungen wahrscheinlicher Resultate eigener Handlungen wesentlich von der Beurteilung der eigenen Fähigkeiten bezüglich dieser Handlungen abhängen (Abhängigkeit der Handlungsergebniserwartungen von

Selbstwirksamkeitserwartungen), und begründet dies damit, dass in der Regel Ergebnisse von der Güte der Durchführung der Handlungen bestimmt werden.[3]

Darüber hinaus wird dem Konstrukt der Selbstwirksamkeitserwartungen der größere Einfluss auf die Handlungsinitiative zugeschrieben (vgl. Tabelle 3.1, Fall 3). Dies wird im Rahmen der Theorie wie folgt begründet: „[I]ndividuals can believe that a particular course of action will produce certain outcomes [Handlungsergebniserwartung], but if they entertain serious doubts about whether they can perform the necessary activities [Selbstwirksamkeitserwartung] such information does not influence their behavior“ (Bandura, 1977a, S. 193, Ergänzungen durch C. M.). Hoch ausgeprägte Selbstwirksamkeitserwartungen werden damit als notwendige Bedingung für die Handlungsaufnahme charakterisiert. Im Gegensatz dazu werden Handlungsergebniserwartungen nicht als einflussreich genug angesehen, um eine Handlungsinitiative grundsätzlich zu verhindern (vgl. Tabelle 3.1, Fall 2). Dies wird einsichtig, wenn man berücksichtigt, dass mit der Abschätzung wahrscheinlicher Konsequenzen eigener Handlungen eine zusätzliche Bewertung einhergeht, denn eine Person beurteilt gleichzeitig, ob die potentiellen Handlungsresultate erwünscht, wichtig oder nützlich sind. Eine weitere Grundannahme der sozial-kognitiven Theorie besteht nun darin, dass Handlungsergebniserwartungen, die auf ein unerwünschtes oder als unwichtig eingestuftes Resultat abheben, nicht unbedingt für die Handlungsinitiative ausschlaggebend sind, auch wenn hoch selbstwirksame Menschen in diesem Fall eher unter Unzufriedenheit leiden: „Efficacious individuals who cannot gain *valued* outcomes through personal accomplishments will not necessarily cease trying“ (Bandura, 1997, S. 21, Hervorhebung durch C. M.). Dieses Postulat manifestiert ebenfalls die herausgehobene Stellung von Selbstwirksamkeitserwartungen gegenüber Handlungsergebniserwartungen. Für diese beschriebene theoretische Annahme gibt es mittlerweile eine Reihe empirischer Belege (vgl. Bandura, 1997, S. 24).

Zusammenfassend gilt, dass das Konstrukt der Selbstwirksamkeitserwartungen als „agent-means-Relation“ sowie in Abgrenzung zum Konstrukt der Handlungsergebniserwartungen, welches als „means-ends“-Relation konzipiert ist, verstanden werden muss. Das Zusammenspiel beider Konstrukte erklärt dabei unter Kontrolle zusätzlicher Faktoren, wie den tatsächlichen Fähigkeiten oder bestimmter Anreize, das Verhalten eines Individuums bestmöglich, wobei Selbstwirksamkeitserwartungen als ausschlaggebend

[3] Bandura (1997, S. 21) gibt das folgende Beispiel an: „People do not judge that they will drown if they jump in deep water and then infer that they must be poor swimmers. Rather, people who judge themselves to be poor swimmers will visualize themselfes drowning, if they jump in deep water.“ Wichtig ist zu betonen, dass der kausale Zusammenhang beider Konstrukte nur dann Sinn macht, wenn sowohl Selbstwirksamkeitserwartungen als auch Handlungsergebniserwartungen als selbstreferentielle Konstrukte interpretiert werden. Werden Handlungsergebniserwartungen dagegen abstrakter, d. h. nicht auf das eigene Verhalten bezogen, definiert, so ist auch eine andere kausale Reihenfolge denkbar. Man könnte beispielsweise annehmen, dass in der Regel ein bestimmtes menschliches Verhalten zu bestimmten Resultaten führt (Ausdauerndes Schwimmen führt dazu, den See zu überqueren.) und sich dann fragen, ob man selbst in der Lage wäre, sich entsprechend zu Verhalten (Kann ich lang genug schwimmen, um den See zu überqueren?).

Tabelle 3.1.: Mögliche Kombinationen der Ausprägung von Selbstwirksamkeits- und Handlungsergebniserwartung und deren Einfluss auf die Handlungsinitiative (vgl. Bandura, 1997, S. 20).

Fall	Selbstwirksamkeits-erwartung	Handlungsergebnis-erwartung	Handlungs-initiative	psychische & emotionale Begleiterscheinungen
1	+	+	+	Engagement, Zufriedenheit
2	+	–	(+)	Protest, Beschwerde, Aktionismus
3	–	+	–	Mutlosigkeit, Selbstabwertung
4	–	–	–	Resignation, Apathie

$^{+}$ ausgeprägt $^{-}$ nicht ausgeprägt
Anmerkung: Da Handlungsergebniserwartungen immer „ausgeprägt sind“, weil zu einem Verhalten immer ein Resultat assoziiert wird, muss +/– daher als Ausprägung im erwünschten (+) oder unerwünschten (–) Sinn interpretiert werden.

für die Handlungsinitiative angesehen werden: „Given appropriate skills and adequate incentives, however, efficacy expectations are a major determinant of people's choice of activities, how much effort they will expend, and of how long they will sustain effort in dealing with stressful situations“ (Bandura, 1977a, S. 194).

Warum das „ends“ in der häufig zitierten Version der Definition steht. Es bleibt die Frage, warum Bandura die Definition so anlegt, dass diese auf „outcomes“ abzielt, also eine „means-ends“-Relation nahelegt und damit unter Umständen leicht fehlinterpretiert werden kann. Bandura selbst erkennt das Problem, dass „attainments“ als „outcomes“ interpretiert werden können. Er stellt klar, dass beide Begriffe voneinander unterschieden werden müssen und legt zunächst fest: „[A]n outcome is a consequence of a performance, not the performance itself“ (Bandura, 1997, S. 22). „Attainments“ müssen daher im Sinne von „Leistungen“, genauer als graduelle Abstufungen einer Handlung („performance levels“) interpretiert und übersetzt werden (vgl. Bandura, 1997, S. 22).

Außerdem genügen sowohl das Handeln als auch auf das Handeln bezogene Selbsteinschätzungen in der Regel keinem Selbstzweck – man fragt sich nicht ständig, ob man diese oder jene Handlung ausüben kann und führt auch keine Handlungen zum Selbstzweck aus. In der Regel spielen Selbstwirksamkeitserwartungen dann eine Rolle, wenn die zu beurteilende Handlung in einem kausalen Zusammenhang mit einem Ziel (Handlungsergebniserwartung) betrachtet wird. Dies verdeutlicht auch die symbiotische Beziehung der Konstrukte der Selbstwirksamkeits- und Handlungsergebniserwartungen im Rahmen der sozial-kognitiven Theorie, die in der eingangs zitierten Definition (vgl. Seite 50) anklingt und ggf. zu Missverständnissen führen kann.

3.2.2. Weitere Eigenschaften des Konstruktes, die sich aus der Definition und der theoretischen Einbettung ergeben

Wie oben ausführlich dargelegt, ist die „agent-means"-Relation ein wesentliches Merkmal des Konstruktes der Selbstwirksamkeitserwartungen. Aus der Definition (vgl. Seite 50) lassen sich jedoch zwei weitere Merkmale des Konstruktes direkt ableiten, und zwar die Subjektivität/Selbstreferentialität und der Handlungsbezug. Drei weitere Eigenschaften ergeben sich indirekt aus dem Handlungsbezug (Schwierigkeitsbezug, Situationsspezifität/Kontextabhängigkeit, Domänenspezifität). Alle Eigenschaften werden in Abbildung 3.4 visualisiert.

Die Eigenschaft der Subjektivität/Selbstreferentialität ergibt sich direkt aus der Konstruktion als *Selbst*einschätzung *eigener* Fähigkeiten bzw. daraus, dass eben als Akteur die Person selbst fungiert und sich zusätzlich auf sich selbst (seine Handlungen) bezieht. Es ist dabei wichtig zu betonen, dass es sich gerade *nicht* um eine (subjektive) Bewertung anderer Personen und deren Handeln oder Können etc. handelt.

Zusätzlich ist die Selbsteinschätzung per definitionem auf (konkrete) Handlungen („courses of action") bezogen, woraus sich die Eigenschaft des Handlungsbezugs ergibt, die in Kombination mit der Selbstreferentialität auch die Grundlage der Charakterisierung als „agent-means"-Relation darstellt und nachfolgend präzisiert wird. Aus der Theorie zum Konstrukt folgt nämlich, dass sich Selbstwirksamkeitserwartungen nicht auf einzelne, u. U. triviale basale Fähigkeiten („motor acts") beziehen, weil eine solche Einschätzung nicht von Belang wäre. Vielmehr handelt es sich um eine Einschätzung eigener Fähigkeiten bezüglich einer komplexen Handlung und zwar in dem Sinne, dass die Orchestrierung mehrerer Fähigkeiten/Handlungen in einer spezifischen Situationen im Vordergrund steht (vgl. Bandura, 1997, S. 36 f.). Als Beispiel führt Bandura (1997, S. 837 f.) das Autofahren an. Es geht eben nicht darum einzuschätzen, ob man sich in der Lage sieht, die Bremse zu betätigen oder das Lenkrad nach links oder rechts zu drehen, sondern beispielsweise um das sichere Manövrieren eines Autos von A nach B im Feierabendverkehr der Stadt X.

Komplexe Handlungen sind dabei nicht notwendigerweise schwierig. Wenn das Autofahren im Feierabendverkehr hinreichend oft durchgeführt wurde, ist davon auszugehen, dass dieses komplexe Verhalten eher als leicht eingeschätzt wird. Einige Autoren weisen in diesem Zusammenhang darauf hin, dass sich die Selbsteinschätzung deshalb jeweils auf eine herausfordernde, d. h. Anstrengung und Ausdauer erfordernde Situation beziehen muss (vgl. Maddux, 2002, S. 278; Schwarzer & Jerusalem, 2002, S. 35). Fragen nach der Selbstwirksamkeitserwartung würden sonst keinen Sinn ergeben (Schmitz & Schwarzer, 2000, S. 12), weil diese dann „triviale" Selbsteinschätzungen adressieren würden, sodass beispielsweise nicht zwischen Personen mit hohen und niedrigen Selbstwirksamkeitserwartungen unterschieden werden könnte: „If there are no obstacles to overcome, the activity is easily performable and everyone is highly efficacious" (Bandura,

2006, S. 311). Bei der Messung des Konstruktes wäre folglich mit Deckeneffekten zu rechnen (vgl. Bandura, 1997, S. 42 f.). Schmitz und Schwarzer (2000, S. 12) stärken deshalb die Eigenschaft des Schwierigkeitsbezugs in ihrer Definition des Konstruktes: „Unter Selbstwirksamkeitserwartung versteht man die subjektive Gewi[ss]heit, eine neue oder schwierige Aufgabe auch dann erfolgreich bearbeiten zu können, wenn sich Widerstände in den Weg stellen."

Eine weitere Eigenschaft des Konstruktes, die aus der Handlungsorientierung resultiert, ist die Situations- bzw. Kontextspezifität. Grundsätzlich können Handlungen auf unterschiedlichen Spezifitätsniveaus beschrieben werden, wobei über das Ausmaß der Situations- und Kontextdarstellung das Spezifitätsniveau geregelt werden kann (Kann ich im Allgemeinen Autofahren? Kann ich einen Ferrari auf dem Übungsgelände XY bei Regen so schnell wie möglich durch den Slalomparcours fahren ohne die aufgestellten Hütchen umzufahren?). Bandura (1997, S. 40) betont, dass es sich bei Selbstwirksamkeitserwartungen nicht um eine globale Persönlichkeitsvariable handelt und plädiert deshalb für eine situationsbezogene Erfassung des Konstruktes. Welches Spezifitätsniveau im Spannungsfeld von Prädiktivität (sehr konkret) und Verallgemeinerbarkeit (eher global) ein optimales Maß für die Operationalisierung von Selbstwirksamkeitserwartungen darstellt, ist Gegenstand einer bereits seit langem geführten und nicht abgeschlossenen Debatte (vgl. Abschnitt 3.6.1).

Aus der Klassifizierungsmöglichkeit von Handlungen ergibt sich die letzte und wenig erklärungsbedürftige Eigenschaft des Konstruktes der Selbstwirksamkeitserwartungen, nämlich die Domänenspezifität. Personen können sich demnach bezüglich unterschiedlicher Handlungen aus unterschiedlichen Domänen ganz unterschiedlich einschätzen.

Zusammenfassend gilt, dass sich Selbstwirksamkeitserwartungen immer auf eine situationsspezifische Selbsteinschätzung eigener Fähigkeiten bezüglich komplexer, hinreichend schwieriger Handlungen einer Domäne beziehen. Selbstwirksamkeitserwartungen sind daher spezifische (handlungsbezogene, schwierige, situationsspezifische) Fähigkeitsselbsteinschätzungen, weshalb die Begriffe Selbstwirksamkeitserwartung und Fähigkeitsselbsteinschätzung nicht synonym verwendet werden dürfen. Zusammenfassend ergeben sich die folgenden Eigenschaften des Konstruktes:

1. Subjektivität/Selbstreferentialität (vgl. Abbildung 3.4, A),
2. (konkreter) Handlungsbezug (vgl. Abbildung 3.4, B),
3. Schwierigkeitsbezug (vgl. Abbildung 3.4, C),
4. Situationsspezifität/Kontextabhängigkeit (vgl. Abbildung 3.4, D),
5. Domänenspezifität (vgl. Abbildung 3.4, implizit enthalten).

3.2.3. Weitere strukturelle Eigenschaften des Konstruktes

Selbstwirksamkeitserwartungen werden von Bandura (1977a, S. 194) als mehrdimensionales Konstrukt konzipiert. Mehrdimensionalität meint in diesem Sinne, dass sich Selbstwirksamkeitserwartungen nach dem Maß ihrer Ausprägung („magnitute/level“), ihrer Stabilität bzw. Festigkeit („strength“) und bezüglich ihres Spezifitätsniveaus („generality“) unterscheiden. Der Aspekt des Spezifitätsniveaus wurde bereits oben erläutert. Unterschiedliche Ausprägungen bei verschiedenen Personen innerhalb einer Domäne resultieren in der Regel aus unterschiedlichen Einschätzungen der (objektiven) Schwierigkeitsgrade der zu beurteilenden Handlungen. Liegen beispielsweise wenige Erfahrungen auf einem entsprechenden Gebiet vor, so sind die Selbstwirksamkeitserwartungen ggf. schwach ausgebildet oder wenig stabil und können leichter durch andere Erfahrungen verändert werden.

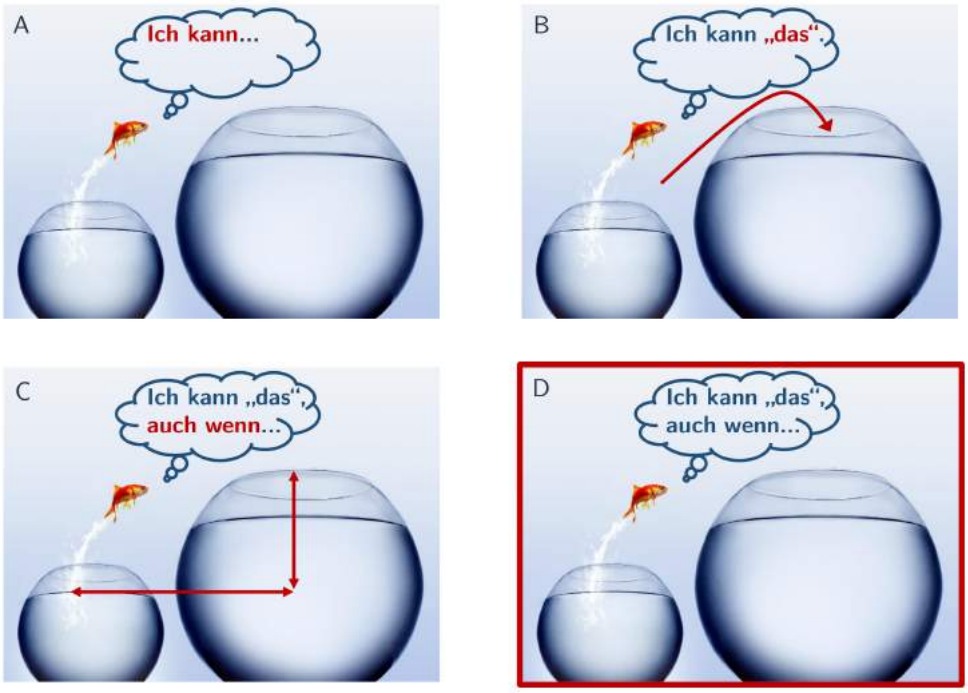


Abbildung 3.4.: Veranschaulichung der Eigenschaften des Konstruktes der Selbstwirksamkeitserwartungen: Subjektivität/Selbstreferentialität (A), Handlungsbezug mit „das“ z. B. als „Springen in ein anderes Aquarium“ (B), Schwierigkeitsbezug (C), Situations-/Kontextabhängigkeit (D), Domänenspezifität (implizit enthalten). Quelle: Mikael Damkier, Shutterstock.com.

Die benannten strukturellen Eigenschaften des Konstruktes sind unter anderem für dessen Messung relevant. In einer experimentellen Studie forderte Bandura die Probanden auf, eine Liste verschiedener Handlungen/Aufgaben in eine Reihenfolge zu bringen, gestaffelt nach ihrer persönlichen Einschätzung darüber, welche der vorgelegten Hand-

lungen sie sich am ehesten zutrauen würden (Messung der Ausprägung). Anschließend wurden die Probanden aufgefordert, für jede Aufgabe anzugeben, wie sicher sie sich hinsichtlich der jeweiligen Einschätzungen waren (Messung der Stabilität). Schließlich mussten die Probanden die Prozedur für ähnliche Aufgaben jedoch bezogen auf eine angrenzende Domäne durchführen, um die Generalität zu messen (vgl. Bandura, 1977a). Der Begriff der Mehrdimensionalität wird von Bandura eigentlich, wie oben dargelegt, bezüglich der Ausprägung, Stabilität und Spezifität des Konstruktes genutzt. Häufig wird er jedoch bezüglich der Eigenschaft der Domänenspezifität verwendet, um zu verdeutlichen, dass Selbstwirksamkeitserwartungen bezüglich unterschiedlicher Bereiche ganz unterschiedlich ausgeprägt sein können und damit nicht eindimensional konzipiert werden dürfen.

Da Selbstwirksamkeitserwartungen bezüglich unterschiedlicher Spezifitätsniveaus (in derselben oder in unterschiedlichen Domänen) erfasst werden können, stellt sich die strukturelle Frage der Hierarchie des Konstruktes: „When different spheres of activity are governed by similar sub-skills there is some interdomain relation in perceived efficacy“ (Bandura, 2006, S. 308). Hinsichtlich dieses domäneninternen Zusammenhangs steht die These im Raum, dass dieser äußerst gering ausfällt, weil die Einschätzung eines Spezialfalles selten aus einem Globalurteil abgeleitet werden kann (vgl. Choi, 2005). Es wird also davon ausgegangen, dass die Hierarchie des Konstruktes eher schwach ausgeprägt ist.

Darüber hinaus stellt sich die Frage, ob es sich bei einer globalen, unspezifischen Erfassung noch um das Konstrukt der Selbstwirksamkeitserwartungen handelt, da diese per definitionem domänen-, handlungs- und situationsspezifisch konzipiert sind. Bandura (2006, S. 307) bemerkt dazu: „There is no all-purpose measure of perceived self-efficacy. The ‘one measure fits all’ approach usually has limited explanatory and predictive value because most of the items in an all-purpose test may have little or no relevance to the domain of functioning. [...] Scales of perceived self-efficacy must be tailored to the particular domain of functioning that is the object of interest.“

3.3. Quellen von Selbstwirksamkeitserwartungen

In den vorangegangenen Abschnitten wurden das Konstrukt der Selbstwirksamkeitserwartungen einbettende theoretische Ansätze dargestellt sowie die Definition des Konstruktes eingehend analysiert und Merkmale aus dieser abgeleitet. Nachfolgend wird beschrieben, auf welcher Grundlage Selbsteinschätzungen der genannten Art überhaupt getroffen werden können, wobei davon ausgegangen wird, dass insbesondere vier Quellen die Einschätzung unterschiedlich stark beeinflussen. Die Analyse der Quellen von Selbstwirksamkeitserwartungen kann in einem nächsten Schritt auch dazu verwendet werden, Unterschiede zu anderen, ähnlichen oder verwandten Konstrukten kriteriengeleitet zu

benennen.

Grundsätzlich sind die fundamentalen kognitiven Fähigkeiten des Menschen zur Selbstreflexion und Selbstregulation, sowie das Verständnis von Ursache-Wirkungs-Beziehungen wichtige Voraussetzungen für die Entwicklung und Ausbildung von Selbstwirksamkeitserwartungen (vgl. Maddux, 2002, S. 279). Mit der Entwicklung dieser Fähigkeiten im frühen Kindesalter bilden sich damit auch Selbstwirksamkeitserwartungen nach und nach aus.

Bandura (1977b) leitet aus seiner Theorie des sozialen Lernens („social learning theory“) vier hauptsächliche Informationsquellen ab, die eine Person (implizit) nutzen kann, um Rückschlüsse auf ihre eigenen Fähigkeiten zu ziehen. Zu diesen Quellen zählen eigene Leistungen oder Könnenserfahrungen („performance accomplishments/mastery experience“), Beobachtungen Anderer/stellvertretende Erfahrungen („vicarious experience“), verbales Überzeugen/Feedback („verbal persuasion“) sowie physiologische Zustände („physiological states“).

Grundsätzlich gilt, dass Effekte gesteigert werden können, wenn unterschiedliche Quellen herangezogen und miteinander kombiniert werden. Es muss jedoch berücksichtigt werden, dass nicht jede Informationsquelle für jede Person gleich ergiebig ist. Ob und inwiefern Selbstwirksamkeitserwartungen durch die genannten Einflussgrößen adressiert werden, hängt stark von der jeweiligen individuellen kognitiven Verarbeitung der Informationen ab, die wiederum durch Kontexteffekte beeinflusst sein können. Verzerrend wirken beispielsweise Diskriminationseffekte oder problematische Kausalattribuierungen (vgl. Bandura, 1977a, S. 200 f.). Positiv wie negativ kann sich der folgende Effekt auswirken: „Once established, enhanced self-efficacy tends to generalize to other [substantially different] situations“ (Bandura, 1977a, S. 195).

Die vier möglichen Quellen werden nachfolgend in Kürze beschrieben. Zusätzlich wird jeweils eine Auswahl an kognitiven Verarbeitungsstrategien angeführt, die zu einer Stärkung oder Schwächung von Selbstwirksamkeitserwartungen führen können. Für eine ausführliche Darstellung der Quellen, sowie eine Aufstellung einer Vielzahl empirischer Studien zur Stützung der theoretischen Annahmen sei auf Bandura (1997, S. 79-115) verwiesen. Es soll an dieser Stelle noch einmal betont werden, dass die nachfolgend beschriebenen Vorgänge meistens unbewusst ablaufen und sich eine Explikation der Einflussfaktoren in der Regel für Befragte schwierig gestaltet.

Direkte Erfahrungen Die direkte Erfahrung zu machen, erfolgreich eine bestimmte Situation bewältigen zu können, gilt als die einflussreichste Quelle für Selbstwirksamkeitserwartungen. Insbesondere das Erleben wiederholten Erfolgs trägt zur stabilen Ausbildung der Überzeugung von den eigenen Fähigkeiten bei: „After strong efficacy expectations are developed through repeated success, the negative impact of occasional failures is likely to be reduced“ (Bandura, 1977a, S. 195). Bei Personen mit wenig

ausgeprägten Selbstwirksamkeitserwartungen geht es in Therapie- oder Ausbildungskontexen häufig darum, Situationen zu kreieren, die Erfolgserlebnisse versprechen. Dazu können gezielt Maßnahmen ergriffen werden wie das Antizipieren von Schwierigkeiten und potentiellen Handlungsoptionen, das schrittweise Abarbeiten eines Problems, das Einräumen realistischer Zeitfenster, das Zulassen von Teamarbeit, das Zulassen von Hilfsmitteln oder schrittweises Erhöhen von Aufgabenschwierigkeiten (vgl. Bandura, 1977a, S. 196). Ob Erfolgserlebnisse tatsächlich zu einer Steigerung der Selbstwirksamkeitserwartungen führen, hängt wesentlich von mehreren Faktoren ab, zum Beispiel davon, ob sich der Erfolg in unterschiedlichen Situationen einstellt. Von entscheidender Bedeutung können Diskriminationseffekte sein, das heißt Erfahrungen und abgeleitete Konsequenzen variieren beispielsweise mit situationalen Faktoren, sodass keine sinnvollen Schlussfolgerungen gezogen werden können. Als besonders einflussreich erweisen sich Kausalattribuierungen. Wenn Erfolgserlebnisse mit äußeren Faktoren und nicht mit eigenen Fähigkeiten begründet werden, dann können diese schwerlich zur Steigerung der Einschätzung eigener Fähigkeiten herangezogen werden. Umgekehrt tragen Misserfolge weniger zu einer Verringerung der Selbstwirksamkeitserwartung bei, wenn diese auf situationale Gegebenheiten zurückgeführt werden statt auf mangelnde Fähigkeiten. Es reicht zudem nicht aus, sich selbst den Erfolg zuzuschreiben. Wird dieser auf Fleiß und nicht auf Fähigkeit attribuiert, so sind ebenfalls keine Steigerungen der Selbstwirksamkeitserwartungen zu erwarten. Darüber hinaus wird der Effekt der Stärkung der Selbstwirksamkeitserwartung über die wahrgenommene Aufgabenschwierigkeit mediiert (vgl. Bandura, 1977a, S. 201).

Stellvertretende Erfahrung Als zweiteffektivste Informationsquelle kann das Beobachten anderer Personen in herausfordernden Situationen angesehen werden. Wenn diese Personen erfolgreich die Aufgaben bewältigen, kann das Zutrauen in die eigenen Fähigkeiten steigen, eine gleiche oder ähnliche Situation ebenfalls gut bewältigen zu können. Wiederum gibt es einige Faktoren, die den Effekt stellvertretender Erfahrungen beeinflussen. Beispielsweise ist es wichtig, dass die beobachtete Person, tatsächlich erfolgreich ist. Darüber hinaus wird der Effekt gesteigert, wenn ein Individuum viele verschiedene Personen beim erfolgreichen Handeln beobachten kann. „If people of widely differing characteristics can succeed, then observers have a reasonable basis for increasing their own sense of self-efficacy“ (Bandura, 1977a, S. 197). Problematisch ist es, wenn Beobachtende den Handelnden spezielle Fähigkeiten zusprechen, denn dann können diese in den Augen der Beobachtenden nicht mehr als Vergleichsmaßstab herangezogen werden (vgl. Bandura, 1977a, S. 201).

Verbales Überzeugen/Feedback Als weitere Einflussgröße auf Selbstwirksamkeitserwartungen gilt verbaler Zuspruch. Für sich genommen kann der Effekt der Überzeugungs-

arbeit auf die Selbstwirksamkeitserwartungen einer Person eher als gering bezeichnet werden, insbesondere dann, wenn die Person tatsächlich über geringe Fähigkeiten verfügt, weil gutes Zureden dann eher einem „Vorgaukeln“ gleichkommt. In Kombination mit den anderen genannten Quellen bilden sie jedoch einen zusätzlichen relevanten Faktor: „That is, people who are socially persuaded that they possess the capabilities to master difficult situations and are provided with provisional aids for effective action are likely to mobilize greater effort than those who receive only the performance aids“ (vgl. Bandura, 1977a, S. 198). Voraussetzung für einen positiven Einfluss ist allerdings, dass Individuen in gewisser Weise an die Kraft der Überzeugung „glauben“, diese zulassen oder ein Feedback als nützlich erachten, da sonst diese Informationsquelle von vornherein als irrelevant eingestuft werden kann (vgl. Bandura, 1977a, S. 201). Letztlich ist der Effekt des verbalen Überzeugens auf die Selbstwirksamkeitserwartungen der Zuhörenden auch davon abhängig, wie viel Glaubwürdigkeit, Seriosität, Prestige, Expertise oder Zuversichtlichkeit dem Feedback-Gebenden von den Betroffenen zuerkannt wird.

Physiologische Zustände Emotionale Erregung in herausfordernden Situationen stellt den letzten Einflussfaktor auf individuelle, spezifische Selbsteinschätzungen dar. Dies resultiert aus der Annahme, dass Individuen aus physiologischen Zuständen eine Art Feedback ableiten können. Schweißausbrüche oder Angst könnten beispielsweise mit fehlenden Fähigkeiten in Zusammenhang gebracht werden, Gelassenheit und Zuversicht dagegen mit vorliegenden Fähigkeiten. „Because high arousal usually debilitates performance, individuals are more likely to expect success when they are not beset by aversive arousal“ (Bandura, 1977a, S. 198). Welcher Informationsgehalt aus einem Erregungszustand abgeleitet wird, hängt wesentlich davon ab, welche Bedeutung diesem zuerkannt wird, denn in unterschiedlichen Situationen kann ein und dasselbe physiologische Merkmal (z. B. Schwitzen) unterschiedlich interpretiert werden (z. B. als Angst, Wut, Ärger, Nervosität, Freude ...). Auch bezüglich emotionaler Erregungszustände spielen Kausalattributionen eine wichtige Rolle. Es ist daher von Relevanz, ob Individuen ihre Nervosität auf fehlende Fähigkeiten oder situationale Gegebenheiten zurückführen (vgl. Bandura, 1977a, S. 202).

3.4. Abgrenzung und Beziehung zu anderen Konstrukten

Für die Charakterisierung des Konstruktes ist es zentral, es von ähnlichen oder verwandten Konstrukten abzugrenzen, also Gemeinsamkeiten und Unterschiede herauszuarbeiten, denn dieses Wissen hilft bei der konkreten, domänenspezifischen Operationalisierung des Konstruktes. Zugleich können anhand der herausgearbeiteten konstruktdifferenzierenden Kriterien bereits existierende Operationalisierungen daraufhin beurteilt werden, ob tatsächlich Selbstwirksamkeitserwartungen im Sinne Banduras vorliegen oder nicht.

Dies wird beispielsweise für Adaptionen des Konstruktes im Bereich der Lehrerbildung relevant (vgl. Kapitel 4). Die Grundlage einer solchen Gegenüberstellung von Konstrukten liefert Bandura, indem er selbst relevante Konstrukte benennt und sie in Relation zu Selbstwirksamkeitserwartungen analysiert (vgl. Bandura, 1997, S. 10-35). Die nachfolgende Zusammenstellung von Konstrukten und ihrer Beziehung zum Konstrukt der Selbstwirksamkeitserwartung orientiert sich an Banduras Liste (Selbstkonzept, Selbstwertgefühl, Kontrollkonstrukte, Effektanzmotivation), wird jedoch um einzelne weitere relevante Konstrukte ergänzt (z. B. Persönlichkeit, Einstellung). Die jeweilige Abgrenzung kann nur in Umrissen erfolgen, denn jedes Konstrukt ist zumeist selbst in einem umfänglichen theoretischen Rahmen verortet. Aus diesem Grund werden die einzelnen Konstrukte lediglich grob umrissen und ein Schwerpunkt auf den Vergleich zum Konstrukt der Selbstwirksamkeitserwartungen gelegt. Das Klassifikationsschema nach Skinner (1996) bietet erneut einen nützlichen Orientierungsrahmen für die Charakterisierung und Abgrenzung der Konstrukte.

3.4.1. Persönlichkeit

Die Persönlichkeitspsychologie bietet vielfältige Zugänge im Rahmen unterschiedlichster Forschungsparadigmen und damit eine Vielzahl an Möglichkeiten, das Konstrukt der Persönlichkeit zu definieren. Mummendey und Grau (2014, S. 24) formulieren eine „Kürzest-Definition“ des Persönlichkeitsbegriffs, die hier nützlich erscheint und vergleichbar mit der von Asendorpf (2016) ist: „Persönlichkeit ist [...] der Inbegriff der überdauernden individuellen Merkmalsausprägungen eines Menschen“ (Mummendey & Grau, 2014, S. 24), wobei „überdauernd“ hier *nicht* meint, dass Veränderung ausgeschlossen ist. Da es eine große Anzahl denkbarer zuschreibbarer Merkmale gibt, „kann jedes Individuum durch eine gewissermaßen einzigartige Konstellation von Merkmalsausprägungen beschrieben werden“ (Mummendey & Grau, 2014, S. 25). Selbstwirksamkeitserwartungen sind ihrer Natur nach veränderbar, erweisen sich jedoch als relativ stabiles Konstrukt (vgl. Abschnitt 3.3). Insofern können Selbstwirksamkeitserwartungen als Persönlichkeitsmerkmal eines Menschen interpretiert werden. Aufgrund der Domänenspezifität des Konstruktes gilt jedoch: „[T]he efficacy belief system is not a global trait [...]“ (Bandura, 2006, S. 307). Vielmehr bilden die Selbstwirksamkeitserwartungen bezüglich unterschiedlicher Domänen für jede Person ein ganz individuelles, dispositionales Muster (vgl. Bandura, 1997, S. 42). Mayr (2011, S. 137) bezeichnet Selbstwirksamkeitserwartungen daher im Gegensatz zu sehr breiten Faktoren wie den klassischen Merkmalen des Fünf-Faktoren-Modells (B5/Big-Five: Neurotizismus, Extraversion, Offenheit, Verträglichkeit und Gewissenhaftigkeit) als spezielles, enges Persönlichkeitsmerkmal.

Ein wesentlicher Unterschied zwischen engen und bereiten Persönlichkeitsmerkmalen ist unter Umständen die Art der Entwicklungsprozesse, die diesen zu Grunde liegen. Wie bereits gesagt, handelt es sich bei Persönlichkeitsmerkmalen nicht um

statische Konstrukte. Allerdings weisen Studien auf substantielle genetische Einflüsse auf Persönlichkeitsmerkmale hin. „Es ist davon auszugehen, dass die Interaktion zwischen Genen und Umweltfaktoren einer komplexen Dynamik unterliegt. So können z. B. Umwelteinflüsse Genaktivitäten auslösen und damit auch auf die grundlegenden Persönlichkeitsmerkmale Einfluss nehmen, die Persönlichkeitsmerkmale führen aber auch zur Wahl einer bestimmten, mit diesen Merkmalen kompatiblen Umwelt, was zur Verfestigung der betreffenden Merkmale beiträgt“ (Mayr, 2011, S. 133). Selbstwirksamkeitserwartungen werden dagegen konzeptionell nicht als genetisch bedingt verstanden und sind durch Lernprozesse veränderbar (vgl. Abschnitt 3.3, Bandura, 1977a, 1997).

Wenn domänenspezifische Selbstwirksamkeitserwartungen eher als spezielle Persönlichkeitsmerkmale konzipiert sind, dann stellt sich die Frage, welche Zusammenhänge zu globalen Persönlichkeitsmerkmalen bestehen. Gemäß des Symmetrieprinzips nach Wittmann, wonach Prädiktor und Kriterium im Sinne einer hohen Vorhersagewahrscheinlichkeit auf einem ähnlichen Spezifitätsniveau erfasst werden sollten, sind keine hohen Korrelationen zwischen globalen Persönlichkeitsmerkmalen und spezifischen Selbstwirksamkeitserwartungen zu vermuten. Eher sind Korrelationen zu unspezifischeren, allgemeinen Selbstwirksamkeitserwartungen erwartbar (vgl. Mayr, 2011, S. 129). „General self-efficacy may be treated as a trait efficacy, yet the utility of this global construct seems to lie in examining its relation with other personality traits“ (Choi, 2005, S. 203). Exemplarisch werden für Lehrer-Selbstwirksamkeitserwartungen Forschungsergebnisse zum Zusammenhang mit den Persönlichkeitsmerkmalen berichtet (vgl. Kapitel 4, Unterabschnitt 4.5.1/Persönlichkeit).

3.4.2. Einstellungen (“attitudes“)

Die Forschung zu Einstellungen füllt einige Regalmeter in universitären Bibliotheken. Zudem wird der Begriff vielfältig verwendet (vgl. Kaiser, 2016; Six, 2016). Deshalb wird auch hier nur die „Kürzest-Definition“ von Mummendey und Grau (2014, S. 26) herangezogen: „Einstellung im Sinne von „attitude“ bezeichnet [...] die Art und Weise, wie sich ein Individuum [...] in seinen Gedanken, Gefühlen, Bewertungen und gegebenenfalls seinen Verhaltensweisen bzw. Intentionen auf ein [...] Objekt richtet.“ Einstellungen gelten als etwas Überdauerndes, jedoch darüber hinaus als grundsätzlich veränderbar (vgl. Mummendey & Grau, 2014, S. 26). Damit können sie ebenfalls als Persönlichkeitsmerkmale interpretiert werden. Unter der Prämisse, dass es sich bei dem Einstellungsobjekt auch um das wertende Individuum selbst handeln kann, können Einstellungen auch auf die eigene Person bezogen sein. In der Regel sind sie jedoch nicht als selbstreferentielles Konstrukt konzipiert. Aufgrund des wertenden Charakters sowie der affektiven Komponenten des Einstellungskonstruktes sind Selbstwirksamkeitserwartungen eher nicht als Einstellung zu sich selbst aufzufassen. Selbstwirksamkeitserwartungen rekurrieren auf eine eher sachliche, unemotionale Selbsteinschätzung, die zunächst ohne Wertung

vorgenommen wird.

3.4.3. Selbstkonzept

> „[I]t is sometimes difficult to distinguish between [...] self-concept and self efficacy."
>
> *Bong & Skaalvik, 2003, S. 31*

Das (akademische) Selbstkonzept ist eines der am meisten beforschten Konstrukte, insbesondere innerhalb der Pädagogischen Psychologie. Eine ausführlichere Analyse des Konstruktes kann allein deshalb hier nicht erfolgen.[4] Nach einer kurzen definitorischen Fassung liegt der Schwerpunkt dieses Abschnittes auf dem Vergleich des Konstruktes mit Selbstwirksamkeitserwartungen.

Möller und Trautwein (2015, S. 178) fassen unter dem Selbstkonzept die kognitive/mentale Repräsentation der eigenen Person, die Einstellungen und Einschätzungen zur eigenen Person enthält. Zum Selbstkonzept einer Person gehören folglich auch Einschätzungen der eigenen Fähigkeiten, was die Verbindung zum Konstrukt der Selbstwirksamkeitserwartungen einsichtig macht. Darüber hinaus legt die Definition nahe, dass das (globale) Selbstkonzept zusätzlich zu den beschreibend/evaluativen Komponenten eigener Fähigkeiten („Ich bin gut/schlecht in ...") auch affektive („Ich mag...") Komponenten umfasst. Bong und Skaalvik (2003, S. 7) weisen darauf hin, dass keine Einigkeit darüber herrscht, ob beide Facetten tatsächlich konstruktkonstituierend sind. Einige Forschende legen auf eine Trennung evaluativer und affektiver Bestandteile wert, während andere bewusst beide Facetten in ihren Instrumenten operationalisieren (vgl. z. B. SDQ I-III, Marsh, 1990a). „The issue of whether or not the competence and affective components of self-concept are empirically distinguishable has not been resolved" (Bong & Skaalvik, 2003, S. 7 f.). Werden Selbstkonzepte als generalisierte Fähigkeitsselbsteinschätzungen konzipiert und affektive Komponenten außen vor gelassen, erhöht sich der Überlapp zum Konstrukt der Selbstwirksamkeitserwartungen, denn Selbstwirksamkeitserwartungen schließen in keinem Fall affektive Urteile ein (vgl. Abschnitt 3.2.1). Gleichzeitig bilden Selbstwirksamkeitserwartungen, wie in Unterabschnitt 3.2.2 dargestellt, sehr spezifische Fähigkeitsselbsteinschätzungen (Handlungsbezug, Schwierigkeitsbezug, etc.), sodass eine totale Übereinstimmung beider Konstrukte zunächst nicht anzunehmen ist.

Der gemeinsame Kern beider Konstrukte liegt also im Bezugspunkt der subjektiven Einschätzung eigener Fähigkeiten. Darüber hinaus sind beide Konstrukte konzeptionell domänenspezifisch angelegt. Allerdings werden Selbstkonzepte im Gegensatz zu

[4] Eine Überblicksdarstellung findet sich z. B. bei Möller und Trautwein (2015).

Selbstwirksamkeitserwartungen in der Regel als sehr viel globalere, zusammenfassende Einschätzungen formuliert, die keinen konkreten Handlungsbezug aufweisen: „Self-concept judgements can be domain specific but are not task specific“ (Pajares, 1996, S. 561). Inhaltlich ist dies damit begründbar, dass sich Personen bezüglich einzelner Aufgaben oder Handlungen durchaus als fähig beurteilen können, auch wenn das Selbstkonzept bezüglich der übergreifenden Domäne dieser Aufgaben eher schwach ausgeprägt ist, wobei auch der umgekehrte Fall denkbar ist.

Dass „Selbstkonzepte generalisierte fachspezifische Fähigkeitseinschätzungen dar[stellen]“ (Möller & Köller, 2004, S. 19) wird auch an konkreten Operationalisierungen deutlich (vgl. dazu die Auswahl an Instrumenten und Itembeispiele im Anhang, Tabelle A.1 auf S. 311). Irreführend ist jedoch der Ansatz, dass die Domäne bzw. das Abstraktionsniveau der Domäne festlegt, ob eher spezifischere oder eher globalere Einschätzungen vorgenommen werden müssen, weshalb Bong und Skaalvik (2003, S. 17) darauf insistieren, das Spezifitätsniveau der Domäne vom Spezifitätsniveau der Messung zu unterscheiden. Sie illustrieren an einem Beispiel (Items nach Marsh, 1990b; Lau, Yeung, Jin und Low, 1999; Pajares, Miller und Johnson, 1999), dass bei festgelegter aufgabenspezifischer Domäne (z. B. „writing“) sowohl das Selbstkonzept (z. B. durch Items wie „Compared to others, I am good at writing.“ oder „I am satisfied with how well I do in writing.“) als auch Selbstwirksamkeitserwartungen (Beispielitem: „How confident are you that you can write a paragraph with details that support the topic sentence or main idea?“) erhoben werden können (Bong & Skaalvik, 2003, S. 19 ff.), weil es sich bei den Selbstkonzept-Items immer um aggregierte Urteile handelt.[5]

Die Domänenspezifität kann als strukturelle Gemeinsamkeit beider Konstrukte bezeichnet werden, die nicht nur theoretisch angelegt (vgl. Shavelson, Hubner & Stanton, 1976; Bandura, 1997), sondern empirisch auch gut belegt ist. So zeigt sich die Domänenspezifität des Selbstkonzeptes darin, dass aufgrund sich aufhebender Effekte sozialer (d. h. externaler) und dimensionaler (d. h. internaler) Leistungsvergleiche Selbstkonzepte gegensätzlicher Domänen (wie z. B. Mathematik und Sprachen) beinahe unkorreliert sind (vgl. Abbildung 3.5). Dieses Phänomen wird mit Hilfe des „internal/external-frame-of-reference“-Modells (I/E-Modell) erklärt (vgl. z. B. Marsh, 1986; Möller &

[5] Ein weiteres einigermaßen gelungenes Beispiel für eine Operationalisierung beider Konstrukte bez. derselben Domäne erfolgte im Rahmen der PISA-Studie 2003 (vgl. OECD, 2005, S. 292 f.; Lee, 2009). Ein Beispielitem für das mathematische Selbstkonzept lautet „I learn mathematics quickly.“, während mathematische Selbstwirksamkeitserwartungen z. B. durch das Item „How confident do you feel about solving an equation like $3x + 5 = 17$?“ erhoben werden. Häufig verschwimmen jedoch die Konstrukte, wenn Items zur Messung von Selbstwirksamkeitserwartungen nicht spezifisch genug formuliert werden. Dies wird deutlich, wenn man konkrete Operationalisierungen miteinander vergleicht. Schwarzer und Jerusalem (1999, S. 18 f.) formulieren beispielsweise bez. der schulbezogenen Selbstwirksamkeitserwartungen Items wie dieses: „Es fällt mir leicht, neuen Unterrichtsstoff zu verstehen.“ Tatsächlich fällt es schwer, den konkreten Handlungs- oder Schwierigkeitsbezug zu erkennen und die Konstrukte von Operationalisierungen des (globaleren) akademischen Selbstkonzeptes abzugrenzen (z. B. „I learn things quickly in most school subjects“ nach Marsh, 1990c, S. 625).

Köller, 2004; Möller, Pohlmann, Köller & Marsh, 2009). Selbstwirksamkeitserwartungen gegensätzlicher Domänen korrelieren dagegen in der Regel positiv miteinander (vgl. ebenfalls Abbildung 3.5), weil man davon ausgeht, dass andere Bezugsgrößen als soziale und dimensionale Vergleiche eine Rolle spielen (vgl. Abschnitt 3.3: eigene Erfahrungen, stellvertretende Erfahrungen, etc.) und sich somit die Effekte nicht aufheben (vgl. u. a. Bong, 1998; Möller & Trautwein, 2015, S. 191). Dieser Befund deckt sich mit der Annahme nach Bandura (1977a), dass sich Selbstwirksamkeitserwartungen bezüglich eines Bereiches auch auf andere Bereiche übertragen können, insbesondere dann, wenn die Bereiche ähnliche Handlungen adressieren. Aufgrund des unterschiedlichen Grades der Korreliertheit der Konstrukte bezüglich verschiedener Domänen nehmen Parker, Marsh, Ciarrochi, Marshall und Abduljabbar (2014, S. 35) an, dass Selbstwirksamkeitserwartungen weniger stark domänenspezifisch sind als entsprechende Selbstkonzepte. In jedem Fall sagen sowohl die Selbstwirksamkeitserwartungen als auch das Selbstkonzept einer Person Leistungen in einer Domäne besser vorher, wenn die Konstrukte entsprechend domänenspezifisch erhoben werden (vgl. z. B. Möller & Köller, 2004; Möller u. a., 2009; bzw. Bong, 2002). Gemeinsam ist beiden Konstrukten daher, dass zwischen eigenen Leistungen/Erfahrungen und den jeweiligen Konstrukten reziproke Zusammenhänge bestehen.

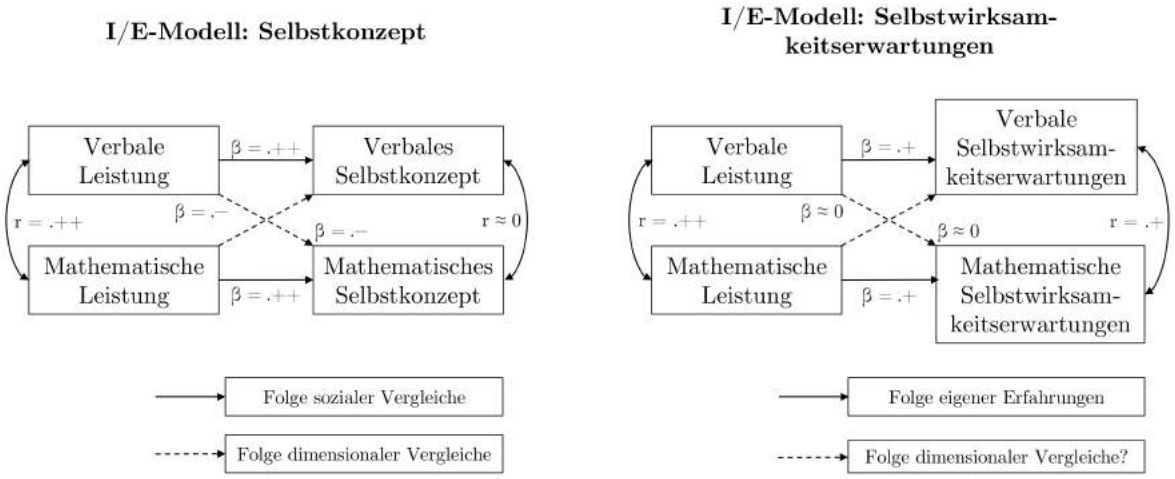


Abbildung 3.5.: Vergleich des I/E-Modells für das Selbstkonzept (vgl. Möller & Trautwein, 2015, S. 190; nach Marsh, 1986, S. 134) und Selbstwirksamkeitserwartungen (vgl. u. a. Marsh, Walker & Debus, 1991, S. 342; Bong, 1998, S. 105).

Auf struktureller Ebene wird jeweils die Hierarchie der beiden Konstrukte aufgrund der Domänenspezifität und der Operationalisierung auf unterschiedlichen Spezifitätsniveaus – insbesondere bezüglich des Selbstkonzepts – rege beforscht und diskutiert (vgl. u. a. Marsh, Byrne & Shavelson, 1988; Marsh, 1990c; Brunner, Keller, Hornung, Reichert & Martin, 2009, 2010). Zusammenfassend geht man davon aus, dass die Hierarchie der Konstrukte unterschiedlich stark ausgeprägt ist, wobei Selbstwirksamkeitserwartungen

im Gegensatz zum Selbstkonzept als weniger bzw. schwach hierarchisch eingestuft werden (vgl. Bong & Skaalvik, 2003, S. 22 f.). Die Frage nach der Hierarchie der Konstrukte adressiert die Frage nach einem optimalen Spezifitätsniveau der Messung mit dem Ziel einer möglichst großen Prädiktivität, welche nicht geklärt ist. Bandura (1997, S. 10 f.) verweist darauf, dass letztlich eine Passung der Niveaus von Zielkriterium und Selbstwirksamkeitserwartung vorliegen muss und dass globalere Messungen in der Regel an prädiktiver Kraft verlieren. Dass auch Selbstwirksamkeitserwartungen teils auf einem relativ unspezifischen Niveau gemessen werden, hängt oft damit zusammen, dass gleichzeitig wenig spezifische Zielkriterien (z. B. Noten) von Interesse sind (vgl. Bong & Skaalvik, 2003, S. 17).

Sowohl Selbstwirksamkeitserwartungen als auch das Selbstkonzept gelten als relativ stabile, allerdings veränderliche Konstrukte. Die Definition von Markus und Kunda (1986, S. 859) beleuchtet diesen Aspekt für das Selbstkonzept anschaulich: „[T]he self is not a unitary structure, or even a generalized average of images and cognitions. Instead, the self-concept encompasses within its scope a wide variety of self-conceptions – the good selves, the bad selves, the hoped-for selves, the feared selves, the not-me selves, the ideal selves, the possible selves, the ought selves [...]. The self-concept at a given moment – the working self-concept – is a subset of this universe of self-conceptions.“ Diese Definition macht weiterhin einsichtig, warum das Selbstkonzept zwar ebenfalls situationsabhängig sein kann, im Vergleich zu Selbstwirksamkeitserwartungen jedoch grundsätzlich als stabiler („cross-situational consistency“) anzusehen ist, denn die Gesamtheit möglicher Selbstkonzepte einer Person („the universe of self-conceptions“) ist sehr konstant, wobei bestimmte Sets möglicher Selbstkonzepte von zentralerer Bedeutung sind, und zwar in dem Sinne, dass sie häufiger aktiviert werden und damit den Kern des Persönlichkeitsbildes konstituieren (vgl. Markus & Kunda, 1986, S. 859). Auch wenn Bandura (1977a, S. 200) darauf hinweist, dass sich ausgeprägte Selbstwirksamkeitserwartungen nicht durch einmalige Erfahrungen ändern lassen, sondern durchaus eine gewisse Änderungsresistenz aufweisen, können im Rahmen von Interventionen Selbstwirksamkeitserwartungen eher adressiert werden als das Selbstkonzept einer Person (vgl. Bong & Skaalvik, 2003, S. 25 ff.). Dieser Unterschied im Grad der Veränderlichkeit beider Konstrukte erscheint intuitiv einsichtig, denn ein Selbstbild, welches eine Vielzahl von Komponenten kognitiver wie affektiver Art umfasst und sich durch eine Vielzahl von Erfahrungen über einen langen Zeitraum entwickelt, ist eben weniger leicht beeinflussbar, als es Überzeugungen bezüglich ganz spezifischer Tätigkeiten sind.

Letztlich ist beiden Konstrukten gemeinsam, dass sie in Zusammenhang mit einer Reihe anderer Konstrukte wie Motivation, Interesse, Wahlentscheidungen sowie (akademischen) Leistungen stehen (vgl. exemplarisch für den Bereich der pädagogischen Psychologie überblicksartig Pajares, 1996; Bong & Skaalvik, 2003, S. 27 f.; Möller &

Trautwein, 2015), wobei die Ergebnisse hinsichtlich der Prädiktivität von Verhaltensvariablen durch die beiden Konstrukte uneinheitlich sind. Je nachdem auf welchem Spezifitätsniveau die Zielvariablen erhoben werden, werden Selbstwirksamkeitserwartungen (vgl. u. a. Pajares & Miller, 1994) oder dem Selbstkonzept (vgl. u. a. Huang, 2012) eine größere prädiktive Kraft zugeschrieben.

Zusammenfassung Im weitesten Sinne können sowohl das Selbstkonzept als auch Selbstwirksamkeitserwartungen als Fähigkeitsselbsteinschätzungen verstanden werden. Selbstwirksamkeitserwartungen unterscheiden sich jedoch vom Selbstkonzept durch die sehr spezifische Art der Fähigkeitsselbsteinschätzung, die u. a. durch einen konkreten Handlungs- und Schwierigkeitsbezug gekennzeichnet ist. Das Selbstkonzept ist dagegen auf einer globaleren Ebene (weniger handlungsorientiert), wenn auch bezüglich derselben Domäne konzipiert. Darüber hinaus werden Selbstwirksamkeitserwartungen im Rahmen der sozial-kognitiven Theorie als rationale Selbsteinschätzung konstruiert, während das Selbstkonzept grundsätzlich auch affektive Komponenten enthalten kann. Weitere Unterschiede weisen die Konstrukte hinsichtlich ihrer Genese, Hierarchie und Stabilität auf. Tabelle 3.2 enthält die Quintessenz des durchgeführten Konstruktvergleichs.

Auf theoretischer Ebene scheint die Unterscheidung beider Konstrukte damit möglich zu sein. Die Schwierigkeit der Differenzierung von Selbstwirksamkeitserwartungen und Selbstkonzept liegt zumeist auf der Ebene konkreter Operationalisierungen. Eine mögliche Begründung ist, dass beide Konstrukte grundsätzlich für unterschiedliche Spezifitätsniveaus operationalisiert werden können (vgl. Bong & Skaalvik, 2003, S. 17) und in dem Sinne kein Niveau als Grenzlinie angegeben werden kann (vgl. z. B. Marsh, 1990a, S. 99). Insbesondere erweist sich die hinreichend spezifische Formulierung von Selbstwirksamkeitserwartungen bei zunehmend globaler werdenden Domänen als problematisch.

Auch wenn formal Distinktionskriterien für die Konstrukte angegeben werden können, so erscheint es unbefriedigend, dass keine Rahmentheorie existiert, in der beide Konstrukte inhaltlich sinnvoll miteinander in Beziehung gesetzt werden, sodass sich ein Forschungsdesiderat ergibt, welches Marsh (1990a, S. 97) bereits vor 27 Jahren andeutet: „An important task that has not been pursued sufficiently is to integrate the different self-related constructs into a unified theory of self.“ Es wird der Versuch unternommen, das fehlende, beide Konstrukte umfassende Theoriegebäude zu kompensieren, indem sich sowohl eine Reihe theoretischer (vgl. u. a. Bong & Clark, 1999; Bong & Skaalvik, 2003) als auch empirischer Arbeiten (vgl. u. a. Pietsch, Walker & Chapman, 2003; Marsh, Dowson, Pietsch & Walker, 2004; Choi, 2005; Ferla, Valcke & Cai, 2009, 2010; Stankov, Lee, Luo & Hogan, 2012; Huang, 2012; Bong, Cho, Ahn & Kim, 2012; Scherer, 2013; Parker u. a., 2014; Jansen, Scherer & Schroeders, 2015) der Beziehung beider Konstrukte widmen.

Tabelle 3.2.: Vergleich von Selbstkonzept und Selbstwirksamkeitserwartungen (adaptiert und ergänzt nach Bong & Skaalvik, 2003, S. 10).

Kriterium	**Selbstkonzept**	**Selbstwirksamkeitserwartung**	**U/G**[1]
Inhalt	Wahrnehmung/Einschätzung eigener Kompetenzen	Wahrnehmung/Einschätzung eigener Kompetenzen	G
	beschreibend/evaluativ, affektiv	beschreibend/evaluativ	U
	gegenwartsorientiert	gegenwartsorientiert	G
Spezifitäts-niveau	domänenspezifisch	domänenspezifisch	G
	globaler (Bereich)	spezifischer (Situation/Kontext/Aufgabe/Handlung)	U
Struktur[2]	mehrdimensional	mehrdimensional	G
	stärker hierarchisch	schwach hierarchisch	U
Genese	eigene Erfahrungen/Leistungen	eigene Erfahrungen/Leistungen	G
	Vergleiche (sozial, dimensional, temporal)	Vergleiche (kriteriale/persönl. Standards)	U
	antizipierte Sicht „relevanter Anderer“	Beobachtungen, verbale Überzeugungen, physiol. Zustände	U
	Kausalattributionen	Kausalattributionen	G
Stabilität	veränderlich, aber stabil	veränderlich, aber stabil	G
	stabiler	veränderlicher	U
Prädiktor	Leistungen, Gedanken, Gefühle, Handlungen	Leistungen, Gedanken, Gefühle, Handlungen	G

Anmerkung: Bong und Skaalvik (2003, S. 10, 24 f.) gehen davon aus, dass es sich beim Konstrukt der Selbstwirksamkeitserwartungen um ein zukunftsgerichtetes und beim Konstrukt des Selbstkonzeptes um ein vergangenheitsorientiertes Konstrukt handelt. Dieser Lesart wird hier mit Verweis auf Bandura (2006, S. 308 f.) nicht gefolgt.

[1] U - Unterschied, G - Gemeinsamkeit in den Konstitutionalisierungen

[2] Der Begriff der Mehrdimensionalität wird unterschiedlich gebraucht und teils synonym zum Begriff der Domänenspezifität verwendet. Bandura (1977a) meint damit die Aspekte der Stärke, Stabilität und Generalität des Konstruktes. Bezüglich des Selbstkonzeptes kann z. B. die Konzeption durch affektive und beschreibend/evaluative Komponenten als mehr-/zweidimensional gedeutet werden. In dieser Lesart wären Selbstwirksamkeitserwartungen eindimensional (vgl. Choi, 2005, S. 198).

3.4.4. Selbstwertgefühl

Das Selbstwertgefühl bezieht sich auf eine affektive Evaluation des eigenen Selbstbildes (Selbstkonzeptes).[6] Es handelt sich dabei um eine Art Gewichtung eigener Interessen, Haltungen, Fähigkeiten, etc. hinsichtlich der individuell wahrgenommenen Angemessenheit dieser Persönlichkeitsfacetten. Ein Individuum kommt im Zuge dieser (unbewussten) Abschätzung zu einem Urteil über den eigenen „Wert". Bergius (2016) definiert das Selbstwertgefühl als „stationäre Gestimmtheit des Selbstseins (Selbst), mit dem der Mensch sich als Träger eines Wertes erlebt. Seine Verneinung ist das Minderwertigkeitsgefühl." Teils wird das Selbstwertgefühl als Komponente des Selbstkonzeptes konzeptualisiert (vgl. z. B. Harter, 1982; Marsh & O'Neill, 1984). Eines der weitverbreitetsten Messinstrumente zur Erfassung des Selbstwertgefühls ist die „Rosenberg Self-Esteem Scale" (RSES vgl. u. a. Rosenberg, 1965; Rosenberg, Schooler, Schoenbach und Rosenberg, 1995).

Bandura (1997, S. 11) stellt heraus, dass es keine fixe Relation zwischen dem Konstrukt der Selbstwirksamkeitserwartung und dem Selbstwertgefühl eines Menschen gibt. Es ist möglich, dass sich eine Person als unfähig bezüglich einer bestimmten Aufgabe ansieht, und trotzdem keine negativen Einflüsse auf das Selbstwertgefühl vorliegen, weil die Person dieser Aufgabe keine Bedeutung beimisst. Die Person könnte sich aber auch stark dafür verurteilen, diese Aufgabe wahrscheinlich nicht bewältigen zu können. Auch hoch ausgeprägte Selbstwirksamkeitserwartungen können individuell sowohl als Quelle von Stolz dienen als auch als nichtig abgetan werden. Es ist jedoch so, dass Menschen eher solche Tätigkeiten ausüben, aus denen sie potentiell eine gewisse Selbstbestätigung ziehen können, sodass dann Korrelationen beider Konstrukte anzunehmen sind.

Selbstwirksamkeitserwartungen stehen in einem reziproken Verhältnis zu (eigenen) Erfahrungen und Leistungen, was für das Konstrukt des Selbstwertgefühls lediglich in abgeschwächter Form gilt. Studien zeigen daher größere Zusammenhänge von Selbstwirksamkeitserwartungen und Zielsetzungen oder (akademischen) Leistungen als zwischen dem Selbstwertgefühl und den genannten Kriterien (vgl. u. a. Mone, Baker & Jeffries, 1995; Bong u. a., 2012). Dieser Befund ist nicht nur aus inhaltlichen, sondern auch aus forschungsmethodischen Erwägungen einsichtig, wenn man bedenkt, dass die meisten Zielkriterien auf einem viel spezifischeren Niveau erfasst werden, als es für das Konstrukt des Selbstwertgefühls sinnvoll wäre (vgl. Swann, Chang-Schneider & Larsen McClarty, 2007, S. 87 f.).

Darüber hinaus unterscheiden sich Selbstwirksamkeitserwartungen und das Selbstwert-

[6] Es ist hilfreich, sich einen idealisierten zweistufigen Prozess vorzustellen. Zunächst nimmt das Individuum Einschätzungen und Wertungen hinsichtlich seiner Fähigkeiten, Interessen etc. vor (Selbstkonzept). In einem zweiten Schritt werden diese Einschätzungen und Wertungen erneut bewertet und zwar in dem Sinne, dass das Individuum entscheidet, ob und inwiefern der sich selbst beigemessene Wert als Mensch durch die identifizierten Fähigkeiten, Neigungen, Interessen, etc. beeinflusst wird.

gefühl auch in den Faktoren, die zur Ausbildung der jeweiligen Konstrukte beitragen. Pelham und Swann (1989, S. 672) identifizieren bezüglich der Genese des Selbstwertgefühls drei Faktoren: die menschliche Fähigkeit, positive wie negative Gefühle wahrzunehmen, eigene Selbstbilder (Selbstkonzepte) sowie Referenzrahmen zur Wertung dieser Selbstbilder z. B. durch individuelle Standards.

3.4.5. Effektanzmotivation ("effectance motivation")

Das Konstrukt der „effectance motivation" geht auf White (1959) zurück und stellt im Rahmen der Selbstbestimmungstheorie nach Ryan und Deci (2000) ein motivationales Grundbedürfnis des Menschen dar. „[A]ll people innately desire to engage in effective interactions with the environment, interactions in which they experience themselves as producing desired effects and preventing undesired effects" (Skinner, 1996, S. 557). Es handelt sich also um ein Wirksamkeitsmotiv (vgl. Schmalt, 2016) und damit ggf. um die Grundlage, auf der sich Selbstwirksamkeitserwartungen überhaupt erst entwickeln können. „It is proposed to designate [...] the experience produced [by effectance motivation] as a feeling of efficacy" (White, 1959, S. 329). Folglich sind Selbstwirksamkeitserwartungen und „effectance motivation" voneinander zu unterscheidende Konstrukte (vgl. Skinner, 1996, S. 557).

Bandura (1997, S. 13 f.) steht der Theorie von White (1959) sehr kritisch gegenüber. Er hält es z. B. für schwierig, die Existenz einer solchen intrinsischen Motivation empirisch überhaupt nachzuweisen. Es müsste ein Messverfahren geben, welches Effektanzmotivation tatsächlich als ursächlich für Verhalten nachweisen könnte. Darüber hinaus diskutiert Bandura die erklärende Kraft einer solchen Motivation, die bis auf Handlungsinitiativen nichts weiter vorhersagen zu können scheint und z. B. an der Vorhersage variablen menschlichen Handelns scheitert (vgl. Bandura, 1997, S. 15). Aus den genannten Gründen konzipiert Bandura *keinen* angeborenen „Trieb" oder eine Art angeborene Motivation als ursächlich für menschliches Handeln, sondern sieht Anreize in Form antizipierter, erwünschter Resultate als ausschlaggebend an: „In social cognitive theory, people exercise control for the benefits they gain by it. [...] [T]he striving for control is not a drive for its own right" (Bandura, 1997, S. 16). Einsichtig wird diese Annahme dadurch, dass Individuen in Situationen bewusst von Handlungen absehen oder Kontrollausübung vermeiden können, was mit einem angeborenen Drang schwerlich zu vereinbaren wäre.

3.4.6. Kontrollüberzeugungen ("locus of control")

Es gibt eine große Anzahl von Konstrukten, die auf eine spezifische Art und Weise Kontrollüberzeugungen definieren (für eine Übersicht vgl. z. B. Skinner, 1996). Wie bereits dargestellt, können auch Selbstwirksamkeitserwartungen zu diesen gezählt werden (vgl.

Abschnitt 3.1.3). Im Folgenden wird das „locus-of-control“-Konstrukt näher beleuchtet, weil insbesondere im Rahmen der Analyse zur Erfassung von „teacher efficacy“ (vgl. Kapitel 4) eine Nähe zu diesem Konstrukt beobachtet wird und es aus diesem Grund wichtig ist, „locus of control“ von Selbstwirksamkeitserwartungen unterscheiden zu können. Auch das „locus-of-control“-Konstrukt wird von mehreren Autoren konzeptionalisiert (vgl. u. a. Rotter, 1966; Levenson, 1973; Lefcourt, 1981, 1982), wobei hier aus den genannten Gründen ein Fokus auf die Definition nach Rotter (1966) gelegt wird. Diese zählt darüber hinaus zu den bekanntesten und am häufigsten in der Forschung aufgegriffenen.

Rotter (1966, S. 1) definiert das Konstrukt als domänen- und situationsübergreifende, d. h. generalisierte Überzeugung einer Person, Ereignisse beeinflussen zu können („internal control“) oder nicht („external control“):

> „When a reinforcement is perceived by the subject as following some action of his own but not being entirely contingent upon his action, then, [...] it is typically perceived as the result of luck, chance, fate, as under the control of powerful others, or as unpredictable because of the great complexity of the forces surrounding him. When the event is interpreted in this way by an individual, we have labeled this a belief in *external control*. If the person perceives that the event is contingent upon his own behavior or his own relatively permanent characteristics, we have termed this a belief in *internal control*“ [Hervorhebungen durch C. M.].

Eine internale Ursachenzuschreibung kann beispielsweise auf das eigene Handeln (Verhalten oder mentale Aktivitäten) bzw. auf bestimmte Attribute (z. B. Persönlichkeit, Fähigkeiten) des Individuums erfolgen (vgl. Skinner, 1996, S. 552). Auch wenn Rotter (1966, S. 4) das Konstrukt als eindimensional und bipolar charakterisiert, so deuten Forschungsergebnisse eher darauf hin, dass es sich hinsichtlich der internalen und externalen Kontrollüberzeugung um separate Dimensionen handelt (vgl. Skinner, 1996, S. 552). Da es sich bei der Konzeptualisierung des Konstruktes im Wesentlichen um die individuelle Einschätzung kausaler Zusammenhänge handelt, kann es nach dem Klassifikationsansatz von Skinner (1996, S. 553) als eine „means-ends“-Relation aufgefasst werden („means“: Verhalten, Attribute, Glück, Schicksal, etc.; ends: Wirkungen, Ereignisse, etc.), was gleichzeitig einen fundamentalen Unterschied zum Konstrukt der Selbstwirksamkeitserwartungen markiert, das als „agent-means“-Relation konzeptualisiert ist. „Beliefs about whether one can produce certain actions (*perceived self-efficacy*) cannot, by any stretch of the imagination, be considered the same as beliefs about whether actions affect outcomes (*locus of control*)“ (Bandura, 1997, S. 20, Hervorhebungen im Original). Während es sich bei Selbstwirksamkeitserwartungen um Fähigkeitsselbsteinschätzungen handelt, heben internale Kontrollüberzeugungen auf eine Einschätzung darüber ab, ob z. B. durch eigene Fähigkeiten Einfluss auf bestimmte Ziele oder Ereignisse genommen werden kann. Man kann daher über stark ausgeprägte internale Kontrollüberzeugungen

verfügen („Es liegt grundsätzlich in meiner Macht, etwas Bestimmtes zu erreichen.“), sich aber (im Moment) die nötigen Fähigkeiten absprechen: „[T]here is no necessary correspondence between self-efficacy and internal control factors [...]“ (Ajzen, 2002, S. 665). Auch erweisen sich Selbstwirksamkeitserwartungen als prädiktiver für Verhalten als Kontrollüberzeugungen (vgl. Bandura, 1997, S. 47).

3.4.7. Handlungsergebniserwartungen („outcome expectancies“)

Wie bereits in Abschnitt 3.2.1 dargestellt sind „self-efficacy beliefs“ (Selbstwirksamkeitserwartungen) und „outcome-expectancy/response-outcome expectation“ (Handlungsergebniserwartungen) zentrale Konstrukte der sozial-kognitiven Theorie Banduras, die in Abgrenzung zueinander definiert sind. Zusammenfassend gilt, dass „self-efficacy beliefs“ eine Einschätzung eigener Fähigkeiten bezüglich konkreter Handlungen widerspiegeln und dass „outcome-expectancies“ eine Einschätzung über wahrscheinliche Konsequenzen dieser individuellen Handlungen darstellen (Bandura, 1997, S. 19-24). Beide Konstrukte beziehen sich auf das individuelle Verhalten einer Person und sind damit selbstreferentiell. Erwartungen über Konsequenzen und Resultate der Handlungen anderer Personen sind ggf. davon abhängig, welche Fähigkeiten diesen Personen zugeschrieben werden. Sie sind jedoch nicht von Selbstwirksamkeitserwartungen der beurteilenden Person abhängig und bilden daher ein anderes Konstrukt und keine Handlungsergebniserwartungen im Sinne Banduras. Um im Bandura'schen Sinn „outcome expectancies“ zu erheben, muss das Folgende beachtet werden:„[I]tems would have to refer to outcomes the individual teacher could expect, given certain actions or means he or she felt capable of delivering“ (Tschannen-Moran & Woolfolk Hoy, 2001, S. 792).

Darüber hinaus werden im Rahmen der sozial-kognitiven Theorie Banduras Handlungsergebniserwartungen durch Selbstwirksamkeitserwartungen einer Person beeinflusst. In Abhängigkeit des individuell beigemessenen Wertes des „outcome“ kann dieser als motivationaler Anreiz dienen, eine Handlung aufzunehmen (vgl. Bandura, 1997, S. 21 ff.). Nach Skinner (1996, S. 553) handelt es sich bei Selbstwirksamkeitserwartungen um eine „agent-means“-Relation, während Handlungsergebniserwartungen als „means-ends“-Relation charakterisiert werden können.

3.5. Die Bedeutung von Selbstwirksamkeitserwartungen

„[F]ully describing the many ways that self-efficacy beliefs are important would take hundreds of pages."

Maddux, 2002, S. 280

Die Bedeutung des Konstruktes der Selbstwirksamkeitserwartung ergibt sich einerseits aus seiner zentralen Stellung innerhalb der sozial-kognitiven Theorie, wobei diese bei Kontrolle der tatsächlichen Fähigkeiten und geeigneter Anreize als entscheidend für die Handlungsaufnahme bzw. Handlungsinitiative angesehen werden (vgl. Bandura, 1977a, S. 194). Andererseits werden von Bandura zusätzlich eine Reihe nützlicher, potentiell leistungssteigernder Faktoren mit Selbstwirksamkeitserwartungen in Verbindung gebracht. Demnach beeinflussen Selbstwirksamkeitserwartungen einer Person nicht nur maßgeblich die Auswahl und Aufnahme einer Handlung, sondern auch wie viel Aufwand und Ausdauer bei der Durchführung der Handlungen von der Person investiert wird. Ebenso wird angenommen, dass durch den Glauben in die eigenen Fähigkeiten das Durchhaltevermögen bei auftretenden Schwierigkeiten gestärkt wird und dass hoch ausgeprägte Selbstwirksamkeitserwartungen im Sinne eines positiven Denkmusters ganz grundsätzlich das Bewältigen äußerer Anforderungen wahrscheinlicher machen. Letztlich geht Bandura davon aus, dass durch die genannten Faktoren wie Ausdauer, Aufwand, Durchhaltevermögen, positives Denken etc. die zu erbringenden Leistungen bzw. die Qualität der Handlungen positiv beeinflusst werden (vgl. Bandura, 1997, S. 3). Selbstwirksamkeitserwartungen haben damit indirekt einen Einfluss auf den Handlungserfolg.

Insgesamt ergibt sich aus den postulierten Zusammenhängen, dass Selbstwirksamkeitserwartungen sowohl für die physische und psychische Gesundheit als auch für selbstregulierende Prozesse von Bedeutung sind. Deutlich wird dies z. B. im Rahmen der Coping-Forschung (Belastungserleben, Burnout, ...): „Individuals' appraisals of whether the stressful situation is potentially controllable and whether their resources are adequate to exercise control have been hypothesized to influence the kind of coping they will show [...]. Appraisals of high control should lead to information seeking, planning, strategizing, preventative efforts, and direct action. Appraisals of low control should lead to confusion, escape, pessimism, and passivity" (Skinner, 1996, S. 556). Im Allgemeinen nimmt Bandura (1997, S. 117-161) an, dass Selbstwirksamkeitserwartungen auf vier Prozessformen „menschlichen Denkens" („human functioning") wirken, die kognitiver, motivationaler, affektiver und selektiver Natur sein und wiederum mediierend auf menschliches Handeln wirken können. Beispiele bezüglich der Prozessformen enthält Tabelle 3.3, wobei diese auch verdeutlicht, dass die vier von Bandura unterschiedenen

Prozesse nicht immer ganz trennscharf sind. Trotzdem verdeutlicht Bandura mit dieser Aufstellung die enorme Bandbreite von Faktoren, auf die Selbstwirksamkeitserwartungen im Rahmen der sozial-kognitiven Theorie wirken können.

Tabelle 3.3.: Wirkungen von Selbstwirksamkeitserwartungen auf andere Konstrukte, systematisiert nach der Art des jeweils hauptsächlich adressierten mentalen Prozeses (Bandura, 1997, S. 117-161).

Prozess	**Konstrukte**
kognitiv	Visualisierung/Antizipation von Szenarien, analytisches/schlussfolgerndes Denken (Abschätzen von Handlungskonsequenzen)
motivational	Ausdauer, Einsatzbereitschaft, Zeitinvestitionen, Kausalattribuierungen, Handlungsergebniserwartungen, Zielsetzungen
affektiv	Stress-, Angst- und Belastungserleben, Emotionen
selektional	Wahlentscheidungen

Selbstwirksamkeitserwartungen spielen aufgrund der angeführten positiven Effekte in vielen Bereichen eine wichtige Rolle – sei es im Sport, im Gesundheitswesen oder im Bildungsbereich (vgl. Bandura, 1997). Zur Überprüfung einer Vielzahl der postulierten Zusammenhänge wurde eine enorme Anzahl an Studien durchgeführt, die zu erheblicher empirischer Evidenz führen. Eine domänenspezifische Darstellung der Forschungsbefunde kann hier nicht geleistet werden. Für einen Einblick sei z. B. auf Skinner, Chapman und Baltes (1988, S. 117), Skinner (1996, S. 549), Bandura (1997), Maddux (2002) oder Schwarzer und Jerusalem (2002) verwiesen.

Auch für den akademischen/schulischen Bereich, dem diese Arbeit im weitesten Sinne zugerechnet werden kann, ist die Bedeutung des Konstruktes schnell erkannt worden, gerade weil Selbstwirksamkeitserwartungen vermittelt über motivationale Prozesse wie Ausdauer, Resilienz, etc. eine Vielzahl von Zielkriterien wie Leistungen, Zielsetzungen, Wahlentscheidungen, Interessen, Lernstrategien etc. von Lernenden beeinflussen können (vgl. für einen Überblick z. B. Schunk, 1991; Pajares, 1996; Zimmerman, 2000; Valentine, DuBois & Cooper, 2004). Hinsichtlich der Bedeutung von Selbstwirksamkeitserwartungen für eine tatsächlich zu erbringende Leistung bemerkt Bandura (1997, S. 35): „A capability is only as good as its execution. The self-assurance with which people approach and manage difficult tasks determines whether they make good or poor use of their capabilities. Insidious self-doubts can easily overrule the best of skills.“ Den Einfluss von Selbstwirksamkeitserwartungen auf die Performanz von Schülerinnen und Schülern haben beispielsweise Bouffard-Bouchard, Parent und Larivée (1991) untersucht. Aufgrund der Bedeutung für diese Arbeit wird für den Bereich der Lehrer-Selbstwirksamkeitserwartungen detaillierter herausgearbeitet, welche Zusammenhänge zu anderen Konstrukten wie dem Unterrichtshandeln, der Lehrergesundheit oder den Schülerleistungen bisher erforscht wurden (vgl. Abschnitt 4.4), weil sich daraus konkreter

die Bedeutung von Selbstwirksamkeitserwartungen von (Physik-)Lehrpersonen ableiten lässt.

3.6. Zur Erfassung des Konstruktes

Sollen Selbstwirksamkeitserwartungen erfasst werden, so gilt es, die dargelegten Eigenschaften des Konstruktes zu berücksichtigen, d. h. im Besonderen die Subjektivität/Selbstreferentialität, den (konkreten) Handlungs- und Schwierigkeitsbezug, die Situationsspezifität/Kontextabhängigkeit sowie die Domänenspezifität des Konstruktes (vgl. Unterabschnitt 3.2.2). Weiterhin müssen Operationalisierungen von denen benachbarter oder ähnlicher Konstrukte dezidiert abgegrenzt werden (vgl. Abschnitt 3.4). Dazu erscheint es ratsam, die „agent-means"-Relation des Konstruktes im Blick zu behalten (vgl. Unterabschnitt 3.1.3), sowie das Spezifitätsniveau der Messung vom inhaltlichen Niveau der interessierenden Domäne zu unterscheiden (vgl. Unterabschnitt 3.4.3). Im Folgenden wird zunächst die Problematik der Wahl des Spezifitätsniveaus sowie der konkreten Operationalisierung des gewählten Niveaus genauer analysiert. Anschließend werden Banduras Empfehlungen zur Messung des Konstruktes dargestellt, die sich im Wesentlichen aus den Ausführungen der letzten Abschnitte des Kapitels ergeben.

3.6.1. Spezifitätsniveau

Auch wenn laut Bandura Selbstwirksamkeitserwartungen auf unterschiedlichen Spezifitätsniveaus („generality") gemessen werden können (vgl. Bandura, 1977a, S. 194; Bandura, 1997, S. 42 f.), unterscheiden sich Messungen auf unterschiedlichen Niveaus hinsichtlich ihrer Prädiktivität und damit Aussagekraft: „[A]nalyses of how efficacy beliefs affect actions rely on microanalytic measures rather than global indices of personality traits or motives of effectance. It is no more informative to speak of self-efficacy in general terms than to speak of nonspecific behavior" (Bandura, 1997, S. 14). Je globaler die Messung angelegt ist, desto größeren Einfluss haben Erhebungskontexte. Wird beispielsweise die allgemeine Selbstwirksamkeitserwartung in einem akademischen Setting erhoben, so werden wahrscheinlich zur Einschätzung eher akademische Erfahrungen herangezogen (vgl. Bandura, 1997, S. 40). In dem Zusammenhang ist tatsächlich fraglich, inwiefern allgemeine Selbstwirksamkeitserwartungen überhaupt erfasst werden können: „Vague items obscure what, in fact, is being measured" (Bandura, 1997, S. 40). Daher ist es von größter Relevanz, die Assoziationen von Probanden zu erfragen, wenn in irgendeiner Form „globale" Einschätzungen erhoben werden sollen, um die Passung zur zu untersuchenden Domäne zu eruieren (vgl. Bandura, 1997, S. 40 f.). Festzuhalten bleibt, dass mit steigender Generalität der Messung, die Erklärungsmacht und Prädiktivität in der Regel abnimmt (vgl. Bandura, 1997, S. 41). Damit einher geht die strukturelle Eigenschaft der sehr lose ausgeprägten Hierarchie globaler und spezifischer

Selbstwirksamkeitserwartungen: „General indices of personal efficacy bear little or no relation to particular activity domain or to behavior“ (Bandura, 1997, S. 42).

Aus den genannten Argumenten soll jedoch nicht abgeleitet werden, dass Selbstwirksamkeitserwartungen nur für sehr spezifische Handlungen und bezüglich spezifizierter situationaler Variablen erfasst werden können (vgl. Bandura, 1997, S. 49), da dies zu einer fragmentarischen Erfassung führen würde. Zu beachten gilt, dass Selbstwirksamkeitserwartungen auf jedem Niveau komplexe Handlungen adressieren. Grundsätzlich unterscheidet Bandura (1997, S. 49) drei geeignete Spezifitätsniveaus zur Erfassung von Selbstwirksamkeitserwartungen:

> „One can distinguish among three levels of generality of assessment. The most specific level measures perceived self-efficacy for a particular performance under a specific set of conditions. The intermediate level measures perceived self-efficacy for a class of performances within the same activity domain under a class of conditions sharing common properties. And finally, the most general and global level measures belief in personal efficacy without specifying the activities or the conditions under which they must be performed.“

Wie bereits erwähnt, steigt die Prädiktivität einer Messung in der Regel mit zunehmendem Spezifitätsniveau der Selbstwirksamkeitsskala, allerdings muss zusätzlich die Passung zum Spezifitätsniveau des Zielkriteriums beachtet werden: „The optimal level of generality at which self-efficacy is assessed varies depending on what one seeks to predict and the degree of foreknowledge of the situational demands“ (Bandura, 1997, S. 49).

Bong (2002) misst für Mathematik und Englisch jeweils „problemspezifische“, „aufgabenspezifische“ und „fachspezifische“ Selbstwirksamkeitserwartungen und kann zeigen, dass diese drei Niveaus unabhängige Faktoren bilden. Alle drei Konstrukte sagen Ergebnisse in Leistungstests (korrespondierendes Spezifitätsniveau zu den problemspezifischen Selbstwirksamkeitserwartungen) verlässlich vorher, wobei die Modelle für problem- und aufgabenspezifische Selbstwirksamkeitserwartungen leicht überlegen sind. Dieses Ergebnis kann vorsichtig dahingehend interpretiert werden, dass unter Umständen auch Skalen auf einem mittleren Spezifitätsniveau verlässliche Prädiktoren spezifischerer Zielvariablen darstellen.[7]

Abschließend sei erwähnt, dass das Spezifitätsniveau der Domäne vom Spezifitätsniveau der Messung unterschieden werden muss (vgl. Unterabschnitt 3.4.3 zum Selbst-

[7] Zusätzlich zeigt die Studie von Bong (2002), dass es von der Domäne abhängt, ob Schülerinnen und Schüler jeweils ihre problem- oder aufgabenspezifischen Selbstwirksamkeitserwartungen höher einschätzten. Für Mathematik schätzen sie sich bezüglich konkreter Probleme schlechter ein als bezüglich bestimmter Aufgabentypen. Ob es sich bei dieser Beobachtung um ein Artefakt handelt oder tatsächlich eine Domänenabhängigkeit des Antwortverhaltens auf unterschiedlichen Spezifitätsniveaus vorliegt, muss zunächst offen bleiben (vgl. Bong, 2002, S. 157).

konzept), wobei sich bisherige Ausführungen auf das Messniveau beziehen. Das Spezifitätsniveau der Domäne legt nicht zwangsläufig das Spezifitätsniveau der Messung fest (vgl. u. a. Bandura, 1997, S. 49 f.; Bong & Skaalvik, 2003). Für globalere Kontexte müssen jedoch gegebenenfalls Urteile bezüglich der Repräsentativität der Auswahl an spezifischen Handlungen gefällt werden. Aus forschungsökonomischen Gründen und weil hochgradig zugespitzte Einschätzungen selten, situationsübergreifende Einschätzungen jedoch meist eher von Interesse sind, werden Selbstwirksamkeitserwartungen häufig auf einem mittleren Spezifitätsniveau der Messung erfasst (vgl. Bandura, 1997, S. 50).

3.6.2. Anleitung zur Konstruktion und Validierung von Selbstwirksamkeitserwartungsskalen

Nachfolgend wird eine (teils kommentierte) Kurzfassung des „Guide for Constructing Self-Efficacy Scales“ (Bandura, 2006) dargestellt.[8] Die Anleitung enthält Hinweise bezüglich der Konstruktion und Validierung von Skalen zur Erfassung von Selbstwirksamkeitserwartungen, die im Folgenden mit Bezug auf die in Abschnitt 3.2.2 herausgearbeiteten Eigenschaften des Konstruktes vorgestellt werden

Allgemeines

- **Abgrenzung zu anderen Konstrukten:** „Perceived self-efficacy should [...] be distinguished from other constructs such as *self-esteem*, *locus of control*, and *outcome expectancies*“ (Bandura, 2006, S. 309). Insbesondere auch bezüglich des Selbstkonzeptes gilt es, das Konstrukt der Selbstwirksamkeitserwartungen entsprechend der dargestellten Kriterien (vgl. Unterabschnitt 3.4.3) abzugrenzen.

Subjektivität/Selbstreferentialität und Gegenwartsbezug

- **Einschätzung eigener Fähigkeiten und keine Absichten widerspiegeln:** „The items should be phrased in terms of *can do* rather than *will do*. *Can* is a judgment of capability; *will* is a statement of intention“ (Bandura, 2006, S. 308, Hervorhebungen im Original).
- **Im Präsens formulieren:** „People are asked to judge their operative capabilities as of now, not their potential capabilities or their expected future capabilities. It is easy for people to imagine themselves to be fully efficacious in some hypothetical future“ (Bandura, 2006, S. 312 f.).

[8] Dieser beruht im Wesentlichen auf etwas älteren Ausführungen, die in seinem Hauptwerk „Self-Efficacy. The Exercise of Control“ abgedruckt sind (vgl. Bandura, 1997, S. 42-54).

Handlungs- und Schwierigkeitsbezug/Auswahl von Handlungen

- **Facettenreiche Skalen konstruieren:** „Behavior is better predicted by people's beliefs in their capabilities to do whatever is needed to succeed than by their beliefs in only one aspect of self-efficacy relevant to the domain“ (Bandura, 2006, S. 310). Zusätzlich muss eine Entscheidung darüber getroffen werden, welche Handlungen eine Domäne in ihrem Kern am besten charakterisieren.
- **Beeinflussbare Verhaltensfaktoren auswählen:** „A comprehensive self-efficacy assessment would be linked to the behavioral factors over which people can exercise some control“ (Bandura, 2006, S. 310).
- **Schwierige Verhaltensweisen und schwierigkeitserzeugende Handlungsbarrieren auswählen:** „Perceived efficacy should be measured against levels of task demands that represent gradations of challenges or impediments to successful performance“ (Bandura, 2006, S. 311). Auch um Deckeneffekte zu vermeiden, schlägt Bandura (1997, S. 43) vor, Hindernisse und Herausforderungen in den Items zu formulieren.
- **Authentische Hürden und Schwierigkeiten aus Interviews ableiten:** „People are asked in open-ended interviews and pilot questionnaires to describe the things that make it hard for them to perform the required activities regularly. The identified challenges or impediments are built into the efficacy items“ (Bandura, 2006, S. 311).

Domänenspezifität

- **Domäne analysieren:** „The construction of sound efficacy scales relies on a good conceptual analysis of the relevant domain of functioning. Knowledge of the activity domain specifies which aspects of personal efficacy should be measured“ (Bandura, 2006, S. 310).
- **Domänenrelevante Verhaltensfaktoren auswählen:** [S]elf-efficacy scales must be tailored to activity domains and assess the multifaceted ways in which efficacy beliefs operate within the selected activity domain. The efficacy scales must be linked to factors that, in fact, determine quality of functioning in the domain of interest“ (Bandura, 2006, S. 310 f.).

Strukturelemente, Format, Durchführung der Messung

- **Mehrstufiges Antwortformat wählen:** „Scales that use only a few steps should be avoided because they are less sensitive and less reliable“ (Bandura, 2006, S. 312).

- **Unipolare Skala verwenden:** „Efficacy scales are unipolar [...]. They do not include negative numbers because a judgment of complete incapability (0) has no lower gradations“ (Bandura, 2006, S. 312).
- **Instruktionstexte der eigentlichen Befragung voranstellen:** „Preliminary instructions should establish the appropriate mindset that participants should have when rating the strength of belief in their personal capability“ (Bandura, 2006, S. 312).
- **Sozial erwünschtem Antwortverhalten vorbeugen:** „Self-efficacy judgments are recorded privately without personal identification to reduce social evaluative concerns. [...] Respondents are informed that their responses will remain confidential and be used only [...] by the research staff. If the scale is labeled, use a nondescript title such as „Appraisal Inventory“ rather than Self-Efficacy. To encourage frank answers, explain to the respondents the importance of their contribution to the research. Inform them that the knowledge it provides will increase understanding and guide the development of programs designed to help people to manage the life situations with which they have to cope“ (Bandura, 2006, S. 314).
- **Items nach ansteigender Schwierigkeit oder zufällig durchmischt darbieten, um Überschätzungen zu vermeiden:** „[A] descending format ordering the items from most to least difficult task demands tended to produce slightly higher self-efficacy appraisals than did an ascending or random order“ (Bandura, 1997, S. 47).

Qualitätssicherung

- **Items pilotieren, überarbeiten & eliminieren:** „Pretest the items. Discard those that are ambiguous or rewrite them. Eliminate items where most people are checking the same response point. Such items do not differentiate among respondents. Items on which the vast majority of respondents check the maximum efficacy category lack sufficient difficulty, challenge, or impediments to distinguish levels of efficacy among respondents. Increase the difficulty level by raising the level of challenge in the item“ (Bandura, 2006, S. 315).
- **Itemkorrelationen und Homogenität der Skala überprüfen:** „The items tapping the same domain of efficacy should be correlated with each other and with the total score. Factor analyses verify the homogeneity of the items“ (Bandura, 2006, S. 315 f.).
- **Interne Konsistenz überprüfen:** „Internal consistency reliabilities should be computed using Cronbach's alpha“ (Bandura, 2006, S. 316).

- **Validität überprüfen:** „[S]elf-efficacy scales should have face validity. They should measure what they purport to measure, that is, perceived capability to produce given attainments. But they should also have discriminative and predictive validity“ (Bandura, 2006, S. 318 f.).

- **Konstruktvalidität prüfen:** „Construct validation is a process of hypothesis testing. People who score high on perceived self-efficacy should differ in distinct ways from those who score low in ways specified by the theory. Verifications of predicted effects provide support for the construct's validity (Bandura, 2006, S. 319).

3.7. Exkurs: Kollektive Selbstwirksamkeitserwartungen

> „People's shared beliefs in their collective power to produce desired results is a key ingredient of collective agency.“
>
> *Bandura, 2006, S. 316*

Auch kollektives Handeln kann mit Hilfe der sozial-kognitiven Theorie beschrieben werden. Dann wird das Modell dadurch erweitert, dass individuelle Kognitionen und Handlungen durch gemeinschaftlich geteilte Überzeugungen/Werte und gemeinsam ausgeführte Handlungen ersetzt werden. Kollektive Selbstwirksamkeitserwartungen definiert Bandura als „people's shared belief in their capabilities to produce effects collectively“ (Bandura, 1997, S. 6). Kollektive Selbstwirksamkeitserwartungen sind jedoch mehr als die Summe der individuellen Selbstwirksamkeitserwartungen der Gruppenmitglieder, nämlich „an emergent group-level attribute that is the product of coordinative and interactive dynamics“ (Bandura, 1997, S. 6). Die Bedeutung des Konstruktes der kollektiven Selbstwirksamkeitserwartungen fasst Bandura (2006, S. 318) wie folgt zusammen: „The findings taken as a whole show that the higher the perceived collective efficacy, the higher the groups' motivational investment in their undertakings, the stronger their staying power in the face of impediments and setbacks, and the greater their performance accomplishments.“ In den letzten Jahren ist diesem Konstrukt daher in Bereichen viel Aufmerksamkeit geschenkt worden, in denen Teamleistungen von besonderer Relevanz sind. Im Bildungsbereich stehen beispielsweise kollektive Lehrer-Selbstwirksamkeitserwartungen im Fokus des Interesses (vgl. u. a. Schwarzer & Schmitz, 1999; Goddard, Hoy & Hoy, 2004).

3.8. Zusammenfassung

Selbstwirksamkeitserwartungen sind im deutschen Sprachraum definiert als „die subjektive Gewissheit, neue oder schwierige Anforderungssituationen auf Grund eigener Kompetenz bewältigen zu können“ (Schwarzer & Jerusalem, 2002, S. 35), wobei diese Definition die Merkmale des Konstruktes (vgl. Unterabschnitt 3.2.2) in besonderer Weise aufgreift. Selbstwirksamkeitserwartungen bilden das zentrale Konstrukt Banduras sozial-kognitiver Theorie und können daher als solide theoretisch eingebettet bezeichnet werden. Das konzeptionelle Framework ist dabei hochgradig formal, auch wenn es dem Ideal Cronbachs und Meehls (1955) im Sinne eines „starken Programms“ nicht völlig entspricht (vgl. Unterabschnitt 2.1.2/Konstruktvalidität). Es stützt sich jedoch auf eine Vielzahl theoretischer Annahmen, aus denen prüfbare Aussagen bezüglich der Genese, der Struktur, der Funktion und der Wirkungen des Konstruktes sowie bezüglich der Einflussfaktoren auf das Konstrukt abgeleitet werden können (vgl. Bandura, 1997, S. 10). Die Bedeutung des Konstruktes ergibt sich dabei nicht nur aus theoretischer Perspektive, sondern auch durch die Vielzahl empirischer Belege bezüglich postulierter Thesen wie dieser: „It is hypothesized that expectations of personal efficacy determine whether coping behavior will be initiated, how much effort will be expended, and how long it will be sustained in the face of obstacles and aversive experiences“ (Bandura, 1977a, S. 191). Die praktische Bedeutung wird damit schnell einsichtig – für den Bereich der Lehrerbildung soll sie im nachfolgenden Kapitel (vgl. insbesondere Abschnitt 4.4) detaillierter betrachtet werden.

Sollen Selbstwirksamkeitserwartungen mit Hilfe von Items erfasst werden, so muss die Natur des Konstruktes insbesondere in Abgrenzung zu ähnlichen Konstrukten für Operationalisierungen reflektiert werden. Aus der Analyse der Definitionen des Konstruktes sowie grundlegender theoretischer Annahmen der sozial-kognitiven Theorie Banduras können eine Vielzahl von Konstrukteigenschaften benannt und erläutert werden, die zur detaillierten Charakterisierung des Konstruktes beitragen. Außerdem erweist sich der Ansatz von Skinner (1996) als besonders tragfähig, wonach Selbstwirksamkeitserwartungen als „agent-means“-Relation aufgefasst werden können. Dadurch ergibt sich nicht nur eine zusätzliche Charakterisierungsmöglichkeit, sondern ein eindeutiges Kriterium zur Abgrenzung von Selbstwirksamkeitserwartungen von anderen Konstrukten. Beispielsweise kann dieses zur Abgrenzung zu Handlungsergebniserwartungen oder zum „locus-of-control“-Konstrukt verwendet werden, da diese jeweils als eine „means-ends“-Relation konzipiert sind. Hinsichtlich der Beziehung zu Handlungsergebniserwartungen sollte beachtet werden, dass diese ebenfalls zur Aufklärung von Verhalten beitragen, auch wenn der Anteil ggf. geringer ist und deren Erhebung daher meist vernachlässigt oder unterlassen wird. Unter Umständen kann jedoch bei der Erhebung von Selbstwirksamkeitserwartungen eine implizite Konfundierung mit dem Konstrukt der Handlungsergebniserwartungen vorliegen: „[W]hen researchers have examined only

one aspect of control, they have tended to do so under conditions in which the unmeasured aspect is already present and high. For example, the effects of noncontingency are assessed in situations in which the actions needed to operate the contingencies are guaranteed to be in everyone's repertoire. Or, the effects of self-efficacy are examined in contexts (e.g., schools) in which response-outcome contingencies are known to be high“ (Skinner, 1996, S. 559 f.). Die Abgrenzung zum Konstrukt des Selbstkonzepts erweist sich als komplexes Unterfangen, weil beide Konstrukte eine Reihe gemeinsamer Charakteristika aufweisen. Selbstwirksamkeitserwartungen beziehen sich jedoch nie auf affektive Einschätzungen eigener Fähigkeiten und adressieren ein konkreteres Niveau als das Selbstkonzept bezüglich derselben Domäne.

Um Selbstwirksamkeitserwartungen angemessen zu erfassen und spezifische Testwertinterpretationen entwickelter Skalen zu validieren, können schließlich eine Vielzahl konkreter Hilfestellungen angegeben werden. Zusammenfassend ergibt sich: „An adequate expectancy analysis [...] requires detailed assessment of the magnitude, generality, and strength of efficacy expectations commensurate with the precision with which behavioral processes are measured“ (Bandura, 1977a, S. 194).

4. Von „teacher efficacy“ zu „Lehrer-Selbstwirksamkeitserwartungen“

„In these days of hard-nosed accountability, teachers' sense of efficacy is an idea that neither researchers nor practitioners can afford to ignore.“

Tschannen-Moran & Woolfolk Hoy, 2001, S. 803

Im vorangegangenen Kapitel wurde das zentrale Konstrukt der sozial-kognitiven Theorie Banduras – das Konstrukt der Selbstwirksamkeitserwartungen („self-efficacy beliefs“) – ausführlich beschrieben, definitorisch gefasst sowie von anderen Konstrukten abgegrenzt. Im nachfolgenden Kapitel wird die Adaption des Konstruktes für den Lehrkontext im Fokus stehen. Dabei wird festgelegt, dass bei einer gelungenen domänenspezifischen Formulierung des Konstruktes die Rede von Lehrer-Selbstwirksamkeitserwartungen („teacher self-efficacy beliefs“) sein soll. Als „gelungen“ kann die Adaption genau dann gelten, wenn die aus der Theorie Banduras abgeleiteten Eigenschaften des Konstruktes (vgl. u. a. Abschnitt 3.2.1) hinreichend berücksichtigt werden.

Beachtet werden muss, dass sich im englischen Sprachraum zunächst das Konstrukt „teacher efficacy“ bzw. „teachers' sense of efficacy“ etablierte, welches zwar formal mit der Theorie Banduras verknüpft wurde, jedoch bei genauerer Analyse nicht den anzulegenden Kriterien genügt. In den weiteren Ausführungen wird daher akribisch zwischen „teacher efficacy“ und „teacher self-efficacy beliefs“ (Lehrer-Selbstwirksamkeitserwartungen) im Sinne Banduras unterschieden. Diese Vorgehensweise geht auf einen Vorschlag von Dellinger u. a. (2008) zurück. Es wird sich zudem herausstellen, dass der Begriff „teacher efficacy“ nicht immer einheitlich verwendet wird.

Trotz der zu konstatierenden Inkompatibilität mit der sozial-kognitiven Theorie Banduras ist das „teacher-efficacy“-Konstrukt in der Forschungslandschaft nicht zuletzt aufgrund seiner historischen Entwicklung, insbesondere aufgrund seines Erfolges in Form von „überzeugenden“ ersten Forschungsergebnissen sowie der in der Folge entwickelten Vielzahl von Operationalisierungen sehr präsent. In diesem Kapitel soll daher zunächst eine Rekonstruktion des historischen Ablaufes der definitorischen Fassung des „teacher-efficacy“-Konstruktes erfolgen, um die Gemeinsamkeiten und Unterschiede beider Konstrukte nachvollziehen zu können. Dieses Vorgehen dient außerdem dazu, potentielle Schwierigkeiten bei der Adaption der theoretischen Ideen Banduras sicht-

bar zu machen und zu benennen, um diesen im Rahmen der Neukonstruktion eines Instrumentes aktiv begegnen zu können. In diesem Sinne ist auch die Überschrift des Kapitels zu verstehen – ausgehend vom Konzept „teacher efficacy“ ist in der Forschungsgemeinschaft ein langwieriger und mühsamer Entwicklungsprozess in Gang geraten, in dem sukzessive auch inhaltliche Problemstellen existierender Instrumente erkannt werden. Die Einsicht, dass Operationalisierungen häufig auf einer missverständlichen theoretischen Fundierung aufbauen, kommt einer Errungenschaft gleich, an deren Ende das Forschungsdesiderat eines Instrumentes, welches tatsächlich „teacher self-efficacy beliefs“ (Lehrer-Selbstwirksamkeitserwartungen) im Sinne Banduras abbildet, formuliert werden muss. Im nachfolgenden Kapitel wird diese Behauptung anhand der Resultate einer eingehenden Prüfung einer Reihe von Instrumenten auf ihre Passung zu den theoretischen Prämissen Banduras anschaulich abgeleitet.

Letztlich ergibt sich aus den Analysen dieses Kapitels, dass viele Studienergebnisse, die standardmäßig als empirische Evidenz zur Veranschaulichung der Bedeutung des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen zitiert werden, weil sie Zusammenhänge zu relevanten Konstrukten wie dem Unterrichtshandeln einer Lehrperson aufdecken, auf Instrumenten beruhen, deren Qualität begründet in Zweifel steht. Es muss daher – entgegen eines ersten Eindrucks, der sich beispielsweise aus dem hohen Publikationsaufkommen speist – festgestellt werden, dass die empirische Basis bezüglich der theoretisch abgeleiteten Bedeutung des Konstruktes nicht ausreichend ist. Für theoretisch plausible Aussagen, wie die nachfolgende, können daher in der Regel nur vereinzelt empirische Indizien angeführt werden:

> „Wenig selbstwirksame Lehrer beispielsweise neigen dazu, einfache aber sichere Unterrichtsaktivitäten zu bevorzugen, da sie sich durch innovative oder komplexe Planungen leicht überfordert fühlen, sie kümmern sich kaum um lernschwache Schüler und sind insgesamt wenig motiviert, guten und verständlichen Unterricht zu halten, da sie sich auch wenig zutrauen. Lehrer mit hoher Selbstwirksamkeit gestalten einen insgesamt herausfordernden Unterricht, sie unterstützen Schüler bei der Erzielung von Lernfortschritten und haben mehr Geduld sowie Zuwendung für lernschwache Schüler, weil sie sich selbst mehr zutrauen, stärker motiviert sind und eine hohe Verantwortung für einen erfolgreichen und verständlichen Unterricht empfinden“ (Schwarzer & Jerusalem, 2002, S. 40).

4.1. Definitionsversuche

„At best the notion of teacher efficacy is complex; at worst it is confused."

Woolfolk & Hoy, 1990, S. 81

Die „Geburtsstunde" des Konstruktes „teacher efficacy" beschreiben Tschannen-Moran, Woolfolk Hoy und Hoy (1998, S. 202) wie folgt: „Twenty years ago researchers from the RAND organization added two items to an already extensive questionnaire [...]. It may have been simply a hunch or a whim, but they got results, powerful results, and the concept of teacher efficacy was born." Mittlerweile sind seit der Veröffentlichung des erwähnten Fragebogens im Jahre 1976 durch die RAND-Forschungsgruppe[1] vier Jahrzehnte verstrichen und man könnte ein wohldefiniertes und strukturiertes Forschungsfeld erwarten. Leider kann von einer begrifflichen und theoretischen Klarheit jedoch nicht die Rede sein. Meines Erachtens resultiert dies aus der wenig elaborierten Verknüpfung des „teacher-efficacy"-Konstruktes mit der sozial-kognitiven Theorie Banduras. Im Wesentlichen wurde die Inbezugsetzung des Konstruktes „teacher efficacy" mit der Theorie Banduras nämlich dazu verwendet, das Konstrukt mit einem bekannten und geschätzten Label („Bandura/self-efficacy") zu versehen. In der Folge wurde dann allzu häufig nach dem Motto verfahren „Wo Bandura draufsteht, steckt auch Bandura drin!" und fälschlicherweise angenommen, dass ein Instrument mit einem solchen Label, diesem auch inhaltlich gerecht wird und Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras erfasst werden.

Tatsächlich orientieren sich die ursprünglichen und von Tschannen-Moran u. a. (1998) erwähnten Items trotz der Verwendung des für Banduras Theorie typischen Begriffs „efficacy" nachweislich und explizit an einem anderen theoretischen Überbau als der sozial-kognitiven Theorie Banduras, nämlich an der *„locus-of-control"-Theorie* nach Rotter (vgl. Abschnitt 3.4.6). Allein aus diesem Grund erscheint es äußerst fragwürdig, das Konstrukt „teacher efficacy" nachträglich und ohne ausreichende theoretische Reflexion mit einer anderen Theorie in Verbindung zu bringen. Die Bemühungen der Vereinigung der beiden konzeptionellen Stränge wurde zunächst positiv aufgenommen und wenig hinterfragt. Dieses Vorgehen führte bedauerlicherweise jedoch *nicht* zu einer klaren Herausarbeitung der Unterschiede und damit zu einer begrifflichen Differenzierung, sondern gewissermaßen zu einer Gleichsetzung beider Theorien und damit zu noch größerer inhaltlicher Konfusion. Festzuhalten bleibt, dass unterschiedliche, teils sehr ähnliche Operationalisierungen des Konstruktes „teacher efficacy" daher auf völlig unterschiedlichen theoretischen Annahmen beruhen.

[1] *R*easearch *AN*d *D*evelopment Corporation; siehe: http://www.rand.org/ (letzter Zugriff: 15.07.2015)

Folgt man den Irrungen und Wirrungen der Forschungstradition, dem Chaos begrifflicher Setzungen und Umdeutungen, sowie etlichen Versuchen der Operationalisierung des Konstruktes, so können am Ende dieses Prozesses tatsächlich deutlicher die (Gemeinsamkeiten und) Unterschiede des etablierten „teacher-efficacy“-Konstruktes und eines Konstruktes, welches Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras repräsentiert, formuliert werden. Die Rekonstruktion der begrifflichen und theoretischen Genese des Konstruktes wird deshalb nachfolgend beschrieben.

Die zahlreich dokumentierten Probleme bei der Erfassung des Konstruktes (vgl. nachfolgender Abschnitt 4.3) werden dagegen erst nach und nach auf unzureichende konzeptionelle Vorarbeiten beispielsweise eine unzureichende Reflexion der tatsächlich zugrunde liegenden Theorie zurückgeführt. Mittlerweile gibt es einige Bemühungen, das Konstrukt neu und präzise zu fassen und dabei eine klare Trennung beider theoretischer Positionen zu Grunde zu legen (vgl. u. a. Tschannen-Moran u. a., 1998; Dellinger u. a., 2008).

4.1.1. Definition nach Rotter

Im Jahr 1976 definieren die bereits von Tschannen-Moran u. a. (1998) angesprochenen Forschenden der RAND-Gruppe „the teachers' sense of efficacy“ explizit mit Bezug auf Rotters Theorie[2] als “the extent to which the teacher believes he or she has the capacity to produce an effect on the learning of students“ (Armor u. a., 1976, S. 23). Die beiden nachfolgenden Items operationalisieren das Konstrukt, wobei der Gesamtscore beider Items ein Maß für dessen Ausprägung liefern soll:

- RAND-Item I: „When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his or her home environment.“
- RAND-Item II: „If I try really hard, I can get through to even the most difficult or unmotivated students.“

Der Bezug zur Rotter'schen Theorie (vgl. Abschnitt 3.4.6) wird insofern deutlich, als dass die Items von Lehrpersonen erfragen, inwiefern sie die Einflussnahme auf ihre schulbezogene Umwelt als in ihrer Macht liegend betrachten oder durch externe Faktoren bestimmt sehen und damit relativ globale Urteile über kausale Zusammenhänge im Fokus des Interesses stehen. Konkret geht es in den Items darum, ob Lehrkräfte der Meinung sind, Einfluss auch auf ihre schwierigsten oder unmotiviertesten Schülerinnen und Schüler nehmen zu können („internal locus of control“, RAND II) oder ob sie der

[2] „The standard discussion of efficacy, on which we based our instruments, is J. B. Rotter (1966)“ (Armor u. a., 1976, S. 23). Dieser Bezug ist insofern bemerkenswert, als bei Rotter (1966) der Begriff „efficacy“ nicht ein einziges Mal Verwendung findet.

Meinung sind, dass im Allgemeinen die Leistungen und die Motivation von Schülerinnen und Schülern eher durch äußere Faktoren wie das Elternhaus („external locus of control“, RAND I) bestimmt werden.

Zu beachten ist, dass RAND-Item I den externalen „locus of control“ mit einer allgemeinen Einschätzung verknüpft (Aussage über Lehrpersonen im Allgemeinen), während RAND-Item II eine Einschätzung ganz persönlicher Möglichkeiten in Kombination mit einem internalen „locus of control“ einfordert. Deshalb können die beiden Items gewissermaßen als zwei Pole auf einem Kontinuum interpretiert werden (von „Lehrpersonen können grundsätzlich fast nichts ausrichten“ hin zu „Es liegt allein in meiner Macht, selbst die schwierigsten Schülerinnen und Schüler zu erreichen“).[3] Insgesamt kann die Operationalisierung als sehr gute Adaption der Rotter'schen Theorie beurteilt werden (vgl. Abschnitt 3.4.6). Konsequenterweise wäre das Konstrukt daher eher als „teacher locus of control“ zu bezeichnen.

Wie es für Operationalisierungen eines „locus-of-control“-Konstruktes typisch ist, bilden die beiden RAND-Items im Sinne Skinners (1996) jeweils „means-ends“-Beziehungen ab, auch wenn die Handlungen („means“) in beiden Fällen wenig konkret formuliert werden („can't do much“ bzw. „can get through“). Lehrer-Selbstwirksamkeitserwartungen werden jedoch entsprechend der Theorie Banduras als „agent-means“-Relation formuliert. Damit unterscheiden sich „teacher efficacy“ in der hier vorgestellten Lesart als „teacher locus of control“ und Lehrer-Selbstwirksamkeitserwartungen fundamental voneinander. Der Unterschied beider Konstrukte wird zusätzlich darin deutlich, dass Lehrer-Selbstwirksamkeitserwartungen per Definition selbstreferentiell sind, während insbesondere das erste RAND-Item, wie die Ausführungen zeigen, dieser Forderung nicht gerecht wird.

4.1.2. Definition nach Bandura

Bandura (1997) selbst nimmt keine dezidierte Anwendung der sozial-kognitiven Theorie auf den Lehrkontext, also auch keine genauere Beschreibung des Konstruktes „teacher self-efficacy beliefs“ vor. Allerdings geht er grundsätzlich davon aus, dass sich das Konstrukt für den Lehrkontext als fruchtbar erweist (vgl. Bandura, 1997, S. 240-243) und spricht sich für eine unterrichtsfachbezogene Erfassung des Konstruktes aus.

Interessant ist, dass er direkt Stellung zur Konzeptualisierung und Erhebung von „teacher efficacy“ nach Armor u. a. (1976) bezieht.[4] Einer seiner Kritikpunkte liegt in der

[3] Entsprechend der Ausführungen in Abschnitt 3.4.6 sprechen Ergebnisse einer Reihe von Studien eher dafür, dass es sich um zwei getrennte Dimensionen handelt: Ashton, Buhr und Crocker (1984) sowie Woolfolk und Hoy (1990) berichten z. B. nicht signifikante Korrelationen der beiden Items. Ashton und Webb (1986, S. 135), Coladarci und Fink (1995), zitiert nach, Tschannen-Moran u. a. (1998, S. 227) sowie Tschannen-Moran und Woolfolk Hoy (2001) berichten zwar signifikante, jedoch eher geringe Korrelationen von $.23^* \leq r(R_I/R_{II}) \leq .45^*$.

[4] Genauer bezieht er sich auf eine Nachfolgestudie der RAND-Gruppe (vgl. Berman, McLaughlin,

Erfassung durch lediglich zwei Items, was der Mehrdimensionalität im Sinne der facettenreichen Natur des Konstruktes widerspräche. Darüber hinaus würde die Erfassung eines Urteils über die *allgemeine* Wirksamkeit von Lehrpersonen, wie im ersten RAND-Item enthalten, der selbstreferentiellen Natur des Konstruktes (Urteil über *eigene* Fähigkeiten) nicht entsprechen (vgl. Unterabschnitt 4.1.1). Zusätzlich scheinen ihm die Items zu global und zu wenig auf die unterrichtliche Domäne bezogen zu sein (vgl. Bandura, 1997, S. 243). Zusammenfassend kann konstatiert werden, dass Bandura die RAND-Items (und auch das im nachfolgenden Unterabschnitt 4.1.3 beschriebene Instrument von Gibson und Dembo, 1984) als nicht theoriekonform und damit als ungeeignet hinsichtlich der Erfassung von Lehrer-Selbstwirksamkeitserwartungen erachtet. Die Gründe liegen in der unzureichenden Berücksichtigung der sich aus der Definition und theoretischen Einbettung ergebenden Eigenschaften des Konstruktes der Selbstwirksamkeitserwartungen für den gewählten Lehrkontext. Zur Beurteilung konkreter Operationalisierungen, die vorgeben, sich an Bandura zu orientieren, können die genannten Eigenschaften des Konstruktes (Selbstreferentialität, Handlungsbezug, „agent-means“-Relation, etc.) daher als Kriterien herangezogen werden (vgl. Abschnitt 3.2).

4.1.3. Theoriekonfusion: Wie Gibson und Dembo (1984) versuchen, „teacher effficacy“ und Banduras sozial-kognitive Theorie zusammenzubringen

Gibson und Dembo (1984) unternehmen den folgenschweren Versuch, die sozial-kognitive Theorie Banduras auf das „teacher efficacy“-Konstrukt, wie es durch die RAND-Items konzeptionalisiert wurde, zu übertragen. Sie beziehen sich auf dieselbe, oben genannte Definition wie die RAND-Forschenden und propagieren die Kompatibilität der RAND-Items mit der Bandura-Theorie: „If we apply Bandura's theory to the construct of teacher efficacy, outcome expectancy would essentially reflect the degree to which teachers believed the environment could be controlled, that is, the extent to which students can be taught given such factors as family background, IQ, and school conditions [RAND-Item I]. Self-efficacy beliefs would indicate teachers' evaluation of their abilities to bring about positive student change [RAND-Item 2]“ (Gibson & Dembo, 1984, 570, Anmerkung durch C. M.). In dieser Lesart misst das erste RAND-Item „outcome expectancy“ und das zweite RAND-Item „self-efficacy“ im Sinne Banduras.

Diese These steht jedoch in direktem Widerspruch zu mehreren analytischen Überlegungen. Erstens wurde soeben dargelegt, dass die RAND-Items sehr gut einen Ausschnitt des Konstruktes „teacher locus of control“ abbilden (vgl. Abschnitt 4.1.1) und dass sich „locus of control“ und „self-efficacy“ maßgeblich anhand der jeweils beschriebenen Relationen der Konstrukte unterscheiden lassen (vgl. Abschnitt 3.4.6). Während im Rahmen des „locus-of-control“-Konstrukts „means-ends“-Relationen dargestellt werden, werden

Bass, Pauly & Zellman, 1977), in der jedoch dieselben Items wie bei Armor u. a. (1976) verwendet werden.

Selbstwirksamkeitserwartungen durch „agent-means“-Relationen ausgedrückt. Zwar könnte das „locus-of-control“-Konstrukt als allgemeine Handlungsergebniserwartung interpretiert werden (vgl. Bandura, 1997), aber auch diese Argumentation ist hier nicht zulässig, da Handlungsergebniserwartungen im Sinne Banduras selbstreferentiell konzipiert sind und dies auf das erste RAND-Item nicht zutrifft (vgl. Abschnitt 3.4.7). Folglich bilden weder das erste RAND-Item Handlungsergebniserwartungen („outcome expectancy“) noch das zweite RAND-Item Selbstwirksamkeitserwartungen („self-efficacy“) im Sinne der sozial-kognitiven Theorie Banduras ab.

Trotzdem entwickeln Gibson und Dembo (1984) ihrer Behauptung und Überzeugung folgend ein Instrument – die „Teacher Efficacy Scale“ (TES) – welches die zwei von ihnen angenommenen Faktoren abbildet und Items enthält, die den RAND-Items inhaltlich und strukturell sehr ähnlich sind (vgl. Tabelle 4.1). Inkonsequenterweise nennen Gibson und Dembo (1984) die erste Komponente jedoch nicht „outcome expectancy“, sondern in Übereinstimmung mit Ashton u. a. (1984) „teaching efficacy (TE)“. Die zweite Komponente wird „personal teaching efficacy (PTE)“ genannt“ (vgl. Gibson & Dembo, 1984, S. 573 f.; Ashton u. a., 1984, S. 31 f.). Beide Faktoren werden in der Regel zu einem Gesamtscore kombiniert, der das Konstrukt „teachers' sense of efficacy“ bzw. „teacher efficacy“ abbilden soll (vgl. Gibson & Dembo, 1984, S. 572).

Tabelle 4.1.: Vergleich der RAND-Items (vgl. Armor u. a., 1976; Berman, McLaughlin, Bass, Pauly & Zellman, 1977) mit Beispielitems der Faktoren der „Teacher Efficacy Scale“ (TES) nach Gibson und Dembo (1984).

Items der „Teacher Efficacy Scale“	**RAND-Items**
Item 6, Teaching Efficacy (TE): „If students are not disciplined at home, they aren't likely to accept any discipline.“	**RAND-Item I:** „When it comes right down to it, a teacher really can't do much because most of a student's motivation and performance depends on his or her home environment.“
Item 14, Personal Teaching Efficacy (PTE): „When a student gets a better grade than he usually gets, it is usually because I found better ways of teaching that student.“	**RAND-Item II:** „If I try really hard, I can get through to even the most difficult or unmotivated students.“

Aufgrund der großen Ähnlichkeiten der TES-Items zu den RAND-Items und der damit einhergehenden Nähe zum „locus-of-control“-Konstrukt gilt entsprechend der obigen Argumentation die Feststellung, dass weder die RAND-Items noch die Items der „Teacher Efficacy Scale“ Handlungsergebnis- oder Selbstwirksamkeitserwartungen im Sinne Banduras sozial-kognitiver Theorie abbilden. Auch die Bildung des Gesamtscores steht im Widerspruch zu theoretischen Annahmen Banduras, da im Rahmen der sozial-kognitiven Theorie „self-efficacy“ und „outcome-efficacy“ als unabhängige, wenn

auch interagierende Konstrukte konzipiert sind, die in keinem Fall einen übergeordneten Faktor bilden. Die Bildung eines Gesamtscores spricht eher für die Theorie nach Rotter (1966), der internale und externale Kontrollüberzeugungen zunächst als Extreme eines Kontinuums, und damit eines Konstruktes konzipiert. Skinner (1996) weist jedoch darauf hin, dass Forschungsresultate auch bezüglich internaler und externaler Kontrollüberzeugungen eher für ein zweidimensionales Konstrukt sprechen (vgl. Abschnitte 3.4.6 und 4.3.2). Damit erscheint die Berechnung eines Gesamtscores sogar hinsichtlich beider Theoriegerüste fraglich, weil dieser inhaltlich nicht interpretierbar ist.

Pointiert formuliert existieren mit dem Instrument nach Gibson und Dembo (1984), der „Teacher Efficacy Scale“ (TES), und den Items der RAND-Gruppe nunmehr zwei inhaltlich fast identische Instrumente unterschiedlicher Itemanzahl, die sich zwar auf dieselbe Definition des zugrunde liegenden Konstruktes stützen (vgl. Armor u. a., 1976, S. 23), jedoch auf unterschiedliche theoretische Grundannahmen rekurrieren. Damit ist das Chaos in die Welt getragen: „Several researchers attempted to draw on both Rotter and Bandura, reconciling the two conceptualizations or simply ignoring the distinctions“ (Tschannen-Moran & Woolfolk Hoy, 2001, S. 788). Obwohl eine Vielzahl scheinbar offensichtlicher Widersprüche vorliegen, etabliert sich die TES schnell und wird über Jahrzehnte zum Standardinstrument zur Erfassung von „teacher efficacy“, wobei „teacher efficacy“ jedoch fälschlicherweise als „teacher self-efficacy beliefs“ bzw. als Lehrer-Selbstwirksamkeitserwartung in Sinne Banduras interpretiert wird. Tschannen-Moran u. a. (1998, S. 202) konstatieren zusammenfassend: „This appealing idea, that teachers' beliefs about their own capacities as teachers somehow matter, enjoyed a celebrated childhood, producing compelling findings in almost every study, but it has also struggled through the difficult, if inevitable, identity crisis of adolescence.“

4.1.4. Das Modell von Tschannen-Moran u. a. (1998): Ausweg aus der Krise?

Ausgehend von einer inhaltlichen Kritik an der „Teacher Efficacy Scale“ und insbesondere an der TE-Komponente des Instrumentes erarbeiten Tschannen-Moran u. a. (1998) Bezug nehmend auf Banduras sozial-kognitive Theorie eine eigene Definition des „teacher efficacy“-Konstruktes, wobei es dadurch im Vergleich zur ursprünglichen Intention von Armor u. a. (1976) eine Bedeutungsverschiebung erfährt. Im Folgenden wird das Konstrukt im Sinne der Autoren zunächst näher erläutert, sodass sich anschließend Unterschiede in der Konnotation aufzeigen lassen. Es muss jedoch festgestellt werden, dass auch die Konzeptualisierung von Tschannen-Moran u. a. (1998) *nicht* dem Konstrukt der Lehrer-Selbstwirksamkeitserwartung nach Bandura entspricht.

Tschannen-Moran u. a. (1998, S. 228) betonen, dass ein Individuum nur dann zu einer Einschätzung seiner „teacher efficacy“ gelangen kann, wenn es Einschätzungen zu Anforderungen der Lehr-/Lernsituation („Analysis of Teaching Task“) mit Einschätzungen eigener Kompetenzen/Fähigkeiten („Assessment of Personal Teaching Competence“)

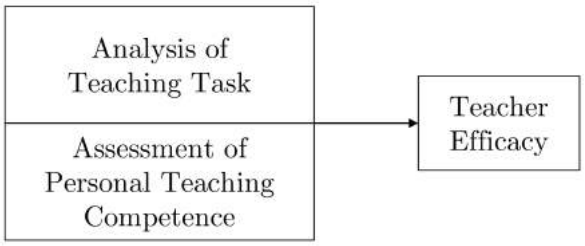


Abbildung 4.1.: Ausschnitt des Gesamtmodells von Tschannen-Moran, Woolfolk Hoy und Hoy (1998, S. 228).

kombiniert (vgl. Abbildung 4.1). Aus ihren Überlegungen ergibt sich folgende definitorische Aussage: „Teacher efficacy is the teacher's belief in his or her capability to organize and execute courses of action required to successfully accomplish a specific teaching task in a particular context" (Tschannen-Moran u. a., 1998, S. 233).

Diese Ausführungen wären mit Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras genau dann kompatibel, würde sich die Anforderungsanalyse lediglich darauf beziehen, welche komplexen Handlungen („means") in der konkreten Situation von der Person unter Umständen ausgeführt werden müssen. In dieser Lesart würde die Definition sogar die Situationsspezifität des Konstruktes in besonderer Weise aufgreifen. Allerdings spitzen Tschannen-Moran u. a. die Anforderungsanalyse derart zu, dass Ziele oder Resultate („ends") in Relation zu den auszuführenden Handlungen („means") abgeleitet werden sollen. Die Anforderungsanalyse entspricht damit einer „means-ends"-Relation (Welche Handlungen führen in der spezifischen Situation die abgeleiteten/angestrebten Ziele oder Resultate herbei?), die ihrer Meinung nach einen integralen Bestandteil des „teacher-efficacy"-Konstruktes bildet: „[W]e agree that a consideration of means-ends relationships (Skinner, 1996) is important for a full understanding of teacher efficacy" (Tschannen-Moran u. a., 1998, S. 212).

Das tatsächlich von Bandura geforderte Merkmal der Situations- bzw. Kontextspezifität (vgl. Abschnitt 3.2.1) wird folglich mit einer Art Zielorientierung gleichgesetzt und damit fehlinterpretiert. Dieses „Missverständnis" führt gerade zu einer symbiotischen Verknüpfung der Konstrukte der Selbstwirksamkeitserwartungen und der Handlungsergebniserwartungen einer Lehrperson. Tschannen-Moran u. a. (1998, S. 233) machen dies wie folgt deutlich: „By conceptualizing teacher efficacy in terms of the confluence of judgments about personal teaching competence and the teaching task, both competence and contingency (i.e., both agent-means and means-ends relations, as described by Skinner, 1996) are considered in an explanation of resultant teacher efficacy." Das interessierende Konstrukt „teacher efficacy" ist damit als „means-ends"-Relation gekennzeichnet und die Inkompatibilität mit der Theorie Banduras offensichtlich, denn es wird keine „agent-means"-Beziehung konzipiert.

Vergleicht man die Konzeption mit den Vorarbeiten von Armor u. a. (1976) und Gibson und Dembo (1984), so entfällt bei Tschannen-Moran u. a. (1998) der inhaltlich kritisierte TE-Faktor des Konstruktes. Der Faktor „external locus of control“ wird damit nicht mehr als konstruktkonstituierend angesehen, woraus die angesprochene Bedeutungsverschiebung des Konstruktes resultiert. Auch die Eigenschaft der Selbstreferentialität wird deshalb nicht mehr verletzt. Einschätzungen darüber, inwiefern eigene Handlungen erwünschte Wirkungen herbeiführen, stehen nun im Zentrum der definitorischen Fassung des Konstruktes. Schließlich muss angemerkt werden, dass die Autoren ihre eigene Definition nicht konsequent bei der Operationalisierung des Konstruktes anwenden. Teils werden keine „ends“ in Form von Handlungsresultaten formuliert (vgl. Unterabschnitt 4.3.5).

Zusammenfassend gilt: Im Modell von Tschannen-Moran u. a. (1998) kann die Komponente der Anforderungsanalyse als Handlungsergebniserwartung und die Komponente der Kompetenzbewertung ggf. als „teacher self-efficacy“ im Bandura’schen Sinne interpretiert werden, da letztere die geforderte „agent-means“-Relation beschreibt. Die Kombination beider Konstrukte zum „teacher efficacy“-Konstrukt definiert jedoch ein neues Konstrukt, welches *keine* „teacher self-efficacy beliefs“ nach Bandura abbildet. Diese Einschätzung kann damit begründet werden, dass die wesentliche strukturelle Eigenschaft und Stärke Banduras Theorie, nämlich die konsequente Unterscheidung zwischen „agent-means“- und „means-ends“-Beziehungen in Form der Trennung der Konstrukte „self-efficacy“ und „outcome-expectancy“ durch die Zusammenführung beider Komponenten gerade nicht berücksichtigt wird. Das von Tschannen-Moran u. a. (1998) definierte Konstrukt unterscheidet sich weiterhin auch vom „teacher-efficacy“-Konstrukt nach Armor u. a. (1976) oder Gibson und Dembo (1984). Alle bisher beschriebenen „teacher-efficacy“-Konstrukte haben jedoch eine Gemeinsamkeit – sie bilden „means-ends“-Relationen ab.

Dass das Konstrukt „teacher efficacy“ von Tschannen-Moran u. a. (1998) trotzdem häufig als konsistent zur Theorie Banduras gewertet wird, liegt unter Umständen daran, dass die Autoren selbst ihr Konstrukt als kompatibel ansehen und es in der theoretischen Tradition Banduras verorten. Zusätzlich integrieren sie in ihr Modell zur Entwicklung und Genese von „teacher efficacy“ die wesentlichen Quellen von Selbstwirksamkeitserwartungen wie eigene Erfahrungen, stellvertretende Erfahrungen etc. (vgl. Abbildung 4.2), die im Rahmen der sozial-kognitiven Theorie von Bandura beschrieben werden (vgl. Abschnitt 3.3) und suggerieren damit eine augenscheinliche Nähe zur Theorie Banduras.

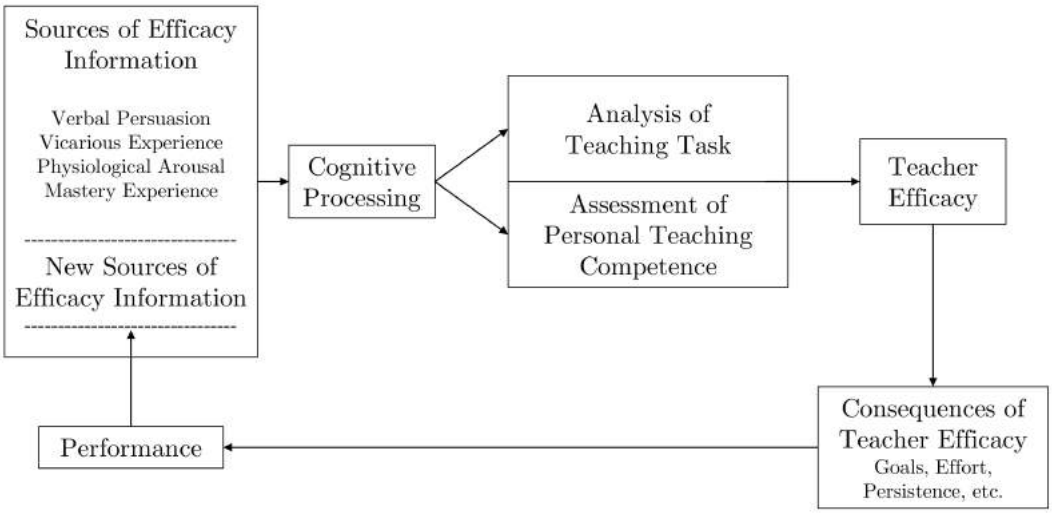


Abbildung 4.2.: Modell von Tschannen-Moran, Woolfolk Hoy und Hoy (1998, S. 228) zur Entwicklung von „teacher efficacy“.

4.1.5. Das Ende der begrifflichen Verwirrung: Dellinger u. a. (2008) bringen Licht ins Dunkel

Dellinger u. a. (2008, S. 751) kommen nach einer eingehenden Analyse der Passung existierender Operationalisierungen zur sozial-kognitiven Theorie Banduras zu dem Schluss: „[E]xisting measures of teacher efficacy do not adequately reflect the theory of self-efficacy.“ Auch wenn die Autoren selbst die Terminologie Skinners (1988, 1996) nicht verwenden, bringen sie zum Ausdruck, dass einer der Hauptgründe für diese Einschätzung in der fehlenden Differenzierung von „agent-means“- und „means-ends“-Relationen zu suchen ist, wie sie beispielsweise in den Konzeptualisierungen von Gibson und Dembo (1984) und Tschannen-Moran u. a. (1998) vorliegt, obwohl dieser Unterscheidung eine fundamentale Bedeutung im Rahmen der sozial-kognitiven Theorie zukommt.

Dellinger u. a. (2008) schlagen als Konsequenz ihrer Analyse vor, klar zwischen dem Konstrukt „teacher efficacy“ (als „means-ends“-Relation), so wie es häufig in Operationalisierungen anzutreffen ist, und dem Konstrukt „teacher self-efficacy“ (als „agent-means“-Relation) zu unterscheiden, wobei letzteres kompatibel mit der sozial-kognitiven Theorie Banduras ist. Sie definieren „teachers' self-efficacy beliefs“ als „[teachers' individual beliefs] about their own abilities to successfully perform specific teaching and learning related tasks within the context of their own classrooms.“ Mit dieser Definition und ihrer theoretischen Einbettung liegt erstmals eine Konzeptualisierung im Sinne der sozial-kognitiven Theorie Banduras vor. „Teacher efficacy“ definieren sie dagegen als die subjektive Überzeugung, Schülerinnen und Schüler sowie deren Leistungen beeinflussen zu können (vgl. Dellinger u. a., 2008, S. 753). Auf den ersten Blick ist ersichtlich, dass sich diese Definition des Konstruktes „teacher efficacy“ an

der Definition von Armor u. a. (1976) orientiert, auf die sich auch andere Autoren berufen, wenn sie „teacher efficacy“ operationalisieren (vgl. z. B. Gibson & Dembo, 1984). Wie aus den vorangegangenen Darstellungen jedoch deutlich geworden sein sollte, gibt es bezüglich konkreter Operationalisierungen durchaus Unterschiede hinsichtlich der jeweiligen Konstruktausschärfungen, sodass im Einzelfall genau geschaut werden muss, wie das Konstrukt „teacher efficacy“ tatsächlich verstanden wird.

An dieser Stelle soll die konzeptionelle Leistung von Dellinger u. a. (2008) noch einmal betont und gewürdigt werden. Nach etwa drei Jahrzehnten entwirren sie erstmals konsequent aus einer theoretischen Perspektive heraus das begriffliche und theoretische Knäuel, das sich im Laufe der Zeit gebildet hat. Die Autoren argumentieren eng an Banduras theoretischen Annahmen, suchen aber auch nach Ursachen für die vorliegenden definitorischen Unklarheiten. Diese machen sie beispielsweise in der bei einer eher oberflächlichen Lektüre der Hauptwerke Banduras möglicherweise leicht fehlinterpretierbaren Konstruktdefinition Banduras (vgl. Abschnitt 3.2.1) sowie in dem Festhalten an nachweislich konträren theoretischen Ansätzen wie der Definition nach Armor u. a. (1976) aus (vgl. Dellinger u. a., 2008, S. 754).

Ihr Hauptanliegen der begrifflichen (und theoretischen) Differenzierung visualisieren Dellinger u. a. in einer modellhaften Abbildung, die sie jedoch nicht (eingehender) erläutern (vgl. Abbildung 4.3). Meiner Meinung nach wird die Hauptbotschaft, nämlich dass die Konstrukte „teacher efficacy“ und „teacher self-efficacy“ voneinander zu unterscheiden sind, weil sie Einschätzungen bezüglich unterschiedlicher Bezugsdimensionen adressieren, durch die Abbildung verständlich, da die beiden voneinander abzugrenzenden Konstrukte farblich hervorgehoben und sie mit ihren jeweiligen (unterschiedlichen) Bezugsgrößen durch Pfeile verbunden sind. Das Konstrukt „teacher self-efficacy“ ist dabei durch die Bezugsgrößen der Lehrkraft („teacher“) und des antizipierten Verhaltens der Lehrkraft („proposed behavior“) gekennzeichnet, während das Konstrukt „teacher efficacy“ ebenfalls die Bezugsgröße Lehrperson aufweist, sich jedoch dahingehend unterscheidet, dass die zweite Bezugsgröße das Schülerverhalten bzw. die Leistungen von Schülerinnen und Schülern („student performance“) darstellt. Durch die Darstellung wird die oben zitierte Definition des Konstruktes „teacher efficiacy“ durch Dellinger u. a. (2008) sogar präzisiert. Indem sie ausschließlich die Bezugsgröße „Lehrperson“ („agent“) festlegen und nicht auch das Verhalten der Lehrperson („means“) als Einflussfaktor auf Schülerverhalten ansehen, konzipieren sie „teacher efficacy“ eher als „agent-ends“-Relation und unterscheiden sich diesbezüglich von anderen Konzeptionen wie denen von Armor u. a. (1976), Gibson und Dembo (1984) oder Tschannen-Moran u. a. (1998). Da das Konstrukt „teacher efficacy“ zunächst keinem theoretischen Überbau zugeordnet ist, ist dessen konzeptionelle Fassung zunächst unabhängig von „einschränkenden“ Theorieanteilen möglich.

Die Abbildung von Dellinger u. a. (2008), die höchstwahrscheinlich vornehmlich als

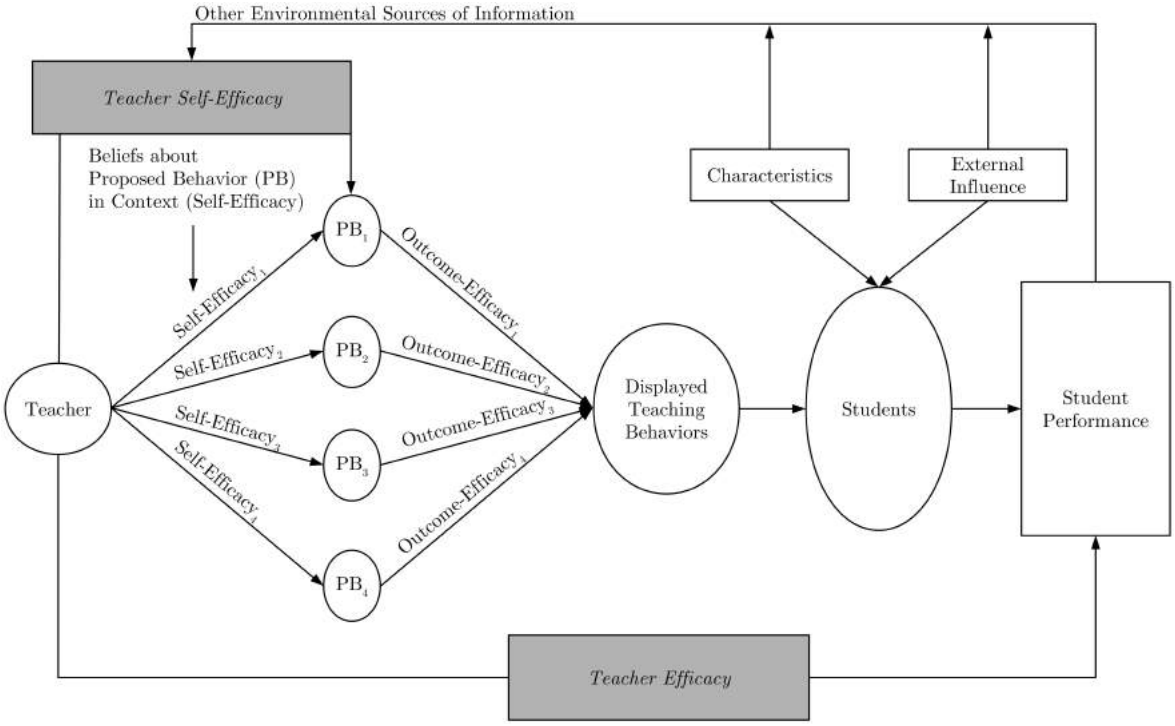


Abbildung 4.3.: Modell von Dellinger, Bobbett, Olivier und Ellett (2008, S. 753) zur Unterscheidung der Konstrukte „teacher efficacy“ und „teacher self-efficacy“.

Synthese ihres Differenzierungsansatzes der Konstrukte gelesen werden muss, soll im Folgenden detaillierter analysiert werden, da diese neben der hauptsächlichen Intention der Autoren, ein hohes Veranschaulichungspotential der Annahmen der sozial-kognitiven Theorie Banduras für den Lehrkontext birgt und damit das Konstrukt der „teacher self-efficacy beliefs“ bzw. der Lehrer-Selbstwirksamkeitserwartungen besser in seinen theoretischen Grundfesten verstanden werden kann.

4.1.6. Vertiefte Auseinandersetzung mit dem Ansatz von Dellinger u. a. (2008): Auf dem Weg zu einem besseren Verständnis des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen

Nachfolgend werden die Abbildung von Dellinger u. a. (2008) sowie die in dieser Darstellung enthaltenen Elemente (vgl. Abbildung 4.3) einer Analyse und kritischen Reflexion unterzogen. Dieses Vorgehen führt zu einer begründeten Optimierung der Darstellung mit dem Ziel, die Annahmen der sozial-kognitiven Theorie für den Lehrkontext deutlicher herauszuarbeiten und somit das Konstrukt der „teacher self-efficacy“ aus der Perspektive der Abgrenzung zum Konstrukt „teacher efficacy“ besser verstehen zu können. Die Analyse erfolgt dabei schrittweise und resultiert in der überarbeiteten Darstellung 4.4, die nachfolgend jeweils als Vergleichsdarstellung herangezogen werden

kann. In einem letzten Schritt wird das Potential der Darstellung dahingehend genutzt, die Kernaussagen Banduras möglichst vollständig für den Lehrkontext abzubilden. Damit verschiebt sich der zunächst intendierte Zweck der potentiellen Nutzung der grafischen Abbildung noch deutlicher, denn es geht nicht mehr vornehmlich darum, die Abgrenzung der Konstrukte zu visualisieren. Die Vervollständigung wird dadurch erreicht, zusätzlich die Quellen von Lehrer-Selbstwirksamkeitserwartungen in das Modell zu integrieren (vgl. Abbildung 4.5).

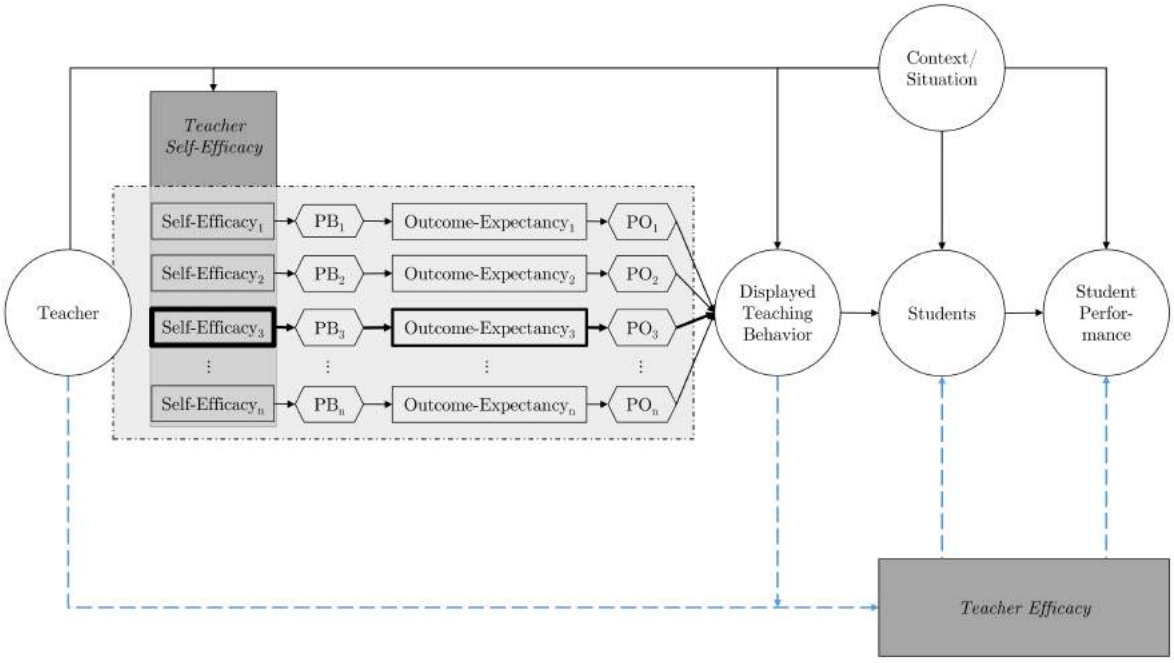


Abbildung 4.4.: Visualisierung der Kernaussagen der sozial-kognitiven Theorie nach Bandura (1997) bezogen auf den Lehrkontext mit dem zentralen Konstrukt der „teacher self-efficacy beliefs“ und seiner Abgrenzung zum verbreiteten „teacher efficacy“-Konstrukt. Grundlegende Überarbeitung des Modells von Dellinger, Bobbett, Olivier und Ellett (2008, S. 753).

Zunächst handelt es sich bei der originalen Abbildung von Dellinger u. a. (2008, S. 753) um die strukturelle Darstellung der Beziehung verschiedener beobachtbarer und nicht beobachtbarer Größen. Genauer wird der Einfluss der zentralen Konstrukte der sozial-kognitiven Theorie Banduras, der Selbstwirksamkeits- und Handlungsergebniserwartungen einer Lehrperson („Self-Efficacy“ bzw. „Outcome Expectancy“) auf das Lehrhandeln der Lehrperson („Displayed Teaching Behavior“), auf die Schülerinnen und Schüler („Students“) sowie auf deren Leistungen bzw. Verhalten („Student Performance“) unter besonderer Berücksichtigung des Kontextes („Characteristics“, „External Influences“, „Environmental Sources of Information“) thematisiert. Damit wird die Grundthese Banduras Theorie, dass nämlich Selbstwirksamkeitserwartungen einen Einfluss auf das Handeln einer Person haben (vgl. Bandura, 1997), auf den

Schulkontext übertragen und dahingehend theoriekonform spezifiziert, dass Lehrer-Selbstwirksamkeitserwartungen nicht nur Einfluss auf die Handlungen einer Lehrperson haben, sondern darüber hinaus und vermittelt über die Handlungen der Lehrperson auch die Schülerinnen und Schüler, das heißt deren Kognitionen und Verhalten beeinflussen (vgl. Bandura, 1997, S. 240-243). Das Ziel der Darstellung liegt in der Abgrenzung der Konstrukte „Teacher Efficacy“ und „Teacher Self-Efficacy“ (vgl. Bildunterschrift von Dellinger u. a., 2008, S. 753: „Differences in teacher self-efficacy beliefs and teacher efficacy.“). Um dieses Anliegen anschaulicher darzustellen, werden zwei Veränderungen vorgenommen:

1. Da sich der Sinn der verwendeten Formen der enthaltenen Faktoren in der Originalabbildung von Dellinger u. a. (2008) dem Betrachter nicht erschließt, wird in einem ersten Schritt anhand der verwendeten Formen zwischen direkt beobachtbaren Faktoren und nicht direkt beobachtbaren Faktoren bzw. Konstrukten unterschieden, um einen besseren Überblick über die Art der in Beziehung stehenden Faktoren zu erhalten. Diesem Gedanken folgend werden alle beobachtbaren Größen (z. B. Akteure, Verhalten der Akteure) durch runde Symbole dargestellt. Da im Modell von Dellinger u. a. (2008) nicht genau geklärt ist, inwiefern „Characteristics“, „External Influences“ und „Environmental Sources of Information“ zu unterscheiden sind, werden diese zu einem Faktor (Context/Situation) zusammengefasst, sodass die Situations- und Kontextbezogenheit weiterhin berücksichtigt ist.

   Nicht direkt beobachtbare Faktoren, d. h. Kognitionen bzw. mentale Prozesse der Akteure (z. B. Selbstwirksamkeitserwartungen) werden dagegen durch eckige Symbole repräsentiert. Farblich hinterlegt ist ein Ausschnitt dieser Kognitionen des Akteurs der Lehrperson („self efficacy“, „outcome-expectancy“), der auf die zentralen Konstrukte der Theorie Banduras fokussiert.

2. Zweitens werden alle dargestellten Beziehungen, die sich aus der Theorie Banduras ergeben, mittels schwarzer Linien dargestellt, während Beziehungen die zur Illustration der Abgrenzung zum „teacher-efficacy“-Konstrukt dienen, blau und gestrichelt dargestellt werden. Um den Fakt aufzugreifen, dass „teacher efficacy“ in der Forschungslandschaft nicht nur wie bei Dellinger u. a. (2008) als „agent-ends“-Relation, sondern auch als „means-ends“-Relation konzeptionalisiert wird (vgl. Armor u. a., 1976; Gibson & Dembo, 1984; Tschannen-Moran u. a., 1998) und damit jeweils auch ganz unterschiedlichen Bedeutungszuschreibungen unterliegt, wird auch eine Verknüpfung des Konstruktes zur Bezugsgröße des Lehrerhandelns („means“) zugelassen. Da außerdem als potentielles „end“ nicht nur das Handeln der Schülerinnen und Schüler im Fokus von Operationalisierungen steht, sondern vielfach auch deren Motivation oder das Lernen (eben der Akteur als Ganzes) wird zusätzlich eine Verbindung des Konstruktes „teacher efficacy“ zum Faktor

„students“ eingefügt. Es ist zu beachten, dass durch das Zulassen zusätzlicher Verknüpfungen die grafische Darstellung allgemeingültiger ist, weil keine spezielle „teacher efficacy“-Definition zu Grunde liegt. Vielmehr entspricht das derart dargestellte „teacher efficacy“-Konzept eher seiner realen Bedeutung in der Forschungslandschaft, in der es als eine Art Sammelbegriff ganz unterschiedlicher definitorischer Auslegungen gekennzeichnet ist. In der überarbeiteten Abbildung dient „teacher efficacy“ damit mehr der Abgrenzung, indem es darstellt, was „teacher self-efficacy“ *nicht* ist.

Dellinger u. a. (2008) adressieren in ihrer Darstellung die Kausalbeziehung der Konstrukte „self efficacy“ und „outcome-expectancy“ (Bandura, 1997, S. 21), wonach die Handlungsergebniserwartungen einer Lehrperson durch ihre Selbstwirksamkeitserwartungen beeinflusst werden (vgl. Abbildung 3.3 in Unterabschnitt 3.2.1 auf Seite 51). Um entsprechend des Zieles der Abbildung die Abgrenzung des Konstruktes „teacher self-effecacy“ sowohl zu „teacher efficacy“ als auch „outcome expectancy“ zu forcieren, führen Dellinger u. a. die Bezugsgröße des „Proposed Behavior“ ein. Bezüglich dieses antizipierten Verhaltens werden die spezifischen Selbsteinschätzungen („teacher self-efficacy beliefs“) entsprechend der Theorie Banduras vorgenommen. Gleichzeitig sind diese antizipierten Verhaltensweisen die Ausgangsgrößen, auf die sich die Handlungsergebniserwartungen beziehen. Diese richten sich bei Bandura jedoch gerade *nicht* auf ein bestimmtes zu beobachtendes Lehrerhandeln („Displayed Teaching Behavior“), wie es ggf. in der Abbildung von Dellinger u. a. (2008) suggeriert wird, sondern beispielsweise auf eine antizipierte Konsequenz des eigenen Handelns. Mit der originalen Abbildung von Dellinger u. a. (2008) wird deshalb unter Umständen eine gängige Fehlinterpretation der theoretischen Annahmen Banduras unterstützt, indem das Verhalten der Lehrperson als „end“ missverstanden werden kann (vgl. Bandura, 1997, S. 22 f.).

3. Aus diesem Grunde wird für das Konstrukt der Handlungsergebniserwartung die zweite zusätzliche Bezugsgröße in Form von antizipierten Handlungskonsequenzen („Proposed Outcomes“) eingeführt. Ein weiterer Vorteil der Darstellung der Bezugsgrößen ist die gleichzeitige Visualisierung der „agent-means“-Realtion für das Konstrukt der Selbstwirksamkeitserwartungen sowie der „means-ends“-Relation für das Konstrukt der Handlungsergebniserwartungen. Durch das Zwischenschalten der Bezugsgrößen wird darüber hinaus auch visuell die notwendige konzeptionelle Trennung beider Konstrukte verdeutlicht.

Dellinger u. a. (2008) wählen in ihrer Darstellung ein realistisches Szenario, indem in einer Situation mehrere Selbstwirksamkeitserwartungen bezüglich unterschiedlicher, potentiell möglicher (antizipierter) Handlungen vorliegen und dementsprechend auch mehrere zugehörige Handlungsergebniserwartungen. Auch wenn es sich nachfolgend um

eine Interpretation der Theorie Banduras handelt, so ist es doch naheliegend, dass ein Individuum in einer konkreten herausfordernden Situation eine Vorauswahl an potentiell sinnvollen (nützlichen, wichtigen) Handlungen vornimmt, also Bewertungen der eigenen Fähigkeiten bezüglich des erfolgreichen Ausführens dieser eingeschränkten Anzahl von Handlungen vorgenommen werden müssen. Es stellt sich jedoch die Frage, welche der Selbsteinschätzungen handlungsrelevant wird. Anhand der grafischen Darstellung von Dellinger u. a. könnte der Schluss gezogen werden, dass für die Wahl des letztlich gezeigten Verhaltens alle Einschätzungen gleich bedeutsam sind, was meines Erachtens mehr als fraglich ist.

4. Entsprechend der Aussagen Banduras (1977a, 1997), dass gerade hoch ausgeprägte und stabile Selbstwirksamkeitserwartungen handlungsleitend sind, wird die Abbildung derart geändert, dass in Form der Breite der Umrandung deutlich wird, welcher Pfad der ausschlaggebende, das heißt der handlungsauslösende ist. Darüber hinaus wird die Rahmung des handlungsleitenden „self-efficacy“-Konstruktes stärker hervorgehoben als die Rahmung des entsprechenden „outcome-expectancy“-Konstruktes, da den Aussagen Banduras folgend gerade die Selbstwirksamkeitserwartungen das zentrale Konstrukt darstellen und einen größeren Einfluss auf die Handlungsentscheidung ausüben (vgl. Bandura, 1997, S. 20).

Will man das Potential der grafischen Abbildung von Dellinger u. a. ausreizen und diese eher zur Veranschaulichung der wesentlichen domänenspezifischen Grundzüge der Theorie Banduras anstatt in erster Linie zur Abgrenzung der Konstrukte „teacher efficacy“ und „teacher self-efficacy“ nutzen, so bietet sich die Erweiterung des Modells um die Quellen von Lehrer-Selbstwirksamkeitserwartungen an. Lehrer-Selbstwirksamkeitserwartungen wirken nämlich nicht nur auf das Handeln der Lehrperson und damit theoretisch auch auf Schülerinnen und Schüler. Sie wirken über das Handeln der Lehrperson in Form einer Rückkopplung und über eine gesonderte kognitive Verarbeitungsschleife auch auf sich selbst, denn Handlungen zählen entsprechend der theoretischen Annahmen Banduras – neben weiteren wie stellvertretenden Erfahrungen in Form von Beobachtungen, etc. – als Quelle der Selbstwirksamkeitserwartungen der Lehrperson (vgl. Abschnitt 3.3).

5. Insofern wird das Modell um den bereits von Tschannen-Moran u. a. (1998) für den Lehrkontext (allerdings für das „teacher-efficacy“-Konstrukt im Sinne der genannten Autoren!) adaptierten Entwicklungsmechanismus von Lehrer-Selbstwirksamkeitserwartungen ergänzt, der sich aus den allgemeinen Überlegungen bezüglich der Quellen von Selbstwirksamkeitserwartungen ergibt (vgl. Bandura, 1977a, 1997). Die daraus resultierende Überarbeitung mündet in Abbildung 4.5. Die zusätzlichen Elemente sind dabei weder eckig noch gänzlich rund dargestellt, was

daran liegt, dass sowohl beobachtbare (Ausdauer, Aspekte von Nervosität) als auch unbeobachtbare Facetten (Zielsetzungen, Aspekte von Nervosität) angesprochen werden.

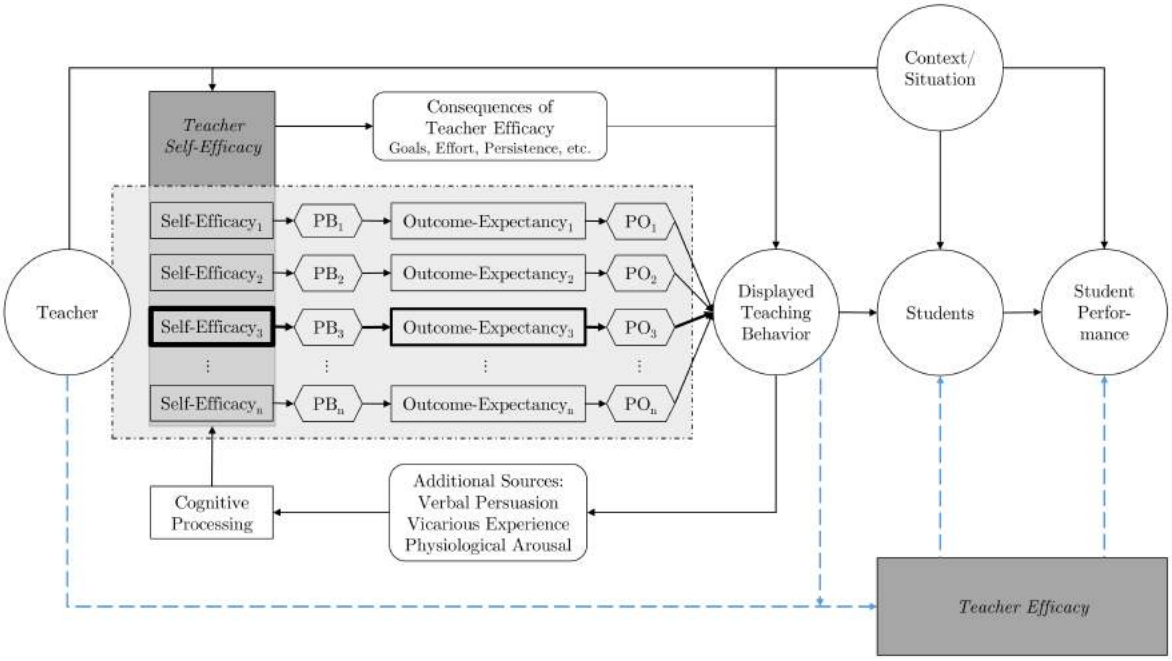


Abbildung 4.5.: Visualisierung der Kernaussagen der sozial-kognitiven Theorie nach Bandura (1997) bezogen auf den Lehrkontext mit dem zentralen Konstrukt der „teacher self-efficacy beliefs“ und seiner Abgrenzung zum verbreiteten „teacher efficacy“-Konstrukt sowie Darstellung des Entwicklungsmechanismus von „teacher self-efficacy beliefs“ aufgrund der postulierten Quellen des Konstruktes nach Bandura (1997). Anlehnung an Dellinger, Bobbett, Olivier und Ellett (2008, S. 753) sowie Tschannen-Moran, Woolfolk Hoy und Hoy (1998, S. 228).

4.1.7. Das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen im deutschen Sprachraum

Im deutschen Sprachraum wurden die Theorie- und Begriffsverwirrungen um das Konstrukt „teacher efficacy“ weitgehend ignoriert. Stattdessen wurde vorwiegend durch englischsprachige Vorarbeiten unbeeinflusst und ausgehend von der sozial-kognitiven Theorie Banduras das Konstrukt der Selbstwirksamkeitserwartung für den Lehrkontext adaptiert. Schwarzer und Jerusalem (2002, S. 40) definieren das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen konsistent zum Theorierahmen Banduras und unter besonderer Berücksichtigung des Merkmals des Schwierigkeitsbezugs (vgl. Abschnitt 3.2.1) wie folgt: „Die Lehrer-Selbstwirksamkeit beinhaltet Überzeugungen von Lehrern, schwierige Anforderungen ihres Berufslebens auch unter widrigen Bedingungen

erfolgreich zu meistern.“ Die formulierte Definition legt zudem eine Konzeptualisierung des Konstrukte als „agent-means“-Relation nahe, was sich in Operationalisierungen auf Grundlage dieser Definition bestätigt (vgl. Abschnitt 4.3.3).

4.1.8. Zusammenfassung

In der „locus-of-control“-Theorie geht es in Form der Ursachenattribuierung gerade um die Verknüpfung von Handlungen und Handlungsresultaten. Daher können konkrete Operationalisierungen dieser Theorie auch als verallgemeinerte Handlungsergebniserwartungen aufgefasst werden (keine „outcome-expectancies“ im Sinne Banduras!). In der sozial-kognitiven Theorie geht es im Gegensatz dazu darum, Einschätzungen eigener Fähigkeiten bezüglich der Planung und Durchführung bestimmter Handlungen von Einschätzungen möglicher Handlungsresultate zu trennen, ohne dabei Zusammenhänge beider Einschätzungen auszuschließen. Je nach Erkenntnisinteresse, Ziel- und Fragestellung können Operationalisierungen bezüglich des einen oder anderen theoretischen Rahmens geeigneter erscheinen. Wichtig ist jedoch, klar zu argumentieren, welche Annahmen hinter definitorischen Adaptionen und konkreten Operationalisierungen stehen, welche Ableitungen möglich sind und welche besser nicht auf Grundlage der gewählten Operationalisierung getroffen werden sollten.

Die Ausführungen haben deutlich gemacht, dass diese konzeptionelle Klarheit jedoch keinesfalls einfach zu erreichen ist. Die Wahl von Begrifflichkeiten (siehe RAND-Gruppe, die den Namen „teacher efficacy“ wählt, der mit einem völlig anderen theoretischen Hintergrund verbunden wird, als sie ihn selbst anlegt) und die Ableitung konkreter Operationalisierungsregeln muss theoriegeleitet und diskursiv erfolgen. Sind erst einmal „falsche“ Annahmen publiziert, so halten sie sich hartnäckig und werden vielfach ungeprüft übernommen – teils stellt sich zudem eine Autoritätsgläubigkeit ein. Anders ist es nicht zu erklären, dass die „Teacher Efficacy Scale“ jahrzehntelang mit Referenz auf Bandura eingesetzt wurde. Und noch einmal: Es spricht formal nichts gegen die Nutzung dieses Instrumentes. Allerdings sollte klar geworden sein, dass mit diesem Instrument keine Selbstwirksamkeitserwartungen im Sinne Banduras erhoben werden, was jedoch behauptet wurde. Insofern ist die Forschungsgemeinschaft gefordert, jeweils kritisch Stellung zu nehmen.

Der Ansatz von Dellinger u. a. (2008) erscheint bisher am geeignetsten, um das Konstrukt „teacher self-efficacy beliefs“ in der Tradition der sozial-kognitiven Theorie zu verstehen. Im Modell von Tschannen-Moran u. a. (1998) werden zwar keine Selbstwirksamkeitserwartungen nach Bandura konzeptualisiert, es hat aber den Charme, wesentliche Entwicklungsmechanismen – wenn auch für das Konstrukt der „teacher efficacy“ – zu berücksichtigen. Eine ausgeschärfte und erweiterte Version des Modells von Dellinger u. a. (2008) beispielsweise um die Quellen- und Entwicklungsmechanismen von „teacher self-efficacy“ wurde unter der kritischen Reflexion der theoretischen Vorgaben

Banduras erarbeitet, was einen Theoriebeitrag dieser Arbeit zur präzisen begrifflichen und definitorischen Fassung des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen darstellt.

4.2. Abgrenzung und Beziehung zu verwandten Konstrukten

Das Konstrukt der Selbstwirksamkeitserwartungen wurde bereits im vorangegangenen Kapitel zu einer Reihe ähnlicher Konstrukte abgegrenzt (vgl. Abschnitt 3.4). Die dort herausgearbeiteten Abgrenzungskriterien gelten gleichermaßen für die jeweiligen auf den Lehrkontext angepassten Konstrukte. Darüber hinaus wurde im letzten Abschnitt 4.1, welcher die definitorische Fassung des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen thematisierte, gleichzeitig die Abgrenzung des Konstruktes zu anderen, insbesondere zum Konstrukt „teacher efficacy“ vorgenommen, welches eine hohe Ähnlichkeit zum „locus-of-control“-Konstrukt aufweist. Nachfolgend werden daher lediglich einige ergänzende Bemerkungen vorgenommen.

4.2.1. Lehrer-Selbstkonzept/“teacher self-concept“

Die Erfassung des Selbstkonzepts von Lehrpersonen hat eine sehr viel kürzere und sehr viel weniger umfangreiche Forschungstradition als die Erfassung des Selbstkonzeptes von Schülerinnen und Schülern. In der Lehrerbildungsforschung wurden traditionell eher die Selbstwirksamkeitserwartungen von Lehrpersonen – hauptsächlich konzeptualisiert als „teacher efficacy“ – erhoben.

Auch wenn das Konstrukt seltener eine Rolle spielt, gelten für das Selbstkonzept von Lehrpersonen bezüglich des Lehrerberufs dieselben Kriterien zur Abgrenzung in Bezug auf das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen, die bereits auf einer allgemeinen Ebene zur Unterscheidung von Selbstwirksamkeitserwartungen und Selbstkonzept herausgearbeitet wurden (vgl. Abschnitt 3.4.3). Diesen Kriterien entsprechend wird das Lehrer-Selbstkonzept auf einer globaleren Ebene, das heißt nicht bezogen auf eine konkrete Handlung konzeptualisiert und kann auch affektive Komponenten enthalten. Zur Veranschaulichung kann eine Reihe von Instrumenten zur Erfassung des Lehrer-Selbstkonzepts, die den genannten Kriterien entsprechen, dem Anhang entnommen werden (vgl. Tabelle A.2, S. 314).

4.2.2. Kontrollüberzeugungen von Lehrpersonen/“teacher locus of control“

Dass Selbstwirksamkeitserwartungen vom „locus-of-control“-Konstrukt konzeptionell abzugrenzen sind, wurde bereits ausführlich dargelegt (vgl. Abschnitt 3.4.6). Dass dies auch für Lehrer-Selbstwirksamkeitserwartungen und Kontrollüberzeugungen von Lehrpersonen („teacher locus of control“) gilt, sollte insbesondere aus den vorangegan-

genen Ausführungen deutlich geworden sein (vgl. z. B. Abschnitt 4.1.1). Neben den Items der RAND-Gruppe, die Kontrollüberzeugungen von Lehrpersonen theoriekonform im Sinne Rotters (1966) abbilden, existieren einige andere Instrumente, die „teacher locus of control“ abbilden (vgl. u. a. Rose & Medway, 1981; Guskey, 1981). Die genannten Instrumente unterscheiden allerdings nicht nur nach internalen und externalen Kontrollüberzeugungen, sondern auch Kontrollüberzeugungen bezüglich positiver und negativer Konsequenzen (vgl. Itembeispiele in Tabelle 4.2). Eine solche zusätzliche Unterscheidung scheint sinnvoll zu sein, da sich empirisch jeweils keine oder lediglich geringe bis mittlere Korrelationen zwischen den Dimensionen zeigen (vgl. u. a. Rose & Medway, 1981; Guskey, 1981; Guskey, 1984; Guskey, 1988; Coladarci & Fink, 1995).

Tabelle 4.2.: Vergleich zweier Instrumente zur Erfassung von „teacher locus of control“ (vgl. Rose & Medway, 1981; Guskey, 1981).

Teacher Locus of Control Scale (TLC): Rose und Medway (1981)	**Responsibility for Student Achievement Questionnaire** (RSA): Guskey (1981)
Control over Student Success: „When the grades of your students improve, it is more likely a. because you found ways to motivate the students, or b. because the students were trying harder to do well.“	**Responsibility for Student Success:** „Suppose you are particularly successful in one class. Would it probably happen a. because you helped them overcome their learning difficulties, or b. because these students usually do well in school?“
Control over Student Failures: „If you couldn't keep your class quiet, it would probably be a. because the students came to school more rowdy than usual, or b. because you were so frustrated that you weren't able to settle them down.“	**Responsibility for Student Failures:** „If a child doesn't well in your class, would it probably be a. because he did not work very hard, or b. because you didn't provide the proper motivation for him?“

* Teacher Locus of Control Scale (TLC).

Obwohl beide Instrumente explizit mit Bezug auf die „locus-of-control“-Theorie entwickelt wurden, ändert Guskey die Nomenklatur und spricht sodann von „teacher efficacy regarding positve/negative results“ (vgl. z. B. Guskey, 1988, S. 66 ff.). Da der Begriff „efficacy“ in der Regel mit Banduras sozial-kognitiver Theorie verknüpft ist, rückt er durch dieses Vorgehen das Konstrukt sprachlich näher an den genannten Theoriekomplex und trägt damit zur Verwirrung darüber bei, welche Konstrukte auf welcher theoretischen Basis durch etablierte Instrumente erfasst werden.

Auch wenn die Konstrukte „teacher self-efficacy“ und „teacher locus of control“ konzeptionell zu unterscheiden sind, so werden trotzdem Zusammenhänge beider Konstrukte vermutet. Bandura (1997, S. 240) geht davon aus, dass hoch ausgeprägte Selbstwirksamkeitserwartungen von Lehrpersonen mit hoch ausgeprägten internalen

Kontrollüberzeugungen einhergehen und dass weniger selbstwirksame Lehrkräfte eher zu hoch ausgeprägten externalen Kontrollüberzeugungen neigen. In Einklang mit diesen Überlegungen finden Skaalvik und Skaalvik (2010) schwache negative Korrelationen zwischen den Dimensionen der Lehrer-Selbstwirksamkeitserwartungen und externalen Kontrollüberzeugungen von Lehrpersonen.

4.3. Zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen

„Besides [...] definition problems, teacher efficacy is a psychometrician's nightmare.“

Guskey, 1998, S. 2

Seit den späten 1970er-Jahren sind nicht zuletzt aufgrund der vielversprechenden Resultate der RAND-Forschungsgruppe eine Reihe von Instrumenten entwickelt, weiterentwickelt und erprobt worden. In der in Abschnitt 4.1 dargestellten Problematik, das zentrale Konstrukt der sozial-kognitiven Theorie Banduras für Lehrpersonen zu adaptieren sowie theoretisch und definitorisch zu fassen, deutet sich bereits die Schwierigkeit an, das Konstrukt der Lehrer-Selbstwirksamkeitserwartung bzw. „teacher self-efficacy“ angemessen zu operationalisieren. Mindestens für die „Teacher Efficacy Scale“ konnte bereits erläutert werden, dass die Items keine Selbstwirksamkeitserwartungen entsprechend der Definition Banduras abbilden. Systematisch sollen im Folgenden Instrumente, die explizit auf Banduras Theorie abheben und damit vorgeben, Lehrer-Selbstwirksamkeitserwartungen zu erfassen, auf ähnliche Inkonsistenzen geprüft werden. Diese Analyse erscheint sinnvoll, um einerseits einen vertieften Einblick in die Forschungslandschaft zu erhalten und damit das neue, zu entwickelnde Instrument besser in die Forschungstradition einbetten oder von ihr abgrenzen zu können. Andererseits können potentielle Defizite oder Stärken existierender Instrumente als Orientierungsrahmen für die eigene Instrumentenentwicklung hilfreich sein.

Die Prüfung der Instrumente wird hauptsächlich entlang inhaltlicher Fragestellungen erfolgen. Letztlich läuft diese Festlegung auf die Teilanalyse der typischerweise zugrunde liegenden, beschreibenden Testwertinterpretation hinaus, die besagt, dass Daten, die mittels der verwendeten Items generiert werden, Auskunft über die Lehrer-Selbstwirksamkeitserwartungen der befragten Probanden geben. Folglich müssen die verwendeten Items kritisch daraufhin betrachtet werden, ob sie tatsächlich Lehrer-Selbstwirksamkeitserwartungen in der Tradition Banduras (!) abbilden. Um bewerten zu können, ob die konstruktkonstituierenden Charakteristika in den Operationalisierungen abgebildet sind, werden als Kriterien die in Abschnitt 3.2.1 herausgearbeiteten Eigenschaften des Konstruktes der Selbstwirksamkeitserwartungen (Selbstreferentiali-

tät, Handlungsorientierung, Schwierigkeitsbezug, Situations- und Domänenspezifität) herangezogen und auf ihre angemessene Übertragung bezüglich des Konstruktes der Lehrer-Selbstwirksamkeitserwartung beurteilt. Ein besonderer Schwerpunkt wird dabei auf die Kriterien des Handlungs- bzw. Schwierigkeitsbezugs gelegt, da die übrigen anhand von Itembeispielen (Selbstreferentialität) oder formalen Angaben wie der Anzahl operationalisierter Dimensionen (Mehrdimensionalität) beurteilt werden können. In Bezug auf die Untersuchung des Handlungsbezugs hat es sich als praktikabel herausgestellt, die Items daraufhin zu prüfen, ob sie im Sinne Skinners (1988, 1996) eine „agent-means"-Relation abbilden, weshalb dies als zusätzliches Prüfkriterium aufgenommen wird. Im Rahmen dieser kriteriengeleiteten Analyse werden automatisch zusätzliche inhaltliche Fragen wie die nach der Abgrenzung zu anderen Konstrukten (Stichwort „teacher efficacy") thematisiert, die auch Bandura in seinen Richtlinien zur Erfassung des Konstruktes einfordert (vgl. Abschnitt 3.6.2).

Zu den formalen Aspekten, die im Rahmen der Analyse ergänzend berücksichtigt werden, gehören die Polarität oder das Antwortformat der verwendeten Skala sowie die Zielgruppe, für die das Instrument ursprünglich entwickelt wurde. Zu betonen ist, dass in den nachfolgenden Abschnitten die Instrumente nicht nach einem festen Schema den genannten Analyseschritten unterzogen werden, sondern vielmehr in ihrer Entwicklungs- und Validierungsgeschichte vorgestellt werden. Dazu gehört es auch, relevante Validierungsergebnisse (Prüfung der Reliabilität, Eindimensionalität, Messinvarianz ...) zu berichten. Dadurch kann nicht nur ein umfassenderes Bild bezüglich der Qualität der dargestellten Instrumente gewonnen werden, sondern – zusätzlich zum eigentlichen Ziel der Analyse – ein Eindruck von gängigen Herangehensweisen der Entwicklung und Validierung von Testinstrumenten.

In die Analyse werden nur solche Instrumente aufgenommen, die explizit auf die sozial-kognitive Theorie Banduras rekurrieren. Darüber hinaus ist es aufgrund des Forschungsanliegens dieser Arbeit von besonderem Interesse, deutschsprachige Instrumente zu recherchieren und auf ihre Qualität zu prüfen. Es werden jedoch nicht alle recherchierten Instrumente eingehender untersucht, sondern nur solche, die potentiell geeignet erscheinen bzw. mindestens teilweise den angelegten Kriterien genügen. Instrumente, die bereits bei einer oberflächlichen Prüfung offensichtliche Mängel aufweisen, werden von einer detaillierten Analyse ausgeschlossen. Eine Liste dieser Instrumente kann dem Anhang entnommen werden (vgl. Tabelle B.1).[5] Auch Instrumente, die explizit mit dem Ziel entwickelt wurden, lediglich einen spezifischen Ausschnitt der Lehrer-

[5]Unberücksichtigt bleiben auch ältere Instrumente, die meist Kurz- oder Kombinationsformen der "Teacher Efficacy Scale" nach Gibson und Dembo (1984) darstellen, und damit ebenfalls als inadäquat zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen eingestuft werden müssen (vgl. u. a. Midgley, Feldlaufer & Eccles, 1989; Newmann, Rutter & Smith, 1989; Lee, Dedrick & Smith, 1991; Raudenbush, Rowan & Cheong, 1992). Einen Einblick in solche Instrumente findet man z. B. bei Tschannen-Moran und Woolfolk Hoy (2001, S. 791).

Selbstwirksamkeitserwartungen zu operationalisieren, werden aufgrund der Vielzahl der in Frage kommenden Instrumente und des damit verbundenen Aufwandes nicht berücksichtigt.[6]

Die Recherche der Instrumente erfolgt einerseits anhand der Durchsicht von Überblicksartikeln zu existierenden Instrumenten (vgl. u. a. Tschannen-Moran u. a., 1998; Tschannen-Moran & Woolfolk Hoy, 2001).[7] Da diese jedoch nicht mehr als aktuell bezeichnet werden können, wird zusätzlich eine umfangreiche Literatursichtung ab dem Jahr 2001 vorgenommen. Insgesamt werden dabei mehr als 100 Publikationen unter dem Themenfeld „Lehrer-Selbstwirksamkeitserwartungen“ („teacher self-efficacy“) bzw. „teacher efficacy“ nach Instrumenten durchsucht. Trotz der Vielzahl der berücksichtigten Veröffentlichungen kann kein Anspruch auf Vollständigkeit erhoben werden.

Das zusammenfassende Ergebnis der detaillierten Analyse von insgesamt zehn Instrumenten kann Tabelle 4.3 entnommen werden. Nachvollziehbar wird diese Essenz durch die Lektüre der nachfolgenden Abschnitte 4.3.1 bis 4.3.10, die sich der ausführlichen Evaluation der Instrumente widmen. Der an Detailkritik weniger interessierte Leser kann diese Abschnitte überspringen und mit der erläuternden Zusammenfassung ab Seite 124 fortfahren. Insgesamt wird sich herausstellen, dass die meisten Instrumente sowohl mit inhaltlichen als auch mit messtheoretischen Problemen behaftet sind, sodass sich das Forschungsdesiderat eines Instrumentes ergibt, dessen Messdaten valide dahingehend interpretiert werden können, dass sie Lehrer-Selbstwirksamkeitserwartungen entsprechend der sozial-kognitiven Theorie Banduras abbilden. Insgesamt können aus der eingehenden Prüfung der Instrumente wertvolle Hinweise für eine Instrumentneuentwicklung gewonnen werden (vgl. Abschnitt 4.3.12).

4.3.1. Personal Teaching Efficacy Vignette Scale

Ashton u. a. (1984) orientieren sich zunächst am Konstrukt der RAND-Gruppe. Sie halten dieses aufgrund der wenigen Items jedoch lediglich für eingeschränkt aussagekräftig. Sie nehmen dies zum Anlass, ein Instrument zu entwickeln,[8] welches entsprechend Banduras Aussagen die Vielzahl der Handlungsfelder einer Lehrperson abdecken soll: „Teachers' sense of efficacy was defined to encompass teachers' confidence in their ability to carry out all the responsibilities of teaching“ (Ashton u. a., 1984, S. 33). Aus Interviews mit Lehrpersonen oder aus Aufsätzen von Lehrpersonen entwickeln sie zunächst 50 Situationsbeschreibungen (Vignetten), die u. a. die Bereiche Motivierung,

[6] Solche Instrumente adressieren z. B. Lehrer-Selbstwirksamkeitserwartungen bezüglich Inklusion/Diversität (vgl. u. a. Siwatu, 2007; Kopp, 2009; Sharma, Loreman & Forlin, 2012) oder bezüglich des Medieneinsatzes (vgl. u. a. Lee & Tsai, 2010; Lee & Lee, 2014).

[7] Eine Sammlung von Instrumenten findet sich auch auf der Website von Anita Woolfolk Hoy: http://anitawoolfolkhoy.com/instruments/ (letzter Zugriff: 20.06.2015).

[8] Ein Teil des Instruments kann unter http://anitawoolfolkhoy.com/wp-content/uploads/2014/09/Ashton-Eff-Vign-1dgs14f.pdf (15.05.2015) eingesehen werden.

Tabelle 4.3.: Charakteristika der Erhebungsinstrumente, die auf Banduras sozial-kognitive Theorie rekurrieren.

Instrumentcharakteristika						**Validierungsaspekte**					
PG[1]	#I[2]	RS[3]	AF[4]	#Sub[5]	Itembeispiele	S[6]	H[7]	S[8]	S/K[9]	D/M[10]	Rel[11]
Personal Teaching Efficacy Vignette Scale: Ashton u. a. (1984), Neuentwicklung											
L	50	✓	7-stufig	–	„A new student has been assigned to your class. Her records indicate that she never does her homework and does not seem to care about her education. Her IQ score is 83 and her achievement scores have been below the 30th percentile. How effective would you be in increasing her achievement test scores?“	✓	∼	✓	∼	∼	a-m a-e
TES – Teacher Efficacy Scale: Gibson und Dembo (1984), Neuentwicklung, orientiert an Armor u. a. (1976)											
L	30/15	✓	6-stufig	2	Personal Teaching Efficacy: „When the grades of my students improve it is usually because I found more effective teaching approaches.“ Teaching Efficacy „If students are not disciplined at home, they aren't likely to accept any discipline.“	∼	x	x	x	x	m-e
TES-Short I: Woolfolk und Hoy (1990), aufbauend auf Gibson und Dembo (1984)											
S	22	✓	6-stufig	2	siehe TES	siehe TES					
TES-Short II: Hoy und Woolfolk (1993), aufbauend auf Gibson und Dembo (1984), Woolfolk und Hoy (1990)											
L	10	✓	6-stufig	2	siehe TES	siehe TES					

Anmerkung: ✓vorhanden, ∼ teils vorhanden, x nicht vorhanden [1] Personengruppe: S (Studierende), R (Referendare), L (Lehrkräfte)
[2] Anzahl Items [3] unipolare Ratingskala [4] Antwortformat [5] Anzahl Subfaktoren/Dimensionen [6] Selbstreferentialität
[7] Handlungsorientierung [8] Schwierigkeitsbezug [9] Situations-/Kontextorientierung [10] Domänenspezifität/Mehrdimensionalität
[11] Relation (vgl. Abschnitt 3.1.3 bzw. Skinner, 1996): a-e (agent-ends), a-m (agent-means), m-e (means-ends)
[*] Einschätzung gilt nicht für einen Subfaktor

Fortsetzung auf der nächsten Seite...

Tabelle 4.3.: *Fortsetzung*

PG[1]	#I[2]	RS[3]	AF[4]	#Sub[5]	Itembeispiele	S[6]	H[7]	S[8]	S/K[9]	D/M[10]	Rel[11]
Instrument zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen: Schmitz und Schwarzer (2000), Neuentwicklung											
L	10	✓	4-stufig	–	„Ich weiß, da[ss] ich es schaffe, selbst den problematischsten Schülern den Stoff zu vermitteln.“ „Ich kann Innovationen auch gegenüber skeptischen Kollegen durchsetzen.“	✓	✓	✓	∼	∼	a-m
TSES – Teacher Self-Efficacy Scale: Bandura (2006), Neuentwicklung											
L	30	✓	9-stufig	7	Instructional Efficacy: „How much can you do to influence class size in your school.“ Efficacy to Create Positive School Climate: „How much can you do to get students to trust teachers.“	✓	∼	x	∼	∼	a-e a-m
OSTES – Ohio State Teacher Efficacy Scale: Tschannen-Moran und Woolfolk Hoy (2001), basierend auf Emmer und Hickman (1991), Bandura (2006)											
L/S	24/12	✓	9-stufig	3	Efficacy for Instructional Strategies: „To what extent can you craft good questions for your students?“ Efficacy for classroom management: „How well can you establish routines to keep activities running smoothly?“ Efficacy for student engagement: „How much can you do to help your students value learning?“	✓	∼	x	∼	∼	m/a-e a-m

Anmerkung: ✓vorhanden, ∼ teils vorhanden, x nicht vorhanden [1] Personengruppe: S (Studierende), R (Referendare), L (Lehrkräfte)
[2] Anzahl Items [3] unipolare Ratingskala [4] Antwortformat [5] Anzahl Subfaktoren/Dimensionen [6] Selbstreferentialität
[7] Handlungsorientierung [8] Schwierigkeitsbezug [9] Situations-/Kontextorientierung [10] Domänenspezifität/Mehrdimensionalität
[11] Relation (vgl. Abschnitt 3.1.3 bzw. Skinner, 1996): a-e (agent-ends), a-m (agent-means), m-e (means-ends)
[*] Einschätzung gilt nicht für einen Subfaktor

Fortsetzung auf der nächsten Seite...

Tabelle 4.3.: *Fortsetzung*

PG[1]	#I[2]	RS[3]	AF[4]	#Sub[5]	Itembeispiele	S[6]	H[7]	S[8]	S/K[9]	D/M[10]	Rel[11]
TCS – Teacher Confidence Scale: Woolfolk Hoy und Spero (2005), Neuentwicklung											
R	32	✓	6-stufig	–	„I am confident in my ability to...: ...evaluate students' work.“ ...facilitate class discussions.“ ...teach science as a co-inquirer with students.“	✓	✓	x	∼	∼	a-m
NTSES – Norwegian Teacher Self-Efficacy Scale: Skaalvik und Skaalvik (2007), Neuentwicklung											
L	24	✓	7-stufig	6	Instruction: „How certain are you that you can: Provide good guidance and instruction to all stu- dents regardless of their level of ability .“ Motivate Students: „How certain are you that you can: Wake the desire to learn even among the lowest achieving students.“	✓	(✓)*	✓	∼	∼	(a-m)*
MSLS – Mehrdimensionale Skala der Lehrer-Selbstwirksamkeitserwartung: Schulte (2008), Neuentwicklung											
S/R	28	✓	4-stufig	5	Unterrichten: „Auch für unterschiedlichste Situationen weiß ich, wie ich verschiedene Medien situationsgerecht einsetze“ Leistungsbeurteilung: „Auch bei heterogenen Schülergruppen kenne ich Prinzipien der adressatenadäquaten Rückmeldung von Leistung.“	✓	∼	∼	∼	∼	a-m

Anmerkung: ✓vorhanden, ∼ teils vorhanden, x nicht vorhanden [1] Personengruppe: S (Studierende), R (Referendare), L (Lehrkräfte)
[2] Anzahl Items [3] unipolare Ratingskala [4] Antwortformat [5] Anzahl Subfaktoren/Dimensionen [6] Selbstreferentialität
[7] Handlungsorientierung [8] Schwierigkeitsbezug [9] Situations-/Kontextorientierung [10] Domänenspezifität/Mehrdimensionalität
[11] Relation (vgl. Abschnitt 3.1.3 bzw. Skinner, 1996): a-e (agent-ends), a-m (agent-means), m-e (means-ends)
* Einschätzung gilt nicht für einen Subfaktor

Fortsetzung auf der nächsten Seite...

Tabelle 4.3.: *Fortsetzung*

PG[1]	#I[2]	RS[3]	AF[4]	#Sub[5]	Itembeispiele	S[6]	H[7]	S[8]	S/K[9]	D/M[10]	Rel[11]
TEBS-Self – Teachers' Efficacy Beliefs System-Self: Dellinger u. a. (2008), Neuentwicklung											
L	30	✓	4-stufig	4/5	Classroom Management/Climate: „Right now in my present teaching situation, the strength of my personal beliefs in my capabilities to maintain a classroom climate that is fair and impartial..“ Individual Differences: „Right now in my present teaching situation, the strength of my personal beliefs in my capabilities to plan activities that accommodate the range of individual differences among my studentsy.“	✓	✓	x	∼	∼	a-m
Self-Efficacy Measure for Student Teachers in Competence-Based Education: van Dinther, Dochy, Segers und Braeken (2013), Neuentwicklung											
S	30	✓	0-100	6	Interpersonal Competence: „How confident are you that you can do the following: I see what happens in the classroom.“ Competence for collaboration with colleagues: „How confident are you that you can do the following: I keep to my agreements.“	✓	∼	x	∼	∼	a-m

Anmerkung: ✓vorhanden, ∼ teils vorhanden, x nicht vorhanden [1] Personengruppe: S (Studierende), R (Referendare), L (Lehrkräfte)
[2] Anzahl Items [3] unipolare Ratingskala [4] Antwortformat [5] Anzahl Subfaktoren/Dimensionen [6] Selbstreferentialität
[7] Handlungsorientierung [8] Schwierigkeitsbezug [9] Situations-/Kontextorientierung [10] Domänenspezifität/Mehrdimensionalität
[11] Relation (vgl. Abschnitt 3.1.3 bzw. Skinner, 1996): a-e (agent-ends), a-m (agent-means), m-e (means-ends)
[*] Einschätzung gilt nicht für einen Subfaktor

Disziplinierung, Instruktion, Planung, Evaluation und Elternarbeit abdecken. Damit versuchen Ashton u. a. der Multidimensionalität des Konstruktes gerecht zu werden, auch wenn zusätzliche Bereiche denkbar sind (Zusammenarbeit mit Kollegen, Reflexion des Unterrichts, Erziehen, ...). Die Autoren adressieren explizit, dass es sich um schwierige Situationen handeln muss und formulieren implizite Hürden, die jeweils im Vignettentext eingebaut sind: „The vignettes are inherently difficult so that a teacher need not feel pressure to report that each situation could be handled expertly“ (Ashton u. a., 1984, S. 33). Darüber hinaus entwickeln Ashton u. a. zwei unterschiedliche Ratingskalen-Formate, die eine Einschätzung sowohl bezüglich einer individuellen als auch einer sozialen Bezugsnorm einfordern.

Ashton u. a. (1984) nehmen durch die Formulierung der Vignetten die Domänen- und Situationsspezifität des Konstruktes ernst. Allerdings sind die meisten Vignetten ohne konkreten Handlungsbezug formuliert, sondern adressieren vielmehr die Konsequenzen oder Resultate nicht näher spezifizierter Handlungen („How effective would you be in motivating this student to get to work?“). Viele Items bilden daher eher „agent-ends“-Relationen ab (vgl. Tabelle 4.3). Weitere Informationen zum Validierungsprozess des Instrumentes liegen nicht vor.

4.3.2. Teacher Efficacy Scale – TES

„TES has come under an increasing amount of fire.“

Roberts & Henson, 2000, S. 5

Aus inhaltlicher Perspektive wurde das Instrument von Gibson und Dembo (1984), die Teacher Efficacy Scale (TES), bereits in Abschnitt 4.1.3, ausführlich analysiert. Es stellte sich heraus, dass die Skala entsprechend der Kriterien weder Selbstwirksamkeitserwartungen noch Handlungsergebniserwartungen im Sinne Banduras abbildet. Das Ergebnis dieser Evaluation enthält Tabelle 4.3. Nachfolgend werden darüber hinaus einige Ergebnisse der empirischen Untersuchung der Skala ergänzt. Die Autoren operationalisieren 30 Items, wobei 16 dieser Items nach einer faktorenanalytischen Untersuchung (Hauptkomponentenanalyse, orthogonale Rotation) die zwei postulierten Faktoren (TE – teaching efficacy & PTE – personal teaching efficacy) abbilden, die zusammen ca. 30 % der Gesamtvarianz erklären. Für die TE-Dimension berichten Gibson und Dembo (1984, S. 574) einen Reliabilitätskoeffizienten bezüglich der internen Konsistenz der Skala von $\alpha_C = .75$ (7 Items), für die PTE-Dimension liegt dieser Wert bei $\alpha_C = .78$ (9 Items).

Es wurde bereits dargestellt, dass aus inhaltlichen Erwägungen die TES aufgrund ihrer Nähe zu den RAND-Items eher das Konstrukt des „teacher locus of control“ abbildet (vgl. Abschnitt 4.1.3). Die inhaltlichen Validitätskriterien werden daher nicht erfüllt (vgl.

zusammenfassend Tabelle 4.3). Auch empirisch gibt es einige Hinweise für die inhaltliche Beziehung zu den RAND-Items. So korreliert in einer Studie von Woolfolk und Hoy (1990, S. 87) das erste RAND-Item lediglich mit der TE-Komponente des Konstruktes ($r = .61^{**}$). Darüber hinaus korreliert das zweite RAND-Item stärker mit der PTE-Komponente des Konstruktes ($r = .53^{**}$) als mit der TE-Facette ($r = .20^{**}$).[9] Darüber hinaus berichten Gibson und Dembo (1984, S. 571) eine geringe Korrelation beider Facetten ($r(TE/PTE) = -.19^{*}$), sodass sie von unabhängigen Faktoren ausgehen, was ebenfalls für die RAND-Items gilt.[10]

Woolfolk und Hoy (1990) überprüfen die von Gibson und Dembo (1984) vorgeschlagene Faktorstruktur und testen das Instrument gleichzeitig für angehende Lehrkräfte. Dazu benutzen sie die von Gibson und Dembo vorgeschlagenen 16 Items des TES, die in deren Analyse die TE- und PTE-Komponenten konstituieren, sowie vier zusätzliche Items, die ihrer Meinung nach auf angehende Lehrkräfte zugeschnitten sind sowie beide RAND-Items. Aufbauend auf dieser 22 Items umfassenden Version des TES nutzen Hoy und Woolfolk (1993) ein noch kürzeres Instrument, welches jeder Dimension die fünf Items mit den höchsten Faktorladungen aus der Studie von Woolfolk und Hoy (1990) zuordnet. Die Dimension der „teaching efficacy“ wird von Hoy und Woolfolk (1993, S. 357) in „general teaching efficacy“ (GTE) umbenannt – die Bezeichnung „personal teaching efficacy“ (PTE) wird beibehalten. Die Bezeichnungen werden in nachfolgenden Forschungsarbeiten weitgehend übernommen. Die inhaltlichen Argumente, die gegen die Abbildung des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen nach Bandura sprechen, gelten auch für die genannten Kurzskalen.

Exkurs: Eingehendere Analyse der Teacher Efficacy Scale Es wurde bereits mehrfach festgestellt, dass die „Teacher Efficacy Scale“ keine Selbstwirksamkeitserwartungen im Sinne Banduras abbildet, sodass eine eingehendere Analyse auf den ersten Blick hier nicht erforderlich erscheint. Die „Teacher Efficacy Scale“ ist jedoch nicht nur das am häufigsten genutzte Instrument zur Erfassung von „teacher efficacy“, sondern auch das am häufigsten hinsichtlich naheliegender Testwertinterpretationen geprüfte (vgl. u. a. Guskey & Passaro, 1994; Deemer & Minke, 1999; Brouwers & Tomic, 2003). Aus dem Nachvollziehen gängiger Kritikpunkte sowie der Wahl von Validierungsmaßnahmen können deshalb grundsätzliche Aspekte bezüglich der Herangehensweise einer Instrumentvalidierung abgeleitet werden, sodass sich eine tiefere Analyse lohnt.

[9] Dieses Ergebnis replizieren beispielsweise Tschannen-Moran und Woolfolk Hoy (2001, S. 802): $r(R_I/TE) = r(R_{II}/PTE) = .65^{**}$, $r(R_{II}/TE) \approx r(R_I/PTE) = .12^{*}$ Zusammenfassend berichten Tschannen-Moran u. a. (1998, S. 213): „When the RAND items were included in the factor analysis with the Gibson and Dembo measure, RAND 1 [...] loaded on the GTE factor, and RAND 2 [...] loaded on the PTE factor.“

[10] Dieses Resultat der Unkorreliertheit der beiden TES-Faktoren replizieren beispielsweise Woolfolk und Hoy (1990), Emmer und Hickman (1991), Rich, Lev und Fischer (1996), Tschannen-Moran und Woolfolk Hoy (2001), Woolfolk Hoy und Spero (2005).

Kritisch betrachtet wird unter anderem die postulierte Zweidimensionalität des Instrumentes. Es steht in Frage, ob diese tatsächlich die theoretische Annahme der zwei Faktoren der Internalität und Externalität widerspiegelt oder aufgrund gewählter Formulierungen hervorgerufen wird. Analysiert man beispielsweise die innere Struktur der Items, die auf die TE-Komponente laden, so rekurrieren allein sieben von neun Items auf den Einfluss des Elternhauses. Lediglich zwei Items adressieren institutionelle Rahmenbedingungen. Es ist daher wenig erstaunlich, dass diese Items faktorenanalytisch zusammenfallen. Die verbleibende Variation in den Items reicht jedoch aus, um auch andere Clusterungen insbesondere des PTE-Faktors für denkbar zu halten, gerade wenn zusätzliche Items das Instrument ergänzen. Eine Reihe von Forschungsanstrengungen werden dahingehend unternommen die Faktorstruktur zu untersuchen; dabei werden zumeist drei- (Woolfolk & Hoy, 1990; Emmer & Hickman, 1991; Soodak & Podell, 1996; Rich u. a., 1996) oder vierdimensionale Modelle geprüft (Lin & Gorrell, 1998). In der Regel verbessert sich die Varianzaufklärung gegenüber dem ursprünglichen Modell von Gibson und Dembo (1984); sie erreicht jedoch maximal 48 % (vgl. Lin & Gorrell, 1998).

Guskey und Passaro (1994) überprüfen die These von Woolfolk und Hoy (1990, S. 89), dass die zweifaktorielle Struktur deshalb auftritt, weil die Items der PTE-Skala fast durchgängig positiv („I can ...“), während die TE-Items fast ausschließlich negativ formuliert („Teachers cannot ...“) sind. Sie variieren dementsprechend systematisch die Formulierungen und finden mittels faktorenanalytischer Verfahren zwei Faktoren, die internale und externale Ursachenzuschreibungen adressieren. Ein Ergebnis dieser Untersuchung scheint zu sein, dass Lehrpersonen bei der Beantwortung des Fragebogens nicht nach dem Referenzrahmen „I vs. Teacher“ unterscheiden (vgl. Guskey & Passaro, 1994). Deemer und Minke (1999, S. 4) greifen die Vorarbeiten von Guskey und Passaro (1994) auf und führen die Argumentationskette weiter, indem sie den Einfluss von Positiv/Negativ-Formulierungen untersuchen. Ihrer Meinung nach sind die Items des externalen Faktors zumeist negativ formuliert. Sie formulieren alle Items bezüglich des Referenzrahmens „I“ und variieren systematisch zu jedem Item eine positive und eine negative Version. Faktorenanalytisch erhalten Deemer und Minke (1999) einen Faktor, der zumeist internale, allerdings positive wie negative Items enthält.

Zusammenfassend scheint die zumeist angenommene zweifaktorielle Struktur mindestens teilweise durch die Formulierungen hervorgerufen zu sein. Einschränkend sei angemerkt, dass die Studien, die den Einfluss von Formulierungen untersuchen, jeweils neue Items ohne gesonderte Validierung in die Fragebogenversionen aufnehmen. Auch Studien, die mittels eines konfirmatorischen Ansatzes die Faktorstruktur der Skala untersuchen, kommen einhellig zu dem Schluss, dass sowohl die ursprüngliche zweidimensionale (vgl. z. B. Kushner, 1993; Brouwers & Tomic, 2003; Denzine, Cooney & McKenzie, 2005), als auch alternative drei- oder vierdimensionale Modelle (vgl. z. B. Brouwers & Tomic, 2003; Denzine u. a., 2005), die Daten nicht ausreichend repräsentie-

ren. Exploratorische Verfahren legen dagegen nie die Ablehnung der Modelle nahe (vgl. z. B. Kushner, 1993).

In den genannten Studien, die die Qualität der „Teacher Efficacy Scale“ untersuchen, zeigt sich eindrücklich, welche Bedeutung Itemformulierungen haben können. Dabei kann sowohl mangelnde Variation als auch eine systematische Variation problematisch sein. Darüber hinaus offenbart sich, dass die Wahl der Analysemethode maßgeblich die Interpretation der Ergebnisse von Validierungsstudien beeinflussen kann. Betont werden sollte, dass statistische Verfahren natürlich Indizien hinsichtlich der Qualität eines Instruments liefern. Eine kritische Betrachtung des theoretischen Fundaments eines Instruments, sowie dessen Implikationen bezüglich der Itementwicklung und der Struktur des Instruments sollte allerdings in jedem Falle einer statistischen Analyse vorausgehen, um nicht den Blick für das Wesentliche – das den Skalen zugrunde gelegte theoretische Konstrukt – zu verlieren, sondern eine Verknüpfung von Theorie und statistischen Methoden zu erreichen. Denn unabhängig davon, welche Faktorstruktur die Daten angemessen abbildet, bleibt sonst die inhaltliche Bedeutung der Faktoren hinsichtlich eines Theorierahmens unklar: „[I]t is also important to note that our study focused primarily on the question, 'What do teacher efficacy scales actually measure?' We did not directly address the potentially more important question, 'What is teacher efficacy?'“ (Guskey & Passaro, 1994, S. 640).

4.3.3. Instrument zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen nach Schmitz und Schwarzer (2000)

Für den deutschsprachigen Raum gilt die Skala zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen von Schmitz und Schwarzer (2000) als die am häufigsten eingesetzte Skala.[11] Orientierend für die Konstruktion des Instrumentes setzen die Autoren die Regel, Selbstwirksamkeitserwartungen strikt von Handlungsergebniserwartungen zu trennen, weil dies als Quintessenz der Bandura-Theorie angenommen wird (vgl. Schmitz & Schwarzer, 2000, S. 13 f.). Tatsächlich werden in den Items hauptsächlich agent-means-Relationen abgebildet. Schmitz und Schwarzer arbeiten zunächst literaturbasiert relevante Facetten des Lehrerberufs heraus, bezüglich derer sich Lehrpersonen potentiell als selbstwirksam erleben können. Dazu zählen die Themenfelder berufliche Leistung, berufliche Weiterentwicklung, soziale Interaktion mit Schülern, Eltern und Kollegen sowie Umgang mit Berufsstress. Zu jedem Bereich entwickeln sie fünf bis neun Items, die auf Konstruktionsregeln (Formulierung in der ersten Person Singular, Adressierung von Fähigkeiten durch Phrasen wie „können“ oder „in der Lage sein“) und einer semantischen Struktur („subjektive Gewissheit, etwas tun zu können, auch wenn sich Hindernisse in den Weg stellen“) basieren, die sie direkt aus ihrer ausgeschärften Definition (vgl. Abschnitt 4.1.7) ableiten. Das Merkmal des Schwierigkeitsbezugs wird von den Autoren

[11] Das Instrument wurde auch ins Englische übersetzt (vgl. Schwarzer, Schmitz & Daytner, 1999).

damit in besonderer Weise berücksichtigt. Schmitz und Schwarzer (2000) geben an, aus diesem relativ großen Itempool eine ökonomische Kurzskala entwickeln zu wollen und sich bei der Itemauswahl weniger auf statistische Kennwerte als auf inhaltliche Argumente zu verlassen. Die Kurzskala soll aus jedem Themenbereich mindestens ein Item enthalten. Fraglich ist, ob tatsächlich alle relevanten Kerntätigkeiten des Lehrerberufs berücksichtigt wurden, beispielsweise fehlen Items zur Planung oder Reflexion/Evaluation des Unterrichts und auch die Durchführung von Unterricht (z. B. Instruktion) wird nur rudimentär behandelt. Gelegentlich wird an dem Instrument bemängelt, dass das Konstrukt eindimensional angelegt ist und damit einer zentralen Forderung Banduras nach Mehrdimensionalität nicht entsprochen wird. Aufgrund dessen wird die Eignung in Forschungskontexten oder der Nutzen für die Schulentwicklung in Frage gestellt (vgl. Skaalvik & Skaalvik, 2007, S. 612).

Positiv hervorzuheben ist die transparente Darstellung theoretischer Vorannahmen und die darauf aufbauende Ableitung konkreter Itemkonstruktionsregeln. Die Umsetzung erfolgt jedoch nicht immer konsequent. So werden nicht für alle Items Hürden formuliert (Itembeispiel: „Ich traue mir zu, die Schülerinnen und Schüler für neue Projekte zu begeistern."). Berücksichtigt werden sollte, dass die Hürden sowohl internale als auch externale Gegebenheiten adressieren (z. B. „Unwohlsein" vs. „skeptische Kollegen"), was ggf. die Vergleichbarkeit der Items einschränkt. Teils stehen die Items in Widerspruch zu Konstruktionsregeln von Bandura, beispielsweise wenn Handlungen in der Zukunft thematisiert werden („Ich bin mir sicher, da[ss] ich mich in Zukunft auf individuelle Probleme der Schüler noch besser einstellen kann."). In der Regel werden in den Items jedoch Selbstwirksamkeitserwartungen bezüglich komplexer Lehrerhandlungen abgebildet. In Frage steht die vornehmlich durch die Items transportierte „agent-means"-Relation lediglich für das oben bereits zitierte Item, welches auf die Begeisterung von Schülerinnen und Schülern abzielt. Begeisterung scheint mehr eine Folge bestimmter Lehrerhandlungen („ends") zu sein, sodass dieses Item eher eine „agent-ends"-Relation abbildet.

Alle 27 Items werden durchmischt an einer relativ großen Stichprobe ($N > 500$) zu zwei Messzeitpunkten (MZP) pilotiert. Welche (inhaltlichen) Kriterien letztlich zur Reduktion auf eine zehn Items umfassende Skala führen, wird nicht dargelegt. Statt dessen werden psychometrische Eigenschaften der Kurzskala berichtet, wie z. B. eine interne Konsistenz von $\alpha_C = .76$ (MZP 1) bzw. von $\alpha_C = .81$ (MZP 2): „Die Beziehungen der einzelnen Items zur Gesamtskala lagen in einem mittleren bis hohen Bereich, was sicherstellt, da[ss] kein Item redundant ist und die Breite des Konstruktes berücksichtigt wird. Die Skala erweist sich als annähernd normalverteilt" (Schmitz & Schwarzer, 2000, S. 22).

Die Zusammenhänge zu anderen Konstrukten wie dem Belastungserleben etc. erweisen sich als hypothesenkonform und stabil über die Messzeitpunkte, sodass diese als

Argumente hinsichtlich der „konvergenten“ Validität interpretiert werden. Schmitz und Schwarzer (2000, S. 19) berichten eine hohe Korrelation von $r \approx .70^{**}$ zur Skala der allgemeinen Selbstwirksamkeitserwartung und darüber hinaus leicht höhere Korrelationen der Lehrer-Selbstwirksamkeitserwartung zu anderen relevanten Konstrukten wie dem Belastungserleben oder Burnoutfaktoren als dieser Variablen zur allgemeinen Selbstwirksamkeitserwartung. Dieser Befund wird als „prädiktive Überlegenheit“ der Skala der Lehrer-Selbstwirksamkeitserwartung gegenüber der Skala der allgemeinen Selbstwirksamkeitserwartung gedeutet (vgl. Schmitz & Schwarzer, 2000, S. 20 f.). Es werden keine Maßnahmen zur Prüfung der Eindimensionalität des Instrumentes durchgeführt.

Ein fundamentales Missverständnis aus messtheoretischer Perspektive liegt vor, wenn Schmitz und Schwarzer (2000, S. 22) keine Unterschiede in den Ausprägungen zu unterschiedlichen Zeiten für verschiedene Gruppen als Indiz für die Messinvarianz des Instrumentes bezüglich dieser Gruppen heranziehen: „Eine universell einsetzbare Skala sollte unterschied[s]los gut in demographisch verschiedenen Populationen messen. Zu beiden Me[ss]zeitpunkten zeigten sich für die Lehrer-Selbstwirksamkeits-Skala in der Messung praktisch keine Unterschiede in Abhängigkeit von Geschlecht, Alter, Berufserfahrung, den Berufsjahren an der derzeitigen Schule oder der Lehre in Unter- bzw. Oberstufe.“ Von der Ausprägung eines Konstruktes in verschiedenen Gruppe kann jedoch nicht darauf geschlossen werden, dass das Instrument tatsächlich für jede Gruppe dasselbe Konstrukt abbildet.

4.3.4. Teacher Self-Efficacy Scale – TSES

Bandura entwickelt ein Instrument, welches zunächst unveröffentlicht bleibt, aber später in leicht modifizierter Weise abgedruckt wird (vgl. Bandura, 1990, 2006). Unter Wissenschaftlerinnen und Wissenschaftlern zirkuliert es jedoch schnell (vgl. Tschannen-Moran & Woolfolk Hoy, 2001, S. 791). Bandura legt sein Hauptaugenmerk auf die Mehrdimensionalität des Konstruktes und entwickelt dementsprechend 30 Items bezüglich der sieben Subskalen „efficacy to influence decision making“, „efficacy to influence school resources“, „instructional efficacy“, „disciplinary efficacy“, „efficacy to enlist parental involvement“, „efficacy to enlist community involvement“ und „efficacy to create a positive school climate“. Die Itemaufteilung ist dabei hochgradig ungleichmäßig (efficacy to create positive school climate: 8 Items, efficacy to influence school resources: 1 Item).

Über den Validierungs- und Entstehungsprozess der Items ist nichts bekannt. Es ist fraglich, ob die in den Items angesprochenen Aufgaben (z. B. „How much can you do to get churches involved in working with your school?“), tatsächlich für den Lehreralltag relevante Tätigkeiten darstellen (vgl. Tschannen-Moran & Woolfolk Hoy, 2001, S. 795; Skaalvik & Skaalvik, 2007, S. 613). Zentrale Tätigkeitsbereiche wie z. B. die Planung oder Reflexion von Unterricht werden nicht berücksichtigt.

Zudem scheinen die Items in ihrer Konstruktion den konstituierenden Merkmalen des

Konstruktes der Selbstwirksamkeitserwartungen teils nicht zu entsprechen. Beispielsweise beginnen alle Items mit der einleitenden Phrase „How much can you do to...“, was soviel bedeutet wie „Wie viel kannst du tun, *um*...“. In den Items geht es also jeweils darum einzuschätzen, wie viel getan werden könnte, um etwas *Bestimmtes* zu erreichen. Damit liegt zum Teil ein Schwerpunkt auf den zu erreichenden Zielen oder Resultaten (Motivation/Erinnerungsleistung/Hausaufgabenquote der Schülerinnen und Schüler steigern) bestimmter, jedoch in den Items nicht weiter spezifizierter Handlungen. Folglich werden durch einen Teil der Items eher „agent-ends“-Relationen transportiert als „agent-means“-Relationen. Einige der formulierten Aktivitäten scheinen zudem außerhalb oder nur zum Teil im Bereich der Kontrolle der Lehrkräfte zu liegen („How much can you do to influence the decisions that are made in the schools?“, „How much can you do to influence the class sizes in your school?“). Darüber hinaus enthalten die Items keine schwierigkeitserzeugenden Merkmale wie die von Bandura selbst geforderten Hürden (vgl. Abschnitt 3.6.2). Zusammenfassend entspricht das Instrument erschreckend wenig den von Bandura formulierten Kriterien. Dies ist ggf. damit zu erklären, dass das Instrument bereits Anfang der 1990er Jahre entwickelt wurde, die Itemkonstruktionsregeln jedoch ggf. erst später ausgeschärft wurden (vgl. Bandura, 1997, 2006).

4.3.5. Ohio State Teacher Efficacy Scale – OSTES

Tschannen-Moran und Woolfolk Hoy (2001) ergänzen auf Grundlage von Gesprächen beispielsweise mit erfahrenen Lehrpersonen die von Bandura vorgeschlagenen Tätigkeitsbereiche (vgl. vorangegangener Abschnitt 4.3.4) um „assessment“, „adjusting the lesson to individual student needs“, „dealing with learning difficulties“, „repairing student misconceptions“ und „motivating student engagement and interest“.

Inhaltlich problematisch ist es, dass sich Tschannen-Moran und Woolfolk Hoy (2001) – entsprechend ihrer theoretischen Vorarbeiten (Tschannen-Moran u. a., 1998, S. 783) und wie viele ihrer Vorgänger – auf die Definition des Konstruktes nach Armor u. a. (1976) beziehen (vgl. Abschnitt 4.1.8). Daraus ergibt sich eine Vermischung von „self-efficacy“ und „outcome-expectancy“, die sich darin äußert, dass mit den Items unter Umständen eher „means-ends“- bzw. „agent-ends“-Relationen (je nach Wertung des Handlungsbezugs) abgebildet werden (vgl. Abschnitt 4.1.4). Die angesprochene Thematik wird in einer Vielzahl der Items der OSTES virulent (Beispielitems: „How much can you do to foster student creativity?“, „How much can you do to help your students value learning?“). Items, die mit der Phrase „How well...“ beginnen, adressieren dagegen häufig eine „agent-means“-Relation („How well can you respond to difficult questions from your students?“). Handlungsbarrieren sind in die Items nicht eingearbeitet, sodass das konstituierende Merkmal des Schwierigkeitsbezugs nicht adressiert wird (vgl. Skaalvik & Skaalvik, 2007, S. 613; Avanzi u. a., 2013, S. 70).

Der erarbeitete relativ große Itempool (52 Items) wird im Zuge dreier Validierungs-

studien reduziert bzw. durch neue Items ergänzt. Ob und wenn ja, welche inhaltlichen Kriterien zum Ausschluss eines Items führen, bleibt unklar. Zumindest werden ausschließlich statistische Kennwerte wie Faktorladungen zur Rechtfertigung berichtet. Der Itempool enthält neben den Neuentwicklungen auch Originalitems von Bandura (vgl. Tabelle 4.3 bzw. Abschnitt 4.3.4), sowie Items, die anderen Instrumenten entnommen wurden (vgl. Emmer & Hickman, 1991).

Schlussendlich präsentieren Tschannen-Moran und Woolfolk Hoy (2001) eine Lang- (24 Items) und eine Kurzversion (12 Items) ihres Instrumentes, welches drei Subfaktoren aufweist („efficacy for student engagement“, „efficacy for instructional strategies“, „efficacy for classroom management“). Die Kennwerte einer Hauptachsenanalyse mit Varimaxrotation (Validierungsstudie 3, 24-Item-Version) sind vielversprechend (Varianzaufklärung: 58 %). Die Faktorladungen liegen zwischen $\lambda = .50$ und $\lambda = .78$ und auch die internen Konsistenzen der Subskalen sind sehr gut: $\alpha_C = .91$ (instruction), $\alpha_C = .90$ (management), $\alpha_C = .87$ (engagement). Die Korrelationen zwischen den Subskalen liegen zwischen $r = .58$ und $r = .70$. Für die Kurzskala ergeben sich vergleichbare Werte (vgl. Tschannen-Moran & Woolfolk Hoy, 2001, S. 798-801). Warum die dreifaktorielle Struktur auch aus inhaltlicher Perspektive sinnvoll ist, wird nicht diskutiert. Als Ausgangspunkt wurden immerhin eine Vielzahl unterschiedlicher, eigens entwickelter oder von Bandura adaptierter Dimensionen angelegt, sodass mit drei trennscharfen Dimensionen nicht unbedingt zu rechnen ist. Die geringe Anzahl von drei extrahierten Subfaktoren widerspricht in gewisser Weise der von Bandura geforderten Mehrdimensionalität des Konstruktes (vgl. Bandura, 1997; Avanzi u. a., 2013).

Die Hauptachsenanalyse mit Varimaxrotation der beiden Instrumentversionen getrennt für Lehramtsstudierende ($N = 111$) und erfahrene Lehrkräfte ($N = 255$) liefert ein differenzierteres Bild. Für die erfahrenen Lehrpersonen können drei Faktoren extrahiert werden, die dann 54 % (Langversion) bzw. 65 % (Kurzversion) der Gesamtvarianz erklären und mit den bereits genannten Faktoren übereinstimmen. Für die Lehramtsstudierenden präferieren die Autoren eine einfaktorielle Lösung, die 57 % (Langversion) bzw. 61 % (Kurzversion) der Gesamtvarianz aufklärt. Das Ergebnis erklären sich Tschannen-Moran und Woolfolk Hoy (2001, S. 801) wie folgt: „Subscale scores may have little meaning for prospective teachers who have yet to assume real teaching responsibilities.“

Wird für das Gesamtsample eine Hauptachsenanalyse mit Varimaxrotation durchgeführt, die auf die Extraktion eines Faktors ausgelegt ist, so erklärt dieser 75 % (Langversion) bzw. 68 % (Kurzversion) der Gesamtvarianz. Die Gesamtskala weist dann eine interne Konsistenz von $\alpha_C = .94$ (Langversion) bzw. von $\alpha_C = .90$ (Kurzversion) auf. Das Ergebnis deuten Tschannen-Moran und Woolfolk Hoy (2001) als Indiz für einen Faktor zweiter Ordnung und nutzen die vermeintliche Existenz eines solchen in Kombination mit den signifikanten, moderaten Korrelationen der Subskalen untereinander (Langversion: $.58 \leq r \leq .70$ bzw. Kurzversion: $.46 \leq r \leq .61$) zur Rechtfertigung

der Berechnung eines Gesamtscores und dreier Teilscores.

Für die verbesserte Qualität des Instruments spricht der große Sprung, der bezüglich der Varianzaufklärung (bis zu 75 % der Gesamtvarianz werden erklärt) im Vergleich zu anderen Instrumenten erreicht wird. Forschungsmethodisch beachtlich ist es, dass Tschannen-Moran und Woolfolk Hoy (2001) ihr Instrument bereits vor der Veröffentlichung von anderen Wissenschaftlern prüfen lassen und direkt auf deren Kritik[12] reagieren und zusätzliche Items erarbeiten, die sie in einer dritten Validierungsstudie prüfen! Aus methodischer Perspektive ist nicht einsichtig, warum wiederholt exploratorische Faktorenanalysen durchgeführt werden, obwohl mindestens im Rahmen der zweiten Studie explizit Hypothesenprüfungen im Zentrum des Interesses stehen (Roberts & Henson, 2001, S. 14 f.).

Mehrere Studien widmen sich der weiteren Validierung des Instrumentes (vgl. u. a. Heneman, Kimball & Milanowski, 2006; Klassen u. a., 2009; Fives & Buehl, 2010; Tsigilis, Koustelios & Grammatikopoulos, 2010; Duffin, French & Patrick, 2012; Mohamadi & Asadzadeh, 2012). Die dreifaktorielle Struktur des Instrumentes kann für Lehrpersonen sowohl bezüglich der Kurzskala (Heneman u. a., 2006; Klassen u. a., 2009; Fives & Buehl, 2010; Mohamadi & Asadzadeh, 2012) als auch bezüglich der Langversion (Fives & Buehl, 2010; Tsigilis u. a., 2010) bestätigt werden. In einer Studie ($N \approx 1000$) gelingt die Replikation nicht ohne stärkere Adaption der Skala (Wolters & Daugherty, 2007). Für Lehramtsstudierende wird zumeist ein einfaktorielles Modell nahegelegt (vgl. Langform: Duffin u. a., 2012; Kurz- & Langversion: Fives & Buehl, 2010). Die Forschenden gehen davon aus, dass die fehlende Lehrerfahrung erklärt, warum Studierende scheinbar nicht zwischen verschiedenen Aspekten des Unterrichtens differenzieren. Lamote und Engels (2010) führen eine Hauptkomponentenanalyse durch und finden eine zweifaktorielle Struktur in den Daten. Die Studie von Fives und Buehl (2010) kann sowohl für die Kurz- als auch die Langversion des OSTES für Lehramtsstudierende eine eindimensionale Struktur replizieren.

Darüber hinaus erweist sich die Skala als messinvariant bezüglich Lehrpersonen unterschiedlicher Kulturkreise. Dieser Befund gilt sogar, wenn die Lehrpersonen an unterschiedlichen Schulformen unterrichten (Kanada/Grund- und Sekundarschullehrpersonen vs. Singapur/Sekundarschullehrpersonen, vgl. Klassen u. a., 2009). Indizien für Messinvarianz bezüglich Subkohorten unterschiedlicher Schulformen finden sich auch in der Studie von Heneman u. a. (2006).

Die Ohio State Teacher Efficacy Scale wird später von den Autoren selbst „Teachers Sense of Efficacy Scale“ (TSES) genannt, weshalb es zu Verwechslungen mit der Skala von Bandura kommen kann (vgl. http://anitawoolfolkhoy.com/instruments/, letzter Zugriff am 20.01.2016). Die Skala wurde in mehrere Sprachen übersetzt (z. B. ins

[12] Die Existenz des dritten Faktors „efficacy for classroom management“ wird auf Grund statistischer Analysen von Roberts und Henson (2001) in Frage gestellt.

Türkische und Spanische) und gilt mittlerweile als „predominant measure of teacher efficacy throughout the world“ (Duffin u. a., 2012, S. 827).

4.3.6. Teacher Confidence Scale – TCS

Die „Teacher Confidence Scale“ von Woolfolk Hoy und Spero (2005) kann als mehr oder weniger ad-hoc, im Zuge eines konkreten Forschungsinteresses konstruiertes Instrument angesehen werden. Die Items entstehen aus Gesprächen mit Ausbilderinnen und Ausbildern über Fähigkeiten, die Referendarinnen und Referendare am Ende der Ausbildungszeit besitzen sollten. Die 32 mit Bezug auf Bandura entwickelten Items decken die große Bandbreite der Ausbildungsmodule ab („music/math/science/literacy methods“, „equity/diversity“, „educational psychology“, etc.). Die Skala ist damit mehrdimensional angelegt. Der Fokus liegt stark auf dem Unterrichten. Andere Facetten werden ausgeblendet (Umgang mit Kollegen, Eltern, Unterrichtsplanung, etc...). Die Spezifitätsniveaus einzelner Items unterscheiden sich teils stark („I am confident in my ability to teach algebra.“ vs. „I am confident in my ability to teach basic concepts of fractions.“). Die Items adressieren jeweils komplexe Handlungen, sodass klare agent-means-Relationen abgebildet werden. Allerdings werden keine schwierigkeitserzeugenden Handlungsbarrieren in die Items inkludiert. Aufgrund der vielen Items erreichen die Werte der internen Konsistenz herausragende Werte von $\alpha_C \approx .95$. Das Instrument wird aufgrund der geringen Stichprobe ($N = 53$) nicht faktorenanalytisch untersucht. Es ist stark anzunehmen, dass sich bei einer höheren Probandenzahl Subfaktoren (z. B. Lehrer-Selbstwirksamkeitserwartungen bezüglich des Unterrichtens von Mathematik) herauskristallisieren.

4.3.7. Norwegian Teachers Self-Efficacy Scale – NTSES

Ausgehend von der Problematik, dass existierende Instrumente wie die TSES oder OSTES zentrale Dimensionen des Lehrerberufs (insbesondere für das norwegische Schulsystem relevante Komponenten wie differenzierenden Unterricht) nicht aufgreifen oder aber konstituierende Merkmale wie Handlungsbarrieren unberücksichtigt lassen, entwickeln Skaalvik und Skaalvik (2007) mit dem Ziel, Banduras Itemkonstruktionshinweise zu berücksichtigen, ein neues Instrument. Aus einer Anforderungsanalyse anhand des Curriculums entwickeln sie die sechs Subskalen „Instruction“, „Adapting Instruction to Individual Students' Needs“, „Motivating Students“, „Keeping Discipline“, „Cooperating with Colleagues and Parents“ und „Coping with Changes and Challenges“. Wie bereits mehrfach erwähnt, kann es sich bei diesen sechs Facetten nicht um eine vollständige Liste relevanter Handlungsfelder für Lehrpersonen handeln.

Die Items der Skalen beziehen sich in der Regel auf eine Einschätzung komplexer Handlungen (agent-means-Relation), die durch gezielte Handlungsbarrieren zugespitzt

werden (Schwierigkeitsbezug). Lediglich für eine Subskala scheint dies nicht gelungen zu sein, da sich diese auf die Motivierung von Schülerinnen und Schülern bezieht und damit jeweils ein Ziel (Motivationssteigerung) besonders angesprochen wird (means-ends-Relation). Die Skalen werden faktorenanalytisch untersucht (konfirmatorisch und explorativ), wobei sich die intendierte sechsfaktorielle Struktur bestätigt. Die interne Konsistenz der Subskalen ist sehr gut ($.80 \leq \alpha_C \leq .91$).

Auch für die adaptierte italienische Version des Instrumentes ergeben sich gute psychometrische Kennwerte sowie eine Replikation der Faktorstruktur. Darüber hinaus erweist sich das Instrument als messinvariant für Kohorten der Länder Norwegen und Italien (vgl. Avanzi u. a., 2013).

4.3.8. Mehrdimensionale Skala der Lehrer-Selbstwirksamkeitserwartung – MSLS

Die Kritik an der eindimensionalen Konzeption der Skala von Schmitz und Schwarzer (2000) aufnehmend, orientiert sich Schulte (2008) an den bildungswissenschaftlichen Standards für die Lehrerbildung (vgl. KMK, 2004) zur Neuentwicklung eines mehrdimensionalen Instrumentes. Die in den Bildungsstandards festgelegten vier Kompetenzbereiche sind durch insgesamt 11 übergeordnete Kompetenzen beschrieben, zu denen Schulte jeweils Items konzipiert. Exploratorische Faktorenanalysen führen letztlich zu einer fünffaktoriellen Struktur mit den Faktoren „Unterrichten“, „Leistungsbeurteilung“, „Diagnostische Kompetenz“, „Kommunikation und Konfliktlösung“ sowie „Anforderungen des Lehrerberufs“. Insgesamt verbleiben nach Itemreduktion aufgrund zu geringer Trennschärfen 28 Items im Instrument (vgl. Schulte, 2008, S. 40 ff.). An die Itemkonstruktion legt die Autorin eine ähnliche semantische Struktur an, wie sie von Schmitz und Schwarzer (2000) verwendet wird: Es solle jeweils um die subjektive Gewissheit gehen, etwas tun zu können, auch wenn Schwierigkeiten aufträten (Schulte, 2008, S. 37).

Diese genannte Struktur findet sich jedoch nur in einem von fünf Faktoren wieder (Unterrichten). Die Folge ist, dass Handlungsbarrieren (Schwierigkeiten) für die Items der restlichen vier Faktoren quasi nicht formuliert werden („Nach Leistungsbeurteilungen kann ich den Schülern Lernperspektiven aufzeigen.“). Wahrscheinlich beeinflusst durch die Orientierung an den Bildungsstandards, die hauptsächlich auf Wissen abzielen (vgl. Abschnitt 6.1.2), rekurriert der Großteil der Items ebenfalls hauptsächlich auf eine Einschätzung des eigenen Wissens bzw. auf die Nutzung dieses Wissens („Unabhängig vom Thema weiß ich, wie ich Schüler in den Unterricht einbeziehe.'", „Ich bin davon überzeugt, bei Problemen in der Schule von meinen Kenntnissen über Kommunikation Nutzen ziehen zu können.“). Der Handlungsbezug ist damit lediglich in Ansätzen erkennbar und zielt in der Regel auf die Berücksichtigung des Wissens im Unterricht („Ich kenne die Vor- und Nachteile von Leistungsbeurteilungen und kann sie im Unterricht berücksichtigen.“). Die vier Items des Faktors „Anforderung an den Lehrerberuf“ stellen Paraphrasierungen der Einschätzung der eigenen Kenntnisse aus

der Stress-/Belastungsforschung dar („Um in der Schule zurecht zu kommen, kenne ich wesentliche Ergebnisse der Stressforschung.“, „Um in der Schule zurecht zu kommen, kenne ich wesentliche Ergebnisse der Belastungsforschung.“), sodass in Frage steht, ob es sich tatsächlich um einen Faktor handelt (vgl. auch die relativ hohen Itemtrennschärfen $.62 \leq r_{it} \leq .78$). Für den durch fünf Items repräsentierten Faktor „Diagnostische Kompetenz“ sind ähnliche Zweifel angebracht, da die Items lediglich auf die Identifikation von Hochbegabung und die Unterscheidung von Lernstörungen abzielen. Die Forderung bzw. das Anliegen, Lehrer-Selbstwirksamkeitserwartungen mehrdimensional zu erfassen, ist damit nur in Ansätzen umgesetzt.

Einige Items sind darüber hinaus als unverständlich einzuschätzen („Wenn ich mal das Gefühl habe, in meiner Unterrichtsplanung nicht weiter zu kommen, kenne ich Verfahren für die Beurteilung von Unterrichtsqualität.“, „Ich vertraue darauf, bei Schülern mit Schwierigkeiten, Arbeitsstörungen unterscheiden zu können.“) oder adressieren eine Einschätzung zukünftiger Fähigkeiten („Im späteren Schulalltag kann ich verschiedene Bezugssysteme der Leistungsbeurteilung gegeneinander abwägen.“), was nach Bandura vermieden werden sollte (vgl. Unterabschnitt 3.6.2).

In einer weiteren Studie überprüft Schulte (2008, S. 79-98) die gefundene Faktorstruktur für erfahrene Lehrpersonen mittels exploratorischer und konfirmatorischer Faktorenanalysen. Es deuten sich ähnliche Faktoren, jedoch teils andere Itemzuordnungen an. Ein sechsfaktorielles Modell beschreibt, auch für die Kohorte der Studierenden, die Daten letztlich am besten. Allerdings fällt die Messinvarianzprüfung negativ aus, weshalb die vorgenommenen Mittelwertvergleiche nicht zulässig und nicht interpretierbar sind.

4.3.9. Teachers' Efficacy Beliefs System-Self – TEBS-Self

Dellinger u. a. (2008, S. 751) reflektieren sehr gründlich die Definition von Armor u. a. (1976), derzufolge „teacher efficacy“ als die Überzeugung einer Lehrperson definiert wird, Schülerleistung beeinflussen zu können und kommen zu dem Schluss, dass diese das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras lediglich ungenügend abbildet (vgl. Abschnitt 4.1.5).

Auf Grundlage ihrer theoretischen Vorarbeiten (vgl. dazu auch Abschnitt 4.1.5) entwickeln Dellinger u. a. (2008, S. 756) ein neues Instrument, welches sie mit einer sehr großen Stichprobe von ca. 2400 Grundschullehrkräften validieren. In einer Pilotstudie ($N = 434$) untersuchen sie unterschiedliche Möglichkeiten der Formulierung des Itemstammes („I am able to.../I can...“ versus“ My belief in my ability to... is...“), der jeweils mit 15 nicht näher beschriebenen „teaching related tasks“ verknüpft wird. Sie berichten lediglich moderate Korrelationen der Skalen, die sich ausschließlich durch den Itemstamm zu unterscheiden scheinen. Die Autoren entscheiden sich letztlich für den „Belief-Itemstamm“, da dieser stärker an ihre theoretischen Überlegungen anzuknüpfen

scheint. Darüber hinaus passen sie den Itemstamm so an, dass er möglichst situationsspezifisch formuliert ist und Befragte dazu angeregt werden, bei der Beantwortung der Items an den eigenen Unterricht zu denken („Right now in my present teaching situation, the strength of my personal beliefs in my capabilities to...“).

Für die anschließende Itemkonstruktion leiten Dellinger u. a. (2008, S. 756) literaturbasiert die sieben Bereiche „long-range planning“, „managing the learning environment“, „classroom climate“, „enhancing and enabling learning“, „enabling thinking“, „classroom-based assessment of student learning“ und „professional responsibilities“ ab, bezüglich derer sie 51 Items entwickeln, die dann durch Expertinnen und Experten bezüglich der Wichtigkeit für das Konstrukt bewertet werden. Nach einer Itemreduktion verbleiben 30 Items in den Subskalen. Sie entscheiden sich für ein 4-stufiges Antwortformat, weil in Vorstudien die Verwendung einer 10-stufigen Skala gezeigt hat, dass lediglich von den Antwortkategorien 7-10 (insbesondere von den Kategorien 8 und 9) Gebrauch gemacht wurde.

Die Items genügen den meisten Kriterien, die aus den Charakteristika der Konstruktdefinition nach Bandura resultieren. Sie thematisieren in der Regel konkrete, komplexe Handlungen („provide students with specific feedback about their learning“, „communicate to students the specific learning outcomes of the lesson“, „communicate to students content knowledge that is accurate and logical“) und drücken damit eine „agent-means“-Relation aus. Lediglich drei von dreißig Items sind hinsichtlich dieser Relation fragwürdig und adressieren eher eine „means-ends“-Relation („provide a positive influence on the academic development of students“, „motivate students to perform to their fullest potential“, „improve the academic performance of students“). Diese Items entsprechen damit eher nicht dem Konstrukt nach Bandura. Dellinger u. a. (2008) konzipieren ihre Items ohne Handlungsbarrieren und vernachlässigen damit das Merkmal des Schwierigkeitsbezugs. Durch den gewählten Itemstamm explizieren die Autoren jedoch in besonderem Maße die Situationsspezifität des Konstruktes. Allerdings wird die Beantwortung damit auch hochgradig individualisiert und ist dadurch wahrscheinlich weniger intersubjektivierbar.

Die Validierung erfolgt in drei unabhängigen Dissertationen, wobei die Items jeweils mittels exploratorischer Faktorenanalysen untersucht werden. Leider werden nicht nur unterschiedliche Rotationsverfahren verwendet, sondern jeweils auch unterschiedlich viele sowie unterschiedlich zusammengesetzte Faktoren extrahiert, sodass die Faktorstruktur des Instruments als ungeklärt bezeichnet werden muss.

4.3.10. Self-Efficacy Measure for Student Teachers in Competence-Based Education

Es gibt nur wenige Instrumente, die explizit zur Erfassung von Selbstwirksamkeitserwartungen *angehender* Lehrpersonen entwickelt wurden (vgl. Tabelle 4.3), was den

Ausgangspunkt der Instrumententwicklung von van Dinther u. a. (2013) darstellt. Diese fußt zudem auf der konsequenten Orientierung an den niederländischen Standards zur Lehrerbildung (vgl. van Dinther u. a., 2013, S. 17). In diesen werden sieben Kompetenzbereiche formuliert, wobei sechs für angehende Lehrpersonen relevant sind und dementsprechend den Orientierungsrahmen für die Itemkonstruktion bilden („interpersonal competence“, „pedagogical competence“, „subject knowledge and methodological competence“, „organisational competence“, „competence for collaboration with colleagues“, „competence for reflection and development“). Dazu werden die angegebenen Kriterien eines jeden Kompetenzbereichs (z. B. „I demonstrate my interest in every child.“) reformuliert und die einleitende Phrase „How confident are you that you can do the following: ...“ jedem Kriterium vorangestellt. Nach dieser Vorgehensweise entstehen 44 Items, die im Rahmen von Pilotierungsstudien (Think-Aloud-Studie mit Lehramtsstudierenden und erfahrenen Lehrkräften sowie Fragebogenerhebung mit anschließender exploratorischer Faktorenanalyse) auf 31 Items reduziert werden.

Die Items genügen dabei trotz expliziter Orientierung an Banduras Vorgaben zur Itemkonstruktion (vgl. van Dinther u. a., 2013, S. 173) nicht den Kriterien. Beispielsweise werden in den Items vornehmlich unterkomplexe Handlungen beschrieben („I observe...“, „I monitor time...“, „I see what happens...“). Entsprechend hoch fallen die Mittelwerte bei einer Antwortskala von 0-100 aus (durchschnittlich 80 auf jeder Subskala). Auch Handlungsbarrieren werden nicht in die Items integriert.

Die Skalenreliabilitäten erweisen sich als gut ($.74 \leq \alpha_C \leq .96$). Anhand von konfirmatorischen Faktorenanalysen wird auf eine sechsfaktorielle Struktur (g-Faktormodell) im Datensatz geschlussfolgert. Der Modellfit erweist sich als gerade noch akzeptabel ($CFI = .92$, $RMSEA = .06$).

4.3.11. Zusammenfassung

„Subsequently, several decades of research followed using some form of the original two items or using these items as the gold standard for developing measures of teacher efficacy“ (Dellinger u. a., 2008, S. 755), wobei hier die beiden RAND-Items gemeint sind. Dieses Phänomen ist damit zu erklären, dass erste Studien vielversprechende Ergebnisse lieferten und sich die RAND-Items und die „Teacher Efficacy Scale“ daraufhin rasant verbreiteten. „[TES] was, after all, published in a leading journal and was developed through recognized and respected methodologies. Unfortunately, the theoretical and psychometric weaknesses were overlooked, and researchers of teacher efficacy prematurely foreclosed on the instrument's developmental identity“ (Henson, 2001, S. 13).

Mittlerweile ist die Problematik der Vermischung unterschiedlicher konzeptioneller Ansätze weitgehend erkannt und eine Vielzahl neuer Instrumente, die explizit mit Bezug auf Banduras sozial-kognitive Theorie sowie seine Hinweise zur Itemkonstruktion entwickelt wurden, bereichern die Forschungslandschaft. Trotzdem wird weiterhin häufig

der definitorische Ansatz gewählt, Lehrer-Selbstwirksamkeitserwartungen als Überzeugungen zu fassen, Schüler(leistungen) beeinflussen zu können („teacher efficacy“). Zudem wird dieser Ansatz als kompatibel zu Banduras konzeptionellen Annahmen bewertet.[13] Damit liegt meines Erachtens ein fundamentales Missverständnis vor (vgl. Abschnitte 3.2.1 und 4.1), welches sich weiterhin und damit auch auf neuere Operationalisierungen des Konstruktes auswirkt und sich darin äußert, dass nach wie vor Items formuliert werden, die eher „means-ends“- oder „agent-ends“-Relationen abbilden (vgl. Tabelle 4.3). Im letzten Abschnitt wurden systematisch Instrumente, die sich in theoretischer Hinsicht auf Bandura berufen bezüglich ihrer Passung zu dessen Grundannahmen, die in vorangegangenen Abschnitten bereits dargelegt wurden, überprüft. Nachfolgend werden die Ergebnisse (vgl. auch Tabelle 4.3) zusammenfassend dargestellt.

Validierungsaspekte Es wurden insgesamt zehn Instrumente auf ihre Passung zur Theorie Banduras geprüft. Alle Instrumente erfassen „teacher efficacy“ bzw. „teacher self-efficacy beliefs“ auf einem mittleren Spezifitätsniveau, weshalb immer ein Kompromiss zur Situations- und Kontextspezifität des Konstruktes eingegangen werden muss. Aus diesem Grund wurden in der Regel alle Instrumente mit „teils vorhanden“ ($\sim$) bewertet. Die meisten Instrumente adressieren das Kriterium der Domänenspezifität und berücksichtigen mehrere, wenn auch unterschiedlich viele Dimensionen (zwischen zwei bis sieben). Drei Instrumente werden eindimensional ausgerichtet, wobei die Items dann sehr heterogen formuliert werden und so der Mehrperspektivität Rechnung getragen werden soll. Das von Bandura geforderte Kriterium der Domänenspezifität/Mehrdimensionalität des Konstruktes wird allerdings nie vollumfänglich berücksichtigt, sodass dies ebenfalls immer mit „teils vorhanden“ ($\sim$) beurteilt wird. Berücksichtigt werden muss unter Umständen, dass die Domäne des Lehrerberufs schlecht definierbar ist und in der Regel kein Konsens über Handlungsfelder oder Tätigkeiten vorliegt (vgl. exemplarisch Kapitel 6).

Am häufigsten (in insgesamt sechs von zehn Fällen) berücksichtigen die Entwickler nicht die von Bandura geforderte Integration von Handlungsbarrieren zur Erzeugung schwieriger Items. Insbesondere in älteren Instrumenten adressieren die Items selten „agent-means“-Relationen, was zumeist an wenig spezifischen Handlungsdarstellungen

[13] Besonders gut lässt sich diese Behauptung an Ausschnitten des Artikels von Klassen, Tze, Betts und Gordon (2011) veranschaulichen. Die Autoren verwenden zum Beispiel zwei unterschiedliche Definitionen synonym. Die Eine betont die Zuversicht der Lehrperson, das Lernen der Schülerinnen und Schüler zu beeinflussen (S. 21), die andere betont die Zuversicht, über Fähigkeiten zu verfügen, bestimmte Handlungen durchführen zu können (S. 22). Zusätzlich weisen Klassen u. a. (2011, S. 37) darauf hin, dass „measures most congruent with self-efficacy theory attended closely to [...] context specificity (judgments related to specific outcomes).“ Mit diesem Statement rekurrieren sie klar auf die erste Definition, die jedoch gerade *nicht* kompatibel mit der Theorie Banduras ist (vgl. Abschnitt 4.1.4). Zwei Sätze später weisen sie verwirrenderweise darauf hin, dass das größte Problem der Operationalisierung des Konstruktes der „teacher self-efficacy“ die Vermischung mit dem Konstrukt der Handlungsergebniserwartung darstelle (S. 37).

und expliziter Zielperspektiven erkennbar ist. Alle Instrumente genügen dem Kriterium der Selbstreferentialität. Die einzige Ausnahme bildet die „Teacher Efficacy Scale“, die unter anderem Einschätzungen zur Wirksamkeit von Lehrpersonen im Allgemeinen einfordert.

Insgesamt kann festgestellt werden, dass keines der untersuchten Instrumente, die angelegten Kriterien vollständig erfüllt. Die Instrumente von Schmitz und Schwarzer (2000) sowie von Skaalvik und Skaalvik (2007) scheinen die inhaltlichen Kriterien am besten zu erfüllen. Ein Instrument, welches erschöpfend die durch Bandura gesetzten Kriterien erfüllt, kann daher als Forschungsdesiderat bezeichnet werden. „[T]here is a need for a new measure to assess teachers' self-efficacy beliefs as existing measures of teacher efficacy do not adequately reflect the theory of self-efficacy“ (Dellinger u. a., 2008, S. 751). Unter Umständen können existierende Instrumente ausgeschärft und überarbeitet werden, insbesondere wenn theoretisch fundiert argumentiert wird, warum die operationalisierten Dimensionen vollumfänglich die gewählte Domäne abdecken. Grundsätzlich muss festgehalten werden, dass für die neueren Instrumente im Vergleich zur vielfach genutzten und lange Zeit als Standard geltenden „Teacher Efficacy Scale“ ein deutlicher Qualitätssprung hinsichtlich der inhaltlichen Angemessenheit der Instrumente zu verzeichnen ist (vgl. Tabelle 4.3).

Instrumentcharakteristika und methodische Herangehensweisen Die meisten Instrumente sind für die Befragungsgruppe der erfahrenen Lehrpersonen entwickelt worden. Lediglich zwei Instrumente wurden für Lehramtsstudierende konzipiert (vgl. Schulte, 2008; van Dinther u. a., 2013). Selten werden Instrumente in nachfolgenden Validierungsstudien auf ihre Übertragbarkeit auf andere Personengruppen überprüft (vgl. in Ansätzen Validierungsstudien zum OSTES: Heneman u. a., 2006; Klassen u. a., 2009). Eine fundamentale Änderung kann bezüglich der statistischen Analyse neu entwickelter Skalen ausgemacht werden. Werden frühere Skalen zumeist mittels exploratorischer Faktorenanalysen untersucht (Gibson & Dembo, 1984; Woolfolk & Hoy, 1990; Tschannen-Moran & Woolfolk Hoy, 2001; Schulte, 2008; Dellinger u. a., 2008), so deutet sich für neuere Instrumente eine Verschiebung zur Nutzung konfirmatorischer Verfahren an (vgl. Skaalvik & Skaalvik, 2007; van Dinther u. a., 2013), die auch verstärkt zur Prüfung entwickelter Instrumente herangezogen werden (Kushner, 1993; Roberts & Henson, 2001; Brouwers & Tomic, 2003; Denzine u. a., 2005; Heneman u. a., 2006; Klassen u. a., 2009; Tsigilis u. a., 2010; Duffin u. a., 2012; Avanzi u. a., 2013). Das am weitesten verbreitete deutschsprachige Instrument wurde zu Validierungszwecken keinen faktorenanalytischen Verfahren unterzogen (vgl. Schmitz & Schwarzer, 2000).

Auch das Antwortformat genutzter Ratingskalen unterscheidet sich stark von Instrument zu Instrument, auch wenn jeweils unipolare Skalen verwendet werden. Die meisten Autoren orientieren sich an Banduras Empfehlung und wählen sechs oder mehr

Antwortkategorien. Drei Instrumente nutzen vierstufige Verfahren (vgl. Schmitz & Schwarzer, 2000; Schulte, 2008; TEBS-Self: Dellinger u. a., 2008). Vier Instrumente favorisieren eine ungerade Anzahl an Antwortkategorien (vgl. Ashton u. a., 1984; TSES: Bandura, 2006; OSTES: Tschannen-Moran & Woolfolk Hoy, 2001; NTSES: Skaalvik & Skaalvik, 2007).

4.3.12. Ableitungen für die Instrumententwicklungen

> „The construct of *teacher efficacy* has been around for more than a quarter century, but measurement of *teachers' self-efficacy beliefs* is still in its infancy."
>
> *Dellinger u. a., 2008, S. 763*

Aus der Analyse existierender Instrumente können einige Anhaltspunkte für die eigene Instrumententwicklung abgeleitet werden. Ein augenscheinliches Ergebnis ist es, dass es nicht ausreichend ist, formal Kriterien an die Entwicklungsprozedur anzulegen. Diese Kriterien können noch so gut begründet und aus der Theorie abgeleitet sein, wenn keine kontinuierliche Reflexion und Überprüfung der Umsetzung dieser Kriterien durchgeführt wird, so verläuft eine Orientierung an diesen ins Leere. Einige Autoren neuerer Instrumente zitieren Banduras Konstruktionsvorschriften als Orientierungsrahmen und verfehlen eine Umsetzung, weil nicht reflektiert wird, inwiefern die Regeln tatsächlich in die Itemkonstruktion eingeflossen sind bzw. welche Regeln aus welchen Gründen ggf. absichtlich nicht berücksichtigt werden. Insbesondere das Formulieren von Hürden wird meist außer Acht gelassen. Darüber hinaus wird nicht konsequent zwischen „teacher efficacy" und „teacher self-efficacy beliefs" unterschieden. Die Analyse der mit dem jeweiligen Item verknüpften Relation nach Skinner (1996) hat sich diesbezüglich als nützliches Tool erwiesen. Es sollte daher auch in der Instrumententwicklung Anwendung finden. Als Standardverfahren der statistischen Prüfung neu entwickelter Instrumente können konfirmatorische Faktorenanalysen identifiziert werden. Als zusätzliche Schwierigkeit bei der Instrumententwicklung erscheint die vollumfängliche Berücksichtigung aller potentiellen Dimensionen des Lehrerhandelns. Aus den Analyseergebnissen könnte folgende Agenda für eine theorie- und kriterienorientierte Instrumententwicklung abgeleitet werden:

- Begründung und Reflexion der zugrunde gelegten Theorie.
- Begründung, Analyse und Reflexion der zugrunde gelegten Definition.
- Darstellung von Charakteristika des Konstruktes unter Berücksichtigung der zugrunde gelegten Definition bzw. Theorie.

- Begründete Ableitung von Itemkonstruktionsregeln aus den theoretischen Annahmen.
- Itementwicklung anhand der Konstruktionsregeln.
- Überprüfung und Reflexion der Umsetzung, ggf. zyklische Überarbeitung der Items.
- Erprobung des Instruments.

Zusammenfassend sollte nach Dellinger u. a. (2008, S. 756) das Folgende bei Neuentwicklungen berücksichtigt werden: „Three issues must be addressed if a measure of teachers' self-efficacy beliefs is to improve the past and current state of assessment in this area. First, the measure should clearly, and accurately, reflect the meaning of self-efficacy. Second, the measure must assess teachers' self-efficacy beliefs in the context in which the beliefs are formed. Third, the specific tasks selected for the measure should be meaningful. For instance, the tasks of interest in teachers' classrooms and schools could reflect those tasks, skills, or abilities that are documented correlates of effective teaching and learning.“

4.4. Empirische Basis zur Relevanz des Konstruktes

> „Clearly, to draw intelligent conclusions about results of studies of teacher efficacy, we must look first at how efficacy is defined in each study.“
>
> *Woolfolk & Hoy, 1990, S. 90*

Lehrer-Selbstwirksamkeitserwartungen wird eine große Bedeutung im schulischen Kontext bezüglich des Lehrerverhaltens und des Lernens von Schülerinnen und Schülern beigemessen, was direkt aus den Annahmen der sozial-kognitiven Theorie ableitbar ist (vgl. Abschnitt 3.1). Eine Zusammenstellung potentieller Zusammenhänge des Konstruktes bezogen auf unterrichtsrelevante Lehrer- und Schülervariablen inkl. einer Plausibilitätsprüfung nehmen Woolfolk Hoy und Davis (2006) vor (vgl. Abbildung 4.6). Darüber hinaus werden weitere, eher indirekt unterrichtsrelevante Zusammenhänge beispielsweise hinsichtlich der Lehrergesundheit (z. B. Burnout, Jobzufriedenheit) angenommen.

In den letzten Jahrzehnten gab es viele Forschungsanstrengungen, um die postulierten Zusammenhänge auch empirisch untermauern zu können. In Publikationen oder Forschungsanträgen wird dann in der Regel auf diese Studien Bezug genommen, um die Bedeutung des Konstruktes zu belegen und zu veranschaulichen. Allerdings wird bei diesen Zusammenstellungen häufig missachtet, welche Instrumente die jeweiligen, als Beleg angeführten Studien nutzen. Im vorangegangenen Abschnitt 4.3

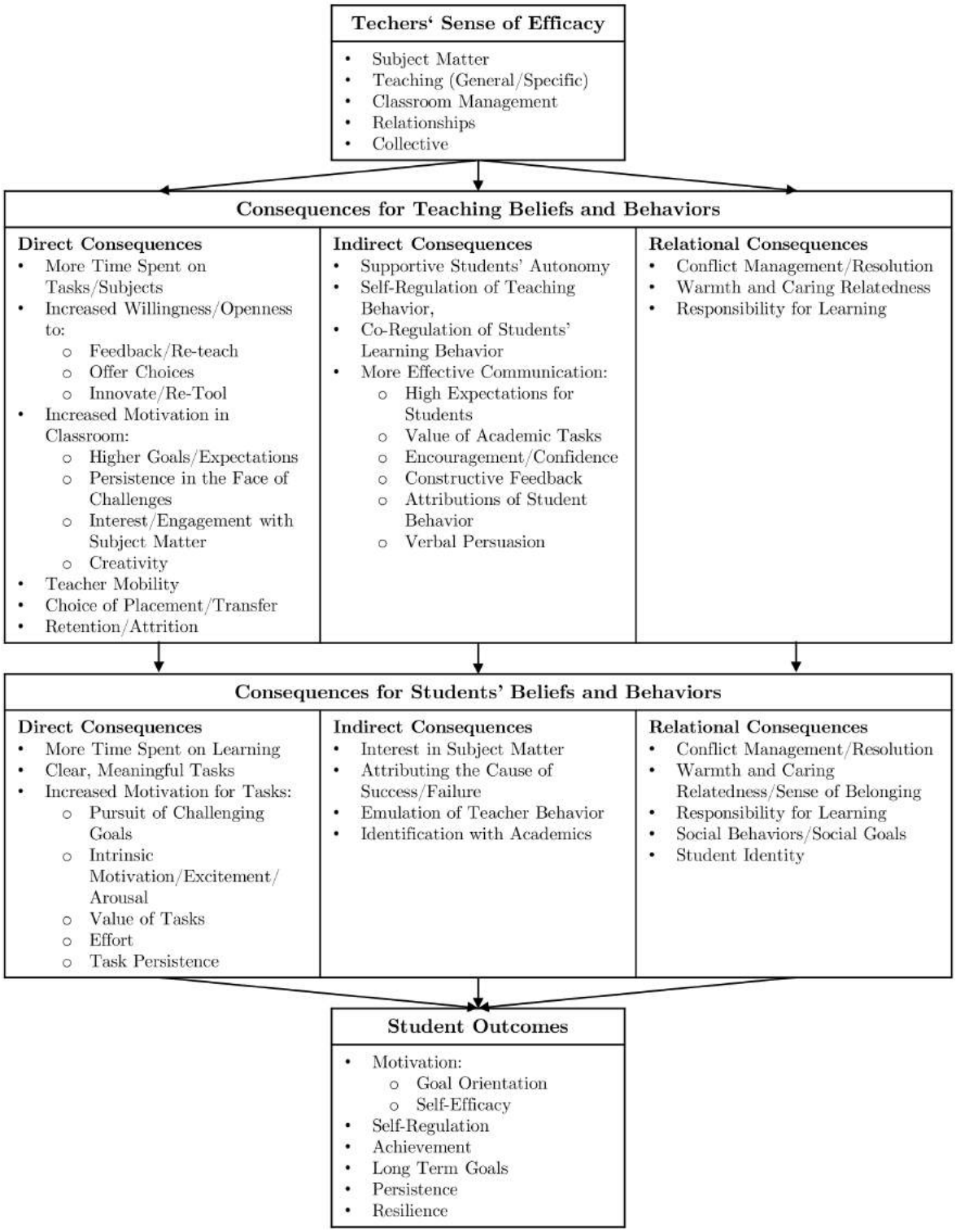


Abbildung 4.6.: Postulierte Zusammenhänge des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen zu Lehrer- und Schülervariablen (Quelle: Woolfolk Hoy & Davis, 2006, S. 125; Woolfolk Hoy, Hoy & Davis, 2009, S. 635).

konnte bereits gezeigt werden, dass Instrumente unter Umständen andere Konstrukte als Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras abfragen: „What is called teacher efficacy in a particular study may actually be teachers' sense of political power within the school, feelings of responsibility for student successes or failures, sense of academic futility, general educational philosophy, belief in their power to influence students, or some composite of these beliefs“ (Woolfolk & Hoy, 1990, S. 90). Zu solchen Erhebungsinstrumentarien zählen die „Teacher Efficacy Scale“ oder die RAND-Items, die jedoch lange Zeit standardmäßig Verwendung fanden. In der Regel kann davon ausgegangen werden, dass Studien, die vor der Jahrtausendwende durchgeführt wurden, auf diesen Instrumenten aufbauen. Es ist einsichtig, dass diese Studien, die nachweislich andere Konstrukte thematisieren als Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras, grundsätzlich nicht herangezogen werden sollten, um Zusammenhänge von Lehrer-Selbstwirksamkeitserwartungen zu interessierenden Konstrukten zu illustrieren: „Although claims about the influence of teachers' self-efficacy beliefs on students achievement, success with curriculum innovation, and so on, may be true statements, one cannot make those claims on the basis on that body of evidence if the instruments are not valid measures of teachers' self-efficacy beliefs“ (Denzine u. a., 2005, S. 690).

Allerdings ist dies allzu häufig trotzdem der Fall. *Exemplarisch* (!) kann diese Behauptung an einer Passage eines aktuellen Artikels von van Dinther u. a. (2013) zu Lehrer-Selbstwirksamkeitserwartungen nachvollzogen werden (vgl. Textausschnitt und Analyse der zitierten Studien im Anhang C, Seite 317). Die Autoren analysieren im Theorieteil ihres Artikels die Bedeutung des Konstruktes „teacher self-efficacy“ und führen zur Untermauerung ihrer Statements insgesamt 17 Studien und drei Sammelwerke bzw. Theorieartikel als Belege für den Zusammenhang des Konstruktes zu Variablen wie Schülerleistungen, Schülermotivation oder Enthusiasmus von Lehrpersonen an, wobei einige Studien mehrfach zitiert werden (vgl. Tabelle C.1 im Anhang). Allein die Hälfte der Studien verwendet die „Teacher Efficacy Scale“ oder eine Kurzversion/Adaption selbiger; in drei Studien beruhen die Aussagen der Studien auf Korrelationen mit den RAND-Items. Weitere zwei Studien nutzen Items, die nicht transparent gemacht werden oder Zusammenstellungen von Items, die als ad-hoc-Konstruktionen bezeichnet werden können und nicht den anzulegenden Standards entsprechen. Eine Studie nutzt explizit den „Responsibility for Student Achievement Questionnaire“ von Guskey (1981), der nachweislich am „locus-of-control“-Konstrukt nach Rotter (1966) orientiert ist (vgl. Abschnitt 4.2.2). Die Ergebnisse einer einzigen Studie beruhen auf der Nutzung einer neueren, adäquateren Skala.

Wenn also die Bedeutung des Konstruktes anhand empirischer Studien dargelegt werden soll, so muss in jedem Fall die Wahl der Instrumente, die in den als Beleg dienenden Studien genutzt werden, berücksichtigt werden. Insbesondere ältere Studien, d. h. Studien, die vor der Jahrtausendwende durchgeführt wurden, stehen dabei im Verdacht, inad-

äquate Erhebungsinstrumente zu nutzen, wenn Lehrer-Selbstwirksamkeitserwartungen von Interesse sind. Aber auch aktuelle Studien nutzen immer noch die „Teacher Efficacy Scale“ (vgl. u. a. Tournaki & Podell, 2005; Eren, 2009; Tuchman & Isaacs, 2011; Woodcock, 2011; Woods & Rhoades, 2013; Reilly, Dhingra & Boduszek, 2014). In ihrem Review geben Klassen u. a. (2011, S. 36) an, dass immerhin in einem Drittel ihrer recherchierten Studien dieses Instrument genutzt worden sei. Grundsätzlich sind daher auch aktuelle Studien im Hinblick auf die Verwendung der Instrumente kritisch zu analysieren. Auch deshalb, weil bei Verwendung neuerer Instrumente als der „Teacher Efficacy Scale“ oder den RAND-Items das Konstrukt zwar häufig angemessener erfasst wird, aber die verwendeten Operationalisierungen unter Umständen trotzdem (teils erhebliche) Diskrepanzen zur Definition des Konstruktes nach Bandura aufweisen (vgl. Abschnitt 4.3). Zusätzlich sind eine Reihe von Instrumenten im Umlauf, die extra zum Zweck der Untersuchung eines bestimmten Zusammenhangs des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen konzipiert wurden, jedoch selten den definitorischen Kriterien genügen (vgl. Klassen u. a., 2011, S. 36). Eine Auflistung von Itembeispielen dieser Instrumente kann im Anhang Tabelle B.1 entnommen werden.

Wenn in den nachfolgenden Abschnitten also die Bedeutung des Konstruktes anhand empirischer Studien dargestellt werden soll, dann werden erstens die in den Studien verwendeten Instrumente transparent gemacht und zweitens nur solche Studien angeführt, die sich auf eines der in Abschnitt 4.3 analysierten Instrumente stützen und gegenüber der „Teacher Efficacy Scale“ einen deutlichen Qualitätssprung aufweisen. Um solche Studien zu identifizieren, wurden im Rahmen einer ausgedehnten Literaturrecherche über 110 nach dem Jahr 2000 veröffentlichte Publikationen gesichtet. Ergebnisse qualitativ ausgerichteter Studien werden aufgrund des hier verfolgten Fokus nicht berichtet.

Durch die nachhaltige Kritik an bestehenden Instrumenten (vgl. u. a. Guskey, 1998; Henson, 2002) und die Bereitstellung neuer Instrumentarien (vgl. u. a. Schmitz & Schwarzer, 2000; Tschannen-Moran & Woolfolk Hoy, 2001) hat um die Jahrtausendwende international ein regelrechter Boom der Erforschung des Konstruktes „teacher efficacy“ bzw. „teacher self-efficacy“ eingesetzt (vgl. Klassen u. a., 2011, S. 27 ff.). Mit Blick auf die beispielhaft in Abbildung 4.6 postulierten Zusammenhänge des Konstruktes zu Lehrer- und Schülervariablen muss jedoch konstatiert werden, dass trotz des hohen Publikationsaufkommens kaum Studien vorliegen, die den genannten Qualitätskriterien entsprechen und den Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen zum Lernen oder zu Leistungen von Schülerinnen und Schülern erforschen. Zudem fallen die Ergebnisse der wenigen Studien uneindeutig bis widersprüchlich aus (vgl. Klassen u. a., 2011, S. 37). Weitaus häufiger werden „within teacher factors“ wie Jobzufriedenheit oder Burnout ins Zentrum des Forschungsinteresses gestellt, sodass diesbezüglich die empirische Evidenz als gut bis sehr gut bezeichnet werden kann (vgl. Klassen u. a., 2011, S. 37).

4.4.1. Lehrervariablen

Zielorientierung Alle nachfolgend angeführten Studien nutzen die „Ohio State Teacher Efficacy Scale“ von Tschannen-Moran und Woolfolk Hoy (2001). Sowohl Cho und Shim (2013) als auch Wolters und Daugherty (2007) berichten mittlere Korrelationen unterrichtsbezogener Zielsetzungen von Lehrerinnen und Lehrern mit deren Selbstwirksamkeitserwartungen ($.17^{*} \leq r \leq .39^{**}$), wobei die Zielsetzung von Lehrpersonen von der Wahrnehmung der gesetzten Ziele der Schule und vermittelt über die Lehrer-Selbstwirksamkeitserwartungen abhängig zu sein scheint (vgl. Cho & Shim, 2013). Auch Chong, Klassen, Huan, Wong und Kates (2010) berichten Korrelationen von Lehrer-Selbstwirksamkeitserwartungen und den wahrgenommenen, durch die Schule gesetzten Zielen (Itembeispiel der Operationalisierung des „academic climate“: „The school sets high standards for academic performance“, $r = .49^{***}$).

Unterrichtsplanung In der Studie von Stender, Brückmann und Neumann (2014) bilden Lehrer-Selbstwirksamkeitserwartungen (erfasst mittels einer Kurzversion des Instrumentes von Schmitz und Schwarzer (2000) aus der COACTIV-Studie, vgl. Baumert u. a., 2008, S. 96) neben dem Enthusiasmus, selbstregulativen Fähigkeiten und physikdidaktischem Wissen einen elementaren Anteil professioneller Kompetenz, der wiederum signifikant die Qualität von Skripten zur Unterrichtsplanung beeinflusst.

Unterrichtsqualität/Lehrerhandeln Der Einfluss von Lehrer-Selbstwirksamkeitserwartungen auf Unterrichtsqualität wird selten direkt miteinander in Beziehung gesetzt. Teils werden selbst berichtete Unterrichtspräferenzen oder andere „within teacher factors“ als Indikatoren für Unterrichtsqualität herangezogen. Klassen und Tze (2014) untersuchen in ihrer Metaanalyse 31 Studien, die unter anderem Unterrichtsqualität direkt (d. h. durch Beobachtungen) und indirekt (durch Einschätzungen der Schülerinnen und Schüler) sowie im Zusammenhang zu Lehrer-Selbstwirksamkeitserwartungen erfassen. Sie berichten eine kleine durchschnittliche Effektstärke von $\bar{r} = .28$. Der Befund ist nicht nur aufgrund der enthaltenen indirekten Erhebungsverfahren äußerst vorsichtig zu interpretieren, sondern auch weil Studien in die Analyse eingeflossen sind, die Lehrer-Selbstwirksamkeitserwartungen nur in Ausschnitten erheben (z. B. „Political Efficacy Scale“ von 1954: Trentham, Silvern & Brogdon, 1985).

Ryan, Kuusinen und Bedoya-Skoog (2015) berichten niedrige Korrelationen ($.22^{*} \leq r \leq .36^{**}$) der OSTES-Facette des „classroom managements“ zu Beobachtungskategorien des Unterrichts wie „Organisation“, sowie „instruktionaler“ und „emotionaler Unterstützung“. Im Gegensatz dazu finden Jamil, Downer und Pianta (2012, S. 125) keine Korrelationen dieser Kategorien zu den OSTES-Subskalen. In einer Studie von Heneman u. a. (2006) korrelieren alle Dimensionen der „Ohio State Teacher Efficacy Scale“ mit vier verschiedenen Dimensionen der Unterrichtsplanung und -durchführung

(.22*$\leq r \leq$.33**), wobei diese Dimensionen durch Fremdbeurteilungen erhoben wurden. In der Studie von Ruys, Van Keer und Aelterman (2011) zeigen Studierende einen Anstieg ihrer Fähigkeiten der Implementation kollaborativer Lernformen in ihren Unterricht (im Rahmen eines Praktikums), der mit ihren Lehrer-Selbstwirksamkeitserwartungen entsprechend der OSTES positiv korreliert, jedoch nicht mit ihrem Fachwissen. Beachtet werden muss, dass die Fähigkeiten hauptsächlich durch Selbsteinschätzungen erhoben wurden. Moè, Pazzaglia und Ronconi (2010) dokumentieren mittlere Korrelationen der von den Lehrpersonen selbst beschriebenen Unterrichtspraxis mit den Subskalen der OSTES ($r \approx .50$*).

Mit einer Kurzversion der Skala von Bandura untersuchen Justice, Mashburn, Hamre und Pianta (2008) den Einfluss von Lehrer-Selbstwirksamkeitserwartungen auf die Qualität der Instruktion in Vorschulklassen. Sie finden sehr schwache Zusammenhänge und in Regressionsanalysen zeigen Lehrer-Selbstwirksamkeitserwartungen keinen Einfluss auf die Qualität der Instruktion.

Holzberger, Philipp und Kunter (2014) können in ihrer Studie zeigen, dass Lehrer-Selbstwirksamkeitserwartungen (gemessen mittels einer Kurzversion des Instrumentes von Schmitz & Schwarzer, 2000) das von den Schülerinnen und Schülern eingeschätzte Unterrichtsverhalten (kognitive Aktivierung, Klassenraummanagement, Lehrer-Schülerverhältnis) vorhersagen können. Insbesondere weisen sie auf einen negativen Interaktionseffekt von Lehrer-Selbstwirksamkeitserwartungen und „intrinsic need satisfaction“ auf das wahrgenommene Lehrer-Schülerverhältnis hin. Wenn durch die Schulumwelt die entsprechende Bedürfnisbefriedigung nicht erfolgt, so haben hohe Lehrer-Selbstwirksamkeitserwartungen entsprechend negative Auswirkungen (vgl. Holzberger u. a., 2014). Es liegt nahe, den Befund so zu werten, dass Unterrichtshandeln sowohl durch externale als auch internale Faktoren beeinflusst ist. Kocher (2014, S. 176 ff.) findet dagegen *keine* Zusammenhänge zwischen Lehrer-Selbstwirksamkeitserwartungen und Unterrichtsqualitätsmerkmalen wie Instruktionseffizienz, Schülerorientierung, kognitive Aktivierung oder Klarheit/Strukturierung. Auch zum sichtbaren Unterrichtsgeschehen wie den verwendeten Sozialformen oder dem Redeanteil der Akteure finden sich keine signifikanten Korrelationen (Kocher, 2014, S. 178 ff.). Gebauer (2013, S. 129 ff.) findet dagegen schwache Zusammenhänge zwischen Lehrer-Selbstwirksamkeitserwartungen und der von den Lehrpersonen eingeschätzten Unterrichtsqualität bezüglich solcher Faktoren wie Schülerpartizipation oder differenzierte Lernkultur der jeweiligen Schule (*nicht* der eigenen Unterrichtsqualität!).

Lehrer-Selbstwirksamkeitserwartungen werden darüber hinaus sowohl mit „commitment“ (vgl. Rots, Aelterman, Vlerick & Vermeulen, 2007; Chan, Lau, Nie, Lim & Hogan, 2008, Skala: OSTES) als auch mit außerunterrichtlichem Engagement in Verbindung gebracht (vgl. Schmitz & Schwarzer, 2000).

Insgesamt ergibt sich ein uneindeutiges Bild. Einerseits werden Unterrichtsqualität

und Unterrichtshandeln der Lehrpersonen ganz unterschiedlich erfasst, was den Vergleich der Studienergebnisse erschwert. Andererseits finden sich für die beiden Instrumente, die am häufigsten verwendeten wurden (Schmitz & Schwarzer, 2000; Tschannen-Moran & Woolfolk Hoy, 2001), widersprüchliche Ergebnisse. Teils lassen sich gar keine Zusammenhänge finden, teils korrelieren nur einzelne Subfacetten oder aber es werden hohe Korrelationen (aller Subfacetten) berichtet. Es zeichnet sich Forschungsbedarf ab.

Fachwissen und Leistungen Zunächst werden Studien angeführt, die das Instrument von Schmitz und Schwarzer (2000) nutzen. Dicke u. a. (2015) finden keinen Einfluss des bildungswissenschaftlichen Wissens (vgl. BilWiss-Studie, Linninger u. a., 2015) auf Veränderungen der Lehrer-Selbstwirksamkeitserwartungen. Auch Schulte, Bögeholz und Watermann (2008) können in ihrer Studie keine Korrelation des pädagogischen Wissens zur Lehrer-Selbstwirksamkeitserwartung nach Schmitz und Schwarzer (2000) feststellen. Riese (2009, S. 157) untersucht Physiklehramtsstudierende und berichtet, dass Lehrer-Selbstwirksamkeitserwartungen weder mit dem physikalischen Fachwissen noch mit dem fachdidaktischen oder pädagogischen Wissen korrelieren. Auch Kocher (2014, S. 184) berichtet keine signifikanten Korrelationen von Lehrer-Selbstwirksamkeitserwartungen zu Dimensionen des Planungswissens von Lehramtsstudierenden oder beginnenden Lehrpersonen.

Auch in einer Studie, die Lehrer-Selbstwirksamkeitserwartungen anhand des Instrumentes von Tschannen-Moran und Woolfolk Hoy (2001) erfasst, werden keine Korrelation zwischen dem Summenscore der Lehrer-Selbstwirksamkeitserwartungen und Noten des Abschlussexamens bzw. erbrachten Leistungen im Studium berichtet (Moulding, Stewart & Dunmeyer, 2014, S. 64). Lediglich Schulte u. a. (2008) berichten für einzelne Dimensionen ihres eigenen Instrumentes – der mehrdimensionalen Skala zur Erfassung der Lehrer-Selbstwirksamkeitserwartung (MSLS) – schwache Korrelationen zum pädagogischen Wissen (Unterrichten/Leistungsbewertung: $r = .14^{*}/.24^{**}$).

Die Befunde deuten an, dass Lehrer-Selbstwirksamkeitserwartungen eher nicht mit dem Wissen oder den Leistungen in Wissenstests zusammenhängen, was theoriekonform interpretiert werden könnte. Demnach stehen vornehmlich eigene oder stellvertretende Erfahrungen im Zusammenhang zu Selbstwirksamkeitserwartungen und nicht das erworbene Wissen. Einschränkend sei angemerkt, dass überprüft werden müsste, inwiefern die Spezifitätsniveaus der jeweils eingesetzten Instrumente zueinander passen, da Ergebnisse ggf. durch eine Inkompatibilität der Niveaus verzerrt sein könnten.

Burnout und Jobzufriedenheit Bei Erhebung der Lehrer-Selbstwirksamkeitserwartung mittels der Skala nach Schmitz und Schwarzer (2000) zeigte sich vielfach, dass diese negativ mit Burnoutfacetten (vgl. z. B. Schmitz, 2001; Schwarzer & Hallum, 2008; Betoret, 2009; Dicke u. a., 2015) und positiv mit der Jobzufriedenheit (vgl.

z. B. Gebauer, 2013) korreliert. In der Studie von Dicke u. a. (2015) sagt emotionale Erschöpfung Änderungen der Lehrer-Selbstwirksamkeitserwartungen vorher, wobei dies umgekehrt nicht gilt. Schwarzer und Hallum (2008) identifizieren Lehrer-Selbstwirksamkeitserwartungen als Prädikator für Burnout. Der Zusammenhang wird über den wahrgenommenen Stress mediiert.

Der negative Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen und Burnout wird auch berichtet, wenn Lehrer-Selbstwirksamkeitserwartungen durch das Instrument von Tschannen-Moran und Woolfolk Hoy (2001) erhoben werden (vgl. u. a. Fives, Hamman & Olivarez, 2007; Mohamadi & Asadzadeh, 2012; Dicke u. a., 2014). Für die Teilfacette „classroom management“ des OSTES zeigen Dicke u. a. (2014), dass diese bei niedriger Ausprägung als Prädiktor für die wahrgenommene emotionale Erschöpfung fungiert, wobei genau genommen ein Mediationseffekt über die wahrgenommenen Störungen des Unterrichts vorliegt. Einige Studien zeigen darüber hinaus, dass Lehrpersonen mit höheren Selbstwirksamkeitserwartungen eine größere Jobzufriedenheit aufweisen (Klassen u. a., 2009; Klassen & Chiu, 2010; Moè u. a., 2010) und sich durch Stress weniger belastet fühlen (Klassen u. a., 2009; Klassen & Chiu, 2010).

Negative Korrelationen von Lehrer-Selbstwirksamkeitserwartungen zu Burnoutdimensionen sowie positive Korrelationen zur Jobzufriedenheit zeigen sich auch bei Erfassung des Konstruktes mittels der Skalen von Skaalvik und Skaalvik (2007, 2010), sodass sich unabhängig vom jeweiligen Erhebungsinstrument ein sehr einheitliches Bild ergibt.

4.4.2. Schülervariablen

Schülerleistungen Wie bereits angedeutet gibt es wenige Studien, die den Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen und Schülerleistungen mittels einigermaßen adäquater Instrumente untersuchen. Teils beruhen die Ergebnisse zwar auf enorm hohen Fallzahlen, gleichzeitig jedoch auch auf ungenügenden Instrumenten.[14] Aus diesem Grunde muss erneut darauf hingewiesen werden, den Überblicksartikel von Klassen und Tze (2014) vorsichtig zu interpretieren. Eine Reihe aufgenommener Studien sind teils nicht ohne Weiteres zugänglich (z. B. einige Dissertationen), nutzen die „Teacher Efficacy Scale“ (z. B. Anderson, Greene & Loewen, 1988; Allinder, 1995; Corkett, Hatt & Benevides, 2011) oder Instrumente, die nicht transparent gemacht werden (Muijs & Reynolds, 2002). Die gefundenen Effektstärken sind zusätzlich äußerst gering. Auch Woolfolk Hoy, Hoy und Davis (2009) rekurrieren in ihrer Überblicksdarstellung zu großen Teilen auf Studien, die die TES nutzen.

Heneman u. a. (2006) finden für die Subskalen der „Ohio State Teacher Efficacy Scale“

[14] Whitley (2010) benutzen beispielsweise ein eigens entwickeltes Instrument, welches jedoch augenscheinlich nicht den geforderten Kriterien entspricht (vgl. Tabelle B.1 im Anhang). Throndsen und Turmo (2013) nutzen dagegen ein Instrument von Midgley u. a. (2000), welches u. a. RAND-Item II enthält.

keinen Zusammenhang zu Schülerleistungen wie Lesen oder Mathematik, während Mohamadi und Asadzadeh (2012) geringe Korrelationen ($.17^{**} \leq r \leq .33^{**}$) für zwei der drei Subskalen zu Noten von Schülerinnen und Schülern berichten.

Guo, Dynia, Pelatti und Justice (2014) messen Lehrer-Selbstwirksamkeitserwartungen mittels einer etwas gekürzten Version der Bandura-Skala und finden keinen Zusammenhang zu den Sprachleistungen von Vorschulkindern. Für Kinder mit Sprachbeeinträchtigungen zeigt sich ein *negativer* Interaktionseffekt von Lehrer-Selbstwirksamkeitserwartungen und dem beobachteten Unterstützungsverhalten der Lehrperson hinsichtlich der Sprachleistung. Das bedeutet, dass Kinder von Lehrpersonen mit *niedrigen* Lehrer-Selbstwirksamkeitserwartungen und einem hohen Unterstützungsverhalten am meisten profitieren (vgl. Guo u. a., 2014, S. 17).

Da nicht viele Studien übrig bleiben, die den selbst auferlegten Kriterien genügen, um hier berichtet zu werden, muss die Forschungslage als äußerst dürftig und daher auch als wenig aussagekräftig eingeschätzt werden. Es ergibt sich daher ein Forschungsbedarf, wobei das Forschungsinteresse diesbezüglich als sehr hoch eingeschätzt werden kann.

4.4.3. Zusammenfassung

Die Evidenz zur Betonung der Bedeutung des Konstruktes erscheint bei Sichtung einiger Überblicksartikel (vgl. Woolfolk Hoy & Davis, 2006; Woolfolk Hoy u. a., 2009; Klassen & Tze, 2014) zunächst erschlagend zu sein. Bei genauerer Betrachtung werden jedoch häufig Studien zitiert, deren Ergebnisse auf eher inadäquaten Instrumenten bezüglich der Messung von Lehrer-Selbstwirksamkeitserwartungen nach Bandura beruhen (vgl. Analyse des Textauszugs von van Dinther u. a. (2013) im Anhang C). Insgesamt erscheint damit die empirische Evidenz in einem anderen Licht. Besonders wenige Studien liegen vor, die den Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen und Lehrerverhalten bzw. Schülervariablen seriös untersuchen. Ich konnte keine Studie finden, die den Einfluss von Lehrer-Selbstwirksamkeitserwartungen auf andere Schülervariablen als deren Leistungen untersucht. Dabei werden gerade bezüglich des Interesses, der Motivation oder den Wahlentscheidungen von Schülerinnen und Schülern Zusammenhänge erwartet (vgl. Abbildung 4.6), für die es damit empirisch keinerlei Belege gibt. In diesem Bereich zeichnet sich damit ein besonders hoher Forschungsbedarf ab. Gegebenenfalls können die genannten Zusammenhänge besser mit einem Instrument untersucht werden, welches Lehrer-Selbstwirksamkeitserwartungen unterrichtsfachspezifisch erfasst (vgl. Kapitel 5).

Auffällig ist, dass sich die „Ohio State Teacher Efficacy Scale“ von Tschannen-Moran und Woolfolk Hoy (2001) international etabliert und damit die „Teacher Efficacy Scale“ von Gibson und Dembo (1984) als Standardinstrument abgelöst hat. Dies ist insofern problematisch, als auch diese Skala einige Defizite aufweist (vgl. Abschnitt 4.3.5) und sich hier ggf. ein ähnliches Phänomen zeigt, wie es für die „Teacher Efficacy

Scale“ berichtet wurde: Die Faktorstruktur des Instruments wird in zahlreichen Studien untersucht und meist belegt. Inhaltlich wird die Skala dagegen kaum noch in Frage gestellt oder überhaupt kritisch diskutiert. Die Übersetzung der Skala von Schmitz und Schwarzer (2000) ins Englische (Schwarzer u. a., 1999) hat dazu beigetragen, dass auch diese Skala international Anwendung findet. Im deutschsprachigen Raum kann sie als Standardinstrument zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen bezeichnet werden. Auch wenn das Instrument nicht mehrdimensional ausgelegt ist, so bilden die Items das Konstrukt im Sinne Banduras besser ab als die Items der „Ohio State Teacher Efficacy Scale“ (vgl. Tabelle 4.3).

4.5. Quellen von Lehrer-Selbstwirksamkeitserwartungen

Insgesamt ist die Frage nach den Quellen der Lehrer-Selbstwirksamkeitserwartung wenig erforscht (vgl. Woolfolk Hoy u. a., 2009, S. 638; Klassen u. a., 2011). Es gelten jedoch für Lehrer-Selbstwirksamkeitserwartungen grundsätzlich dieselben Annahmen bezüglich der Genese und der Quellen, wie sie Bandura ganz allgemein für Selbstwirksamkeitserwartungen beschrieben hat (vgl. Abschnitt 3.3).

Korrelationsanalysen bestätigen zumeist die Annahme, dass eigene Unterrichtserfahrungen die Hauptquelle für Lehrer-Selbstwirksamkeitserwartungen bilden und emotionale Erregung in einem umgekehrten Verhältnis zu Lehrer-Selbstwirksamkeitserwartungen steht und den geringsten Einfluss ausübt (vgl. Mohamadi & Asadzadeh, 2012; O’Neill & Stephenson, 2012).[15] Mohamadi und Asadzadeh (2012, S. 431) berichten beispielsweise Korrelationen der drei Subskalen der „Ohio State Teacher Efficacy Scale“ zur Wahrnehmung der Einflussfaktoren „eigene Erfahrungen“ ($.26^{**} \leq r \leq .36^{**}$), „Beobachtungen Anderer“ ($.18^{**} \leq r \leq .21^{**}$), „verbale Überredung“ ($.23^{**} \leq r \leq .29^{**}$) sowie „emotionale Erregung“ ($-.11 \leq r \leq -.20^{**}$), die der genannten These entsprechen. Korrelationen ähnlicher Größenordnung werden auch in anderen Studien angegeben (Poulou, 2007, S. 209; O’Neill & Stephenson, 2012, S. 540). Häufig führen Studierende als Hauptquelle jedoch ihre eigenen Fähigkeiten bzw. ihre Persönlichkeit an (vgl. z. B. Poulou, 2007).

Obwohl der Großteil der Forschungsbemühungen durch eine quantitative Herangehensweise geprägt ist, widmen sich mittlerweile auch einige Studien in Form von qualitativen Ansätzen der Frage nach den Quellen oder der Beeinflussung von Lehrer-Selbstwirksamkeitserwartungen. Puchner und Taylor (2006) sowie Chong und Kong (2012) beleuchten beispielsweise in Fallstudien den Einfluss von Kooperation unter Lehrpersonen bei der Stundenplanung auf deren Lehrer-Selbstwirksamkeitserwartungen.

[15] Diese Studien nutzen das „Teacher Source Inventory“ von Poulou (2007).

4.5.1. Einflussfaktoren auf Lehrer-Selbstwirksamkeitserwartungen

Im Zuge der bereits mehrfach erwähnten umfangreichen Literaturrecherche (vgl. Abschnitte 4.3 und 4.4) konnten eine Reihe von Studien ausfindig gemacht werden, die Einflussfaktoren auf Lehrer-Selbstwirksamkeitserwartungen thematisieren. Zudem wird häufig der Zusammenhang globaler Persönlichkeitsmerkmale zum Konstrukt der Lehrer-Selbstwirksamkeitserwartung untersucht.

Schulort und andere Kontextvariablen Bis auf zwei Ausnahmen erfassen alle nachfolgend angeführten Studien Lehrer-Selbstwirksamkeitserwartung mit Hilfe der „Ohio State Teacher Efficacy Scale“ von Tschannen-Moran und Woolfolk Hoy (2001). Richter, Kunter, Lüdtke, Klusmann und Baumert (2011) und Gebauer (2013) nutzen dagegen das Instrument von Schmitz und Schwarzer (2000).

Der Standort der Schule erweist sich in einigen Studien als relevanter Kontexteffekt. So zeigen Studierende in Praxisphasen („student teaching“), deren Schule eher im städtischen bzw. ländlichen Raum gelegen ist, signifikant niedrigere Lehrer-Selbstwirksamkeitserwartungen als Studierende an einer Schule im suburbanen Raum (vgl. u. a. Knoblauch & Woolfolk Hoy, 2008; Knoblauch & Chase, 2015). Gegensätzliche Ergebnisse werden in der Studie von Tschannen-Moran und Hoy (2007, S. 950) berichtet, wonach der Schulstandort keinen Einfluss auf die Lehrer-Selbstwirksamkeitserwartung hat.

Moulding u. a. (2014) finden höhere Lehrer-Selbstwirksamkeitserwartungen für Studierende, die ihre mehrwöchige Praxisphase („student teaching“) an Schulen absolvieren, deren Schülerinnen und Schüler höhere Leistungen in staatlichen Vergleichsarbeiten zeigen. Zusätzlich berichten die Autoren, dass Studierende höhere Lehrer-Selbstwirksamkeitserwartungen aufweisen, wenn sie sich durch Mentoren unterstützt fühlen (schwache Korrelation: $r = .24^{*}$). Auch scheint die Zunahme von Lehrer-Selbstwirksamkeitserwartungen über den Zeitraum einer Praxisphase durch die wahrgenommene Unterstützung beeinflusst zu sein (vgl. Woolfolk Hoy & Spero, 2005; Fives u. a., 2007; Richter u. a., 2011). Tschannen-Moran und Hoy (2007) identifizieren für Lehramtsstudierende die wahrgenommene Unterstützung z. B. durch Kollegen als schwachen negativen Prädikator für deren Lehrer-Selbstwirksamkeitserwartungen und erklären diesen Effekt damit, dass eher weniger selbstwirksame Studierende auf Unterstützung angewiesen sind. Für erfahrene Lehrpersonen erweist sich die wahrgenommenen Unterstützung nicht als prädiktiv für deren Lehrer-Selbstwirksamkeitserwartungen.

Einen Zusammenhang des wahrgenommenen Schulklimas zu Lehrer-Selbstwirksamkeitserwartungen von Studierenden ($.20^{**} \leq r \leq .40^{**}$) finden Meristo und Eisenschmidt (2014) in ihrer Studie. Tschannen-Moran und Hoy (2007) berichten, dass insbesondere für Lehramtsstudierende die Einschätzung der zur Verfügung stehenden Ressourcen wie Materialien etc. einen Einfluss auf die Ausprägung der Lehrer-Selbstwirksamkeitserwartungen hat. Ein weiterer Einflussfaktor auf Lehrer-Selbstwirksamkeitser-

wartungen scheint in Abhängigkeit der Lehrerfahrung der Lehrperson das Verhalten des Schulleiters zu sein, wobei einige Verhaltensweisen wie Kommunikation erst mit steigender Berufserfahrung bedeutsam werden (vgl. Walker & Slear, 2011). Auch Gebauer (2013) berichtet (schwache) Korrelationen der Lehrer-Selbstwirksamkeitserwartung zu Faktoren wie der Qualität der Schulleitung, aber auch zum wahrgenommenen Arbeitsklima sowie der Güte der Kooperation mit Kolleginnen und Kollegen.

Stevens, Aguirre-Munoz, Harris, Higgins und Liu (2013) berichten einen Einfluss des Fachwissens auf die Zunahme der Lehrer-Selbstwirksamkeitserwartungen durch Weiterbildungsmaßnahmen. Lehrpersonen mit hohem fachlichen Wissen profitieren dabei in besonderem Maße.

Informelle soziale Erfahrungen (Nachhilfe, Babysitting, etc.) scheinen zunächst keinen Einfluss auf die Lehrer-Selbstwirksamkeitserwartungen zu haben, allerdings steigt der Einfluss bei vorliegenden Erfahrungen in mehr als einem Bereich (vgl. O'Neill & Stephenson, 2012, S. 540). Dagegen berichten Tuchman und Isaacs (2011) von Korrelationen informeller Erfahrungen mit Lehrer-Selbstwirksamkeitserwartungen (.13* $\leq r \leq$ .30**), wenn diese positiv konnotiert sind, wobei die Dauer der Erfahrung keine Rolle spielt.

Persönlichkeit Jamil u. a. (2012) berichten zwischen den Subskalen der „Ohio State Teacher Efficacy Scale" und dem Persönlichkeitsmerkmal der Extraversion geringe Korrelationen (.29* $\leq r \leq$ .33*) sowie niedrige negative Korrelationen der Subskalen zum Persönlichkeitsmerkmal des Neurotizismus (-.23* $\leq r \leq$ -.27*). Die anderen Persönlichkeitsmerkmale werden in der Studie nicht erhoben.

Djigić, Stojiljković und Dosković (2014) erfassen Lehrer-Selbstwirksamkeitserwartungen mit dem Instrument von Skaalvik und Skaalvik (2007) und geben für Grundschullehrkräfte die größten Zusammenhänge zwischen der Lehrer-Selbstwirksamkeitserwartung und den Dimensionen Offenheit ($r = .39^{**}$) und Gewissenhaftigkeit ($r = .43^{**}$) an, wobei Gewissenhaftigkeit sogar als Prädiktor für Lehrer-Selbstwirksamkeitserwartungen identifiziert wird.

Aus Studie 2 von Schulte (2008) kann ein Vergleich der Korrelationen von Persönlichkeitsmerkmalen zu Lehrer-Selbstwirksamkeitserwartungen, erfasst durch die beiden deutschsprachigen Instrumente, beschrieben werden. Die Autorin dokumentiert mittlere Korrelationen der Lehrer-Selbstwirksamkeitserwartungen nach Schmitz und Schwarzer (2000) zu Neurotizismus, Extraversion, Verträglichkeit und Gewissenhaftigkeit ($r = -.48^{**}/.54^{**}/.31^{**}/.34^{**}$), sowie keine signifikante Korrelation zu Offenheit. Extraversion und Neurotizismus erweisen sich als Prädiktoren für Lehrer-Selbstwirksamkeitserwartungen. Die Korrelationen der Persönlichkeitsmerkmale zu den Subskalen der Lehrer-Selbstwirksamkeitserwartungen, erfasst durch das eigene Instrument von Schulte (2008), fallen jeweils geringer aus und keines der Merkmale erweist sich als prädiktiv.

Auch Kocher (2014) findet signifikante Korrelationen von Lehrer-Selbstwirksamkeitserwartungen nach Schmitz und Schwarzer (2000) zu den Persönlichkeitsdimensionen Neurotizismus ($r = -.46^*$) und Extraversion ($r = .37^*$), die in der Größenordnung den berichteten Korrelationen von Schulte (2008) ähneln. Mayr (2011, S. 138) erklärt die hohen Korrelationen von Lehrer-Selbstwirksamkeitserwartungen nach Schmitz und Schwarzer (2000) zu Neurotizismus und Extraversion mit der Ähnlichkeit der Skala zu den B5-Facetten der Ängstlichkeit (Neurotizismus) und Durchsetzungsfähigkeit (Extraversion).

4.6. Zur Entwicklung von Lehrer-Selbstwirksamkeitserwartungen

Auch um die Entwicklung von Lehrer-Selbstwirksamkeitserwartungen anhand von solchen Studien zu beschreiben, die Instrumente nutzen, die ein Qualitätsmindestmaß aufweisen, kann die bereits mehrfach angesprochene sehr umfangreiche Literaturrecherche genutzt werden (vgl. Abschnitte 4.3, 4.4 und 4.5). Nachfolgend wird zusätzlich danach unterschieden, ob Studien längs- oder querschnittlich ausgerichtet sind.

4.6.1. Ergebnisse querschnittlich angelegter Studien

Sowohl Studien, die die „Ohio State Teacher Efficacy Scale“ nutzen (vgl. u. a. Heneman u. a., 2006; Wolters & Daugherty, 2007; Tschannen-Moran & Hoy, 2007; Cheung, 2008; Fives & Buehl, 2010; Tuchman & Isaacs, 2011), als auch Studien, die das Instrument von Schmitz und Schwarzer (2000) verwenden (vgl. u. a. Chan, 2008), berichten einen positiven Einfluss der Anzahl der Jahre an Lehrerfahrung auf die Ausprägung der Lehrer-Selbstwirksamkeitserwartungen. Schulte (2008, Studie 1) berichtet zum Beispiel höhere Lehrer-Selbstwirksamkeitserwartungen (erfasst durch MSLS) für Referendare im Vergleich zu Studierenden im ersten Studienjahr.

Diesem scheinbar eindeutigen Befund widersprechen einige Studien, die keine Unterschiede in den Ausprägungen von Lehrer-Selbstwirksamkeitserwartungen nach Schmitz und Schwarzer (2000) zwischen beginnenden und erfahrenen Lehrpersonen (vgl. Hecht, 2013) bzw. zwischen Studierenden und Referendaren (vgl. Studie 1 von Schulte, 2008) messen. Auch Gebauer (2013) findet keinen Zusammenhang von Dienstjahren und Lehrer-Selbstwirksamkeitserwartung. Aber auch in Studien, die die OSTES nutzen, zeigt sich teils kein Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen und den Unterrichtserfahrungen in Jahren (vgl. z. B. Moè u. a., 2010).

In der Studie von Klassen und Chiu (2010) zeigt sich für ein sehr großes Sample ($N \approx 1400$) ein nichtlinearer Zusammenhang von Unterrichtserfahrung in Jahren und Lehrer-Selbstwirksamkeitserwartungen (vgl. Abbildung 4.7). Den Autoren zufolge steigen diese in den ersten 23 Berufsjahren an, um dann wieder abzufallen. Unter Umständen kann das Ergebnis von Klassen und Chiu (2010) den Befund einiger Studien relativieren,

die keinen Zusammenhang von Lehrer-Selbstwirksamkeitserwartung und Dienstjahren bzw. keine Unterschiede in der Ausprägung der Lehrer-Selbstwirksamkeitserwartung von beginnenden und erfahren Lehrpersonen finden (Moè u. a., 2010; Gebauer, 2013; Hecht, 2013), weil Korrelationen in der Regel einen linearen Zusammenhang abbilden.

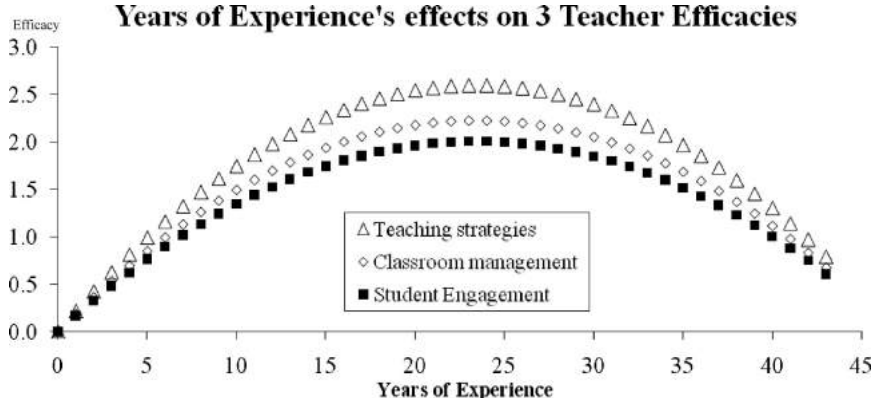


Abbildung 4.7.: Zusammenhang von Unterrichtserfahrung in Jahren und Lehrer-Selbstwirksamkeitserwartungen getrennt für die drei Dimensionen der „Ohio State Teacher Efficacy Scale“ (Klassen & Chiu, 2010, S. 747).

Lamote und Engels (2010) weisen zudem darauf hin, dass der grundsätzliche Trend des Anstiegs der Lehrer-Selbstwirksamkeitserwartungen unter Umständen von Abfällen während oder nach Praxisphasen unterbrochen wird. Die Autoren vergleichen querschnittlich Studierende mit und ohne Praxiserfahrungen (erstes vs. zweites/drittes Semester) und finden für die Subskala „classroom management“ der OSTES niedrigere Lehrer-Selbstwirksamkeitserwartungen für Studierende, die bereits Praxisphasen wie Praktika im Studium absolviert haben. Damit könnte ggf. auch der oben berichtete Befund von Schulte (2008, Studie 1) erklärt werden, falls der Erhebungszeitpunkt für die Referendare in dem Sinne unglücklich gewählt wurde, dass dieser nicht zu Beginn des Referendariats lag, sondern nach einer ausgedehnten „Praxisphase“, d. h. in der Mitte oder zum Ende des Referendariats. Zusätzlich zu dem genannten Effekt trifft wahrscheinlich das Phänomen zu, dass sich Studienanfänger (Ein Großteil der Befragten befindet sich im dritten Semester!) in der Regel überschätzen.

Längsschnittliche Erhebungen sollten ein klareres Bild bezüglich der Entwicklung von Lehrer-Selbstwirksamkeitserwartungen in unterschiedlichen Ausbildungsstadien zeichnen.

4.6.2. Ergebnisse längsschnittlich angelegter Studien

Studium Pendergast, Garvis und Keogh (2011) verzeichnen einen Abfall der Lehrer-Selbstwirksamkeitserwartungen nach Tschannen-Moran und Woolfolk Hoy (2001) im

ersten Studienjahr für Studierende des Vorschul- und Grundschullehramts, während die Lehrer-Selbstwirksamkeitserwartungen Studierender des Lehramtes für den Sekundarschulbereich keinen Veränderungen unterliegen. Lamote und Engels (2010) berichten dagegen leichte Anstiege der Lehrer-Selbstwirksamkeitserwartungen für Studierende des Lehramts für Sekundarschulen zu Beginn des Studiums.[16]

Einige Autoren berichten auch einen Anstieg der Lehrer-Selbstwirksamkeitserwartungen für das letzte Studienjahr, welches eine ausgedehnte Praxiserfahrung („student teaching“) umfasst (vgl. u. a. Fives u. a., 2007; Woolfolk Hoy & Spero, 2005; Knoblauch & Woolfolk Hoy, 2008; Knoblauch & Chase, 2015). Für den sich in der querschnittlichen Untersuchung von Lamote und Engels (2010) abzeichnende „Praxisschock“ (vgl. Unterabschnitt 4.6.1) scheinen sich hier keine Indizien finden zu lassen. Allerdings wird die Praxisphase von einigen Autoren auch als sehr gut betreut beschrieben (vgl. Woolfolk Hoy & Spero, 2005, S. 353), sodass der Mentoreneffekt (vgl. Unterabschnitt 4.5.1) den Abfall der Lehrer-Selbstwirksamkeitserwartungen ggf. verhindert und sogar ins Gegenteil verkehrt.

Auch Schulte (2008, Studie 2) berichtet Anstiege der Lehrer-Selbstwirksamkeitserwartungen (erhoben mittels des MSLS) nach Praxisphasen im Studium. Für eine sehr kleine Stichprobe ($N = 24$) und unter Verwendung der Skala von Schmitz und Schwarzer (2000) finden Krofta und Nordmeier (2014) keinen Anstieg der Lehrer-Selbstwirksamkeitserwartungen über eine Praxisphase im Studium.

Referendariat und erstes Berufsjahr Nachfolgend angeführte Studien nutzen das Instrument von Schmitz und Schwarzer (2000). Sowohl Richter u. a. (2011) als auch Dicke u. a. (2015) berichten Anstiege der Lehrer-Selbstwirksamkeitserwartungen während des Referendariats. Dem widersprechend findet Hecht (2013) keinen Anstieg der Lehrer-Selbstwirksamkeitserwartungen im ersten Jahr des Vorbereitungsdienstes; sie erweisen sich vielmehr als stabil. Kocher (2014, S. 158) untersucht die Entwicklung der Lehrer-Selbstwirksamkeitserwartungen von Junglehrerinnen und -lehrern in der Schweiz über den Zeitraum vom Ende des Studiums bis zum Ende des ersten Berufsjahres und findet keinen Anstieg der Lehrer-Selbstwirksamkeitserwartungen. Allerdings sind die Fallzahlen gering und methodische Zweifel angebracht ($N(t_1) = 21 < N(t_2) = 39$). Allerdings findet auch Keller-Schneider (2008) im ersten Berufsjahr keine Veränderungen der Lehrer-Selbstwirksamkeitserwartungen.

Erfahrene Lehrpersonen Unter Verwendung der „Ohio State Teacher Efficacy Scale“ berichten einige Forschende einen Anstieg der Lehrer-Selbstwirksamkeitserwartungen durch Weiterbildungsprogramme im Beruf (vgl. u. a. Karimi, 2011; Stevens u. a., 2013).

[16] Sie verwenden die Items der Kurzversion des OSTES, allerdings bilden sie zwei anstatt drei Faktoren.

4.6.3. Zusammenfassung und Kommentar zur Problematik der Interpretation der Studienergebnisse aufgrund der Erfassung des Konstruktes mittels defizitärer Instrumente

Woolfolk Hoy und Spero (2005) erfassen längsschnittlich zu drei Messzeitpunkten (zu Beginn der Praxisphase „student teaching“, Phase I, zu dessen Ende, Phase II und am Ende des ersten Schuljahres als „fertige“ Lehrperson, Phase III) „teacher efficacy“ unter Verwendung unterschiedlicher Instrumente (TES-Short II, TSES, TCS). Zwei dieser Instrumente (TSES, TCS) sind empirisch kaum überprüft (vgl. Abschnitte 4.3.4 und 4.3.6). Auf allen Skalen nimmt das jeweils gemessene Konstrukt zunächst zu (zwischen Phase I und II, vgl. Abbildung 4.8).

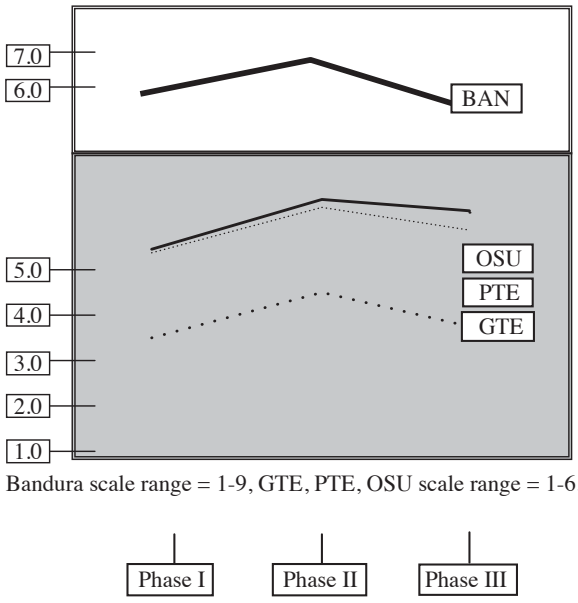


Abbildung 4.8.: Entwicklung der Lehrer-Selbstwirksamkeitserwartungen im Referendariat und ersten Jahr als Lehrperson erfasst mit BAN ≙ TSES, OSU ≙ TCS, PTE/GTE ≙ TES (Woolfolk Hoy & Spero, 2005, S. 351).

Dass die Praxisphase weniger als Praxisschock erlebt wird, wurde oben bereits diskutiert. Vielmehr scheint der Praxisschock mit dem Berufsbeginn einzutreten, denn für zwei von drei Skalen fallen zu Messzeitpunkt 3 die Werte auf das Ausgangsniveau zurück – lediglich auf der TCS wird das Niveau gehalten (vgl. Abbildung 4.8). Da die TCS am ehesten „agent-means“-Relationen abbildet, also ein Fokus auf Einschätzungen eigener Fähigkeiten liegt, während die anderen Skalen darüber hinaus Einschätzungen der Wirksamkeit auf Schülervariablen abfragen, liegt die Interpretation nahe, dass der Abfall auf diesen Skalen dadurch zustande kommt, dass Referendare zunächst

ihren Einfluss auf Schülervariablen überschätzen und diese Einschätzung mit steigender eigener Erfahrung korrigieren.

Insofern sind die berichteten Studien hinsichtlich der Entwicklung von Lehrer-Selbstwirksamkeitserwartungen vorsichtig zu interpretieren. Dies gilt insbesondere für Ergebnisse von Studien, die die OSTES nutzen, da diese teils den Einfluss auf Schülervariablen erhebt (vgl. Abschnitt 4.3.5). Insgesamt scheinen sich Lehrer-Selbstwirksamkeitserwartungen im Verlauf der Berufsausübung zu verändern. Insbesondere lassen die empirischen Befunde den Schluss zu, von einem Anstieg über die Zeiten der Berufsausbildung sowie die ersten 20 Jahre im Lehrerberuf auszugehen. Bezüglich eines Praxisschocks im Zuge des Übergangs zwischen Ausbildungsphasen oder aufgrund von Praxiserfahrungen im Studium können bislang lediglich Plausibilitätsargumente bzw. Indizien angeführt werden. Es werden weitere Studien benötigt, um die Entwicklung von Lehrer-Selbstwirksamkeitserwartungen unter Berücksichtigung unterschiedlicher Einflussfaktoren besser beschreiben zu können.

4.7. Zur Ausprägung von Lehrer-Selbstwirksamkeitserwartungen in unterschiedlichen Gruppen

Die erwähnte Literaturrecherche wurde zusätzlich genutzt, um die Studien auf interessierende Gruppenunterschiede zu sichten. Die Ergebnisse einer solchen Analyse sind für die Instrumentneuentwicklung von Relevanz, weil Mittelwertunterschiede in unterschiedlichen Gruppen im Rahmen der theoretischen Fundierung hinsichtlich ihrer Passung zum Theoriegebäude interpretiert werden müssen.

Schulform Die nachfolgend angeführten Studien nutzen, soweit nicht anders angegebenen, die „Ohio State Teacher Efficacy Scale“. In der Regel werden die höchsten Lehrer-Selbstwirksamkeitserwartungen für Lehrpersonen der Vorschule berichtet, gefolgt von Selbstwirksamkeitserwartungen von Lehrpersonen an Grundschulen, die wiederum in der Regel höher ausfallen als Selbstwirksamkeitserwartungen von Lehrkräften an Sekundarschulen (vgl. z. B. Wolters & Daugherty, 2007; Tschannen-Moran & Hoy, 2007; Klassen & Chiu, 2010; Fives & Buehl, 2010; Pendergast u. a., 2011; Meristo & Eisenschmidt, 2014; Ryan u. a., 2015). In der Studie von Tschannen-Moran und Hoy (2007, S. 952 f.) zeigt sich dieses Phänomen jedoch nur für erfahrene Lehrpersonen, nicht für Lehramtsstudierende.

Einzelne Studien differenzieren auch für den Sekundarschulbereich. So dokumentieren Chong u. a. (2010) höhere Lehrer-Selbstwirksamkeitserwartungen für Lehrpersonen an „high-track schools“ (Gymnasien) als an „middle schools“ (Realschulen) und erklären dies durch die durchschnittlich besseren Leistungen der Gymnasiasten. Dieser Befund wäre zudem mit den in Unterabschnitt 4.5.1 beschriebenen Kontexteffekt kompatibel,

wonach sie Lehrpersonen als selbstwirksamer einschätzen, wenn sie Schülerinnen und Schüler mit durchschnittlich höheren Leistungen unterrichten (vgl. auch Moulding u. a., 2014). Im Gegensatz dazu finden Wolters und Daugherty (2007) keine Unterschiede in der Ausprägung der Lehrer-Selbstwirksamkeitserwartung von Lehrkräften an „middle“ und solchen an „high schools“. Dieses Ergebnis wird im deutschen Sprachraum von Gebauer (2013) und unter Einsatz des Instrumentes von Schmitz und Schwarzer (2000) bestätigt, denn die Autorin findet ebenfalls keine Unterschiede in den Ausprägungen der Lehrer-Selbstwirksamkeitserwartung für Lehrpersonen an Gymnasien und Hauptschulen.

Geschlecht Nicht immer wird geprüft, ob Unterschiede in der Ausprägung der Lehrer-Selbstwirksamkeitserwartung aufgrund des Geschlechts vorliegen, auch weil aus theoretischer Perspektive dafür keine plausible Argumentation vorliegt. Es ist vielmehr davon auszugehen, dass bei der Beurteilung der eigenen Fähigkeit eine bestimmte Handlung auszuführen, das Geschlecht – entsprechend der Annahmen der sozial-kognitiven Theorie – keine Rolle spielen sollte. Diese Annahme resultiert daraus, dass als Hautquellen von Selbstwirksamkeitserwartungen nicht soziale Vergleiche, sondern eigene, individuelle Erfahrungen angenommen werden. Theoriekonform wird operationalisierungsübergreifend kein Einfluss des Geschlechts dokumentiert (vgl. z. B. NTSES: Djigić u. a., 2014; OSTES: Tschannen-Moran & Hoy, 2007; Pendergast u. a., 2011; O’Neill & Stephenson, 2012; MSLS: Schulte, 2008, Studie 3; Instrument von Schmitz und Schwarzer: Schmitz, 2001; Schwarzer & Hallum, 2008; Richter u. a., 2011).

Eine Ausnahme bildet die Studie von Cheung (2008), in der das Geschlecht als Prädiktor für Lehrer-Selbstwirksamkeitserwartungen identifiziert wird, wobei Frauen höhere Lehrer-Selbstwirksamkeitserwartungen aufweisen. Teils werden auch für einzelne Subdimensionen der „Ohio State Teacher Efficacy Scale“ geschlechtsspezifische Unterschiede festgestellt. So finden Djigić u. a. (2014) höhere Lehrer-Selbstwirksamkeitserwartungen im Umgang mit Disziplin und Klassen und Chiu (2010) bezüglich des „classroom managements“ für männliche Lehrpersonen.

4.8. Zusammenfassung

Im vorangegangenen Kapitel wurde die Notwendigkeit der definitorischen Klarheit sowie der präzisen Bezeichnung des interessierenden Konstruktes der Lehrer-Selbstwirksamkeitserwartungen dargelegt und veranschaulicht. Teils werden mit derselben Begrifflichkeit höchst unterschiedliche Konstrukte gemeint oder erfasst. Insbesondere ist es wichtig, zwischen „teacher efficacy“ und „teacher self-efficacy beliefs“ bzw. Lehrer-Selbstwirksamkeitserwartungen zu differenzieren, wobei Ersteres die Überzeugung adressiert, eine Wirkung auf das Lernen und Leisten von Schülerinnen und Schülern zu erzielen, während letztere die Überzeugung beschreiben, Handlungen im unterrichtlichen/schulischen

Kontext erfolgreich planen und ausführen zu können, auch wenn sich dabei Widerstände in den Weg stellen. Das Konzept der „teacher self-efficacy beliefs“ bzw. der Lehrer-Selbstwirksamkeitserwartung ist darüber hinaus an der sozial-kognitiven Theorie Banduras (1986, 1997) orientiert. Das Konzept der „teacher-efficacy“ mag ein relevanter Faktor hinsichtlich der Vorhersage von Unterrichtshandeln oder Schülerleistungen sein, jedoch sollten Gemeinsamkeiten und Unterschiede zu theoretischen Bezugspunkten wie den Theorien Rotters (1966) oder Banduras (1997) transparent gemacht werden. Dass dies häufig nicht der Fall ist, zeigt sich in den Operationalisierungen der Konstrukte „teacher efficacy“ und „teacher self-efficacy beliefs“, die häufig auf Bandura rekurrieren, aber theoretische Setzungen, wie die strikte Trennung von „agent-means“-Relationen (Selbstwirksamkeitserwartungen) und „means-ends“-Relationen (Handlungsergebniserwartungen), nicht oder unvollständig berücksichtigen.

Eine der Skalen, die beansprucht, im Sinne Banduras Lehrer-Selbstwirksamkeitserwartungen zu erfassen, ist die „Teacher Efficacy Scale“ (Gibson & Dembo, 1984), die jedoch nach Abgleich der herausgearbeiteten Kriterien als inadäquat bezeichnet werden muss. Trotzdem wurde diese Skala lange Zeit standardmäßig (und fälschlicherweise) zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen genutzt. Mittlerweile übernehmen diese Rolle die Instrumente von Tschannen-Moran und Woolfolk Hoy (2001) und Schmitz und Schwarzer (2000), wobei auch diese unter Berücksichtigung der anzulegenden Kriterien das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen lediglich mit Einschränkungen erfassen. Die detaillierte Analyse existierender Instrumente schärft den Blick dafür, wie Items formuliert sein müssen, sodass sie den Kriterien (Selbstreferentialität, Handlungsbezug, Schwierigkeitsbezug, Kontextspezifität, Mehrdimensionalität) genügen, die aus der sozial-kognitiven Theorie abgeleitet werden können. Häufig und wie im Falle des Instrumentes von Tschannen-Moran und Woolfolk Hoy (2001), der „Ohio State Teacher Efficacy Scale“, wird die Formulierung von Handlungsbarrieren und damit der Schwierigkeitsbezug sowie die konsequente Handlungsorientierung missachtet. Letzt genannter Punkt führt dabei in der Regel zu einer Fokussierung auf Handlungsziele oder -ergebnisse wie sie für das „teacher-efficacy“-Konstrukt typisch ist. Hauptkritikpunkt am Instrument von Schmitz und Schwarzer (2000) ist die fehlende Berücksichtigung des Merkmals der Mehrdimensionalität durch die Erfassung des Konstruktes durch eine zehn Items umfassende Skala. Bei aller Kritik muss festgehalten werden, dass die genannten Instrumente im Vergleich zur „Teacher Efficacy Scale“ dem Konstrukt im Sinne Banduras um einiges näher kommen (vgl. Tabelle 4.3).

Wenn die sich aus der Theorie ergebende Bedeutung oder die Entwicklungsmechanismen des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen empirisch untermauert werden sollen, so werden häufig Studien zitiert, die das Konstrukt inhaltlich eher weniger angemessen repräsentieren, da sie zur Erhebung auf Instrumente wie die erwähnte „Teacher Efficacy Scale“ zurückgreifen. Um die Bedeutung bzw. Entwicklung

von Lehrer-Selbstwirksamkeitserwartungen darstellen zu können, wurden im Rahmen einer umfangreichen Literaturrecherche über 110 Studien gesichtet, die im Zeitraum zwischen 2001 und 2015 erschienen sind. Dabei wurde nach solchen Publikationen gesucht, die im Vergleich zur „Teacher Efficacy Scale“ angemessenere Instrumente zur Erfassung des Konstruktes benutzen. Insgesamt liegt weit weniger belastbare empirische Evidenz bezüglich postulierter Zusammenhänge des Konstruktes zu anderen relevanten Konstrukten vor, als beispielsweise vor dem Hintergrund der langen Forschungstradition zu erwarten wäre. Darüber hinaus sind die Forschungsbefunde teils widersprüchlich. Die Nutzung unterschiedlicher Erhebungsinstrumente scheint jedoch nicht systematisch die Ergebnisse zu verzerren. Ungereimtheiten gehen wahrscheinlich eher auf unterschiedliche Stichproben(größen) oder die unterschiedliche Operationalisierungen der zusätzlich erhobenen Konstrukte zurück.

Zusammenfassend lässt sich konstatieren, dass insbesondere der Zusammenhang von Lehrer-Selbstwirksamkeitserwartungen zum Unterrichtshandeln von Lehrpersonen, zur Unterrichtsqualität oder zu Schüleroutputvariablen bisher unzureichend erforscht wurde. Der Zusammenhang zu Burnoutindikatoren wie der Berufszufriedenheit oder dem Stressempfinden wurde dagegen häufig in Studien untersucht und es ergibt sich das einheitliche Bild, dass Lehrer-Selbstwirksamkeitserwartungen als Ressource zur Verhinderung eines Burnouts eingestuft werden können. Bezüglich der Entwicklung des Konstruktes scheint sich die Tendenz zu ergeben, dass Lehrer-Selbstwirksamkeitserwartungen im Verlauf der Berufsausbildung und -ausübung eher anwachsen. Indizien weisen darauf hin, dass nach ca. 20-jähiger Berufserfahrung dieser Trend ins Stoppen gerät und sich ggf. umkehrt. Dies würde auch erklären, weshalb in querschnittlichen Erhebungen bei erfahrenen Lehrpersonen teils keine Korrelationen zur Berufserfahrung in Jahren auftreten. Für den vielzitierten Praxisschock können aus den durchsuchten Studien dagegen wenig mehr als erste Hinweise abgeleitet werden. Unter Umständen sind jedoch auch die gewählten Forschungsdesigns nicht dazu geeignet, diesen tatsächlich in den Daten aufzuspüren. Auch diesbezüglich kann also ein Forschungsbedarf abgeleitet werden. Entsprechend der theoretischen Vorhersage ergeben sich zumeist keine geschlechtsspezifischen Unterschiede in der Ausprägung von Selbstwirksamkeitserwartungen, allerdings scheinen eine Reihe anderer Faktoren die Ausprägung des Konstruktes zu beeinflussen. Dazu zählen beispielsweise das spezifische Lehramt (Primarstufe oder Sekundarstufe) sowie eine Reihe von Kontextfaktoren wie die wahrgenommene Unterstützung durch Mentoren oder die Schulleitung sowie die Kooperation mit Kolleginnen und Kollegen oder der Schulstandort. Unter Umständen ist es ebenfalls von der Berufserfahrung abhängig, welche Faktoren jeweils eine größere Rolle hinsichtlich der Beeinflussung der Lehrer-Selbstwirksamkeitserwartungen spielen. Es ist davon auszugehen, dass sich die Ergebnisse auf Lehrer-Selbstwirksamkeitserwartungen von Physiklehrpersonen bzw. auf Lehrer-Selbstwirksamkeitserwartungen, die auf einem spezifischeren Niveau wie

beispielsweise bezüglich physikdidaktischer Handlungsfelder erhoben werden, übertragen lassen. Dieser Annahme soll im folgenden Kapitel nachgegangen werden.

5. Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern ("science teaching self-efficacy beliefs")

„[T]eachers sense of instructional efficacy is not necessarily uniform across different subjects. [...] Therefore, teacher efficacy scales should be linked to various knowledge domains."

Bandura, 1997, S. 243

Wie durch obiges Zitat angedeutet, erscheint es sinnvoll, das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen für Unterrichtsfächer auszudifferenzieren, da davon auszugehen ist, dass jedes Unterrichtsfach eigene Spezifika und damit typische Handlungen von Lehrpersonen aufweist (z. B. Planung von Schülerexperimenten im Physikunterricht, Strukturierung einer Textinterpretation im Fremdsprachenunterricht), die in einer allgemeinen, fachübergreifenden Beschreibung nicht zum Tragen kommen, weil beispielsweise der Umgang mit Kolleginnen und Kollegen oder Eltern im Fokus des Interesses steht. Für das Unterrichtsfach Physik sind daher Lehrer-Selbstwirksamkeitserwartungen bezüglich physikdidaktischer Handlungsfelder von Relevanz. Im englischsprachigen Raum wird weniger häufig Physik, dafür das Fach Naturwissenschaften („science") unterrichtet, sodass sich die Lehrer-Selbstwirksamkeitserwartungen dann entsprechend auf naturwissenschaftsdidaktische Handlungsfelder beziehen. Nachfolgend wird von Lehrer-Selbstwirksamkeitserwartungen in naturwissenschafts- oder physikdidaktischen Handlungsfeldern gesprochen, wenn das Konstrukt im Sinne der sozial-kognitiven Theorie Banduras konzeptualisiert und operationalisiert wird. Im englischen Sprachraum wird häufig die Begrifflichkeit "science teaching self-efficacy beliefs" verwendet.

Im folgenden Kapitel sollen die jeweiligen Operationalisierungen des Konstruktes mit dem Label "science teaching self-efficacy beliefs" (oder einem ähnlichen Label) daraufhin beleuchtet werden, ob sie dem Konstrukt der Lehrer-Selbstwirksamkeitserwartungen in physik- bzw. naturwissenschaftsdidaktischen Handlungsfeldern gleichzusetzen, das heißt kompatibel mit den theoretischen Vorgaben Banduras sind. Dabei wird sich herausstellen, dass erstens nur sehr wenige Instrumente existieren und sich diese zweitens – wie eine Vielzahl von Adaptionen bezüglich verschiedener Unterrichtsfächer (vgl. für Mathematik z. B. Enochs, Smith & Huinker, 2000; für Umweltbildung z. B. Sia, 1992) – zumeist explizit auf die bereits ausführlich beschriebene und kritisierte „Teacher

Efficacy Scale“ nach Gibson und Dembo (1984) beziehen. Einige der Hauptkritikpunkte gelten daher auch für die hier dargestellten Skalen. Insgesamt wird deshalb festgestellt werden, dass die Instrumente nicht zur Erfassung des interessierenden Konstruktes der Lehrer-Selbstwirksamkeitserwartungen in physik- bzw. naturwissenschaftsdidaktischen Handlungsfeldern geeignet erscheinen, woraus sich das Forschungsdesiderat der Entwicklung eines neuen Instrumentes ergibt, welches die Vorgaben Banduras umsetzt. Schließlich muss festgestellt werden, dass aufgrund der fehlenden Möglichkeiten der adäquaten Erfassung des Konstruktes keine seriöse Darstellung empirischer Evidenz hinsichtlich der Interaktion mit anderen interessierenden Konstrukten wie dem Unterrichtshandeln von Physiklehrpersonen oder den Entwicklungsmechanismen der Lehrer-Selbstwirksamkeitserwartungen in physik- bzw. naturwissenschaftsdidaktischen Handlungsfeldern vorgenommen werden kann. Die Notwendigkeit der Entwicklung eines neuen Instrumentes, dessen Testwerte valide als Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern (im Sinne Banduras sozial-kognitiver Theorie) interpretiert werden können, wird damit erneut ersichtlich.

5.1. Zur Definition des Konstruktes

Das Konstrukt der Lehrer-Selbstwirksamkeitserwartungen in physik- oder naturwissenschaftsdidaktischen Handlungsfeldern wird im Rahmen von Forschungsanstrengungen zur Instrumententwicklung lediglich andeutungsweise definitorisch gefasst (vgl. u. a. Riggs & Enochs, 1990; Roberts & Henson, 2000; Ritter, Boone & Rubba, 2001). Die Forschenden beziehen sich zumeist auf die theoretische Basis der sozial-kognitiven Theorie Banduras und damit auf die Definition von Selbstwirksamkeitserwartungen nach Bandura (vgl. Abschnitt 3.2.1). Sie geben dann an, das Konstrukt der Selbstwirksamkeitserwartungen unterrichtsfachspezifisch zu operationalisieren ohne jedoch die Definition diesbezüglich auszuschärfen bzw. die Ansprüche, die aus den theoretischen Vorgaben resultieren und im Zuge einer konkreten Operationalisierung berücksichtigt werden müssen, in einer definitorischen Anpassung aufzugreifen und damit zu reflektieren. Theoretische Reflexionen beschränken sich auf die Feststellung, Selbstwirksamkeitserwartungen von Handlungsergebniserwartungen abgrenzen zu müssen. Dass Definitionen lediglich auf globaleren Spezifitätsniveaus angegeben werden, ist einerseits nachvollziehbar, weil das Konstrukt der (Lehrer-)Selbstwirksamkeitserwartungen Forschenden durch die Vielzahl an Vorarbeiten intuitiv zugänglich und damit auch leicht adaptierbar erscheint. Bleibt eine präzise und tiefgreifende Fassung des zu operationalisierenden Konstruktes jedoch aus, so kann eine solche definitorische Basis auch nicht als Kriterium für eine gelungene oder verfehlte Operationalisierung herangezogen werden. Cakiroglu, Capa-Aydin und Woolfolk Hoy (2012, S. 449) geben in ihrem Überblicksartikel eine relativ unspezifische Definition/Beschreibung des Konstruktes an,

die augenscheinlich eine Reihe theoretischer Eckpunkte nicht aufgreift. Sie verstehen unter „science teaching efficacy beliefs [the teachers] confidence in their abilities to teach science and help students learn." Aus den Ausführungen wird deutlich, dass eine begriffliche Ausschärfung und präzise definitorische Fassung des Konstruktes ein Forschungsdesiderat darstellt. Als Grundlage der Instrumentenentwicklung wird in Kapitel 7 unter Bezugnahme auf die sozial-kognitive Theorie Banduras der Versuch einer Definition des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern unternommen.

5.2. Zur Erfassung des Konstruktes

„The TES provided the seed for growing *subject-specific* scales."

Pruski, 2013, S. 1138

Ausgehend von dem Überblicksartikel von Cakiroglu u. a. (2012) werden im Rahmen einer Literaturrecherche Erhebungsinstrumente des Konstruktes gesucht. Dazu werden Publikationen der letzten zehn Jahre systematisch gesichtet. Entsprechend der Vorgehensweise in Abschnitt 4.3 werden die Instrumente anschließend insbesondere daraufhin analysiert, ob sie den in Abschnitt 3.2.2 herausgearbeiteten Eigenschaften des Konstruktes der Selbstwirksamkeitserwartungen (Selbstreferentialität, Handlungsorientierung, Schwierigkeitsbezug, Situations- und Domänenspezifität) entsprechen. Es wurden solche Instrumente zur Prüfung herangezogen, die in der Forschungsgemeinschaft häufig genutzt werden, einen längeren Validierungsprozess durchlaufen haben und vollständig publiziert wurden. Zusätzlich wurde nach deutschsprachigen Instrumenten Ausschau gehalten. Zwei Instrumente, die mit dem Ziel der Erfassung von Lehrer-Selbstwirksamkeitserwartungen in physik- oder naturwissenschaftsdidaktischen Handlungsfeldern hinsichtlich der Implementation forschend-entdeckenden Lernens entwickelt wurden (Smolleck, Zembal-Saul & Yoder, 2006; Walan & Chang Rundgren, 2014), werden aufgrund unvollständiger Publikation bzw. augenscheinlicher Defizite von der Analyse ausgeschlossen. Ein Überblick über die nicht in die Analyse einbezogenen Instrumente findet sich im Anhang D in Tabelle D.1. Itembeispiele dieser Instrumente verdeutlichen zudem, dass die Instrumente nicht den durch Bandura formulierten Ansprüchen genügen.

Die Zusammenfassung der Prüfung, ob und inwiefern die recherchierten Instrumente den Vorgaben Banduras entsprechen und tatsächlich Selbstwirksamkeitserwartungen in der Tradition der sozial-kognitiven Theorie erfassen, kann Tabelle 5.1 entnommen werden. Insgesamt konnten lediglich drei englischsprachige Instrumente, sowie eine deutschsprachige Adaption zur Erfassung von „science teaching efficacy beliefs" gefunden

Tabelle 5.1.: Charakteristika der Erhebungsinstrumente zur Erfassung von „science teaching self-efficacy beliefs“.

Instrumentcharakteristika						**Validierungsaspekte**					
PG[1]	#I[2]	RS[3]	AF[4]	#Sub[5]	Itembeispiele	S[6]	H[7]	S[8]	S/K[9]	D/M[10]	Rel[11]
STEBI-A – Science Teaching Efficacy Beliefs Instrument/Form A: Riggs und Enochs (1990), Adaption nach Gibson und Dembo (1984)											
L	25	✓	5-stufig	2	Personal Science Teaching Efficacy: „When I teach science, I typically welcome student questions.“ Science Teaching Outcome Expectancy Scale: „When the science grades of students improve, it is most often due to their teacher having found a more effective teaching approach.“	∼	∼	x	x	x	m-e a-m
STEBI-B – Science Teaching Efficacy Beliefs Instrument/Form B: Enochs und Riggs (1990), Adaption nach Gibson und Dembo (1984)											
S	23	✓	5-stufig	2	Personal Science Teaching Efficacy: „When I teach science, I will typically welcome student questions.“ Science Teaching Outcome Expectancy: siehe STEBI-A	∼	∼	x	x	x	m-e a-m

Anmerkung: ✓vorhanden, ∼ teils vorhanden, x nicht vorhanden [1] Personengruppe: S (Studierende), R (Referendare), L (Lehrkräfte) [2] Anzahl Items [3] unipolare Ratingskala [4] Antwortformat [5] Anzahl Subfaktoren/Dimensionen [6] Selbstreferentialität [7] Handlungsorientierung [8] Schwierigkeitsbezug [9] Situations-/Kontextorientierung [10] Domänenspezifität/Mehrdimensionalität [11] Relation (vgl. Abschnitt 3.1.3 bzw. Skinner, 1996): a-e (agent-ends), a-m (agent-means), m-e (means-ends)

Fortsetzung auf der nächsten Seite...

Tabelle 5.1.: *Fortsetzung*

PG[1]	#I[2]	RS[3]	AF[4]	#Sub[5]	Itembeispiele	S[6]	H[7]	S[8]	S/K[9]	D/M[10]	Rel[11]
SETAKIST – Self-Efficacy Teaching and Knowledge Instrument for Science Teachers: Roberts und Henson (2000), aufbauend auf Riggs und Enochs (1990)											
L	16	✓	5-stufig	2	Teaching Efficacy: „I feel anxious when teaching science content that I have not taught before.“ Knowledge Efficacy „After I have taught a science concept once, I feel confident teaching it again.“	∼	∼	x	x	x	a-m
SEBEST – Self-Efficacy Beliefs about Equitable Science Teaching and Learning: Ritter u. a. (2001), aufbauend auf Enochs und Riggs (1990)											
S	34	✓	5-stufig	2	personal self-efficacy: „I will be able to effectively teach science to children whose first language is not English.“ outcome expectancy: „Children who are English Language Learners can be successful in learnig science if the teaching is effective.“	∼	x	∼	x	x	a/m-e a-m
Selbstwirksamkeitserwartung bezogen auf das Physik-Lehren: Riese (2009), Kurzversion nach Enochs und Riggs (1990), Bleicher (2004)											
S	6	✓	4-stufig	2	personal self-efficacy: „Wenn ich die Wahl hätte, würde ich meinen künftigen Schulleiter lieber zur Beurteilung meines weiteren Unterrichtsfachs und nicht zur Beurteilung meines Physikunterrichts einladen.“	∼	x	x	x	x	a-m

Anmerkung: ✓vorhanden, ∼ teils vorhanden, x nicht vorhanden [1] Personengruppe: S (Studierende), R (Referendare), L (Lehrkräfte)
[2] Anzahl Items [3] unipolare Ratingskala [4] Antwortformat [5] Anzahl Subfaktoren/Dimensionen [6] Selbstreferentialität
[7] Handlungsorientierung [8] Schwierigkeitsbezug [9] Situations-/Kontextorientierung [10] Domänenspezifität/Mehrdimensionalität
[11] Relation (vgl. Abschnitt 3.1.3 bzw. Skinner, 1996): a-e (agent-ends), a-m (agent-means), m-e (means-ends)

werden. Es zeigt sich, dass sich die Instrumente jeweils stark an der „Teacher Efficacy Scale“ von Gibson und Dembo (1984) orientieren, sodass Kritikpunkte zu großen Teilen auch für die spezifischeren Instrumente gelten. Nachvollziehbar wird diese Einschätzung anhand der detaillierten Itemevaluation in den nachfolgenden Abschnitten.

5.2.1. Science Teaching Efficacy Beliefs Instrument – STEBI

Enochs und Riggs (1990) entwickeln basierend auf den Vorarbeiten von Gibson und Dembo (1984) ein Instrument zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen bezüglich des Unterrichtens der Naturwissenschaften („science teaching efficacy beliefs instrument (STEBI-A)“ für Lehrpersonen im Elementarbereich. Gleichzeitig entwickeln Riggs und Enochs (1990) eine Version für Primarstufen-Lehramtsstudierende (STEBI-B).

In der Tradition der „Teacher Efficacy Scale“ (TES) halten sie an der Operationalisierung zweier Faktoren mit dem Argument fest, dass nach Bandura sowohl Selbstwirksamkeitserwartungen als auch Handlungsergebniserwartungen Verhalten vorhersagen. Entsprechend entwickeln sie ein Instrument mit den beiden Faktoren „personal science teaching efficacy beliefs (PSTEB)“ und „science teaching outcome expectancy (STOE)“. Obwohl sich die Autoren stark am Instrument von Gibson und Dembo (1984) anlehnen, kritisieren Enochs und Riggs (1990, S. 627) an diesem, dass der Faktor der „personal teaching efficacy“ bereits eine Kombination der von Bandura strikt getrennten Konstrukte darstelle. Tatsächlich differieren die Items der jeweiligen Selbstwirksamkeitsskalen beider Instrumente (PTE/PSTEB), während die Faktoren der „teaching efficiacy“ (TES) bzw. „science teaching outcome expectancy“ (STEBI) nahezu identisch sind (vgl. Tabelle 5.2).[1]

Tabelle 5.2.: Itemvergleich der beiden Faktoren der „Teacher Efficacy Scale“ (TES) nach Gibson und Dembo (1984) und des „Science Teaching Efficacy Beliefs Instrument“ (STEBI) nach Riggs und Enochs (1990).

TES	**STEBI**
Personal Teaching Efficacy (PTE): „When a student does better than usual, many times it is because I exerted a little extra effort.“	**Personal Science Teaching Efficacy Beliefs (PSTEB:)** „I am typically able to answer students’ science questions.“
Teaching Efficacy (TE:) „Even a teacher with good teaching abilities may not reach many students.“	**Science Teaching Outcome Expectancy (STOE:)** „Even teachers with good science teaching abilities cannot help some kids learn science.“

Für den „outcome-expectancy“-Faktor (STOE) können daher dieselben Argumente

[1] Weitere Items beider Skalen werden in Tabelle E.1 des Anhangs auf Seite 320 verglichen.

angeführt werden, die bereits bezüglich der inhaltlichen Adäquatheit des „teaching-efficacy“-Faktors (TE) der „Teacher Efficacy Scale“ diskutiert wurden (vgl. Unterabschnitt 4.1.3). Demzufolge handelt es sich aufgrund der fehlenden Selbstreferentialität der Aussagen nicht um Handlungsergebniserwartungen im Sinne Banduras (vgl. Abschnitt 3.4.7). Entsprechend ihrer Kritik konzeptualisieren Riggs und Enochs (1990) die PSTEB-Komponente eher orientiert an „agent-means“-Relationen (vgl. Tabelle 5.2). Dennoch werden eher keine Selbstwirksamkeitserwartungen im Sinne Banduras abgebildet. Dies liegt daran, dass teils eher auf Wissen anstatt auf Handlungen fokussiert wird („I know the steps necessary to teach science concepts effectively.“, „I understand science concepts well enough to be effective in teaching elementary science.“). Wenn Handlungen im Fokus stehen, dann sind diese jedoch sehr global formuliert („I generally teach science ineffectively.“, „I wonder if I have the necessary skills to teach science.“), sodass eher Facetten des Selbstkonzepts erfragt werden. Weiter bleibt festzustellen, dass keine Handlungsbarrieren in den Items enthalten sind („I am typically able to answer students' science questions.“). Einige Items erfragen gar keine spezifischen Einschätzungen eigener Fähigkeiten, sondern vielmehr Vorlieben oder Gewohnheiten zur Unterrichtspraxis („When teaching science, I usually welcome student questions.“, „Given a choice, I would not invite the principal to evaluate my science teaching.“). Aus den genannten Defiziten resultiert, dass die Items unterschiedliche Spezifitätsniveaus adressieren (globaler: „I generally teach science ineffectively.“, spezifischer: „I find it difficult to explain to students why science experiments work.“).

Das Instrument wurde mit Hilfe mehrerer exploratorischer Faktorenanalysen untersucht und enthält in seiner finalen Version 25 Items. Die beiden Faktoren korrelieren schwach ($r = .19^{**}$).

Alle genannten Kritikpunkte gelten auch für die Version des Instruments, welches für Lehramtsstudierende entwickelt wurde (STEBI-B), da die Itemformulierungen weitestgehend identisch sind (Enochs & Riggs, 1990). Die Versionen unterscheiden sich lediglich dadurch, dass zwei Items entfernt wurden und alle Items in Futur umformuliert wurden („When teaching science, I *will* usually welcome student questions.“). Damit widersprechen die Formulierungen der aus inhaltlichen Gründen motivierten Forderung Banduras, Items zur Erfassung von Selbstwirksamkeitserwartungen im Präsens zu verfassen (vgl. Abschnitt 3.6.2).

Ähnlich dem Prozedere nachfolgender Validierungsstudien der „Teacher Efficacy Scale“ werden die Instrumente STEBI-A/B in der Folge kaum inhaltlich diskutiert. Wenn sie überhaupt hinterfragt werden, dann wird jeweils ihre Faktorstruktur erneut überprüft (vgl. z. B. für STEBI-B: Bleicher, 2004; Mulholland, Dorman & Odgers, 2004), die sich größtenteils replizieren lässt. Beide Formen des STEBI etablieren sich über Jahrzehnte als Standardinstrumente zur Erfassung von „science teaching efficacy beliefs“.

5.2.2. Self-Efficacy Teaching and Knowledge Instrument for Science Teachers – SETAKIST

Ausgehend von einer Kritik an der als inakzeptabel gering eingestuften Varianzaufklärung existierender Instrumente sowie am STOE-Faktor des STEBI, der nach Roberts und Henson (2000) eher externe Einflüsse oder Attribuierungen auf den Erfolg oder Misserfolg von Schülerinnen und Schülern erheben würde, entwickeln die genannten Autoren ein eigenes Instrument (SETAKIST: Self-Efficacy Teaching and Knowledge Instrument for Science Teachers). Auch wenn der „outcome-excepctancy"-Faktor des STEBI ggf. eher keine Handlungsergebniserwartungen erfasst, so ist dessen grundsätzliche Operationalisierung jedoch durch Banduras sozial-kognitive Theorie gerechtfertigt und eine Überarbeitung naheliegend. Roberts und Henson (2000) ersetzen diesen jedoch durch einen neuen Faktor, den sie „knowledge efficacy" nennen und der an den Arbeiten Shulmans orientiert sein soll. Inwiefern er sich jedoch in die Theorie Banduras einfügt, wird von den Autoren nicht dargelegt. Der Faktor wird zum Teil aus Items der PSTEB-Komponente des STEBI gebildet, teils aus neu konstruierten Items. Acht weitere Items des ursprünglichen PSTEB-Faktors des STEBI bilden den „neuen" Faktor „teaching efficacy".

Insgesamt bleibt aus inhaltlicher Perspektive völlig unklar, wie sich die Faktoren „knowledge" und „teaching efficacy" voneinander unterscheiden (teaching efficacy: „I wish I had a better understanding of the science concepts I teach.", knowledge efficacy: „After I have taught a science concept once, I feel confident teaching it again."). Da sich beide Faktoren gänzlich oder zu Teilen aus dem Itempool des PSTEB-Faktors des STEBI zusammensetzen, gelten für diese ebenfalls die oben erläuterten Kritikpunkte (vgl. Unterabschnitt 5.2.1). Zusammenfassend kann gesagt werden, dass beide Faktoren eher keine Selbstwirksamkeitserwartungen im Sinne Banduras erfragen.

Erneut wird hier ein Instrument nicht aus einer theoretischen Perspektive heraus analysiert oder Überarbeitungen unterzogen. Pruski u. a. (2013) überprüfen für eine in einzelnen Formulierungen leicht geänderte Version (SETAKIST-Revised) die Faktorstruktur und das Ratingskalenformat und stellen zusammenfassend fest: „The SETAKIST-R, in its current form, cannot be recommended as a measure of science teacher self efficacy" (Pruski u. a., 2013, S. 1133).

5.2.3. Self-Efficacy Beliefs about Equitable Science Teaching and Learning – SEBEST

Aufgrund der sehr wenigen Instrumente zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern soll auch das Instrument von Ritter u. a. (2001) nicht unerwähnt bleiben, obwohl es nur auf ein potentielles Handlungsfeld (Diversität/Gleichberechtigung: „equitable science teaching and learning") abhebt und

damit dem Kriterium der Mehrdimensionalität nicht entspricht. Die Instrumententwickler orientieren sich explizit an den Vorarbeiten von Riggs und Enochs (1990). Entsprechend werden Items bezüglich der Dimensionen „personal self-efficacy“ und „outcome expectancy“ jeweils in den Themenfeldern Ethnie, Geschlecht, Sprachminderheiten und Sozioökonomie konstruiert.

Sowohl die Items zur Erfassung von Selbstwirksamkeitserwartungen als auch die Items zur Erfassung von Handlungsergebniserwartungen entsprechen nicht den von Bandura formulierten bzw. sich aus der sozial-kognitiven Theorie ergebenen Kriterien an die Konstrukte. Die Items der „self-efficacy“-Komponente sind teils im Futur formuliert („I will ...“) und adressieren eher globale als konkrete Handlungen („effectively teach science“, „can do a great deal“), sodass die Items eher als Selbstkonzept-Items aufgefasst werden müssen. Sie fragen grundsätzlich die Überzeugung der Befragten ab, über die Fähigkeit zu verfügen, bestimmte Minderheiten beim Lernen zu unterstützen oder für diese Personengruppen guten Unterricht anbieten zu können. Einige Items adressieren darüber hinaus je nach Bewertung des Fokus „means-ends“- bzw. „agent-ends“-Relationen („I can do a great deal as a teacher to increase the science achievement of children...“) oder zielen auf eine Einschätzung des eigenen Wissens anstatt auf eine Einschätzung eigener Fähigkeiten bezogen auf unterrichtliche Handlungen ab („I do not know teaching strategies to help children...“). Die Items, die Handlungsergebniserwartungen abbilden sollen, rekurrieren jeweils auf Einschätzungen darüber, ob potentiell benachteiligte Personengruppen durch guten naturwissenschaftlichen Unterricht gute (Lern)Leistungen erzielen können. Die Items sind daher nicht selbstreferentiell („Girls have the ability to compete academically with boys in science when they receive quality science instruction.“).

5.2.4. Selbstwirksamkeitserwartung bezogen auf das Physik-Lehren

Riese (2009, Anhang D, S. xviii) nutzt in seiner Studie ein Instrument, welches sechs übersetzte Items der STOE-Facette des STEBI nach Bleicher (2004) beinhaltet. Auch für diese deutschsprachige Version gelten dieselben, bereits oben ausgeführten Kritikpunkte (vgl. Unterabschnitt 5.2.1).

5.2.5. Zusammenfassung

„Rather than attempting to revamp a scale, it might be better to go back to the „master" and beginn again."

Pruski, 2013, S. 1151

Insgesamt muss festgestellt werden, dass lediglich wenige Instrumente existieren und diese darüber hinaus *nicht* zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen bezüglich physik- oder naturwissenschaftsdidaktischer Handlungsfelder entsprechend der sozial-kognitiven Theorie Banduras geeignet sind. Vielfach bauen die Instrumente aufeinander auf oder orientieren sich aneinander, ohne inhaltliche Problemstellen ausreichend zu diskutieren. Für den deutschen Sprachraum konnte lediglich ein Instrument gefunden werden, welches jedoch keine Neuentwicklung, sondern eine verkürzte und übersetzte Version des STEBI-B darstellt und deshalb in keinem Fall empfohlen werden kann. Die zusammenfassenden Analyseergebnisse können Tabelle 5.1 entnommen werden.

5.3. Abschließende Bemerkungen

„By trying to improve or revamp each others' scales, we, the research community, may have created a type of 'in-breeding' that clouds better thinking about efficacy item construction."

Pruski, 2013, S. 1151

Im Rahmen der angesporchenen Literaturrecherche wurden insgesamt 55 englischsprachige Studien im Themenfeld der „science teacher efficacy" mit einem Veröffentlichungsdatum zwischen 2005 und 2015 gesichtet. Allein 39 dieser Studien nutzen die Instrumente STEBI-A oder STEBI-B zur Erhebung von „science teaching efficacy beliefs". Lediglich drei Studien konnten gefunden werden, die das Instrument von Ritter u. a. (2001) heranziehen (Settlage, Southerland, Smith & Ceglie, 2009; Cone, 2009a, 2009b). Nur wenige Studien nutzen eher unbekannte Instrumente (Powell-Moman & Brown-Schild, 2011; Smolleck & Mongan, 2011; Buss, 2010; Avery & Meyer, 2012; Walan & Chang Rundgren, 2014), die teils auf Lehrer-Selbstwirksamkeitserwartungen bezüglich „inquiry teaching" abheben (vgl. u. a. Smolleck & Mongan, 2011; Avery & Meyer, 2012). Darüber hinaus gibt es einige qualitativ angelegte Studien (vgl. u. a. Bencze & Upton, 2006; Brand & Wilkins, 2007; Carrier, 2009; Gunning & Moore Mensah, 2011; Southerland, Sowell & Enderle, 2011; Kazempour, 2013; Moseley, Bilica, Wandless & Gdovin, 2014). Außerdem werden entsprechend der Zielgruppen des STEBI-A/B überwiegend Studierende des Primarstufenlehramts (vgl. u. a. Brígido, Borrachero,

Bermejo & Mellado, 2013; Cartwright & Atwood, 2014; Velthuis, Fisser & Pieters, 2014; Flores, 2015) oder Grundschullehrkräfte (vgl. u. a. Liu, Jack & Chiu, 2007; Palmer, 2011; Lumpe, Czerniak, Haney & Beltyukova, 2012; McKinnon, Moussa-Inaty & Barza, 2014) befragt.

Insgesamt bestätigt die Literaturrecherche den Eindruck, dass STEBI-A/B die Standardinstrumente zur Erfassung von naturwissenschaftsspezifischen Lehrer-Selbstwirksamkeitserwartungen sind. Da diese Instrumente stark defizitbehaftet sind, werden an dieser Stelle konsequenterweise keine Forschungsbefunde zu Zusammenhängen mit anderen Konstrukten sowie zur Entwicklung des Konstruktes berichtet. Es verdichtet sich damit das Desiderat eines neuen Instrumentes zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern, insbesondere für den deutschsprachigen Raum sowie für die Kohorte der Studierenden und Lehrkräfte eines Lehramts für die Sekundarstufe: „[W]e echo the need for a new or revised measure(s) that would reliably assess science teaching efficacy and its components“ (Cakiroglu u. a., 2012, S. 458).

6. Von der „guten“ Lehrperson – Selbstwirksamkeitserwartungen im Forschungsfeld der Lehrerprofessionalisierung

„People's level of motivation, affective states, and actions are based more on what they believe than on what is objectively true.“

Bandura, 1997, S. 2

Das Zitat erinnert an die in Kapitel 3 ausführlich dargelegte Bedeutung des Konstruktes der Selbstwirksamkeitserwartungen im Rahmen der sozial-kognitiven Theorie Banduras, in der Selbstwirksamkeitserwartungen neben anderen, zusätzlichen Faktoren (Fähigkeiten, Handlungsergebniserwartungen, Anreize, etc.) eine entscheidende Rolle bezüglich der Vorhersage von Handlungen, Zielsetzungen usw. einnehmen. Eine Vielzahl von Studienergebnissen bezüglich unterschiedlicher Domänen stützt dabei die postulierten Annahmen (vgl. Bandura, 1997). Die grundsätzlich aus der Theorie ableitbare Bedeutsamkeit des Konstruktes wird auch für das für den schulischen Kontext adaptierte Konstrukt der Lehrer-Selbstwirksamkeitserwartungen angenommen, auch wenn festgestellt werden muss, dass diesbezüglich weniger empirische Evidenz vorliegt als zunächst anzunehmen wäre (vgl. Abschnitt 4.4), was zu großen Teilen auch an den verwendeten, weniger geeigneten Messinstrumenten liegt (vgl. Abschnitt 4.3). In den vorangegangenen Kapiteln 4 und 5 zeigte sich ganz deutlich die Notwendigkeit und Relevanz der Entwicklung eines neuen Testinstruments zur Erforschung des Konstruktes der (Physik)Lehrer-Selbstwirksamkeitserwartungen. Im nachfolgenden Kapitel wird der Frage nachgegangen, wie sich die Erforschung von Selbstwirksamkeitserwartungen von (angehenden) Physiklehrpersonen in der Forschungslandschaft (der Physikdidaktik) verorten lässt.

Allgemein gilt, dass Lehrer-Selbstwirksamkeitserwartungen unabhängig von bestimmten Forschungstraditionen oder -paradigmen erforscht werden können. Eine leitende Forschungsfrage, die in vielfältigen Forschungsrichtungen bearbeitet wurde und wird und zu deren Beantwortung Selbstwirksamkeitserwartungen einen Beitrag leisten können, ist die Frage danach, wann eine Lehrkraft als eine *sehr gute* Lehrperson bezeichnet werden kann. Beispielsweise können Lehrer-Selbstwirksamkeitserwartungen im Rahmen des Persönlichkeitsansatzes als eine von vielen Persönlichkeitsvariablen aufgefasst werden, die unter Umständen gute von weniger guten oder gar schlechten Lehrpersonen zu

trennen vermag und potentiell das Lernen und Leisten von Schülerinnen und Schülern positiv beeinflusst (vgl. z. B. Mayr, 2011).

Ein weiterer Anknüpfungspunkt ergibt sich bezüglich einer aktuell vermutlich vorherrschenden Forschungsrichtung, der Kompetenzorientierung im Zuge der beabsichtigten Professionalisierung der Lehrerbildung.[1] Dies liegt einerseits darin begründet, dass Selbstwirksamkeitserwartungen als spezieller, subjektiver Kompetenzindikator angesehen werden können (vgl. Abschnitt 3.2.2). Andererseits können Selbstwirksamkeitserwartungen von Lehrpersonen im Rahmen der Kompetenzforschung sowohl als konstituierendes als auch determinierendes Merkmal von Lehrerkompetenzen angesehen werden. Trotz dieses fundamentalen Zusammenhangs spielen Selbstwirksamkeitserwartungen bei der Erhebung von Kompetenzen derzeit eine untergeordnete Rolle oder werden vernachlässigt, da der Fokus häufig auf der Erfassung von Wissen oder anderer Kompetenzfacetten liegt. Da Kompetenzen häufig domänenspezifischer, das heißt im Lehrkontext fachspezifischer Natur sind, wird im Rahmen dieses Forschungsansatzes der Bezug zur Erfassung physikdidaktischer Selbstwirksamkeitserwartungen besonders einsichtig.

Nachfolgend wird zunächst der Kompetenzbegriff näher beleuchtet, um potenzielle Schnitt- und Berührungspunkte zum Konstrukt der Selbstwirksamkeitserwartung sowie überlappende Bereiche deutlicher hervortreten zu lassen. Es werden zwei konkurrierende Interpretationsmöglichkeiten der Definition des Kompetenzkonstruktes beschrieben (determinierender vs. konstituierender Ansatz), die dahingehend maßgebliche Bedeutung erlangen, dass sich im Zuge von Forschungsanstrengungen je nach definitorischer Auslegung Operationalisierungen und Erfassungsmethoden wesentlich voneinander unterscheiden und auch Selbstwirksamkeitserwartungen konzeptionell unterschiedliche Rollen spielen. Exemplarisch werden anschließend zwei unterschiedlich abstrakte Ansätze zur Modellierung von Lehrerkompetenz dargestellt, wobei sich die jeweils unterschiedliche definitorische Basis dahingehend auswirkt, dass affektiv-motivationalen Dispositionen des Kompetenzkonstruktes – zu denen auch Selbstwirksamkeitserwartungen zählen – (implizit) unterschiedliche Bedeutungen beigemessen werden. Auch die Erfassung und Entwicklung von Kompetenzen werden nachfolgend thematisiert.

Im zweiten Teil des Kapitels wird die Kompetenzmodellierung von (angehenden) Physiklehrpersonen betrachtet. Es stellt sich heraus, dass ebenfalls häufig Wissen erfragt wird bzw. Wissen und Kompetenz in vielen Projekten als beinahe identisch konzeptionalisiert werden. Bezüglich der Anschlussfähigkeit des zu entwickelnden Instruments zur Erfassung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern können insbesondere aus den Konzeptionen des physikdidaktischen Wissens Bezugspunkte hinsichtlich der Operationalisierung des Konstruktes sowohl auf inhaltlicher

[1] Dabei handelt es sich durchaus um eine politisch gewollte Fokussierung, was unter anderem an den Forschungsfinanzierungen entsprechender Großprojekte deutlich wird (vgl. z. B. https://www.bmbf.de/de/qualitaetsoffensive-lehrerbildung-525.html, letzter Zugriff 21.11.2015).

Ebene als auch bezüglich der Wahl des Spezifitätsniveaus gewonnen werden.

6.1. Kompetenzorientierung in der Lehrerbildung

Als Reaktion auf die Ergebnisse deutscher Schülerinnen und Schüler in den PISA-Studien zur Jahrtausendwende kam es mit der Fokussierung auf Kompetenzen im Rahmen der Etablierung nationaler Bildungsstandards (vgl. u. a. Klieme u. a., 2007) zu einer weitreichenden Veränderung im Bildungswesen. Mit dem Ziel der Qualitätssicherung schulischer Bildung wurden Standards auch in der Lehrerbildung eingeführt (vgl. z. B. Bromme, 1997; Oser, 1997; KMK, 2004, 2010; Frey & Jung, 2011).[2] Nachfolgend soll der üblicherweise zugrunde gelegte Kompetenzbegriff kurz dargestellt und kritisch andiskutiert werden.

6.1.1. Kompetenz – Annäherung an eine Definition

Die meisten Arbeiten im Bereich der Lehrerprofessionalisierung beziehen sich mindestens implizit auf die Kompetenzdefinition von Weinert (2014, S. 27 f.):

> „[M]an [versteht] unter Kompetenzen, die bei Individuen verfügbaren oder durch sie erlernbaren kognitiven Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen, sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten um die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können.“

Kompetenz wird gemäß der Definition situativ, d. h. bezogen auf Handlungen in einem Problemlösekontext und mehrdimensional, als Zusammenspiel von Kognition, Volition sowie Motivation etc. beschrieben. Im tatsächlichen Handeln einer Person (deren Performanz) wird deutlich, ob diese die Situation angemessen bewältigen kann und damit ggf. über Kompetenz verfügt oder nicht. Klieme und Hartig (2007, S. 21) betonen in ihrer Definition des Kompetenzbegriffs neben der Mehrdimensionalität des

[2] Inwiefern es sich bei der Kompetenzorientierung grundsätzlich um ein tragfähiges Konzept handelt, kann im Rahmen dieser Arbeit nicht vollumfänglich diskutiert werden. Zumindest liegt das hehre Anliegen zugrunde, weniger träges Wissen (dafür praxisrelevante Kompetenzen) in der Ausbildung zu „produzieren“ (vgl. Klieme & Hartig, 2007, S. 13). Klassische Kritikpunkte an der Kompetenzorientierung zielen auf die Zweckorientierung oder eine verengte Bildungsperspektive ab (vgl. u. a. Frey & Jung, 2011, S. 561-567; Terhart, 2007a). In jedem Fall handelt es sich bei der Kompetenzorientierung mindestens bezüglich der Forschungsperspektive durchaus auch um einen pragmatischen Zugang. Dabei geht es um „den Anschluss an die internationale Diskussion über professionelle Standards im Lehrerberuf [...] und die Sicherung einer qualitätsvollen Lehrerausbildung“ (Baumert & Kunter, 2006, S. 469). Teils fehlt es „an empirischer Evidenz hinsichtlich der Bedeutung professioneller Kompetenzen für die Qualität von Unterricht sowie den Lernfortschritt und die Persönlichkeitsentwicklung von Schülerinnen und Schülern“ (Baumert & Kunter, 2006, S. 469). In der Regel wird diese Feststellung als Anlass für weitreichende Forschungsmaßnahmen genommen.

Konstruktes den dispositionalen Charakter: „Kompetenzen sind Dispositionen, die im Verlauf von Bildungs- und Erziehungsprozessen erworben (erlernt) werden und die Bewältigung von unterschiedlichen Aufgaben bzw. Lebenssituationen ermöglichen. Sie umfassen Wissen und kognitive Fähigkeiten, Komponenten der Selbstregulation und sozial-kommunikative Fähigkeiten wie auch motivationale Orientierungen."

Nimmt man die aufgeführten Definitionen ernst, so sind Wissensstrukturen keine Kompetenzen im eigentlichen Sinne – sie können lediglich als „kognitive" Facette einer Kompetenz interpretiert werden. Gerade diese Interpretation der Definition, sowohl kognitive (z. B. Wissen) als auch nicht kognitive Dispositionen (z. B. Volitionen, Motivationen) als kompetenzkonstituierend anzusehen, ist jedoch umstritten. Es resultieren Interpretationsansätze, die eine dominierende Stellung kognitiver Facetten präferieren und nicht kognitive Dispositionen eher als Einflussgrößen konzeptualisieren (determinierender Interpretationsansatz, vgl. Abbildung 6.1). Zlatkin-Troitschanskaia und Seidel (2011, S. 223, Hervorhebungen im Original) fassen den Kern der Diskussion treffend zusammen:

> „Während die kognitivpsychologisch orientierte (Bildung-)Forschung Kompetenz als auf spezifische Kontexte bezogene kognitive Leistungsdisposition definiert, die von affektiven Merkmalen *beeinflusst* wird, verstehen andere Forschungsvertreter Kompetenzen als (summative oder multiplikative) Verknüpfung von kognitiven, affektiven und volitionalen Elementen. Auffällig ist hierbei, dass die Einnahme einer der beiden Positionen häufig weder theoretisch noch anhand empirischer Untersuchungen begründet wird."

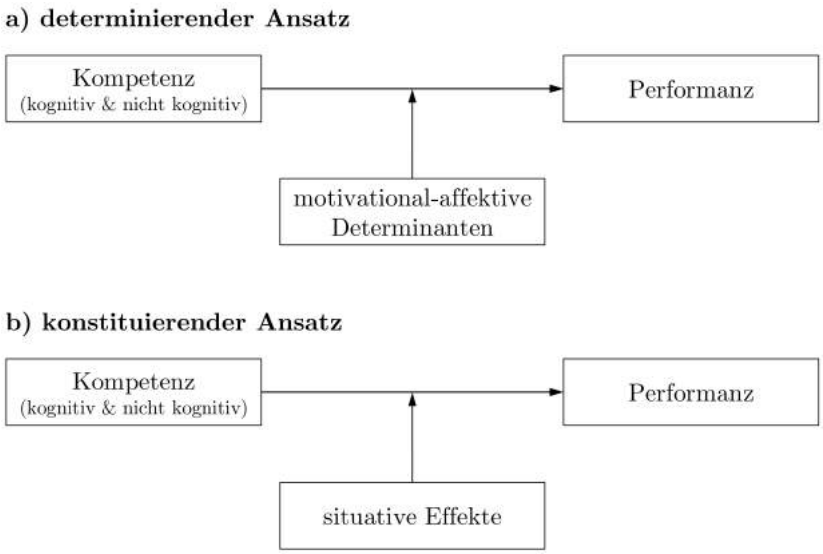


Abbildung 6.1.: Visualisierung unterschiedlicher Interpretationsansätze des Kompetenzbegriffs.

Tatsächlich geht aus den Definitionen nicht eindeutig hervor, in welchem Verhältnis kognitive und nicht kognitive Kompetenzkomponenten zueinander stehen. Es könnte daher argumentiert werden, dass nicht kognitive Bestandteile in Bezug auf die Kompetenz der Person eher eine untergeordnete Rolle spielen. Die Expertiseforschung fokussiert nach Krauss (2011, S. 181) beispielsweise (bewusst) auf kognitive Dimensionen wie Wissen oder Überzeugungen. Teils wird zudem argumentiert, dass nicht kognitive Facetten schwieriger zu erheben seien, weil sie zeitlichen Schwankungen unterlägen, subjektiv verzerrt und unter Umständen nicht explizierbar wären. Dem kann entgegen gehalten werden, das Teile dieser Argumentation auch auf kognitive Kompetenzfacetten zutreffen (vgl. u. a. Vogelsang & Reinhold, 2013) und die Bedeutsamkeit nicht kognitiver Komponenten damit nicht widerlegt wird. Zlatkin-Troitschanskaia und Seidel (2011, S. 229) argumentieren konsequent für den konstituierenden Interpretationsansatz und kommen zu folgendem Schluss:

> „Zwar kann nicht ausgeschlossen werden, dass personale (und situative) Faktoren [...] die Performanz beeinflussen; es ist sogar sehr wahrscheinlich. Die vorliegenden Befunde lassen jedoch vermuten, dass eben jene Faktoren bereits Kompetenz konstituierend wirken und eine multiplikative Verknüpfung aufweisen. Demnach wären ausgezeichnete kognitive Dispositionen nicht in der Lage, fehlende motivationale oder volitionale Elemente zu kompensieren. Damit eine Person kompetent handeln kann, muss also zumindest eine Minimalausprägung bei allen konstituierenden Elementen der Kompetenz vorliegen.“

Der konstituierende Interpretationsansatz der vorgestellten Kompetenzdefinitionen ist insofern mit Banduras sozial-kognitiver Theorie (vgl. Kapitel 3) kompatibel, als dass neben Wissen und Motivationen auch nicht kognitive Dispositionen wie Handlungsergebnis- oder Selbstwirksamkeitserwartungen relevante Faktoren in Bezug auf die Aufnahme von Handlungen und für die Bewältigung von Situationen darstellen. Gleichzeitig wird im Rahmen der sozial-kognitiven Theorie den nicht kognitiven Faktoren – besonders den Selbstwirksamkeitserwartungen – eine entscheidende Rolle zuerkannt, was einen Unterschied zu den genannten Kompetenzdefinitionen darstellt, die wie erwähnt, keine Gewichtung der Kompetenzfacetten vornehmen. In der Lesart des konstituierenden Ansatzes können Selbstwirksamkeitserwartungen als nicht kognitive Kompetenzdimension verstanden werden, womit der zentrale Anknüpfungspunkt des Konstruktes der Selbstwirksamkeitserwartung an die Kompetenzforschung beschrieben wäre.

6.1.2. Modellierung der Kompetenzen von (angehenden) Lehrpersonen

Wie bereits angedeutet, ist die Lehrerbildung – zumindest in Deutschland – kompetenzorientiert ausgerichtet. Nachfolgend werden exemplarisch zwei Kompetenzmodellierungsansätze detaillierter vorgestellt, die sowohl aufgrund ihrer Wirkung auf die

Ausbildungs- als auch auf die Forschungslandschaft relevant erscheinen. Zum einen handelt es sich um die Modellierung von Lehrerkompetenzen der Kultusministerkonferenz (KMK, 2004, 2010), die die rechtlich verbindliche und damit normative Grundlage der Lehrerbildung in Deutschland manifestiert. Außerdem wird der Modellierungsansatz des COACTIV-Forschungsprogramms (Baumert & Kunter, 2011) vorgestellt, auf dem aufgrund seiner breiten Resonanz eine Reihe von Forschungsanstrengungen beruhen.[3] Allein im Bereich der Physikdidaktik können an dieser Stelle nicht alle Arbeiten aufgezählt werden, die das Modell als theoretische Basis nutzen (vgl. Abschnitt 6.2.1).[4] Beide Ansätze sollen insbesondere daraufhin beleuchtet werden, ob eher der determinierende oder konstituierende Interpretationsansatz zum Tragen kommt.

Modellierungsansatz der Kultusministerkonferenz (KMK)

Aus den Schulgesetzen der Länder und den darin enthaltenen Bildungs- und Erziehungszielen leitet die KMK (2004, S. 3) in den *Standards für die Lehrerbildung/Bildungswissenschaften* zunächst Anforderungen ab, aus denen dann die vier Kompetenzbereiche „Unterrichten“, „Erziehen“, „Beurteilen“ und „Innovieren“ synthetisiert werden.

Jedem der vier Kompetenzbereiche werden in einem nächsten Schritt jeweils zwei bis drei globaler formulierte, den Bereich ausschärfende Kompetenzen zugeordnet – insgesamt werden elf Kompetenzen formuliert. Die folgende Kompetenz gehört beispielsweise zum Kompetenzbereich „Unterrichten“: „Lehrerinnen und Lehrer planen Unterricht sach- und fachgerecht und führen ihn sachlich und fachlich korrekt durch“ (KMK, 2004, S. 7). Jeder der elf benannten Kompetenzen werden wiederum Teilkompetenzen in Form von Standards zugewiesen; bezüglich der zitierten Kompetenz des Kompetenzbereichs „Unterrichten“ beispielsweise die Kenntnis von Bildungstheorien, unterschiedlicher Unterrichtsmethodiken oder Verfahren der Leistungsbewertung. Die Standards werden danach unterteilt, ob sie eher im Rahmen der theoretischen (Studium) oder praktischen Ausbildung (Vorbereitungsdienst) erworben werden.

Die angeführten Standards rekurrieren in der Regel auf den Erwerb oder die Anwendung von Wissen (vgl. auch Blömeke, Felbrich & Müller, 2008, S. 17). Einige wenige Standards – meist solche, die für die praktische Ausbildung angeführt werden (Reflektieren) – gehen ggf. über rein kognitive Aspekte hinaus. Es bleibt jedoch unklar, inwiefern auch motivational-affektive Dispositionen zu den Standards dazu gehören (können). Zudem wird an keiner Stelle des Beschlusses expliziert, welcher Kompetenzbegriff zugrunde gelegt wird, sondern lediglich angemerkt, dass es sich bei Kompetenzen um „Fähigkeiten, Fertigkeiten und Einstellungen“ (KMK, 2004, S. 4) handele. Daher

[3] Andere Ansätze können z. B. bei Terhart (2002), Oser (1997), Oser und Oelkers (2001) nachgelesen werden. Einen Überblick sowohl über deutschsprachige als auch englischsprachige Modellierungen liefern z. B. Frey und Jung (2011).

[4] Zlatkin-Troitschanskaia und Seidel (2011, S. 226) kritisieren in dem Zusammenhang die vielfach „wenig reflektierte und elaborierte Übernahme“ theoretischer Rahmenmodelle.

scheint eher davon ausgegangen zu werden, dass „Wissenskompetenzen“ die Grundlage für spätere schulnahe „Handlungskompetenz“ bilden. Zudem weist der gesamte Duktus eher auf eine determinierende Auslegung des Kompetenzbegriffs hin.

Eine Weiterentwicklung des KMK-Modellierungsansatzes kann dem KMK-Beschluss von 2008 entnommen werden. Dieser enthält eine Spezififkation der Anforderungen an (zukünftige) Lehrpersonen hinsichtlich der jeweiligen Fachwissenschaften und Fachdidaktiken, wobei ein ähnliches Muster zum oben analysierten KMK-Beschluss vorliegt: Grundlegende im Studium zu erwerbende Kompetenzen werden im Wesentlichen mit dem Erwerb von Wissen gleichgesetzt. Kompetenzen eher praktischer Natur (gestalten, planen, bewältigen, fördern, beherrschen, ...) sollen dagegen eher im Vorbereitungsdienst erworben werden (vgl. KMK, 2010, S. 2-4).[5]

Beispielhaft können die Einlassungen anhand des physikspezifischen *Kompetenz*profils einsichtig gemacht werden, welches – wie gesagt – ebenfalls eher als Wissensprofil verstanden werden kann. Physiklehramtsstudierende sollen am Ende ihres Studiums im Wesentlichen über physikalisches Fachwissen, Wissen über Erkenntnis- und Arbeitsmethoden, Wissen über die Natur der Physik und physikdidaktisches Wissen verfügen. Darüber hinaus sind Fertigkeiten (!), also keine Kompetenzen im Bereich des Experimentierens und in der Handhabung schultypischer Geräte sowie einige reflektierte Unterrichtserfahrungen anzustreben (vgl. KMK, 2010, S. 38). Das physikdidaktische Wissen sollte sich in der Regel auf die nachfolgenden Themenbereiche beziehen (vgl. KMK, 2010, S. 39):

- fachdidaktische Positionen und Konzeptionen,
- Motivation und Interesse,
- Lernprozesse, Diagnose von Lernschwierigkeiten,
- Planung und Analyse von Physikunterricht,
- Aufgaben, Experimente und Medien sowie
- fachdidaktische Forschung.

Modellierungsansatz von COACTIV

Baumert und Kunter (2006, S. 479) bewerten das Kompetenzmodell der KMK als „begrifflich und kategorial nur schwer an psychologische Handlungstheorien und empirische Forschung anschließbar.“ Aus diesem Grund entwickeln sie selbst ein heuristisches Modell professioneller Handlungskompetenz von Lehrpersonen (vgl. Abbildung 6.2), das sie selbst als nichthierarchisches, generisches Strukturmodell charakterisieren (vgl. Baumert & Kunter, 2006, S. 481). Dazu greifen sie unter anderem die Vorarbeiten Shulmans

[5] „Bei den Studieninhalten zur jeweiligen Didaktik eines Faches wird in der Regel nicht zwischen beiden Sekundarstufen unterschieden“ (KMK, 2010, S. 4).

(1986, 1987) auf und orientieren sich explizit am Kompetenzbegriff Weinerts. Nach ihrem Modell kann professionelle Kompetenz von (angehenden) Lehrkräften als Zusammenspiel der vier Komponenten Wissen, Überzeugungen und Werthaltungen, motivationalen Orientierungen und selbstregulativen Fähigkeiten verstanden werden (vgl. Baumert & Kunter, 2006, S. 479 ff.). Zentral ist ihre, wiederum an Shulman (1986, 1987) orientierte Topologie der Wissenskomponente (Aufteilung in Fachwissen, fachdidaktisches Wissen, pädagogisches Wissen, etc.). Generell ist es möglich, das Modell domänenspezifisch, d. h. für Lehrpersonen unterschiedlicher Fachbereiche auszudifferenzieren.

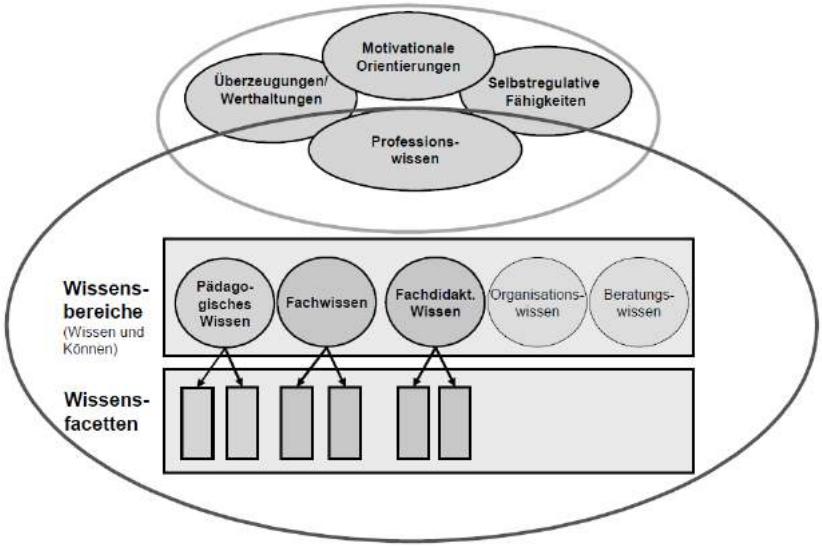


Abbildung 6.2.: Modell der professionellen Handlungskompetenz von Lehrpersonen (Baumert & Kunter, 2006, S. 482).

Die von Baumert und Kunter (2006) beschriebenen vier Facetten/Komponenten spiegeln sowohl kognitive als auch nicht kognitive Kompetenzaspekte in Anlehnung an Weinerts Kompetenzdefinition wider, sodass die Modellierung der Autoren mit dem konstituierenden Paradigma kompatibel ist. Unklar bleibt jedoch, wie genau das Zusammenspiel der benannten vier Facetten/Komponenten funktionieren muss, damit kompetentes (bzw. professionelles) Handeln beobachtet werden kann. Auch eine Betonung gerade dieses Zusammenspiels nehmen die Autoren nicht vor. Vielmehr nehmen die Beschreibungen der einzelnen Wissenskomponenten einen sehr großen Raum ein (rund 13 Seiten), während den verbleibenden drei Komponenten vergleichsweise wenig Beachtung geschenkt wird (zusammen rund neun Seiten), sodass der Eindruck einer impliziten Gewichtung der Facetten entsteht.[6] Darüber hinaus wird die Situationsspezifität des Konstruktes, die in den Kompetenzdefinitionen anklingt, im COACTIV-Modell

[6] Weitere, spezifischere Kritikpunkte am Modell oder an den theoretischen Überlegungen findet sich u. a. bei Zlatkin-Troitschanskaia und Seidel (2011, S. 226) oder bei Helsper (2007).

vernachlässigt. Da das Modell als *Rahmen*modell zur Beschreibung professioneller Handlungskompetenz von Lehrkräften konzipiert ist, muss es für das konkrete Handeln der Lehrpersonen weiter spezifiziert werden (vgl. Baumert & Kunter, 2006, S. 481).

Vergleichende Anmerkungen

Nachfolgend werden einige vergleichende Anmerkungen angeführt, die beispielsweise dann von Relevanz sind, wenn konkrete fachbezogene Kompetenzmodellierungen oder Studien zur Kompetenzerfassung eine Rolle spielen bzw. beurteilt werden müssen (vgl. z. B. Abschnitt 6.2.1).

Zunächst sollte betont werden, dass sich die beiden vorgestellten Modellierungen hinsichtlich ihres Abstraktionsgrades unterscheiden. Wie bereits erwähnt, muss die Konzeption von Baumert und Kunter (2006) für das Lehrerhandeln konkretisiert werden, während die Kompetenzen der KMK (2004, 2010) bereits ausformuliert sind. Baumert und Kunter (2006) nehmen selbst erste Konkretisierungen vor, indem sie beispielsweise darauf hinweisen, dass für die Profession des Lehrerberufs bestimmte motivationale Orientierungen wie zum Beispiel *Lehrer*-Selbstwirksamkeitserwartungen relevant seien oder Wissensdomänen wie die des *pädagogischen* Wissens. Im Prinzip ist es so, dass die Spezifizierung/Konkretisierung des Modells von Baumert und Kunter (2006) auf zwei verschiedene Arten erfolgen kann. Diese sollen beispielhaft bezüglich der professionellen Handlungskompetenz von Physiklehrpersonen verdeutlicht werden.

Die erste Möglichkeit besteht darin, die vier beschriebenen Kompetenzfacetten für Physiklehrpersonen zu spezifizieren. Welche Überzeugungen und Werthaltungen, welches physikalische Wissen, etc. benötigen also (angehende) Physiklehrkräfte? Das Kompetenzmodell würde folglich schrittweise mit relevanten selbstregulativen Fähigkeiten, Überzeugungen, Motivationen und Wissensbestandteilen „aufgefüllt“ werden. Im Idealfall könnte jede denkbare Problemsituation des Lehrerberufs dann mit Hilfe eines spezifischen Bündels – jeweils bestehend aus Elementen aller vier Facetten – „kompetent“ bewältigt werden.

Der zweite Weg sieht im Sinne eines Anforderungsprofils vor, relevante Teilkompetenzen zu identifizieren (z. B. „Erklärkompetenz“: die Kompetenz, Schülerinnen und Schülern etwas zu erklären) und diese dann bezogen auf das Rahmenmodell in ihre Komponenten zu zerlegen. Das würde konkret bedeuten, dass für jede einzelne Teilkompetenz (wie die „Erklärkompetenz“) relevante Wissensbestände, motivationale Orientierungen etc. präzisiert werden müssten. Durch die Identifikation aller relevanten Teilkompetenzen, die (angehende) Physiklehrkräfte erwerben müssen, entstünde eine Systematisierung, die in ihrer Gesamtheit die professionelle Handlungskompetenz von (angehenden) Physiklehrpersonen darstellen würde.

In dem zuletzt skizzierten Sinne präsentiert die KMK eine derartige Konkretisierung, indem eine Reihe von Kompetenzen bzw. Kompetenzbereichen aufgezählt werden.

Allerdings werden auf einem eher allgemeinen Niveau für jede formulierte Teilkompetenz lediglich relevante Wissensbestände als Standards identifiziert und die übrigen Kompetenzfacetten nicht berücksichtigt. Forschungsanstrengungen innerhalb der Physikdidaktik nutzen beide dargestellten Wege zur Modellierung und Erfassung von Kompetenzen (vgl. Unterabschnitt 6.2.1).

Abschließend sei bemerkt, dass bei Analysen von Kompetenzmodellen unterschiedlichen Abstraktionsgrades, Begrifflichkeiten schnell ihre Bedeutung wechseln können. Im Sinne der Kompetenzdefinition nach Weinert und dem Rahmenmodell Baumerts und Kunters handelt es sich bei Kompetenzfacetten oder -komponenten um die kognitiven (z. B. Wissen) oder nicht kognitiven Aspekte (Selbstwirksamkeitserwartungen, Volition, etc.), also Dimensionen der Kompetenz, während bezüglich anderer Kompetenzmodelle die Teilkompetenzen als Komponenten oder Facetten bezeichnet werden. Auf die uneinheitliche Verwendung von Begrifflichkeiten weisen auch Frey und Jung (2011) hin.

6.1.3. Kompetenzentwicklung von Lehrpersonen

Als sichere Annahme gilt, dass „die Handlungskompetenz von Lehrkräften grundsätzlich Veränderungs- und Entwicklungsprozessen unterliegt und in allen Phasen der beruflichen Laufbahn professioneller Entwicklung und Vervollkommnung offen steht“ (Baumert & Kunter, 2006, S. 507). Dies zeigen Studien beispielsweise zur Entwicklung einzelner Kompetenzdimensionen, die jedoch nicht immer im Rahmen des Kompetenzparadigmas durchgeführt worden sind.[7] Andere Untersuchungen fokussieren auf Kompetenzentwicklungen bezüglich einzelner Ausbildungsabschnitte, wie die Kompetenzentwicklung im Studium, im Vorbereitungsdienst oder in der dritten Phase der Lehrerbildung (vgl. z. B Baer u. a., 2007; Lipowsky, 2010). Teils sind die untersuchten Zeiträume sogar noch spezifischer und es werden Entwicklungen über eine Lehrveranstaltung oder ein Praktikum hinweg untersucht (vgl. z. B. Hascher, 2006; oder für einen Überblick zur Wirksamkeit von Praxisphasen z. B. Gröschner & Schmitt, 2010). Die einzige mir bekannte Studie, die Entwicklungsverläufe sowohl kognitiver als auch nicht kognitiver Kompetenzdimensionen in Anlehnung an Baumert und Kunter (2006) für die Ausbildungsphase des Studiums nachzeichnet, ist das Panel zum Lehramtsstudium (PaLea, vgl. Bauer u. a., 2010).

Die Kompetenzentwicklung wird wahrscheinlich begünstigt, wenn systematische Reflexionsgelegenheiten, Coaching und diskursives Feedback den Prozess begleiten. Zusätzlich wird davon ausgegangen, dass ein grundsätzliches Streben nach Selbstverbesserung und die Fähigkeit zur Selbstregulation wichtige Rollen einnehmen (vgl. Baumert & Kunter, 2006, S. 505 f.; Krauss, 2011, S. 182).

[7] Zur Entwicklung von Persönlichkeitsmerkmalen vgl. zusammenfassend beispielsweise Mayr (2011, S. 132 f., 140 ff.); zur Entwicklung von Wissensbeständen vgl. z. B. Schulte u. a. (2008), König und Seifert (2012); zur Entwicklung von Lehrer-Selbstwirksamkeitserwartungen, vgl. Abschnitt 4.6. Darüber hinaus wird auch im Rahmen des berufsbiografischen Forschungsparadigmas die Entwicklung von Lehrpersonen beschrieben (vgl. z. B. Terhart, Czerwenka, Ehrich, Jordan & Schmidt, 1994).

Bezüglich der allgemeinen, fachunspezifischen Kompetenzentwicklung von (angehenden) Lehrkräften gibt es relativ wenige Erkenntnisse, was zum Teil fehlenden Kompetenzentwicklungsmodellen geschuldet ist oder der Unklarheit über adäquate Erhebungsmethoden (vgl. nachfolgender Abschnitt). Einzelne Versuche der theoretischen Beschreibung der Kompetenzentwicklung finden sich beispielsweise bei Frey (2006) oder Lipowsky (2010). Darüber hinaus werden sogenannte Stufenmodelle diskutiert, die unterscheidbare Ausprägungsgrade der Kompetenz mit steigender Berufserfahrungen postulieren (Beispiel: Novizen, fortgeschrittene „Beginner“, kompetent Handelnde, fortgeschrittene kompetent Handende, Experten; vgl. Baumert und Kunter, 2006, S. 506; Krauss, 2011, S. 182). Bezüglich subjektiver Theorien sowie Werthaltungen und Überzeugungen wird die relative Stabilität der Facetten thematisiert, die eine rasche Veränderbarkeit oder Weiterentwicklung hinsichtlich angemessenerer oder erwünschterer Vorstellungen im Rahmen der Aus- und Weiterbildung unter Umständen erschweren (vgl. u. a. Baumert & Kunter, 2006, S. 506).

6.1.4. Messung von Lehrerkompetenzen

Es gibt eine rege Diskussion darüber, wie Kompetenzen sinnvoll erfasst werden können (vgl. u. a. Oser, Curcio & Düggeli, 2007; Kunter & Klusmann, 2010; von Aufschnaiter & Blömeke, 2010; Maag Merki & Werner, 2011). Dabei wird auch thematisiert, welche Teilkompetenzen überhaupt erhoben werden sollen und welche Gewichtung einzelnen Komponenten oder Facetten (wie z. B. Wissenselementen) zuerkannt werden soll. „Insbesondere zeichnen sich Forschungsdefizite ab in Bezug auf die Entwicklung von validen Erhebungsinstrumenten zur empirischen Erfassung der einzelnen Kompetenzdimensionen sowie in Bezug auf die Modellierung des Zusammenhangs zwischen dem Beurteilungsergebnis und möglichen Konsequenzen bzw. den Professionalisierungsmaßnahmen für die betroffenen Lehrpersonen“ (Maag Merki & Werner, 2011, S. 576). Auch Kunter und Klusmann (2010, S. 70) nennen neben weiteren Herausforderungen, die Problematik fehlender standardisierter bzw. „valider“ Instrumente.[8] Außerdem wird kritisiert, dass die häufig zugrunde gelegte Kompetenzdefinition nach Weinert in der Regel nur unzureichend in Operationalisierungen adressiert wird (motivational-affektive Komponenten fehlen, kognitive Facetten werden überbetont), dass Operationalisierungen häufig unkonkret und vage bleiben, was unter Umständen mit dem fehlenden Konsens über relevante Kompetenzen erklärt werden kann und dass sehr viele unterschiedliche Operationalisierungen vorliegen (vgl. von Aufschnaiter & Blömeke, 2010).

Maag Merki und Werner (2011, S. 576 f.) fordern daher die mehrdimensionale Erfassung von Kompetenzen, d. h. die Berücksichtigung von kognitiven und nicht kognitiven Kompetenzdimensionen sowie die theoretisch fundierte Ableitung geeigneter dimensionsspezifischer Kompetenzindikatoren. Weiterhin muss begründet dargelegt werden, wann

[8] Frey (2006) listet verfügbare Instrumentarien zur Messung verschiedener Kompetenzfacetten auf.

aus einer bestimmten Handlung auf eine vorliegende Kompetenz geschlossen werden darf (Performanz vs. Kompetenz). Diese Problematik zielt auf das Interaktionsparadigma, wonach

> „[...] individuelles Verhalten die Funktion eines kontinuierlichen Prozesses multidirektionaler Interaktionen oder Rückkopplungen zwischen der Person und der Situation ist, in der sie sich befindet. An diesem Interaktionsprozess ist das Individuum aktiv beteiligt und beeinflusst diesen. Situationen weisen zudem eine objektive und subjektive Seite auf, wobei die subjektive Interpretation der Situation ein wesentlicher, determinierender Faktor für das konkrete Handeln ist“ (Maag Merki & Werner, 2011, S. 577).

Für den Lehrerberuf gilt dabei in der Regel, dass Situationsspezifika eine große Auswirkung auf das gezeigte Verhalten der Lehrpersonen haben, dass also zum Beispiel „in unterschiedlichen Situationen unterschiedliche Verhaltensweisen für das gleiche Ziel“ (Maag Merki & Werner, 2011, S. 577) genutzt werden. Inwiefern dann jedoch von Performanz auf Kompetenz geschlossen werden kann, bleibt fraglich. Unter Umständen zeigt sich gerade in der angesprochenen situativen Flexibilität des Verhaltens die eigentliche Kompetenz der Lehrperson. „Die Herausforderung von Beurteilungsverfahren besteht darin, beide Seiten, die personale wie auch die situationale, bei der Erfassung und Bewertung von Kompetenzen unter Beachtung ihrer Wechselbeziehung heranzuziehen“ (Maag Merki & Werner, 2011, S. 577).

Klassische Verfahren der Kompetenzmessung berücksichtigen selten die Situativität der Erhebungssituation. Zu den häufig genutzten Verfahren zählen unter anderem Tests und Fragebögen, aber auch Beobachtungen oder Interviews, wobei jede Herangehensweise sowohl auf Selbst- als auch auf Fremdeinschätzungen beruhen kann (vgl. Kunter & Klusmann, 2010, S. 74 ff.; Maag Merki & Werner, 2011, S. 579-584). Selbsteinschätzungen der eigenen Kompetenz werden teils als subjektiver Ansatz der Kompetenzmessung verstanden und Selbstwirksamkeitserwartungen gelten im Rahmen dieser Auffassung als spezielle subjektive Kompetenzindikatoren. Selbstauskünfte eignen sich für die Kompetenzmessung allerdings nur bedingt, denn diese sind in der Regel zwar gute Verhaltensprädiktoren, unterliegen aufgrund ihrer konzeptionellen Andersartigkeit unter Umständen aber größeren Fehleinschätzungen (vgl. dazu Kapitel 3).[9]

[9] „Perceived self-efficacy is not a measure of skills one has [...]. Perceived self-efficacy is a contributor of performance accomplishments, whatever the underlying skills might be“ (Bandura, 1997, S. 37). Im Rahmen der COACTIV-R-Studie konnten beispielsweise nur geringe bis keine Zusammenhänge zwischen Selbsteinschätzung und „direkt“ erhobenen Kompetenzen gefunden werden (vgl. Kunter & Klusmann, 2010, S. 76 ff.), wobei zwecks fehlender Einsichtsmöglichkeiten die Prämisse gelten muss, dass die verwendeten Instrumente tatsächlich Wissen und Selbsteinschätzungen bezogen auf dieses Wissen messen (Stichprobe: Referendare). In anderen Studien werden subjektive Kompetenzeinschätzungen als verlässliche Kompetenzindikatoren gewertet, da vergleichende Kompetenzfremdeinschätzungen in der Regel noch positiver ausfielen und daher als unrealistischer eingestuft wurden (vgl. Hascher, 2006; Baer u. a., 2007).

In der Praxis werden Tests insbesondere zur Erhebung von Kompetenzfacetten wie Wissen eingesetzt, während Fragebögen eher zur Erhebung nicht kognitiver Kompetenzkomponenten genutzt werden. Beobachtungen zum Beispiel durch Videoanalysen werden vielfach zur Erhebung von Teilkompetenzen herangezogen. Um ein differenziertes Bild zu erhalten, sollte vermehrt in Betracht gezogen werden, verschiedene methodische Ansätze zu einem komplexen Bewertungsinstrumentarium oder -system zusammenzufassen.

6.2. Kompetenzmodellierungen von (angehenden) Physiklehrkräften

> „Effective personal functioning is not simply a matter of knowing what to do and being motivated to do it. [...] Effective functioning requires both skills and the efficacy beliefs to use them.“
>
> *Bandura, 1997, S. 36*

Wenn Frey (2006, S. 42) bei seiner Recherche von Instrumentarien zur Erfassung der Kompetenz von Lehrpersonen noch bemängelt, dass nur wenige zur Erfassung fachbezogener Kompetenz vorlägen, so kann für den Bereich der Physikdidaktik mittlerweile ein anderes Bild gezeichnet werden. In den letzten Jahren kann eine Vielzahl von Forschungsbemühungen verzeichnet werden, die sich mit der Konzeption und Erfassung professioneller (fachlicher) Kompetenz von (angehenden) Lehrpersonen der Physik oder Naturwissenschaften beschäftigen. Die Projekte werden im Folgenden kurz dargestellt. Offensichtlich wird dabei der bereits in Abschnitt 6.1.4 konstatierte Fokus, dass nämlich die meisten Projekte auf kognitive Kompetenzdimensionen wie Wissen oder Überzeugungen abheben. Aus den Operationalisierungen physikdidaktischen Wissens können jedoch wichtige Hinweise hinsichtlich der Operationalisierung des in dieser Arbeit im Vordergrund stehenden Instrumentes zur Erfassung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern gewonnen werden. Eine grundsätzliche Kompatibilität des neu zu entwickelnden Instrumentes zu bereits existierenden Erhebungsinstrumenten des physikdidaktischen Wissens wäre auch aus der Forschungsperspektive wünschenswert, wenn zukünftig im Sinne des konstituierenden Ansatzes der Kompetenzauffassung Teilkompetenzen erhoben werden sollen. Auch aus diesem Grund werden die existierenden deutschsprachigen Instrumente zur Erfassung physikdidaktischen Wissens näher betrachtet.

6.2.1. Projekte mit dem Schwerpunkt der Kompetenzmodellierung von (angehenden) Physiklehrkräften

Im Bereich der domänenspezifischen Expertise- und Kompetenzforschung kann die Mathematikdidaktik aufgrund der Vielzahl von großen Forschungsprojekten zur Modellierung, Erfassung und Entwicklung professioneller Kompetenzen von Lehrkräften als wegbereitend bezeichnet werden. Zu diesen Forschungsaktivitäten zählen die Studien COACTIV (vgl. Krauss u. a., 2008; Kunter u. a., 2011), MT21 (vgl. Blömeke, Kaiser & Lehmann, 2008) oder TEDS-M (vgl. Blömeke, Kaiser & Lehmann, 2010a, 2010b). In der physikdidaktischen Kompetenzforschung zum Lehrerberuf sind in den letzten Jahren eine Reihe von größeren Forschungsarbeiten und Projekten zu verzeichnen. Diese können überblicksartig Tabelle 6.1 entnommen werden, wobei Vollständigkeit hier nicht gewährleistet werden kann und soll. Zu berücksichtigen ist weiterhin, dass eine Reihe weiterer Studien und Projekte hier thematisch einzuordnen wäre, die auf die Entwicklung, Validierung oder den Einsatz der Erhebungsinstrumente fokussieren (vgl. z. B. Vogelsang, 2014; Stender, 2014; Cauet, Liepertz, Kirschner, Borowski & Fischer, 2014).

Tabelle 6.1.: Systematisierung der Projekte nach befragten Kohorten, erfassten Kompetenzdimensionen und theoretischer Basis.

untersuchte Kohorte			**Kompetenzdimension**	**Kompetenzdimension tB**[1]	
AP[2]	Schulform[3]	Fach[4]	im Fokus	zusätzlich	
Paderborner Instrumentarium zur Messung der professionellen Kompetenz angehender Physiklehrkräfte (vgl. Riese, 2009)					
S	H, R, Gy	P	physikdidaktisches Wissen physikalisches Wissen pädagogisches Wissen Überzeugungen motivationale Orientierungen		BK
PLUS – Professionswissen von Lehrkräften, naturwissenschaftlicher Unterricht und Zielerreichung seitens der Schüler im Übergang von der Primar- zur Sekundarstufe (vgl. u. a. Ohle, Fischer & Kauertz, 2011; Lange, Kleickmann, Tröbst & Möller, 2012)					
L	P	SU	physikdidaktisches Wissen physikalisches Wissen	(SWE)	S

Fortsetzung auf der nächsten Seite...

Tabelle 6.1.: *Fortsetzung*

untersuchte Kohorte			Kompetenzdimension	Kompetenzdimension tB	
AP	Schulform	Fach	im Fokus	zusätzlich	
QuIP – Quality of Instruction in Physics: Comparing Finnland, Germany and Switzerland (vgl. u. a. Fischer, Labudde, Neumann & Viiri, 2014)					
L	G, R, Gy	P	physikdidaktisches Wissen Klassenmanagement Enthusiasmus konstrukt. Überzeugungen	physikalisches Wissen Interesse Motivation	S
NUK – Naturwissenschaften unterrichten können (vgl. u. a. Brovelli, Bölsterli, Rehm & Wilhelm, 2013; Brovelli, 2014)					
S	?	N, P, C, B	fachdidaktisches Wissen	fachliches Wissen (P/B/C) pädagogisches Wissen Selbstregulation	S
ProwiN – Professionswissen in den Naturwissenschaften (vgl. u. a. Borowski, Neuhaus u. a., 2010)					
L	H, R, G, Gy	B, C, P	physikdidaktisches Wissen physikalisches Wissen pädagogisches Wissen		BK/S
Ko-WADiS – Kompetenzmodellierung und -erfassung zum Wissenschaftsverständnis über naturwissenschaftliche Arbeits- und Denkweisen bei Studierenden (Lehramt) in den drei naturwissenschaftlichen Fächern Biologie, Chemie und Physik (vgl. u. a. Straube & Nordmeier, 2012; Stiller u. a., 2014)					
S	?	B, C, P	Erkenntnisgewinnungs-kompetenz [Wissen]		BK/S
ProfiLe-P – Professionswissen in der Lehramtsausbildung Physik (vgl. u. a. Kulgemeyer u. a., 2012)					
S	H, R, Gy	P	physikdidaktisches Wissen physikalisches Wissen Erklärenskompetenz		S
KiL – Messung professioneller Kompetenzen in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen (vgl. u. a. Kröger, Neumann & Petersen, 2013)					
S	Gy*	P	physikdidaktisches Wissen physikalisches Wissen pädagogisches Wissen		BK/S

Fortsetzung auf der nächsten Seite...

Tabelle 6.1.: *Fortsetzung*

untersuchte Kohorte			Kompetenzdimension	Kompetenzdimension tB	
AP	Schulform	Fach	im Fokus	zusätzlich	
proΦ – Professionelle Kompetenz von angehenden Physiklehrkräften (vgl. u. a. Oettinghaus, Lamprecht & Korneck, 2014; Oettinghaus, 2015)					
R	H, R, Gy	P	physikdidaktisches Wissen physikalisches Wissen Überzeugungen	Persönlichkeitsmerkmale [Determinanten] Selbstregulation	BK
Φactio – Professionelles Handeln von angehenden Physiklehrkräften (vgl. u. a. Korneck, Kohlenberger, Oettinghaus, Kunter & Lamprecht, 2013, 2014; Krüger, Korneck, Oettinghaus & Kunter, 2015)					
S, R	H, R, G	P	physikdidaktisches Wissen physikalisches Wissen Überzeugungen Selbstwirksamkeitserwartungen		BK
FALKO – Fachspezifische Lehrerkompetenzen (vgl. u. a. Göhring & Schödl, 2015)					
L	H, R, G	P	physikdidaktisches Wissen physikalisches Wissen		BK/S
EKoL – Effektive Kompetenzdiagnose in der Lehrerbildung (vgl. www.ekol-fun.de, letzter Zugriff: 11.01.2016)					
S, R	?	B, C, P	Experimentier- & Modellkompetenz [fachdid. Wissen]		?
KEMΦ – Kompetenzentwicklung Physik in der Studieneingangsphase (aufbauend auf Woitkowski, 2015, vgl. https://physik.uni-paderborn.de/reinhold/forschung, letzter Zugriff: 14.01.2016)					
S	?	P	physikalisches Wissen	affektive Merkmale [Determinanten]	?

[1] theoretischer Bezug: BK - Baumert und Kunter (2006), S - Shulman (1986, 1987)
[2] Ausbildungsphase: S - Studierende, R - Referendare, L - Lehrkräfte
[3] Schulform: G - Gesamtschule, Gy - Gymnasium, H - Hauptschule, P - Primarstufe, R - Realschule, * - hauptsächlich
[4] studiertes/unterrichtetes Fach: B - Biologie, C - Chemie, N - Naturwissenschaften, P - Physik, SU - Sachunterricht
weitere Abkürzungen: ? - Angabe unklar, [...] - Anmerkungen durch C. M.
Anmerkungen: Auch wenn sich einige Projekte auf mehrere Domänen beziehen, so werden hier jeweils, wenn möglich, die Kompetenzdimensionen berichtet, die sich auf die Physik beziehen.

Die Schwerpunktsetzung hinsichtlich des professionellen Wissens kann bereits aus einigen Projekttiteln abgelesen werden (PLUS, ProwiN, ProfiLe-P). Andere Titel verweisen dagegen explizit auf die Messung professioneller Kompetenz. Bei genauerer Betrachtung dieser Projekte wird jedoch schnell klar, dass ebenfalls häufig die Kompetenzdimensionen einzelner Wissensfacetten im Fokus stehen (Ko-WADiS, NUK, KiL, proΦ, FALKO). Der Paderborner Ansatz, das QUiP-Projekt sowie Φactio erheben zwar

nicht vollumfänglich, dafür aber zumindest mehrdimensional Kompetenz angehender oder erfahrener Physiklehrkräfte. Die am wenigsten untersuchte Kohorte ist dabei die der Referendare. Anzumerken ist, dass fast alle Arbeiten auf die Vorarbeiten von Shulman (1986, 1987) und/oder Baumert und Kunter (2006) rekurrieren. Die häufigste Form der Erfassung der einzelnen Kompetenzdimensionen ist der schriftliche Test für die kognitiven Komponenten bzw. der Fragebogen für die Erfassung von Werthaltungen oder Überzeugungen. Vereinzelt werden auch Videobeobachtungen (z. B. QUiP, Φactio) oder Vignettentests (vor allem NUK) eingesetzt.

Darüber hinaus widmen sich einige Forschungsanliegen der Erfassung von Kompetenz explizit ausschnitthaft (vgl. Unterabschnitt 6.1.2/Vergleichende Anmerkungen/zweite Möglichkeit), und zwar derart, dass eben nicht der Anspruch erhoben wird, die *gesamte* professionelle Handlungskompetenz einer (angehenden) Physiklehrperson operationalisieren und messen zu wollen. So befassen sich z. B. Stender (2014) sowie Klös, Riese und Reinhold (2015) mit der Unterrichtsplanung, während Cappell und von Aufschnaiter (2011) die Diagnosekompetenz (bez. Schülerkompetenzen) und Rath und Reinhold (2014) sowie Draude und Wodzinski (2015) die Diagnosekompetenz (bez. Schülervorstellungen/Lernschwierigkeiten) untersuchen. Im Rahmen des Projektes ProfiLe-P messen Kulgemeyer und Tomczyszyn (2015) Erklärenskompetenz Physiklehramtsstudierender. Auch die reflektiven (vgl. z. B. Klös u. a., 2015) und argumentativen Fähigkeiten (vgl. Gromadecki-Thiele & Priemer, 2015) von Physiklehramtsstudierenden werden untersucht.

Zudem wird im Rahmen der zuletzt genannten Projekte und im Unterschied zu den in Tabelle 6.1 vorgestellten Projekten häufig nicht aus der erhobenen Performanz (z. B. aus einem Test oder einer Beobachtung) auf die Ausprägung einer Kompetenzdimension wie „Wissen“ oder „Überzeugung“ geschlossen, sondern direkt die Ausprägung der jeweiligen Teilkompetenz anhand bestimmter Indikatoren bewertet. Über die Ausprägung unterschiedlicher Dimensionen dieser Kompetenz, beispielsweise das Wissen über Diagnose, Erklären, Planen, etc. oder zugehörige Selbstwirksamkeitserwartungen wird keine Aussage gemacht.

Zusammenfassend kann konstatiert werden, dass sowohl im Rahmen der tabellarisch dargestellten Projekte, als auch im Zuge der Teilkompetenzen erfassenden Studien kaum affektiv-motivationale Kompetenzdimensionen eine Rolle spielen, d. h. selten bis gar nicht erfasst werden.

6.2.2. Physikdidaktisches Wissen

Aus der Existenz einer Vielzahl von Tests zum physikdidaktischen Wissen (teils in Anlehnung an Shulman als *PCK* – „pedagogical content knowledge“ – konzeptualisiert; zu den Nuancen vgl. Gramzow, Riese und Reinhold, 2013; Gramzow, 2015) können, wie oben bereits angedeutet, eine Reihe von Anknüpfungspunkten zum eigentlichen

Ziel dieser Arbeit, der Entwicklung eines Instruments zur Erfassung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern, abgeleitet werden. Im besten Fall kann ein Konsens bezüglich relevanter physikdidaktischer Handlungsfelder sowie deren Ausdifferenzierung für die eigene Instrumententwicklung genutzt werden. Darüber hinaus ist es hinsichtlich der Anschlussfähigkeit von Belang zu erfahren, für welche Spezifitätsniveaus die Instrumente konzipiert wurden.

Insgesamt konnten acht deutschsprachige Tests[10] zum fachdidaktischen Wissen/PCK recherchiert werden (vgl. Tabelle 6.2). Diese unterscheiden sich hinsichtlich der Konzeptualisierung des Konstruktes sowohl formal als auch inhaltlich stark voneinander. Auf der formalen Ebene werden beispielsweise unterschiedlich viele Dimensionen in das Modell aufgenommen (vgl. dreidimensionales Itementwicklungsmodell bei Gramzow, 2015 vs. eindimensionales Modell bei Brovelli u. a., 2013). Inhaltliche Differenzen können z. B. anhand der in jeder Modellierung vorkommenden Dimension der „Facetten“ aufgezeigt werden. Es werden wiederum nicht nur unterschiedlich viele Facetten operationalisiert, sie variieren auch im formulierten Abstraktionsgrad (z. B. Instruktionsstrategien vs. lernendengerechte/fachgerechte Instruktionsstrategien). Darüber hinaus werden die Modellierungen auf ganz unterschiedliche Spezifitätsniveaus „reduziert“, was an der Dimension der „Inhaltsbereiche“ sichtbar wird. Diese reicht von „Physik“ (NUK), über „Mechanik“ (z. B. Profile-P) bis hin zu „Aggregatzuständen“ (PLUS).[11] Gramzow u. a. (2013, S. 13) bemerken dahingehend, dass „[d]ie Notwendigkeit einer solchen Dimension darin begründet [liegt], dass es keine hinreichend belastbaren Indizien für inhaltsunabhängige fachdidaktische Wissensbereiche (die ohne weiteres von einem physikalischen Inhalt auf einen anderen übertragen werden können) gibt.“

Zusätzlich weist Gramzow (2015, S. 85 f.) darauf hin, dass selbst bei gleicher Benennungen der Facetten keine Vergleichbarkeit der Konzeptualisierungen gegeben sein muss, da diese ganz unterschiedlich operationalisiert sein können. Weiterhin bemängelt die Autorin, dass Operationalisierungen häufig nicht offen gelegt werden, sodass Unterschiede nicht dezidiert erfasst werden können. Zusammenfassend ergibt sich aus der tabellarischen Darstellung, dass kein Konsens darüber besteht, was *das* fachdidaktische Wissen sein soll. „Innerhalb der deutschsprachigen Tradition basieren die Konzeptionen zwar in der Regel auf einem einheitlichen Rahmenmodell professioneller Handlungskompetenz. Sie unterscheiden sich aber häufig in der inneren Struktur und der Dimensionierung fachdidaktischen Wissens“ (Gramzow u. a., 2013, S. 9). Auch Orientierung bietende Standardwerke der Physikdidaktik (vgl. u. a. Mikelskis, 2006;

[10] Im englischen Sprachraum gibt es eine Vielzahl zusätzlicher Instrumente. Aufgrund der Ziele dieser Analyse werden diese jedoch nicht betrachtet.

[11] Borowski, Fischer, Olszewski, Reinhold und Riese (2010, S. 379) korrelieren Tests zum physkdidaktischen Wissen unterschiedlicher Inhaltsbereiche, nämlich den Mechanik-Test von Riese (2009) und den E-Lehre-Test aus dem QuIP-Projekt von Olszewski (2010). Sie finden eine mittlere Korrelation beider Tests ($r = .64^{***}$) bzw. faktorenanalytisch eine latente Korrelation von $r = .83^{***}$, wobei sich diese signifikant von eins unterscheidet.

Tabelle 6.2.: Facetten des physikdidaktischen Wissens in ausgewählten deutschsprachigen Instrumenten.

Projekt[1]	**physikdidaktische Facetten**	**Inhaltsbereiche**
Paderborn (FDW)	– allg. Aspekte physikal. Lernprozesse, – Einsatz von Experimenten, – Gestaltung & Planung von Lernprozesssen, – Beurteilung, Analyse & Reflexion von Lernprozessen, – adäquate Reaktion in kritischen Unterrichtssituationen	SP[2] Experimentieren im Physikunterricht, Mechanik
PLUS (FDW)	– Lernbedingungen (Schülervorstellungen, Lernschwierigkeiten) – Instruktion (geeignete Versuche, inhalt. Sequenzierung)	Aggregatzustände
QuIP (PCK)	– Schülervorstellungen (vorhersagen, erkennen, ändern), – Curriculum (Inhalte für Jahrgangsstufen auswählen), – inhaltl. Schwierigkeiten berücksichtigen, einschätzen, ...	Elektrizitätslehre
NUK (PCK)	– Umgang mit Schülervorstellungen, – kognitive Aktivierung (Verändern von Schülervorstellungen) – lernendengerechte Instruktionsstrategien, – fachgerechte Instruktionsstrategien, – Umgang mit originalen Begegnungen und Experimenten	Physik, Chemie, Biologie
ProwiN (PCK/FDW)	– Experimente, – Modelle/Konzepte, – Schülervorstellungen	SP Mechanik
Profile-P (FDW)	theoretisches Modell/*Itementwicklungsmodell*: – *Instruktionsstrategien*, – *Schülervorstellungen*, – *Experimente/Wissenschaftsverständnis*, – Kontext/Interesse, – Curriculum/Bildungsstandards, – (digitale) Medien, – *fachdid. Konzepte*, – Aufgaben	Mechanik
KiL (PCK/FDW)	– Schülerkognition, – Instruktionsstrategien, – Curriculum, Assessment	Mechanik, Elektrizitätslehre, Optik, etc.
FALKO-Physik (PCK)	– Instruktion (Erklären/Präsentieren von Sachverhalten) – Schülerkognition (Schülerfehler/Lernschwierigkeiten), – Messen und Experimentieren	unbekannt

[1] Angaben in Klammern geben Auskunft darüber, ob im Rahmen des Projektes fachdidaktisches Wissen (FDW) oder „pedagogical content knowledge“ (PCK) operationalisiert wurde, oder ob beide Begrifflichkeiten synonym verwendet werden.
[2] Schwerpunkt
Quellen: Paderborner Instrument: Riese (2009), PLUS: Lange, Kleickmann, Tröbst und Möller (2012), QuIP: Olszewski (2010), NUK: Brovelli (2014), ProwiN: Kirschner (2013), ProfiLe-P: Gramzow (2015), KiL: Kröger, Neumann und Petersen (2015), FALKO: Göhring und Schödl (2015).

Kircher, Girwidz & Häußler, 2010; Hopf, Schecker & Wiesner, 2011) oder Empfehlungen bezüglich der Lehrerausbildung beispielsweise der DPG-Quereinsteigerstudie[12] (vgl. Korneck, Lamprecht, Wodzinski & Schecker, 2010) oder der DPG-Studie zum Lehramtsstudium[13] (DPG, 2014) scheinen die Problematik nicht aufzulösen (vgl. Tabelle F.1 im Anhang).

Eine zusammenfassende Sichtung vieler relevanter Studien und Dokumente nimmt Gramzow u. a. (2013, S. 21-25) vor. Ihr Ergebnis kumuliert in nachfolgender Aufzählung und kann als eine übersichtliche Orientierung gelten, wobei sich der Umgang mit Heterogenität in jeder der Kategorien als Teilaspekt wiederfinden soll (vgl. Gramzow u. a., 2013, S. 22):

- Instruktionsstrategien,
- Schülervorstellungen,
- Experimente/Wissenschaftsverständnis,
- Kontext/Interesse,
- Curriculum/Bildungsstandards/Ziele,
- (Digitale) Medien,
- Fachdidaktische Konzepte,
- Aufgaben.

Die im Anhang F enthaltene Tabelle F.2 liefert eine ausführlichere Darstellung und Ausdifferenzierung der hier aufgeführten Themenfelder nach Gramzow u. a. (2013, S. 22). Diese Aufstellung kann als Vergleichsmaßstab dienen, wenn das neu zu entwickelnde Instrument zur Erfassung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern hinsichtlich des Spezifitätsniveaus und der durch das Instrument abgebildeten Inhalte in Relation zu potentiell relevanten physikdidaktischen Themenfeldern eingeschätzt werden soll.

6.3. Zusammenfassung

Im Rahmen des derzeit vorherrschenden Kompetenzparadigmas in der Lehrerbildung wird unter anderem versucht, einen Beitrag zu der Frage zu leisten, was eigentlich gute von schlechten Lehrkräften unterscheidet. Selbstwirksamkeitserwartungen können in

[12] Die DPG-Quereinsteigerstudie beschreibt einen Konsens unter Physikdidaktikerinnen und Physikdidaktikern und zwar in der Art, dass Module eines Weiterbildungsprogramms zum Erwerb fachdidaktischer Kompetenzen im Sinne physikdidaktischer Basisqualifikation festgelegt und ausgeschärft werden (vgl. auch Korneck & Lamprecht, 2010).

[13] An der DPG-Studie zum Lehramtsstudium haben ebenfalls eine Reihe von Physikdidaktikerinnen und Physikdidaktikern mitgearbeitet, die Empfehlungen zu fachdidaktischen Inhalten zusammenfassen.

diesem Paradigma sowohl als konstituierende Dimension oder als Kompetenz determinierende Variable aufgefasst werden. In beiden Lesarten kommt der Komponente eine hohe Bedeutung zu, die sich beispielsweise aus der sozial-kognitiven Theorie Banduras ergibt und für die eine Vielzahl empirischer Belege angeführt werden kann (vgl. Kapitel 3).

Allerdings konnte festgestellt werden, dass affektiv-motivationale Kompetenzdimensionen, zu denen Selbstwirksamkeitserwartungen zu zählen sind, selten erfasst (vgl. exemplarisch Tabelle 6.1) oder in Modellierungen berücksichtigt werden (KMK, 2004, 2010). Teilweise werden sie als determinierende Größen mit erhoben. Wenn Cauet, Liepertz, Borowski und Fischer (2015) fragen: „Does ist matter, what we measure?“, weil Zusammenhänge von Wissensfacetten zum Lehrerhandeln und zu Unterrichtsvariablen einigermaßen ernüchternd ausfallen (vgl. u. a. Vogelsang & Reinhold, 2013; Vogelsang, 2014; Cauet, Borowski & Fischer, 2015; Cauet, Liepertz, Borowski & Fischer, 2015), dann könnte vorsichtig angemerkt werden, dass Selbstwirksamkeitserwartungen in Kompetenzmodellierungen und -erhebungen unter Umständen eine bisher vernachlässigte Stellschraube darstellen könnten. Einige Studien weisen auf die Relevanz motivationaler und selbstregulativer Facetten bei der Kompetenzmessung hin (vgl. u. a. Stender, 2014). Zudem erscheint es intuitiv einsichtig, dass komplexe (Unterrichts-)Situationen selten allein mit Fachwissen bewältigt werden können.

Bezüglich der Integration des Konstruktes der Selbstwirksamkeitserwartungen existieren Modellierungsansätze (vgl. Rahmenmodell von Baumert und Kunter, 2006), die als Grundlage genutzt werden können. Für den Bereich der Physikdidaktik wird aus dieser Perspektive das Fehlen eines sorgfältig entwickelten Instrumentes zur Erfassung von Selbstwirksamkeitserwartungen auf einem passenden Spezifitätsniveau besonders bedeutsam. Die recherchierten Instrumente zur Erhebung physikdidaktischen Wissens können genutzt werden, um Informationen zur Konstruktion eines solchen Instruments abzuleiten.

Allerdings unterscheiden sich die beschriebenen Instrumente stark nach den operationalisierten Inhalten und dem Spezifitätsniveau (vgl. Tabelle 6.2). „Fast allen Ansätzen ist gemein, dass sie das Wissen über Schülerlernprozesse und -vorstellungen und das Wissen über Lehrstrategien und Darstellungsformen im naturwissenschaftlichen Unterricht als zentralen Inhaltsbereich beinhalten“ (Fischer & Borowski, 2010, S. 696). Von Aufschnaiter und Blömeke (2010, S. 363) weisen jedoch darauf hin, dass „sich zwar u. U. auf der Basis von Schlagwörtern (wie „Schülervorstellungen“, „Nature of Science“ oder „Experimente“) Einigkeit über die wesentlichen Elemente der Lehrerbildung [...] herstellen [lässt], dies jedoch in keiner Weise zu inhaltlich ähnlicher Schwerpunktsetzung führen [muss].“ Für die Entwicklung eines Instrumentes zur Erfassung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern bedeutet dies, dass es keine einfache Lösung hinsichtlich der Festlegung und Spezifizierung relevanter Hand-

lungsfelder geben kann. Trotzdem bieten die genannten Dokumente und Instrumente hinreichend Orientierung und Ansätze für eine begründete Auswahl (vgl. Kapitel 7). Dass physikdidaktisches Wissen jeweils inhaltsspezifisch gedacht werden muss, erscheint nachvollziehbar. Für Selbstwirksamkeitserwartungen muss dies jedoch nicht in gleicher Weise gelten. Personen können sich auf verschiedenen Ebenen selbst einschätzen. Ein Instrument auf einem mittleren Spezifitätsniveau kann dabei möglicherweise rasch für verschiedene Inhaltsbereiche angepasst werden, ohne für globalere Fragestellungen nutzlos zu sein (vgl. Abschnitt 7.2.1).

Abschließend soll auf die eingangs gestellte Frage nach der „guten“ Lehrperson Bezug genommen werden. Es bleibt zu konstatieren, dass die Kompetenzorientierung wenig zur Aufklärung selbiger beigetragen hat, denn es bleibt offen, „welche Kompetenzen in welcher Zusammensetzung und mit welcher Gewichtung einen kompetenten Lehrer ausmachen und ob diese notwendige oder hinreichende Voraussetzungen sind“ (Frey & Jung, 2011, S. 562). Terhart (2007b, S. 20) betont ganz allgemein: „Es ist illusorisch, ein festes Set von Eigenschaften und Verhaltensweisen zu benennen, das den perfekten Lehrer ausmacht.“ Insofern müssen jeweils situationsspezifisch adäquatere von weniger adäquaten Handlungsweisen von Lehrpersonen im Unterricht begründet identifiziert werden. Für die Lehrerbildung bedeutet dies, weiterhin in dem Dilemma zu agieren, aufgrund mehr oder weniger allgemeiner und konsensfähiger Standards Inhalte setzen und konkret ausgestalten zu müssen, ohne in diesem Zusammenhang auf die erhoffte „Sicherheit“ wissenschaftlicher Evidenz bezüglich der Relevanz und Wirksamkeit rekurrieren zu können. Bromme (1997, S. 203) betont die Dialektik der Fokussierung auf das Lehrerhandeln (im Rahmen der Kompetenzorientierung), die im Einklang mit Banduras sozial-kognitiver Theorie verstanden werden kann:

> „Lehrer [haben], je nach Blickwinkel, einerseits nur eine recht geringe Wirkung auf den Lernerfolg ihrer Schüler, [spielen] andererseits aber eine wichtige Rolle für die Gestaltung der Voraussetzungen für die Lernaktivitäten der Schüler. [...] Aus unterrichtspraktischen Gründen sollten Lehrer von ihrer Wirkung überzeugt sein und sich für den entscheidenden Faktor des Schülerlernens halten. Ihr fachlicher und pädagogischer Enthusiasmus beeinflusst die Aufmerksamkeit, das Interesse und die Motivation der Schüler. Andererseits müssen sie zugleich ihren relativ geringen Anteil an dem Zustandekommen von Lernleistungen realistisch beurteilen, um den beruflichen Belastungen standzuhalten und nicht ‚auszubrennen‘.“

Teil II.

Instrumententwicklung, Forschungsfragen und Studiendesign

7. Theoriebasierte Instrumentenentwicklung – Erste Schritte

„Deciding how to measure teacher efficacy presents thorny issues."

Tschannen-Moran & Hoy, 2001, S. 795

Die vorangegangenen Kapitel verdeutlichen den Anspruch an eine umfassende Entwicklung und Validierung eines neuen Testinstrumentes (vgl. u. a. Abschnitt 2.1.1). Die vorliegende Arbeit hat das Ziel, sich diesem formulierten Anspruch zu stellen, das heißt die beschriebenen Standards zu erfüllen oder offenzulegen, warum ggf. einzelne Standards nicht eingehalten werden (können). Die geforderte gründliche theoretische Einbettung wurde in den Kapiteln 3, 4 und 5 hinsichtlich des Konstruktes selbst sowie in Kapitel 6 hinsichtlich einer möglichen Einbettung des Forschungsprojektes in einen Gesamtzusammenhang (z. B. im Rahmen der Kompetenzforschung zum Lehrerberuf) vorgenommen. Im nachfolgenden Kapitel soll nun die Reflexion des theoretischen Kenntnisstandes in Bezug auf die konkrete Testentwicklung erfolgen. Das bedeutet, dass Entscheidungen bezüglich der Testentwicklung begründet, also möglichst aus den theoretischen Vorüberlegungen abgeleitet werden müssen.

Grob wird wie folgt vorgegangen: Zunächst werden relevante theoretische Aspekte, die sich aus Banduras Arbeiten ergeben, als Basis für die Testentwicklung transparent gemacht. Dazu gehört eine aus dem Kenntnisstand abgeleitete Arbeitsdefinition des Konstruktes, aus der sich wiederum konkrete Regeln für die Itemkonstruktion ableiten lassen (Grundlage zur Prüfung auf Inhaltsvalidität). Anschließend wird die inhaltliche Ausrichtung des Instrumentes begründet dargelegt. Erst dann werden konkrete Facetten der Testplanung diskutiert. Aus den Vorbetrachtungen ergibt sich sodann eine zu validierende Testwertinterpretation, die ausführlich dargestellt und von konkurrierenden alternativen Interpretationsansätzen abgegrenzt wird. Im nachfolgenden Kapitel können dann Forschungsfragen und Hypothesen auf Grundlage der festgelegten intendierten Testwertinterpretation formuliert, sowie methodische Vorgehensweisen zur Beantwortung der Fragen, also zur Evaluation des Instrumentes abgeleitet werden (vgl. Kapitel 8).

7.1. Theoretische Ableitungen

Zlatkin-Troitschanskaia und Seidel (2011, 228, Hervorhebung im Original) sehen in der „Zusammenhangslosigkeit von Theorieentwicklung und methodischen Orientierungen in der empirischen Bildungsforschung *das* akute Problem der aktuellen Kompetenzforschung“. Um diesem Problem zu begegnen, ist in einem ersten Schritt die theoretische Basis bereits ausführlich dargestellt worden. Im nächsten Schritt muss diese Basis konsequent und nachvollziehbar auf das eigene Projekt bezogen und angewendet werden. Im anschließenden Kapitel 8 soll die Verzahnung der theoretischen Überlegungen und der zu wählenden methodischen Herangehensweisen dezidiert erfolgen.

7.1.1. Arbeitsdefinition

Die Kritik an existierenden Instrumenten zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen bzw. Lehrer-Selbstwirksamkeitserwartungen in physik- oder naturwissenschaftsdidaktischen Handlungsfeldern wurde bereits dargestellt (vgl. Abschnitte 4.3 und 5.2). Diese bezieht sich zumeist auf eine fehlende oder inkonsequente Verankerung innerhalb eines theoretischen Grundgerüsts, eine ungenügende theoretische Abgrenzung des Konstruktes zu anderen, z. T. ähnlichen Konstrukten wie Handlungsergebniserwartungen, dem „locus-of-control“-Konstrukt oder dem Selbstkonzept sowie auf eine fehlende Inbeziehungsetzung von zugrunde gelegten definitorischen Ansätzen und konkreten Ausformulierungen der Items.

Die Angabe der nachfolgenden Arbeitsdefinition soll als eine Maßnahme verstanden werden, um die genannten Probleme möglichst auszuschließen. Die Arbeitsdefinition beruht auf den theoretischen Arbeiten Banduras (vgl. u. a. Bandura, 1977a, 1997) und kann als eine domänenspezifische Ausschärfung seiner Definition verstanden werden. Weiterhin rekurriert die Arbeitsdefinition explizit auf die in Abschnitt 3.2 herausgearbeiteten Eigenschaften des Konstruktes (Subjektivität/Selbstreferentialität, Handlungsbezug, Schwierigkeitsbezug, Situations- & Kontextabhängigkeit, Domänenspezifität), um die Abgrenzung zu anderen Konstrukten einsichtiger zu machen.

> *Arbeitsdefinition:*
> Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern angehender und erfahrener Physiklehrkräfte beziehen sich auf die Selbsteinschätzung genannter Personen, komplexe unterrichtliche Handlungen planen und durchführen zu können, auch wenn sich Widerstände in den Weg stellen.

Die Arbeitsdefinition greift damit explizit die Subjektivität/Selbstreferentialität des Konstruktes auf, indem Selbstwirksamkeitserwartungen als Selbsteinschätzungen konzipiert werden. Weiterhin wird gefordert, dass sich die Selbsteinschätzung auf eine komplexe unterrichtsbezogene Handlung beziehen muss (Handlungsbezug). Dadurch

werden Globaleinschätzungen zudem eher ausgeschlossen und ein Mindestanforderungsniveau formuliert. Das Merkmal des Schwierigkeitsbezugs wird direkt adressiert, indem die Selbsteinschätzung an eine Hürde gekoppelt wird. Die Domänenspezifität wird expliziert, indem zum einen die möglichen Handlungen auf den (Physik-)Unterricht und der Personenkreis auf Physiklehrkräfte eingeschränkt und zum anderen das Konstrukt domänenkonform benannt wird („Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern“). Die Situationsabhängigkeit ist implizit in der Definition enthalten.

7.1.2. Konstruktionsregeln bez. des Itemstammes

Aus den Eigenschaften des Konstruktes können explizite Konstruktionsregeln für den Itemstamm abgeleitet werden (vgl. Tabelle 7.1). Die Itemkonstruktionsregeln sind in ähnlicher Form bereits veröffentlicht worden (vgl. Meinhardt, 2011; Rabe, Meinhardt & Krey, 2012, S. 304 f.). Sie berücksichtigen die von Bandura formulierten Konstruktionshinweise für Selbstwirksamkeitserwartungsskalen (vgl. Abschnitt 3.6.2). Als Vorgriff muss darauf hingewiesen werden, dass aufgrund der Pilotstudien einzelne Konstruktionsregeln abgewandelt werden müssen (vgl. dazu Abschnitt 11.1 auf Seite 245).

Tabelle 7.1.: Aus der Theorie abgeleitete Itemkonstruktionsregeln.

Eigenschaft des Konstruktes	**abgeleitete Konstruktionsregel**
Subjektivität/Selbstreferentialität: Selbsteinschätzung eigener (aktueller) Fähigkeiten	Formulierung in der ersten Person Singular, Präsens
Handlungsorientierung: Selbsteinschätzung eigener Fähigkeiten bez. komplexer Handlungen	Verwendung bestimmter Phrasen wie „Ich kann...“, „Ich bin in der Lage...“, etc. gefolgt von nicht trivialen bzw. komplexen Handlungen
Domänenspezifität: Selbsteinschätzung eigener Fähigkeiten bez. komplexer unterrichtsbezogener Handlungen	Handlungen müssen spezifisch für die Planung oder Durchführung von Physikunterricht sein
Schwierigkeitsbezug: Selbsteinschätzung eigener Fähigkeiten bez. komplexer Handlungen bei auftretenden Schwierigkeiten	Formulierung passender Handlungsbarrieren durch Verwendung von Phrasen wie „auch wenn“, „obwohl“, etc. Zuspitzung/Erhöhung der Schwierigkeit durch Verwendung von Phrasen wie „immer“, „in jedem Fall“, etc.

Handlungsbarrieren und Zuspitzungen Barrieren können auf internale (z. B. Nervosität) oder externale (z. B. wenig Zeit) Gegebenheiten rekurrieren (vgl. Ajzen, 2002, S. 676 f.). Beide Beschränkungen erfordern eine zusätzliche subjektive Interpretation

auf Seiten des Lesers. Sowohl internale als auch externale Hürden werden individuell unterschiedlich interpretiert, das heißt, an einem individuellen Referenzrahmen gemessen. Eine bestimmte Zeitspanne kann beispielsweise für den einen Probanden wenig Zeit und für den nächsten Probanden viel Zeit bedeuten. Trotzdem kann sich jeder Proband eine Zeitspanne denken, die individuell „Zeitdruck“ bedeutet. Insofern ist davon auszugehen, dass authentische Hürden Items in der Regel schwieriger erscheinen lassen, auch wenn das zusätzliche Schwierigkeitsmaß nicht intersubjektivierbar ist. Handlungsbarrieren bilden daher ein Tool zur Steigerung der subjektiv wahrgenommenen Schwierigkeit in Items und helfen damit, Deckeneffekte zu vermeiden (vgl. Abschnitt 3.6.2).

Für die Itemkonstruktion wird festgelegt, ausschließlich externale Hürden zu formulieren, da davon ausgegangen wird, dass damit das Risiko minimiert wird, dass eine Hürde überhaupt nicht als Hürde wahrgenommen wird. Zusätzlich wird angenommen, dass externale Hürden besser intersubjektivierbar sind. Wichtig ist es, die Authentizität der Items durch die Hürden nicht zu gefährden, sodass Probanden die Items in ihrer Gesamtheit (Itemstamm und zugehörige Hürde) ernst nehmen können. Zudem dürfen die Hürden nicht zu schwer angesetzt sein; für Befragte müssen immer Handlungsmöglichkeiten erkennbar sein. Werden Situationen beschrieben, in denen Probanden zumindest scheinbar keine Optionen besitzen und damit keine Kontrolle ausüben können, so wird die Erhebung von entsprechenden Selbstwirksamkeitserwartungen ad absurdum geführt. Da die Hürden jeweils externaler Natur sind, ist eine Einflussmöglichkeit auf diese zumeist nicht gegeben. Dies verdeutlicht umso mehr die Wichtigkeit des grundsätzlichen Vorhandenseins möglicher Handlungsalternativen.

Die durch Handlung und Hürde insgesamt dargestellte Situation ist im Allgemeinen spezifischer als die Situation/Handlung ohne entsprechende Hürde und nimmt damit die Eigenschaft der Situationsspezifität des Konstruktes in Teilen auf. Wichtig ist, dass die beschriebene Gesamtsituation des Items für die Befragten vorstellbar und verständlich ist, sodass sich diese in die Situation hineinversetzen können und auf Grundlage existierender Vorerfahrungen oder anderer Quellen eine Beurteilung vornehmen können.

Zuspitzungen sollen ebenfalls als Mittel der Schwierigkeitserzeugung genutzt werden, auch wenn deren Verwendung in der Literatur kritisch betrachtet wird (vgl. Itemkonstruktionsregeln in Abschnitt 2.2.4).

Itemstruktur In der Regel ergibt sich aus den Konstruktionsregeln eine spezifische Satzstruktur aus Haupt- und Nebensatz. Die Items beginnen demnach mit einer Phrase wie „Ich bin in der Lage“ und adressieren sodann eine konkrete unterrichtsbezogene Handlung. Der sich anschließende, durch ein Komma getrennte Nebensatz wird dann durch eine Phrase wie „auch wenn“ oder „obwohl“ eingeleitet und beinhaltet eine konkrete, authentische Handlungsbarriere. Unter Umständen enthält ein Item ein zusätzliches schwierigkeitserzeugendes Merkmal, indem Phrasen wie „immer“ oder

„in jedem Fall“ eingebaut werden. Die Itemstruktur kann wie folgt zusammengefasst werden:

„Ich bin [immer] in der Lage ... [Handlung], auch wenn ... [Handlungsbarriere].“

Die Satzstruktur eignet sich dabei im Besonderen, um „agent-means“-Relationen zu formulieren, da durch die Verwendung von Phrasen in der ersten Person Singular der Akteur („agent“) direkt benannt ist und zur Komplettierung des Satzes jeweils eine konkrete Tätigkeit/Handlung („means“) spezifiziert werden muss (vgl. auch „Dont's“ der Itemformulierung im nachfolgenden Abschnitt 7.1.3).

Bemerkungen Es ergeben sich strukturell relativ komplexe Items. In der Regel wird davon abgeraten, solch komplexe Items in Fragebogenerhebungen zu formulieren, da die Komplexität unter Umständen negativ mit der Verständlichkeit und damit der Zuverlässigkeit der Items korreliert. Als Grund wird meist genannt, dass Lesern der Items ein höheres Maß an Aufmerksamkeit und Lesefähigkeit abverlangt wird. Es muss jedoch berücksichtigt werden, dass „[s]yntaktische Komplexität [...] auf ein semantisches Niveau verweisen [kann], das bei Reduktion von Komplexität womöglich bedeutsam verändert bzw. unterschritten wird“ (Mummendey & Grau, 2014, S. 162). An dieser Stelle erscheint die syntaktische Komplexität aus Haupt- und Nebensatz zwingend erforderlich zu sein, um die notwendigen Eigenschaften des Konstruktes in der Operationalisierung berücksichtigen zu können. Es muss in den Validierungsstudien jedoch eruiert werden, ob und ggf. inwiefern der Grad der Komplexität für die Befragten problematisch ist.

Entsprechend der sozial-kognitiven Theorie Banduras ist davon auszugehen, dass Probanden die Befragung nur dann als sinnvoll erleben, wenn sie die Items bezüglich eines bestimmten Ziels als relevant erachten (vgl. Abschnitt 3.2). Insofern müssen solche Items konstruiert werden, die einerseits das zu Grunde liegende Handlungsfeld in seinem Kern abdecken, aber gleichzeitig für eine große Anzahl von Befragten sinnvoll erscheinen. Das bedeutet konkret, dass solche Handlungen und Handlungsbarrieren durch die Items adressiert werden müssen, die als schulrelevant und authentisch wahrgenommen werden.

Entsprechend der Empfehlungen, die in Abschnitt 2.2 herausgearbeitet wurden, werden die Items so formuliert, dass sie keine Verneinungen enthalten.

7.1.3. Itemkonstruktion unter Berücksichtigung des Bezugs zu anderen Konstrukten

„Serious conceptual problems are created when a performance is misconstrued as the outcome of itself [...]."

Bandura, 1997, S. 22

Bandura fordert bezüglich der Itemkonstruktion eine klare Abgrenzung zu ähnlichen Konstrukten (vgl. Abschnitt 3.6.2). Als Quintessenz aus der intensiven Analyse der Unterschiede und Gemeinsamkeiten von Selbstwirksamkeitserwartungen zu anderen Konstrukten (vgl. Abschnitt 3.4) können diese Forderung konkretisiert und einige Unterlassungsregeln für die Itemkonstruktion abgeleitet werden:

1. Wie das einleitende Zitat andeutet, dürfen in keinem Fall Handlungen mit Zielen/Resultaten/Wirkungen/Konsequenzen gleichgesetzt werden, um eine Abgrenzung zu Handlungsergebniserwartungen sicher zu stellen und tatsächlich eine „agent-means"-Relation zu beschreiben (vgl. Unterabschnitte 3.2.1 und 3.4.7).
2. Darüber hinaus sollten in Abgrenzung zum Konstrukt des Selbstkonzepts keine affektiven Elemente in den Items enthalten sein (vgl. Unterabschnitt 3.4.3).
3. Außerdem sollten keine Ursachenzuschreibungen adressiert werden, weil damit Konfundierungen z. B. zum „locus-of-control"-Konstrukt forciert werden (vgl. Unterabschnitt 3.4.6).

7.1.4. Wahl des Antwortformates

Um Vergleichbarkeit und Anschlussfähigkeit zu gewährleisten, wird zunächst auf die vierstufige Ratingskala (stimmt nicht/kaum/eher/genau) von Schmitz und Schwarzer (2000) zurückgegriffen. Die Autoren berichten keine Probleme bei der Verwendung dieser Skala, die auch die von Bandura geforderte und inhaltlich sinnvolle Eigenschaft der Unipolarität aufweist (vgl. Abschnitt 3.6.2). Allerdings weist die Skala relativ wenige Stufen auf, sodass zu befürchten ist, dass extreme Ausprägungen selten auftreten und insgesamt weniger differenzierte Urteile abgegeben werden (vgl. Abschnitt 2.2.4). Ob sich die Stufigkeit der Skala als geeignet erweist, ist zunächst eine offene Frage, die im Zuge der Validierungsstudien aufgegriffen wird.

7.1.5. Reflexion des formalen Status zugrunde liegender Theorien

Banduras sozial-kognitive Theorie kann als fundiert und formal charakterisiert werden. Aus dem Theorierahmen ergeben sich eine Vielzahl überprüfbarer Hypothesen. Insbesondere die innere Struktur des Konstruktes wird durch Bandura detailliert und

nachvollziehbar dargestellt, sodass Eigenschaften hinsichtlich der Hierarchie oder der Dimensionalität des Konstruktes abgeleitet und überprüft werden können (vgl. Abschnitt 3.2). Zusätzlich wird das Konstrukt der Selbstwirksamkeitserwartung in Relation zu einer Vielzahl anderer Konstrukte gestellt. Die dezidierte Beschreibung der Wechselwirkung der Konstrukte wird aber letztlich nur für das Konstrukt der Handlungsergebniserwartung hinreichend ausführlich dargestellt. Dies ist unter Umständen damit zu erklären, dass die anderen Konstrukte nicht im Rahmen der sozial-kognitiven Theorie konzeptionalisiert sind. Wie sich die benachbarten Konstrukte untereinander verhalten, kann daher nicht stringent aus der Theorie abgeleitet werden (vgl. Bandura, 1997, S. 10-24). Insofern kann nur ein sehr eingeschränktes nomologisches Netz abgeleitet werden, welches durch zusätzliche Analysen – beispielsweise Ergebnisse empirischer Forschung – ergänzt werden muss. Zudem bleibt häufig unklar, in welcher Größenordnung Zusammenhänge zu anderen Konstrukten zu erwarten sind. Darüber hinaus ist das nomologische Netz in dem Sinne unvollständig, als keine strenge Axiomatik vorliegt.

Es erscheint damit sinnvoll, im Rahmen erster Validierungsmaßnahmen insbesondere auf die innere Struktur des domänenspezifisch formulierten Konstruktes (z. B. Hierarchie, Dimensionalität, etc.) und nicht auf seine Beziehungen zu anderen Konstrukten zu fokussieren.

7.2. Struktur des Konstruktes

„There is no all-purpose measure of perceived efficacy."

Bandura, 2006, S. 307

7.2.1. Festlegung des Spezifitätsniveaus

Selbstwirksamkeitserwartungen können auf verschiedenen Spezifitätsniveaus erfasst werden (vgl. Abschnitte 3.2.3 und 3.6.1), weshalb für die Itemkonstruktion ein Spezifitätsniveau festgelegt werden muss. Da ein Fragebogen entwickelt werden soll, ist es einigermaßen schwierig, das spezifischste Niveau nach Bandura umzusetzen, da dann in den Items ganz konkrete Situations- und Handlungsbeschreibungen enthalten sein müssen, die die Items umständlich und lang werden lassen. Globale Messungen werden nicht nur von Bandura hinsichtlich ihrer Prädiktivität und Aussagekraft angezweifelt. Daher wird ein mittleres Niveau gewählt, auf dem situationsübergreifende Handlungsklassen beschrieben werden: „The intermediate level measures perceived self-efficacy for a class of performances within the same activity domain under a class of conditions sharing common properties" (Bandura, 1997, S. 49). Items auf einem mittleren Niveau

haben ggf. den Vorteil, leicht auf ein spezifischeres Niveau anpassbar zu sein, sodass auch aus einer forschungsökonomischen Perspektive ein mittleres Niveau plausibel ist.

Selbstwirksamkeitserwartungen von Lehrpersonen auf ein bestimmtes Unterrichtsfach zu beziehen, empfiehlt auch Bandura (1997, S. 243): „[T]eacher efficacy scales [should be] tailored to domains of instructional functioning." Den Grund dafür sieht er darin, dass die Selbstwirksamkeitserwartungen eines Lehrers bezüglich des Unterrichtens verschiedener Fächer ganz unterschiedlich ausgeprägt sein können.

7.2.2. Inhalte/Handlungsfelder

Ziel dieser Arbeit ist es, ein Instrument zu entwickeln und zu validieren, welches Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern erfasst. In Abschnitt 6.2 konnte festgestellt werden, dass kein Konsens darüber herrscht, was *die* physikdidaktischen Handlungsfelder sein sollen. Für die Instrumentenentwicklung kann daher nicht auf ein festgelegtes Set an Handlungsfeldern zurückgegriffen werden.

Es können jedoch Themenfelder benannt werden, die eine Art Minimalkonsens darstellen, weil sie in fast jeder Konzeption physikdidaktischen Wissens, den Standardwerken der Physikdidaktik oder den Empfehlungen zum Physiklehramtsstudium enthalten sind. Zu diesen gehören die Handlungsfelder *Schülerlernprozesse/Schülervorstellungen*, *Lehrstrategien/Darstellungsformen/Instruktion* und das *Experimentieren*. Diese drei Handlungsfelder (sowie ein viertes relevant erscheinendes) werden für die Instrumentenentwicklung aufgegriffen (vgl. Rabe u. a., 2012, S. 304) und wie folgt benannt:

1. Umgang mit Schülervorstellungen (SV),
2. Experimentieren (EX),
3. Elementarisieren (EL),
4. Umgang mit Aufgaben (A).[1]

Folglich werden Skalen zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen (SWE) bezüglich der genannten vier Handlungsfelder konstruiert und wie folgt bezeichnet: SWE-SV, SWE-EX, SWE-EL, SWE-A. Es ergibt sich ein Strukturmodell für Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern (SWE-PHYDID), welches in Abbildung 7.1 dargestellt ist. Dieses Strukturmodell bildet Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern nicht vollständig ab.

Die entwickelten Items für jedes Handlungsfeld können dem Anhang G.1 entnommen werden (Stand Sommer 2011). Eine Auswahl an Themenfeldern, die in den Items adressiert werden, enthält Tabelle 7.2.

[1] Dieses Handlungsfeld wurde erst im Anschluss an die erste Pilotierung (vgl. Meinhardt, 2011) operationalisiert.

Tabelle 7.2.: Auswahl an Themenfeldern der Skalen (Skalenversion V1: Sommer 2011, vgl. Anhang G.1).

SWE-EL-P
Strukturierung physikalischer Inhalte Vereinfachung/Elementarisierung physikalischer Inhalte Planung von Unterrichtssequenzen, die Alltagsphänomene adressieren
SWE-EL-D
Anpassung von Erklärungen im Unterricht inhaltl. Strukturierung und Akzentuierungen von Unterrichtsstunden Herstellen thematischer Querverbindungen/Alltagsbezüge
SWE-SV-P
Erkennen von Schülervorstellungen Adressierung/Einbeziehung von Schülervorstellungen Strategien zum Umgang mit Schülervorstellungen
SWE-SV-D
Erkennen/Beurteilen von Schülervorstellungen Adressierung/Einbeziehung von Schülervorstellungen Strategien zum Umgang mit Schülervorstellungen
SWE-EX-P
Auswahl von Experimenten Zusammenstellung von Experimenten Entwicklung von Experimenten
SWE-EX-D
Durchführung/Variation von Experimenten Reflexion über naturwissenschaftliche Arbeitsweisen Unterstützung beim Experimentieren Inszenierung von Experimenten

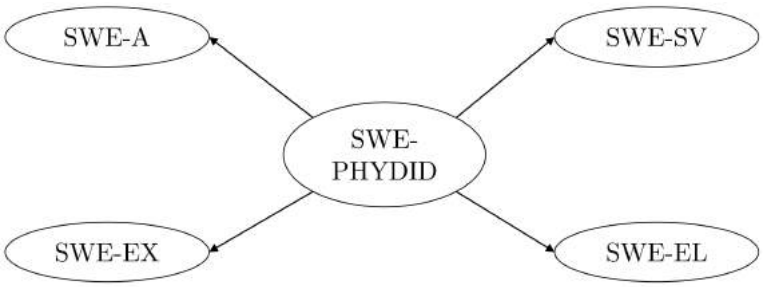


Abbildung 7.1.: Strukturmodell für Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern (SWE-PHYDID) mit den Komponenten Selbstwirksamkeitserwartungen bez. des Umgangs mit Schülervorstellungen (SWE-SV), des Experimentierens (SWE-EX), des Elementarisierens (SWE-EL) und des Umgangs mit Aufgaben (SWE-A).

7.2.3. Dimensionen

Es erscheint nicht nur intuitiv einsichtig, dass die Planung und Durchführung von Unterricht – neben erzieherischen und anderen Tätigkeiten – die zentrale Aufgabe von Lehrerinnen und Lehrern ist. Auch die KMK (2004, S. 3) nennt als Kernaufgabe von Lehrpersonen „die gezielte und nach wissenschaftlichen Erkenntnissen gestaltete Planung, Organisation und Reflexion von Lehr- und Lernprozessen sowie ihre individuelle Bewertung und systemische Evaluation." Diese Einschätzung teilen eine Vielzahl Forschender im Bereich der Lehrerprofessionalisierung (vgl. Bromme, 1997; Baumert & Kunter, 2006; Helsper, 2007). Von Aufschnaiter und Blömeke (2010, S. 365) dokumentieren die Diskussion eines wissenschaftlichen Workshops zur Erfassung professioneller Kompetenz und formulieren als Ergebnis, dass angenommen werde, dass sowohl Fähigkeiten zur Unterrichtsplanung als auch zur Unterrichtsdurchführung Aspekte der Kompetenz professioneller Lehrpersonen seien. Es erscheint damit sinnvoll, die festgelegten Handlungsfelder jeweils nach den Dimensionen Planung und Durchführung von Physikunterricht zu unterscheiden. Es handelt sich dabei um eine Festlegung aus Plausibilitätsgründen, die für jedes Handlungsfeld zu zweidimensionalen Strukturmodellen führt (vgl. Abbildung 7.2).

7.2.4. Hierarchie

Auf der Modellebene sind mehrere strukturell, das heißt hierarchisch unterschiedliche Versionen möglich und denkbar. Aus der Kombination der Operationalisierung mehrerer Handlungsfelder und deren Unterteilung in Planung und Durchführung ergibt sich das Modell in Abbildung 7.2, wobei die Handlungsfelder als strukturgebend konzeptualisiert sind.

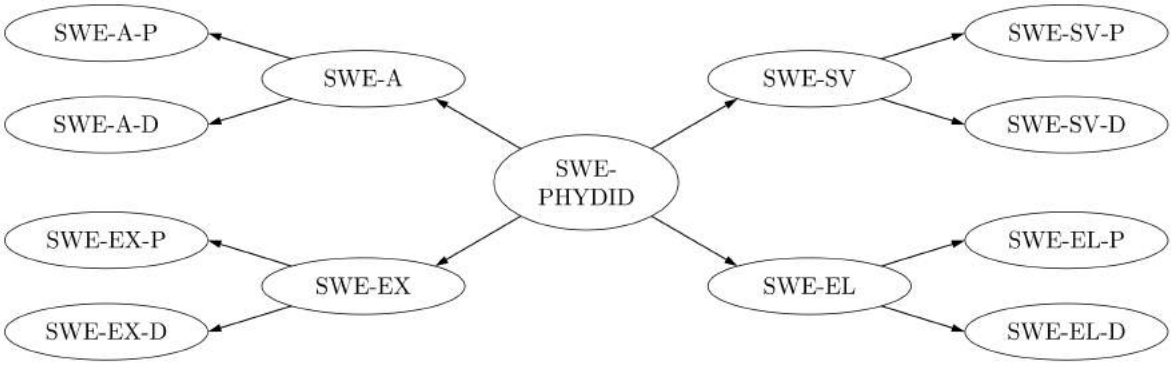


Abbildung 7.2.: Mögliches Strukturmodell zur Beschreibung physikdidaktischer Lehrer-Selbstwirksamkeitserwartungen – Handlungsfelder hauptsächlich strukturgebend.

Eine andere Möglichkeit besteht darin, dass die Dimensionen Planung und Durchführung strukturell überlegen sind. In diesem Fall ergäbe sich das Modell in Abbildung 7.3. Welches Modell die Domäne besser beschreibt, muss letztlich empirisch geprüft werden.

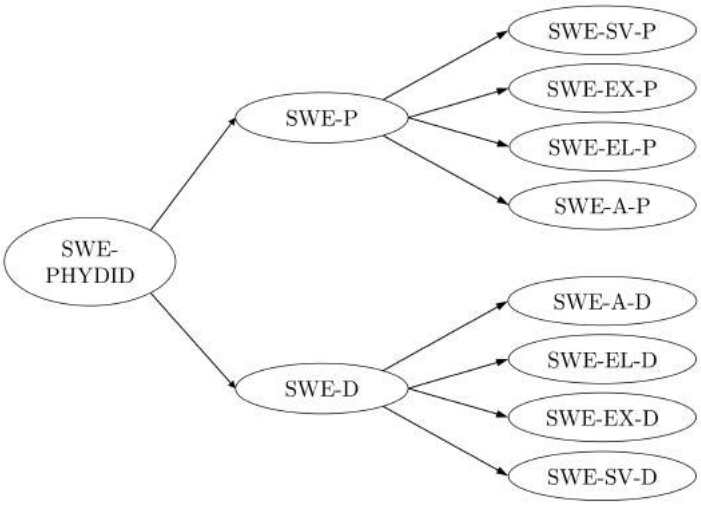


Abbildung 7.3.: Mögliches Strukturmodell zur Beschreibung physikdidaktischer Selbstwirksamkeitserwartungen – Dimensionen hauptsächlich strukturgebend.

7.2.5. Zusammenfassung

Es werden insgesamt acht Skalen zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen (SWE) in physikdidaktischen Handlungsfeldern entwickelt. Diese beziehen sich auf die vier Handlungsfelder Umgang mit Schülervorstellungen (SV), Experimentieren (EX), Elementarisieren (EL) und Umgang mit Aufgaben (A), jeweils in den Dimensionen Planung (P) sowie Durchführung (D) von Physikunterricht: SWE-EX-P, SWE-EX-D,

SWE-EL-P, SWE-EL-D, SWE-A-P, SWE-A-D, SWE-SV-P und SWE-SV-D (Abkürzungsschema: Selbstwirksamkeitserwartungen-Handlungsfeld-Dimension).

7.3. Vorgehen bei der Itemkonstruktion

Die Items wurden im Jahr 2011 in einem Zweier- bis Dreier-Team bestehend aus Thorid Rabe, Olaf Krey und Claudia Meinhardt intuitiv-erfahrungsbasiert, an den formalen Konstruktionsregeln orientiert und mit dem Ziel der kontinuierlichen Weiterentwicklung erarbeitet. Eine zentrale Maxime bestand darin, durch die Items die Breite der Handlungsfelder abzudecken, wobei zu berücksichtigen ist, dass naturgemäß einige Handlungsfelder breiter angelegt sind (z. B. das Handlungsfeld Experimentieren) als andere (z. B. das Handlungsfeld Umgang mit Schülervorstellungen). Elementare Bestandteile des Konstruktionsprozesses bestanden in der Assoziation für den Physikunterricht relevanter, komplexer Handlungen, der Assoziation authentischer Handlungsbarrieren sowie der Diskussion der jeweiligen Passungsverhältnisse von Handlungen und Hürden, der Einhaltung der Konstruktionsvorschriften und alternativer Formulierungsvorschläge. Der gesamte Itemkonstruktions- und -überarbeitungsprozess nahm mehrere Jahre in Anspruch und war durch ein konstruktives Ringen um Präzision, Verständlichkeit, Knappheit, Authentizität und Vollständigkeit gekennzeichnet. Dass dabei die inhaltliche Ausschärfung der einzelnen Handlungsfelder immer besser gelang, illustriert der Vergleich der Tabellen 7.2 und 11.3 auf den Seiten 191 bzw. 249.

7.4. Testplanung

Nachfolgend werden die in Abschnitt 2.2.3 dargestellten Aspekte, die für die Testplanung berücksichtigt werden müssen, schrittweise reflektiert und konkrete, im Rahmen der Validierungsstudien zu beachtende Problemfelder aufgezeigt.

Betrachtungen bez. der Merkmalsart Bei den beschriebenen und zu erfassenden Konstrukten (SWE-EX-P, SWE-EX-D, SWE-EL-P, SWE-EL-D, SWE-SV-P, SWE-SV-D, SWE-A-P, SWE-A-D) handelt es sich um quantitative Merkmale (denn sie variieren nach Bandura (1997) in „magnitude“), deren Ausprägungen mindestens auf einem Ordinalskalen-Niveau erfasst werden können.

Jedes Konstrukt wird durch mehrere Items repräsentiert (Skala) und zunächst formal als eindimensional betrachtet, wobei die Eindimensionalitätshypothese Gegenstand der Validierungsstudien sein wird. Darüber hinaus bilden die jeweilige „Planungs-“ und „Durchführungsskala“ eines Handlungsfeldes das zweidimensionale Konstrukt einer Selbstwirksamkeitserwartung bezogen auf dieses jeweilige Handlungsfeld (vgl. Abbildung 7.2). Insgesamt können die Skalen verschiedene Modelle bezüglich der Selbst-

wirksamkeitserwartungen in physikdidaktischen Handlungsfeldern bilden (vgl. Abschnitte 7.2.3 und 7.2.4). Die Dimensionalität des Konstruktes muss empirisch geprüft werden.

Die Skalen erfassen jeweils ein grundsätzlich überdauerndes bzw. stabiles Konstrukt. Die Itemformulierungen entsprechen einer „state“-Auffassung der Konstrukte. Die Situativität der Konstrukte wird zu einem gewissen Maße durch die Itemformulierungen und die in ihnen enthaltenen Beschreibungen spezifischer Situationen kontrolliert. Im Zuge der Validierungsstudien muss auf eine zusätzliche Kontrolle der Erhebungssituation geachtet werden.

Betrachtungen bez. der Testart Es handelt sich um einen spezifischen Persönlichkeitstest.

Betrachtungen bez. des Geltungsbereiches und der Zielgruppe Der Geltungsbereich wird nachfolgend genauer betrachtet (vgl. Abschnitt 7.5). Die Zielgruppe umfasst Physiklehramtsstudierende, Referendarinnen und Referendare des Faches Physik sowie Physiklehrkräfte. Für die Testentwicklung ergibt sich die Schwierigkeit der Herstellung einer Passung des Instrumentes für diese drei Personengruppen. So nimmt beispielsweise die Unterrichtserfahrung über die genannten Gruppen zu. Der jeweilige aktuelle Erfahrungshintergrund beeinflusst ggf. die Wahrnehmung und Deutung der Items, sodass diese jeweils als „kleinste gemeinsame Vielfache“ zwischen diesen Gruppen konzipiert werden müssen. Analysestichproben sollten Probanden aus allen drei Gruppen enthalten.

Betrachtungen bez. des Testaufbaus und der Testadministration Die Struktur des Testes (Testaufbau) ergibt sich nicht zwangsläufig aus den entwickelten Items. Diese muss jeweils in Bezug auf die der Validierungsstudie zugrunde liegenden Forschungsfrage begründet festgelegt werden.

Berücksichtigt werden muss, dass es sich bei der Erfassung von Selbstwirksamkeitserwartungen um spezifische Selbstauskünfte bzw. Selbsteinschätzungen handelt. Diese sind ihrer Natur nach subjektiv und daher besonders anfällig für potentielle bewusste oder unbewusste Verfälschungen. Insbesondere der unter Umständen durch die Anwesenheit von Ausbilderinnen und Ausbildern (Dozentinnen und Dozenten) im Datenerhebungsprozess auftretenden Tendenz zur sozialen Erwünschtheit muss aktiv entgegen getreten werden, um verzerrende Effekte zu vermeiden. Spezifische Instruktionen sind in jedem Fall zu berücksichtigen (zum Beispiel: Zusage absoluter Anonymität; Bitte um wahrheitsgetreue Auskünfte; Hinweis, die Fragen als Selbstreflexion zu begreifen; Hinweis darauf, dass konkrete Erfahrungen zur Beantwortung nicht zwingend erforderlich sind, ...).

Betrachtungen bez. der Testzeit und der Testlänge Es ist anzunehmen, dass die Konstrukte mit fünf bis zehn Items zufriedenstellend abgebildet werden können, sodass auch eine akzeptable Testzeit (max. 15 Minuten) für einzelne Skalen gewährleistet werden kann. Die tatsächliche Dauer der Beantwortung aller entwickelten Items muss zunächst erprobt werden.

7.5. Festlegung der zu validierenden Testwertinterpretation

> „Validation will progress in proportion as we collectively do our damnedest – no holds barred – with our minds and our hearts."
>
> *Cronbach, 1988, S. 14*

Nachfolgend soll entsprechend der Ausführungen in Kapitel 2 eine Testwertinterpretation festgelegt werden, deren Gültigkeit im Zuge von Validierungsstudien untersucht werden soll. Da Zeit und Ressourcen eines Promotionsprojektes begrenzt sind, kann nur ein Ausschnitt potentieller Validierungsfragen gestellt und (vorläufig) beantwortet werden. Ein konkreter Testnutzen kann im Zuge des Projektes nicht dezidiert untersucht werden. Deshalb wird dem Vorschlag Kanes gefolgt und ein Validitätsargument erarbeitet, welches sich ausschließlich auf den „descriptive part" (Kane, 2001, S. 337) bezieht, also keine entscheidungsbasierten oder extrapolierenden Schlüsse inkludiert, die für eine spezifische Testnutzung relevant wären (vgl. Abschnitt 2.1.4). Zur festgelegten Testwertinterpretation werden darüber hinaus mit ihr potentiell konkurrierende Interpretationsmöglichkeiten angegeben, deren Gültigkeit im Zuge der Validierung widerlegt werden soll. Aus den Annahmen, die der intendierten Testwertinterpretation zugrunde liegen, können Maßnahmen abgeleitet werden, die zur Prüfung selbiger geeignet sind und entsprechend der Vorgaben der Standards bzw. der Empfehlungen Messicks verschiedene Validitätsaspekte (Messick, 1995, S. 744 ff.) bzw. Quellen möglicher Evidenz (AERA, APA, NCME, 2014, S. 14-20) berücksichtigen (vgl. Abschnitt 2.1.3 und Abschnitt 2.1.4/abschließende Bemerkungen). Aus dem Procedere können Forschungsfragen zusammenfassend abgeleitet und im nachfolgenden Kapitel 8 dargestellt werden.

7.5.1. Intendierte Testwertinterpretation

Wie bereits angedeutet, wird in der zu validierenden Testwertinterpretation *kein* spezifischer Testnutzen angegeben, sodass im Rahmen des Promotionsprojektes lediglich beschreibende Elemente des Validitätsargumentes berücksichtigt werden (vgl. Kane, 2001, S. 337). Die intendierte Testwertinterpretation lautet wie folgt:

Intendierte Testwertinterpretation:
Die auf den entwickelten Skalen erzielten Skalenwerte spiegeln die Lehrer-Selbstwirksamkeitserwartungen bez. des skalenspezifischen Handlungsfeldes und der skalenspezifischen Dimension der befragten Physik-Lehramtsstudierenden, Physik-Lehramtsanwärterinnen und -anwärter sowie Physiklehrpersonen wider.

Die derart vorgenommene Fokussierung der Testwertinterpretation ist dabei ein zulässiges Vorgehen und mit den Vorgaben der Standards vereinbar, die zwar bei festgelegter Testnutzung eine Validierung selbiger einfordern, aber nicht festlegen, dass jede Validierung zwangsläufig auf eine Testnutzung spezifiziert sein muss (vgl. AERA, APA, NCME, 2014, S. 14). „It is possible to appraise the construct validity of a test interpretation without considering test use so long as no use is intended. As soon as a use is specified, then the validity investigation, including analysis of effects, must be tailored to the particular application“ (Shepard, 1997, S. 6).

Die hier angegebene Testwertinterpretation kann damit als Basis für spezifischere Testwertinterpretationen[2] (bez. konkreter Testnutzungen) angesehen werden. Dabei sind folgende Nutzungen des Instrumentes grundsätzlich denkbar:

- Selbstevaluation,
- Fremdevaluation in Vorbereitung von Interventionen (z. B. Ausrichtung von Lehrveranstaltungen)
- Zuweisung zu Kursen,
- Selektion (Studienwahl),
- Untersuchung von Entwicklungsverläufen.

7.5.2. Zugrunde liegende Annahmen der intendierten Testwertinterpretation

Der formulierten Testwertinterpretation liegen bestimmte Annahmen zugrunde, die nachfolgend und entsprechend des vorgeschlagenen Vorgehens nach Kane (2001, S. 30) transparent gemacht werden (vgl. Unterabschnitt 2.1.4). Das Ziel sich anschließender Validierungsmaßnahmen ist es dann, die Gültigkeit der dargelegten Annahmen zu prüfen. Wenn die Gültigkeit der nachfolgend genannten Annahmen argumentativ einsichtig gemacht werden kann, dann können diese Annahmen als Argumente für die Rechtfertigung der Gültigkeit bewertender als auch generalisierender Schlüsse dienen.

Damit die intendierte Testwertinterpretation zulässig ist, muss grundsätzlich verlangt werden, dass es sich bei den Items jeweils um *angemessene* Operationalisierungen von Lehrer-Selbstwirksamkeitserwartungen dem gewählten Handlungsfeld und der gewählten Dimension entsprechend handelt, sodass die jeweiligen Skalen in ihrer Gesamtheit *angemessene* Operationalisierungen der intendierten Konstrukte (SWE-SV-P, SWE-SV-D,

[2]Ein Beispiel für eine Formulierung einer konkreten Testwertinterpretation findet sich bei Kane (1992).

SWE-EX-P, etc.) darstellen und Ausprägungen der Konstrukte in bestimmten Gruppen zuverlässig bestimmt werden können. Was dabei jeweils unter *Angemessenheit* zu verstehen ist, soll die nachfolgende Ausdifferenzierung in elf Annahmen erläutern.
Damit also die intendierte Testwertinterpretation zulässig ist, muss auf Itemebene u. a. verlangt werden, dass die konstruierten Items ...

1. ... Selbstwirksamkeitserwartungen im Sinne der Arbeitsdefinition abbilden.
2. ... bei Probanden intendierte kognitive Prozesse hervorrufen.
3. ... von Probanden inhaltlich wie intendiert verstanden werden.
4. ... für Probanden authentisch sind.
5. ... gruppenübergreifend inhaltlich ähnlich verstanden werden.

Zusätzlich muss für die Zulässigkeit der intendierten Testwertinterpretation gewährleistet werden, dass ...

6. ... eine Positionierung zu den Items ausreichend differenziert möglich ist.

Damit die intendierte Testwertinterpretation zulässig ist, muss darüber hinaus auf Skalenebene u. a. verlangt werden, dass ...

7. ... die intendierten Konstrukte jeweils treffend und repräsentativ durch Items abgebildet werden.
8. ... die konstruierten Skalen das jeweils intendierte Konstrukt zuverlässig abbilden.
9. ... die intendierten Konstrukte gegeneinander eindeutig abgrenzbar sind.
10. ... die intendierten Konstrukte zu anderen/ähnlichen Konstrukten eindeutig abgrenzbar sind. (Es gilt die Prämisse, dass für die „Referenzkonstrukte“ eine valide Testwertinterpretation möglich ist.)
11. ... die intendierten Konstrukte gruppenübergreifend ähnlich wahrgenommen werden.

Die Liste ist nicht vollständig, da nur solche Annahmen angeführt werden, die im Zuge des Validierungsprozesses durch geeignete Maßnahmen tatsächlich überprüft werden können (vgl. nachfolgender Abschnitt 7.5.3). Die dargestellten Annahmen verdeutlichen darüber hinaus, dass ein Validitätsargument nur auf argumentativem Wege und nicht algorithmisch oder durch simple Anwendung von Routinen abgeleitet werden kann. Die jeweiligen Annahmen sind nicht trennscharf und nicht unabhängig voneinander, sondern fokussieren jeweils auf bestimmte Aspekte der Angemessenheit der Operationalisierungen.

Zusätzlich zu den dargestellten elf Annahmen, die notwendige Voraussetzungen für die Gültigkeit der Testwertinterpretation darstellen, können weitere Annahmen angegebenen werden, die im Falle ihrer Gültigkeit jedoch lediglich als Plausibilitätsargumente im Argumentationsprozess angeführt werden können.
Wenn also die intendierte Testwertinterpretation zulässig ist, dann können zum Beispiel ...

12. ... Mittelwertunterschiede auf den Skalen bezüglich verschiedener Probandengruppen entsprechend theoretischer Überlegungen plausibel gemacht werden.

7.5.3. Gewählte Verfahren zur Prüfung der Gültigkeit getroffener Annahmen

> „Validators should indeed do what the detached scientist would do, as best they can within the constraints usually imposed by a short time horizon."
>
> *Cronbach, 1988, S. 14*

Nachfolgend werden Möglichkeiten zur Prüfung der Gültigkeit der Annahmen, die der zu validierenden Testwertinterpretation zugrunde liegen, dargestellt. Dabei handelt es sich um keinen vollständigen Katalog aller überhaupt in Frage kommenden Maßnahmen, sondern um eine *Auswahl*, deren Umsetzung im Zeitrahmen eines Dissertationsprojekts als realistisch eingestuft wird. Insofern handelt es sich auch um pragmatische Entscheidungen. Trotzdem wurde versucht, eine große Anzahl von Maßnahmen in den Validierungsprozess zu integrieren, um abschließend ein möglichst stichhaltiges Validitätsargument ableiten zu können sowie um ein breites Fundament für sich anschließende Validierungsschritte, die beispielsweise auch mögliche Testnutzungen adressieren, zu schaffen. Viele der Möglichkeiten wurden bereits in Kapitel 2 vorgestellt.

Die nachfolgend dargestellten Maßnahmen (vgl. Tabelle 7.3) werden den nach Messick (1995, S. 744 ff.) zu berücksichtigenden (inhaltsbezogenen, substantiellen, strukturellen, generalisierenden und externen) Validitätsfacetten zugeordnet (vgl. Unterabschnitt 2.1.3), wobei konsequenzenbezogene Maßnahmen aufgrund der Natur der Testwertinterpretation (vgl. Unterabschnitt 7.5.1) nicht durchgeführt werden. Damit wird den Empfehlungen der Standards gefolgt, Belege für die Gültigkeit bestimmter Annahmen aus unterschiedlichen Quellen zu generieren (vgl. AERA, APA, NCME, 2014, S. 14).[3]

[3] An dieser Stelle könnte der Eindruck entstehen, dass die Validierung erst mit dem Abarbeiten der dargestellten Maßnahmen beginnt. Die Grundlagen in Bezug auf die Sinnhaftigkeit der Durchführung der dargestellten Prüfmaßnahmen bildet jedoch eine gewissenhafte theoretische Fundierung (Arbeitsdefinition erarbeiten und theoretisch stützen, Konstruktionsvorschriften theoretisch begründet ableiten...), die damit essentieller Teil der Validierung ist (vgl. Abschnitt 2.1).

Tabelle 7.3.: Maßnahmen zur Prüfung der Gültigkeit der Annahmen, die der intendierten Testwertinterpretation zugrunde liegen.

Maßnahme	**Annahme**	**Validitätsaspekt nach Messick**
(1) Experten befragen, inwiefern die explizierten Konstruktionsvorschriften angemessen und richtig umgesetzt sind	1	inhaltlich
(2) Beantwortungsprozess der Probanden beobachten und prüfen, ob Selbsteinschätzungen vorgenommen, Handlungen als komplex umschrieben, die Hürden als schwierigkeitserzeugend wahrgenommen sowie domänenrelevante Erfahrungen assoziiert werden	1, **2**, 3	substantiell
(3) Experten fragen, inwiefern die konstruierten Items (für unterschiedliche Probandengruppen) verständlich sind	3, 5, 11	inhaltlich
(4) Probanden (unterschiedlicher Probandengruppen) beim Paraphrasieren der Items beobachten und prüfen, inwiefern intendierte domänenspezifische Einheit aus Handlung und Handlungsbarriere beschrieben sowie intendierte Dimensionen assoziiert wird	1, 2, **3**, 5, 11	inhaltlich, generalisierend
(5) Experten und Probanden fragen, inwiefern die konstruierten Items authentisch sind	**4**	inhaltlich
(6) Experten und Probanden fragen, für wie schwierig sie die konstruierten Items halten	**5**	inhaltlich
(7) Passung der Stufigkeit der Ratingskalen (statistisch) überprüfen	**6**	strukturell
(8) Probanden nach Eignung des Antwortformates befragen	**6**	strukturell
(9) Experten und Probanden fragen, inwiefern Handlungsfelder treffend und repräsentativ abgebildet sind	**7**	inhaltlich
(10) Eindimensionalität der Skalen (statistisch) prüfen	**7**	strukturell
(11) Reliabilitätsanalysen durchführen	**8**	generalisierend
(12) Experten fragen, inwiefern Planungs- und Durchführungsdimension eines Handlungsfeldes zu trennende Konstrukte darstellen	9	strukturell
(13) Zusammenhänge der entwickelten Skalen untereinander prüfen (Korrelationsanalyse)	**9**	strukturell
(14) Dimension der Handlungsfelder statistisch prüfen	**9**	strukturell
(15) Zusammenhänge der entwickelten Skalen zu Skalen prüfen, die Selbstwirksamkeitserwartungen auf anderen Spezifitätsniveaus abbilden (Korrelationsanalyse)	**10**	strukturell, external
(16) Zusammenhänge der entwickelten Skalen zu Skalen prüfen, die andere Konstrukte abbilden (Korrelationsanalyse)	**10**	external
(17) Messinvarianz durch Mehrgruppenvergleiche prüfen	**11**	generalisierend
(18) Mittelwertvergleiche durchführen	**12**	strukturell

7.5.4. Konkurrierende Testwertinterpretationen

Zur intendierten Testwertinterpretation sind mehrere alternative Interpretationen denkbar, die sich aus der dargestellten Nähe des Konstruktes der Selbstwirksamkeitserwartungen zu anderen Konstrukten (vgl. Abschnitt 3.4) ergeben. Deren Widerlegung wird im Validierungsprozess besondere Aufmerksamkeit geschenkt werden. Folgende mit der intendierten konkurrierende Testwertinterpretationen werden betrachtet:

1. Die erhobenen Daten spiegeln das Selbstkonzept in physikdidaktischen Handlungsfeldern der befragten Probanden wider.
2. Die erhobenen Daten spiegeln Handlungsergebniserwartungen in physikdidaktischen Handlungsfeldern der befragten Probanden wider.
3. Die erhobenen Daten spiegeln Kontrollüberzeugungen in physikdidaktischen Handlungsfeldern der befragten Probanden wider.

7.5.5. Abschätzung potentieller Konsequenzen

> „At best, test developers can be sensitive to social issues and try to support responsible test use.“
>
> *Reckase, 1998, S. 16*

Das Abschätzen potentieller Konsequenzen aufgrund des Testeinsatzes gestaltet sich insofern schwierig, als das Instrument, wie bereits mehrfach erwähnt, zunächst zu keinem spezifischen Zweck entwickelt wird, auch wenn mehrere Anwendungskontexte denkbar sind (vgl. Unterabschnitt 7.5.1). Dem Zitat folgend können unter der Prämisse eines positiv verlaufenden Validierungsprozesses jedoch einige Anwendungsgebiete eher empfohlen werden als andere. Insbesondere Forschungsanliegen, die auf Verhaltensvorhersagen angewiesen sind, d. h. auf extrapolierende Schlüsse im Zuge der Stützung der Testwertinterpretationen zurückgreifen (z. B. Kurszuweisung, Selektion), müssen zusätzliche weitreichende Validierungsmaßnahmen anstreben, da diesbezüglich in dieser Arbeit kein Beitrag geleistet wird.

Ob das Konstrukt im Rahmen der genannten Anwendungskontexte überhaupt als sinnvolles Kriterium eingesetzt werden kann, steht außerdem zunächst in Frage. Dies ist unter anderem damit zu begründen, dass im Sinne erklärender Schlüsse bisher nicht erschöpfend dargestellt werden kann, woraus konkrete Ausprägungsunterschiede des Konstruktes resultieren. Der exakte Bedeutungsgehalt unterschiedlicher Skalenwerte bleibt damit zunächst vage.

Eine Konsequenz, die aus der gewissenhaften Beantwortung des Fragebogens resultiert, ist die direkte Auseinandersetzung der Probanden mit ihren eigenen Fähigkeiten.

Insofern stellt das Ausfüllen eine Reflexionsgelegenheit für die Befragten dar. Die Auseinandersetzung kann dabei zu positiven Effekten führen; beispielsweise motivierend wirken, sich selbst weiterzuentwickeln. Auch eine demotivierende Wirkung muss jedoch in Betracht gezogen werden, insbesondere wenn die eigenen Fähigkeiten als eher gering eingeschätzt werden und keine Möglichkeiten gesehen werden, diese in absehbarer Zeit zu verbessern, weil in der subjektiven Betrachtungsweise die Lerngelegenheiten fehlen oder Anderes Priorität genießt.

Die Befragten setzen sich jedoch nicht nur mit ihren eigenen Fähigkeiten auseinander, sondern auch mit einem durch die Items implizit wie explizit vermittelten Bild von Physikunterricht. Dies kann zu bestätigenden, aber auch ablehnenden oder verunsichernden Gedanken und Gefühlen führen.

Mögliche Konsequenzen sollten potentielle Anwender beim Einsatz des Instrumentes in Forschungs- oder Lehrkontexten beachten. Zusammenfassend scheint der Anwendungsbereich der Selbstevaluation/Reflexion nach erfolgreicher Validierung als der am ehesten zu empfehlende und zu vertretende. Nicht alle denkbaren Konsequenzen können an dieser Stelle antizipiert werden (vgl. Reckase, 1998).

7.6. Zusammenfassung

> „[Test developers] are responsible for the uses that they advertise and that are closely implied by the name of their test."
>
> *Shepard, 1997, S. 13 f.*

Die Itemkonstruktion wird an verschiedenen zu berücksichtigenden Aspekten ausgerichtet; beispielsweise an den sich aus den Definitionen ergebenden Charakteristika des Konstruktes (vgl. Abschnitt 3.2), den Kriterien zur Abgrenzung zu anderen, benachbarten Konstrukten (vgl. Abschnitt 3.4) sowie an den bekannten Problematiken bei der Konstruktion von Items zum Konstrukt der Lehrer-Selbstwirksamkeitserwartung (vgl. Abschnitt 4.3). Damit kann die Itementwicklung bzw. die Instrumententwicklung im Allgemeinen als theoretisch fundiert bezeichnet werden. Die herausgearbeiteten Konstruktionsregeln können zusätzlich als Kriterien im Zuge des Validierungsprozesses genutzt werden, an denen die Angemessenheit der konkreten Umsetzung beurteilt werden kann.

Insgesamt werden acht Skalen zu vier Handlungsfeldern und bezüglich zweier Dimensionen entwickelt. Die Wahl der Handlungsfelder erfolgt dabei nach Sichtung einer Vielzahl von Dokumenten, die im Rahmen des Kompetenzansatzes vorgestellt wurden (vgl. Unterabschnitt 6.2.2). Die Dimensionen ergeben sich aus Überlegungen, die jedoch keiner etablierten Theorie entstammen. Diese eher intuitive Festlegung wird jedoch

von einer Vielzahl von Autoren und Wissenschaftlern im Forschungsbereich als sinnvoll erachtet. Die inhaltliche Ausgestaltung der Handlungsfelder und Formulierung der Items erfolgt orientiert an den genannten Anhaltspunkten, ansonsten jedoch erfahrungsbasiert.

Im Rahmen dieser Dissertation kann die Validierung nicht vollumfänglich geleistet werden (was auch grundsätzlich nicht möglich ist, vgl. Kapitel 2). Der Schwerpunkt liegt auf dem beschreibenden Teil des Validitätsargumentes (vgl. Kane, 2001), welches bezüglich der dargestellten Testwertinterpretation abgeleitet werden soll. Die Annahmen, die der intendierten Testwertinterpretation zugrunde liegen, sind transparent gemacht worden. Entsprechend der Vorgaben der Standards (AERA, APA, NCME, 2014) soll deren Gültigkeit im Zuge von nachfolgend detaillierter beschriebenen Validierungsstudien geprüft werden.

8. Forschungsfragen, Hypothesen, Studiendesign

Die in diesem Kapitel vorgestellten Forschungsfragen ergeben sich aus der zu validierenden intendierten Testwertinterpretation (vgl. Unterabschnitt 7.5.1) bzw. direkt aus den dieser Testwertinterpretation zugrunde liegenden Annahmen (vgl. Unterabschnitt 7.5.2). Aufgrund der detailliert vorgenommenen theoretischen Fundierung des Item- und Skalenentwicklungsprozesses können begründete Hypothesen abgeleitet werden, die es in Validierungsstudien zu überprüfen gilt. Diese Studien werden im Abschnitt 8.2 vorgestellt. Sie rekurrieren auf die in Unterabschnitt 7.5.3 abgeleiteten Maßnahmen zur Prüfung der Gültigkeit der erwähnten Annahmen.

8.1. Forschungsfragen und Hypothesen

Es können eine leitende, übergeordnete Forschungsfrage sowie eine Vielzahl untergeordneter Teilforschungsfragen und zugehörige Hypothesen formuliert werden.

8.1.1. Übergeordnete Forschungsfrage

Die übergeordnete Forschungsfrage (üFF) resultiert aus dem Ziel dieser Arbeit, das Forschungsdesiderat der Neuentwicklung eines Instrumentes zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern angehender und erfahrener Physiklehrpersonen aufzugreifen.

üFF Inwiefern können mit dem neu entwickelten Instrument gewonnene Testwerte entsprechend der formulierten Testwertinterpretation valide gedeutet werden? Inwiefern geben die Testwerte also auf dem gewählten Spezifitätsniveau (d. h. in physikdidaktischen Handlungsfeldern bezogen auf die Planung und Durchführung von Physikunterricht) valide Auskunft über die Selbstwirksamkeitserwartungen von angehenden und erfahrenen Physiklehrpersonen?

8.1.2. Teilforschungsfragen und zugehörige Hypothesen

FF1 Inwiefern bilden die Items Selbstwirksamkeitserwartungen entsprechend der abgeleiteten Arbeitsdefinition ab? (vgl. Annahme 1, Unterabschnitt 7.5.2)

Hypothese:
1.1. Die Items entsprechen im Wesentlichen aufgrund der sorgfältigen und theoriegeleitet vorgenommenen, d. h. an theoretisch abgeleiteten Konstruktionsregeln

orientierten Itemkonstruktion, den Charakteristika von Selbstwirksamkeitserwartungen.

FF2 Inwiefern rufen die Items bei Probanden intendierte kognitive Prozesse hervor? (vgl. Annahme 2, Unterabschnitt 7.5.2)

Hypothese:
2.1. Bei der Bearbeitung der Items werden aufgrund der theoretischen Fundierung der Itemkonstruktion entsprechend der Merkmale des intendierten Konstruktes (vgl. Abschnitt 3.2.2) erwartete kognitive Prozesse bei den Probanden hervorgerufen, d. h. es werden (1) Selbsteinschätzungen bez. (2) domänenspezifischer, (3) komplexer Handlungen unter (4) Berücksichtigung der als schwierigkeitserzeugend wahrgenommenen Handlungsbarrieren vorgenommen.

FF3 Inwiefern werden die Items von Probanden inhaltlich wie intendiert verstanden? (vgl. Annahme 3, Unterabschnitt 7.5.2)

Hypothese:
3.1. Die Items werden aufgrund der gewissenhaft vorgenommenen Itemkonstruktion in der Regel von Probanden wie intendiert verstanden, d. h. es werden jeweils wie erwartet entsprechend der intendierten Domäne zum Handlungsfeld und zur Dimension passende Einheiten aus Handlung und Handlungsbarriere assoziiert.

FF4 Inwiefern sind die Items für Probanden authentisch? (vgl. Annahme 4, Unterabschnitt 7.5.2)

Hypothese:
4.1. Die in den von Experten entwickelten Items beschriebenen Situationen (Einheit aus Handlung und Handlungsbarriere) werden als relevant für die Planung und Durchführung von Physikunterricht erachtet (Authentizität).

FF5 Inwiefern können die spezifischen Lehrer-Selbstwirksamkeitserwartungen der Befragten mit Hilfe der Stufigkeit der verwendeten Ratingskala differenziert erfasst werden? (vgl. Annahme 5, Unterabschnitt 7.5.2)

Hypothese:
5.1. Da die verwendete vierstufige Skala bereits in einer Reihe von Studien eingesetzt wurde und keine Indizien publiziert wurden, die für ein Nichtfunktionieren sprechen (vgl. Unterabschnitt 7.1.4), ist zunächst davon auszugehen, dass diese ein ausreichend differenziertes Urteil der Befragten zulässt.

FF6 Inwiefern sind die intendierten Konstrukte durch die entwickelten Items treffend und repräsentativ abgebildet? (vgl. Annahme 7, Unterabschnitt 7.5.2)

Hypothesen:
6.1. Es ist davon auszugehen, dass aufgrund der theoretischen Vorarbeiten (vgl. Abschnitt 7.5.2) die intendierten Konstrukte durch die konstruierten Items größtenteils treffend abgebildet werden, d. h. durch eindimensionale Skalen repräsentiert werden.

6.2. Es ist davon auszugehen, dass aufgrund der theoretischen Vorarbeiten (vgl. Abschnitt 7.5.2) in der Regel die intendierten Konstrukte durch die konstruierten Items repräsentativ abgebildet werden. In Einzelfällen müssen ggf. aufgrund nicht beachteter oder überbetonter Facetten der jeweiligen Domäne seitens der Itemautoren Items ergänzt oder gestrichen werden.

FF7 Inwiefern bilden die konstruierten Skalen das jeweils intendierte Konstrukt zuverlässig ab? (vgl. Annahme 8, Unterabschnitt 7.5.2)

Hypothese:
7.1. Aufgrund der sorgfältigen, theoriegeleiteten Instrumententwicklung ist davon auszugehen, dass die konstruierten Skalen das jeweils intendierte Konstrukt zuverlässig abbilden.

FF8 Inwiefern sind die intendierten Konstrukte gegeneinander eindeutig abgrenzbar? (vgl. Annahme 9, Unterabschnitt 7.5.2)

Hypothesen:
8.1. Da die Literaturrecherche aufzeigt, dass Experten die Planung und Durchführung von Unterricht in der Regel als zwei Dimensionen der Kerntätigkeit des Lehrerberufs ansehen (vgl. Abschnitt 7.2.3), ist davon auszugehen, dass sich die Differenzierung der Handlungsfelder in die Dimensionen Planung und Durchführung von Physikunterricht als tragfähig erweist.

8.2. Nach Bandura (2006, S. 308) korrelieren Selbstwirksamkeitserwartungen ähnlicher Domänen positiv miteinander, insbesondere dann, wenn ähnliche Teilhandlungen oder Fähigkeiten in den Domänen eine Rolle spielen (vgl. Abschnitt 3.2.3). Deshalb wird unter der Prämisse, dass Planungs- und Durchführungsdimension voneinander zu trennende Konstrukte abbilden, angenommen, dass die Zusammenhänge zwischen den entwickelten Skalen im mittleren Bereich liegen (1). Zusätzlich wird angenommen, dass die Zusammenhänge am höchsten sind, wenn die Skalen ein und demselben Handlungsfeld angehören (2), weil dann der größte Überlapp ähnlicher Tätigkeiten zu erwarten ist. Darüber hinaus wird ebenfalls mit Bezug auf Bandura angenommen, dass die Planungsskalen der jeweiligen

Handlungsfelder untereinander stärker korrelieren als mit Durchführungsskalen anderer Handlungsfelder (3), weil Planungstätigkeiten unterschiedlicher Handlungsfelder einander ähnlicher sind als Planungs- und Durchführungshandlungen unterschiedlicher Handlungsfelder. Analoges gilt für die Korrelationen der Durchführungsskalen untereinander.

FF9 Inwiefern sind die intendierten Konstrukte zu anderen/ähnlichen Konstrukten wie allgemeinen Selbstwirksamkeitserwartungen, Lehrer-Selbstwirksamkeitserwartungen, dem physikalischen Selbstkonzept, internalen Kontrollüberzeugungen sowie allgemeinen Persönlichkeitsmerkmalen abgrenzbar? (vgl. Annahme 10, Unterabschnitt 7.5.2)

Hypothesen:
9.1. Den theoretischen Vorüberlegungen entsprechend (vgl. Abschnitte 3.2.3 und 3.6.1) ist davon auszugehen, dass globalere Messungen der Selbstwirksamkeitserwartungen (allgemeine Selbstwirksamkeitserwartung, Lehrer-Selbstwirksamkeitserwartung) eher gering mit den spezifischen, domänenbezogenen Lehrer-Selbstwirksamkeitserwartungen korrelieren.

9.2. Da Lehrer-Selbstwirksamkeitserwartungen auf einem spezifischeren und domänenähnlicheren Niveau erfasst werden als allgemeine Selbstwirksamkeitserwartungen, ist davon auszugehen, dass deren Korrelationen mit den neu entwickelten Skalen der Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern etwas höher ausfallen, jedoch trotzdem lediglich im geringen bis mittleren Bereich liegen, da kaum Gemeinsamkeiten in den jeweils operationalisierten Handlungsbezügen auszumachen sind.

9.3. Das physikalische Selbstkonzept sollte im Sinne einer diskriminanten Validierung kaum Zusammenhänge zu Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern aufweisen, da die Konstrukte jeweils unterschiedliche Domänen aufgreifen und das physikalische Selbstkonzept darüber hinaus durch affektive Komponenten erfasst wird. Etwas höhere Korrelationen des physikalischen Selbstkonzepts sind ggf. zu den Selbstwirksamkeitserwartungsskalen zu erwarten, deren Handlungsfelder einen größeren Bezug zur Fachwissenschaft aufweisen. Dies könnte im besonderen Maße für das Handlungsfeld Elementarisieren gelten, welches u. a. das Strukturieren und Vereinfachen physikalischer Inhalte für Schülerinnen und Schüler adressiert.

9.4. Die Beziehung von Selbstwirksamkeitserwartungen zu (internalen) Kontrollüberzeugungen wurde in Abschnitt 3.4.6 thematisiert, woraus sich ergibt, dass mittlere Korrelationen beider Konstrukte zu erwarten sind.

9.5. Theoretischen Erwartungen und bisherigen empirischen Ergebnissen entspre-

chend sollten allgemeine Persönlichkeitsmerkmale mit allgemeinen Selbstwirksamkeitserwartungen in einem höheren Maße korrelieren als mit Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern, wobei lediglich geringe Korrelationen zu erwarten sind (vgl. Unterabschnitte 3.4.1 und 4.5.1/Persönlichkeit).

FF10 Inwiefern ist die gruppenübergreifende Operationalisierung der intendierten Konstrukte gelungen? (vgl. Annahmen 5 & 11, Unterabschnitt 7.5.2)

Hypothese:
10.1. Da das Itemkonstruktionsteam das Erfahrungsspektrum der zu befragenden Kohorten widerspiegelt (eher erfahrene Lehrpersonen: Thorid Rabe und Olaf Krey; eher lehrunerfahren: Claudia Meinhardt), also bereits im Konstruktionsprozess potentiell unterschiedliche Lesarten antizipiert werden können, ist davon auszugehen, dass die Items zumeist in jeder der drei Befragungsgruppen ähnlich verstanden (vgl. Hypothese 3.1), d. h. auch als ähnlich schwierig wahrgenommen werden.

FF11 Inwiefern bilden die Skalen theoretische Erwartungen bez. der Entwicklung von Lehrer-Selbstwirksamkeitserwartungen oder Unterschiede aufgrund der unterrichteten Schulform oder des Geschlechts ab? (vgl. Annahme 12, Unterabschnitt 7.5.2)

Hypothesen:
11.1. Entsprechend der Ergebnisse der umfangreichen Literaturrecherche in Abschnitt 4.7/Geschlecht und der fehlenden theoretischen Begründbarkeit sind keine geschlechtsspezifischen Unterschiede zu erwarten.

11.2. Unterschiede in den Ausprägungen der Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern sind ggf. aufgrund der gewählten Schulform zu erwarten, weil in der Regel höhere Lehrer-Selbstwirksamkeitserwartungen für Lehrpersonen an Gymnasien dokumentiert werden konnten (vgl. Abschnitt 4.7/Schulform) und zunächst davon auszugehen ist, dass sich diese Tendenz auch für spezifischer erhobene Lehrer-Selbstwirksamkeitserwartungen zeigt.

11.3. Da es sich bei nachfolgend näher beschriebenen Validierungsstudien lediglich um Querschnitterhebungen handelt, können keine Aussagen über Entwicklungsverläufe von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern getätigt werden. Die Daten können jedoch als Quasilängsschnitt interpretiert werden, da Probanden unterschiedlicher Erfahrungshorizonte befragt werden. Aufgrund der Ergebnisse der Literaturrecherche (vgl. Abschnitt 4.6) ist damit zu rechnen, dass für Probanden mit mehr Praxiserfahrung (mehr Dienstjahre) höhere Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern vorliegen (1). Mit einem Praxisschock im Studium ist nicht zu rechnen (2), da Praxisphasen wie Schulpraktische Studien in der Regel gut betreut werden (vgl.

Unterabschnitt 4.5.1/Mentoreneffekt). Für den Übergang Studium-Referendariat ist eher mit einem Praxisschock zu rechnen (3), da im Referendariat die Betreuung zumeist weniger intensiv ausfällt.

8.2. Projektdesign

„[T]he mindless application of any method invites confusion."

Luppescu, 1993, S. 285

Im nächsten Abschnitt soll dargelegt werden, inwiefern die abgeleiteten Prüfmaßnahmen im Rahmen des Dissertationsprojektes realisiert werden. Dazu werden der grobe Ablauf des dreiteiligen Projektes sowie die (methodische) Ausrichtung der einzelnen Studien beschrieben. Die einzelnen methodischen Verfahren sollen an dieser Stelle jedoch nur benannt werden. Im Zuge der ausführlichen Darstellung der Studien in den sich anschließenden Kapiteln werden wichtige methodische Entscheidungen transparent gemacht und begründet.

8.2.1. Übersicht des Projektablaufs

Ausgehend von den bereits abgeleiteten Maßnahmen zur Prüfung der Gültigkeit der Annahmen, die der zu validierenden Testwertinterpretation zugrunde liegen (vgl. Abschnitt 7.5.3), werden mehrere Validierungsstudien geplant, die die Durchführung der Maßnahmen und damit die Prüfung der formulierten Hypothesen (vgl. vorangegangener Abschnitt 8.1.2) zum Ziel haben. Es ist davon auszugehen, dass die Qualität der Items/Skalen maßgeblich gesteigert werden kann, wenn mehrere Überarbeitungszyklen in den Validierungsprozess integriert sind. Die Umsetzung der Maßnahmen wird deshalb in mehrere Pilotstudien aufgeteilt, um eine iterative Überarbeitung der Skalen zu ermöglichen (vgl. Tabelle 8.1). Im Rahmen der letzten Studie, die die Hauptstudie des Dissertationsprojektes bildet, sollen ca. 1000 Probanden mit dem Instrument befragt und statistische Analysen durchgeführt werden. Insgesamt sind drei Überarbeitungsphasen geplant (vgl. Tabelle 8.1). Aufgrund der mehrfachen Itemüberarbeitungen resultiert eine relativ zeitintensive Anlage des Projektes, weil für jeden Projektabschnitt jeweils Planungs-, Erhebungs-, Auswertungs-, Interpretations-, Überarbeitungs- und Dokumentationsphasen durchlaufen werden müssen. Im nächsten Abschnitt werden die einzelnen Projektteile detaillierter erläutert.[1] Vertiefende Informationen zu den einzelnen Studien enthalten die nächsten Kapitel (vgl. Pilotstudie I: Kapitel 9, Pilotstudie II & III: Kapitel 10, Hauptstudie: Kapitel 12).

[1] Das grobe Projektdesign wird auch bei Rabe, Krey und Meinhardt (2013b) vorgestellt.

Tabelle 8.1.: Grober Ablauf des Dissertationsprojektes.

Jahr(e)	Projektteil, Maßnahmen, Veröffentlichungen
2010/2011*	Projektplanung, Pilotstudie I & Itemüberarbeitung Maßnahmen: 10, 11, 13, 15, 16, 18 vgl. Meinhardt (2011), Rabe, Meinhardt und Krey (2012, 2013a, 2013b)
2012	Pilotstudie II, III & Itemüberarbeitung Maßnahmen: 1, 2, 3, 4, 5, 6, 8, 9, 10, 12 vgl. Meinhardt, Rabe und Krey (2014a, 2014b)
2013/2014	Hauptstudie & Itemüberarbeitung Maßnahmen: 7, 10, 11, 13, 14, 15, 16, 17, 18 vgl. Meinhardt, Rabe und Krey (2015, 2016a, 2016b)
2015	Auswertung & Dokumentation (Skalenhandbuch) vgl. Meinhardt, Rabe und Krey (2016c)
2016/2017**	Dissertationsschrift

* Realisierung des Projektteils in Form einer Masterarbeit, vgl. Meinhardt (2011).
** Die Niederschrift wurde aufgrund der Absolvierung des Vorbereitungsdienstes unterbrochen.

8.2.2. Vorstellung einzelner Projektteile

Tabelle 8.2 fasst zusammen, inwiefern die benannten Studien im Rahmen des Projektes Beiträge zu den formulierten Forschungsfragen liefern können und damit zur Prüfung der Annahmen beitragen, die der intendierten Testwertinterpretation zugrunde liegen. Zusätzlich enthält die Tabelle Angaben zur konkreten methodischen Ausgestaltung der einzelnen Studien, die im Folgenden für die einzelnen Projektteile etwas näher erläutert werden.

Pilotstudie I Im Rahmen einer ersten eher quantitativ ausgerichteten Pilotstudie soll zunächst eruiert werden, ob es grundsätzlich möglich ist, auf dem gewählten Spezifitätsniveau Skalen zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern zu konstruieren. Dazu sollen die erarbeiteten Skalen in einer der in Frage kommenden Kohorten, bei der prinzipiell die meisten Schwierigkeiten aufgrund der Unerfahrenheit zu erwarten sind (Studierende), eingesetzt und anschließend einigen *statistischen* Analysen unterzogen werden, die zum Beispiel Indizien bezüglich der Eindimensionalität der Skalen liefern (z. B. konfirmatorische Faktorenanalysen). Es werden dabei ausschließlich Methoden der klassischen Testtheorie angewandt. Welche konkreten Maßnahmen durchgeführt werden, kann Tabelle 8.1 entnommen werden. Es ist sehr wahrscheinlich, dass durch den ersten Einsatz der Items bereits ersichtlich wird, welche Items überhaupt nicht funktionieren oder welche einer Umformulierung bedürfen, sodass eine erste Itemüberarbeitung sinnvoll erscheint. Diese erste Pilotstudie kann als Vorarbeit zum eigentlichen Dissertationsprojekt gewertet

Tabelle 8.2.: Studien im Rahmen des Validierungsprojektes, deren methodische Ausgestaltung sowie deren Beiträge zur Beantwortung der Forschungsfragen.

Studie	Methode	Bezug Forschungsfragen										
		1	2	3	4	5	6	7	8	9	10	11
Pilotstudie I	Fragebogenerhebung (S)											
	Konfirmatorische Faktorenanalyse						x	x	x			
	Korrelationsanalyse (KTT)								x	x		
	Varianzanalyse (KTT)											x
Pilotstudie II	Interviews (S, R, L)		x	x	x	x			x		x	
Pilotstudie III	Fragebogenerhebung (E)	x		x	x		x		x			
Hauptstudie	Fragebogenerhebung											
	Konfirmatorische Faktorenanalyse						x	x	x		x	
	Korrelationsanalyse (KTT)								x	x		
	Varianzanalyse (KTT)											x
	Raschanalyse					x		x	x		x	

x: Indizien ableitbar
S: Studierende (Physiklehramt), R: Referendare (Physiklehramt), L: Lehrpersonen (Physiklehramt), E: Experten (Psychologie, Physikdidaktik)

werden. Sie wurde im Rahmen der Masterarbeit der Autorin realisiert (vgl. Meinhardt, 2011).

Pilotstudie II & III Im Falle akzeptabler, ermutigender Ergebnisse der ersten Pilotstudie sollen weitere Pilotstudien geplant und durchgeführt werden, die sich dann vornehmlich den dargestellten Annahmen auf Itemebene widmen sowie den zugehörigen Maßnahmen (vgl. Tabelle 8.1). Letztere bestimmen dabei aufgrund ihrer Formulierung („Experten und Probanden fragen ...“) maßgeblich das methodische Vorgehen (Interviewstudie, Expertenbefragung). Die konzipierten Pilotstudien sind folglich eher *qualitativer* Natur. Eine Überarbeitung der Items im Anschluss an diese Studien verspricht ergiebig zu werden, da die Items im Rahmen der Gespräche und Befragungen intensiv analysiert werden, sodass inhaltliche Nuancen und Problemstellen sichtbar werden, die in der Itementwicklergruppe aufgegriffen und diskutiert werden können.

Hauptstudie Nach einer zweifachen Überarbeitung kann mit einer hinreichend großen Wahrscheinlichkeit davon ausgegangen werden, dass mit den Items tatsächlich Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern erfragt werden. Im Zuge dieser abschließenden größer angelegten Querschnittstudie in allen interessierenden Befragungsgruppen sollen vornehmlich *statistische* Analysen vorgenommen werden, um auch die Gültigkeit der verbleibenden Annahmen der intendierten Test-

wertinterpretation, die auf Skalenebene angesiedelt sind, mit Indizien untermauern zu können.

Für den Bereich der statistischen Auswertung ergeben sich mehrere methodische Möglichkeiten. Faktorenanalytische Verfahren können durchaus als Standardverfahren zur Analyse von Tests und Fragebögen bezeichnet werden. Lange Zeit dominierten dabei exploratorische Ansätze die Forschungslandschaft (vgl. auch Unterabschnitt 4.3.11). Insbesondere für die Validierung von Instrumenten zur Erfassung von Selbstwirksamkeitserwartungen wurde zunehmend die Anwendung konfirmatorischer Faktorenanalysen gefordert (vgl. u. a. Dellinger u. a., 2008, S. 763; Roberts & Henson, 2001, S. 22). Mittlerweile werden diese standardmäßig angewandt, weshalb auch in diesem Projektteil auf konfirmatorische Faktorenanalysen zurückgegriffen wird. Zusätzlich werden verstärkt auch Methoden der Item-Response-Theory zu Validierungszwecken genutzt (vgl. u. a. Ritter u. a., 2001; Boone u. a., 2011; Head, 2012; Pruski u. a., 2013), auf die deshalb in dieser Arbeit ebenfalls nicht verzichtet werden soll.

Durch die Anwendung einfacher Strukturgleichungsmodelle (konfirmatorische Faktorenanalysen) und des Raschmodells werden in dieser Arbeit zwei paradigmatisch grundsätzlich verschiedene methodische Ansätze aufgegriffen (klassische vs. probabilistische Testtheorie, vgl. Moosbrugger, 2012b, 2012a), die in der Fachdidaktik zwar etabliert sind, sich jedoch häufig auch unversöhnlich gegenüberstehen. Trotzdem stellen beide Ansätze geeignete Verfahren zur Verfügung, um jeweils eine Vielzahl der benannten Annahmen bzw. Hypothesen zu prüfen (z. B. konfirmatorische Mehrgruppenvergleiche bzw. DIF-Analyse zur Prüfung der Messinvarianz, vgl. Annahme 11 bzw. FF10). Insofern stellt diese Arbeit einerseits ein Angebot an die Forschungsgemeinschaft dar, weil beide Paradigmen bedient werden. Andererseits kann die Anwendung beider Verfahren als eine Art methodische Kreuzvalidierung verstanden werden, die im besten Fall zur Stärkung des Validitätsargumentes beiträgt. Die detaillierte Abwägung beider theoretischer Ansätze hinsichtlich der Angemessenheit der Anwendung im Rahmen der Auswertung konkreter Validierungsmaßnahmen steckt noch in den Kinderschuhen und kann in dieser Arbeit nicht geleistet werden.

Offensichtliche Vorteile der Anwendung des Raschmodells liegen beispielsweise in der zusätzlichen Möglichkeit der Analyse der Angemessenheit von Antwortformaten (vgl. Annahme 6 bzw. FF5) sowie darin, dass eine Intervallskalierung der erhobenen Daten nicht vorausgesetzt wird. Dafür ist die Eindimensionalität der Konstrukte eine notwendige Bedingung und diese kann damit streng genommen nicht überprüft werden, so wie es mittels konfirmatorischer Faktorenanalysen möglich ist (vgl. Annahme 7 bzw. FF6).

Es ist davon auszugehen, dass aus der beschriebenen methodischen Herangehensweise eine Reihe von Hinweisen zur Verbesserung der Items bzw. Skalen resultieren, die in einem weiteren Überarbeitungszyklus adressiert werden können.

Projektabschluss Die Endphase des Dissertationsprojektes ist durch umfangreiche Auswertungs- und Dokumentationsmaßnahmen gekennzeichnet, die unter anderem in einem Skalenhandbuch (vgl. Meinhardt u. a., 2016c) zusammengefasst wurden.

Teil III.

Empirie

9. Exkurs: Pilotstudie I

Die erste Pilotstudie soll an dieser Stelle nicht erneut ausführlich dargestellt werden, da dies bereits in einer eigenen Qualifikationsarbeit geschehen ist (vgl. Meinhardt, 2011) und wesentliche Ergebnisse auch anderweitig veröffentlicht wurden (vgl. Rabe u. a., 2012). Um jedoch für den Leser dieser Arbeit den Argumentationsstrang nicht zu unterbrechen, werden nachfolgend die Kernaussagen der Studie zusammenfassend dargestellt und erläutert.

Wie bereits in Unterabschnitt 8.2.2 zum Ausdruck gebracht wurde, besteht das grobe Ziel dieser ersten Studie darin, zu eruieren, ob eine Erfassung des Konstruktes auf dem gewählten Spezifitätsniveau grundsätzlich möglich und eine weitere Überarbeitung der Skalen sinnvoll ist (vgl. auch Meinhardt, 2011, S. 28 f.). Dazu sollen die entwickelten Skalen einer kleineren Stichprobe vorgelegt werden und anschließend statistisch (faktorenanalytisch) untersucht werden. Auch die Untersuchung von Zusammenhängen zu anderen Konstrukten und die Auswertung von Varianzanalysen interessierender Subgruppen sollen erste Rückschlüsse dazu erlauben, inwiefern die Testwerte tatsächlich als Ausprägungen von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern interpretiert werden können.

Zu beachten ist, dass zum Zeitpunkt der Durchführung dieser Pilotstudie natürlich entsprechend der Zielstellung nicht die hier in Abschnitt 8.1 dargelegten Forschungsfragen im Mittelpunkt standen. Eine Zuordnung zu diesen wird in Abschnitt 9.6 rückwirkend vorgenommen, auch um die Verknüpfung der Studien darzustellen. Die in dieser Studie letztlich im Vordergrund stehenden Gültigkeitsprüfungen von Annahmen auf Skalenebene (vgl. Unterabschnitt 7.5.2), die in der Hauptstudie erneut, aber umfassender in den Fokus rücken (vgl. Kapitel 12), können mit Blick auf die Ableitung eines übergeordneten Validitätsargumentes nur eine untergeordnete Rolle spielen, da sich nachfolgende Prüfungen auf überarbeitete Itemversionen beziehen.

9.1. Stichprobe

Insgesamt wurden 84 Studierende der Universität Potsdam befragt, die mindestens im dritten Semester immatrikuliert waren und erste physikdidaktische Lehrveranstaltungen belegten (Einführungsvorlesung Physikdidaktik). Ein Großteil der Befragten gab an, männlichen Geschlechts zu sein ($N = 54$, ca. 56%) und studierte das Lehramt für Gymnasien ($N = 76$, ca. 90%). Insgesamt hatten 40 Studierende bereits Praxiserfahrungen durch die Schulpraktischen Übungen gesammelt, während $N = 44$ Studierende angaben,

über keine Praxiserfahrungen zu verfügen. Detailliertere Angaben zur Stichprobe finden sich bei Meinhardt (2011, S. 35 f.).

9.2. Fragebogen

Der Fragebogen enthielt die bis dahin entwickelten sechs Skalen zur Erfassung der Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern (Version 1: SWE-SV-P, SWE-SV-D, SWE-EX-P, SWE-EX-D, SWE-EL-P, SWE-EL-D, vgl. auch Anhang G.1). Darüber hinaus wurden neben demografischen Aspekten die allgemeine Selbstwirksamkeitserwartung nach Schwarzer und Jerusalem (1999), die Lehrer-Selbstwirksamkeitserwartung nach Schmitz und Schwarzer (2000) sowie das physikalische Selbstkonzept in einer Adaption von Hoffmann, Häussler und Lehrke (1998) erhoben. Die Items wurden skalenweise im Fragebogen dargeboten. Der vollständige Fragebogen ist bei Meinhardt (2011, S. 82 ff.) abgedruckt.

9.3. Methodische Anmerkungen

Die methodische Herangehensweise wurde bereits detailliert erläutert (vgl. Meinhardt, 2011, S. 36-42). Für Hinweise bezüglich der verwendeten Kriterien zur Beurteilung der Ergebnisse sei auch auf Meinhardt u. a. (2016c, S. 7-11) bzw. auf Abschnitt 12.6 ab Seite 261 dieser Arbeit verwiesen.

9.4. Analyseverfahren und benutzte Analysesoftware

Folgende Software wurde zur Auswertung der erhobenen Daten verwendet:

- Reliabilitäts-, Korrelations- & Varianzanalyse: SPSS 19.0 (IBM Corp. Released, 2010),
- Konfirmatorische Faktorenanalyse: Amos 19.0 (Arbuckle, 2010).

9.5. Ergebnisse

Wie bereits mehrfach erwähnt, sei für Details zu den Ergebnissen und deren Interpretation auf Meinhardt (2011) verwiesen.

9.5.1. Reliabilitätsanalyse

Die in Tabelle 9.1 dargestellten Reliabilitätskoeffizienten Cronbach's Alpha α_C liegen für die neu entwickelten Skalen im mittleren Bereich ($.65 \leq \alpha_C \leq .79$). Die Reliabilitätskoeffizienten für die zusätzlich eingesetzten Skalen liegen im Bereich $.67 \leq \alpha_C \leq .83$.

9.5.2. Konfirmatorische Faktorenanalyse

Bezüglich der konfirmatorischen Faktorenanalyse können in vier von sechs Fällen die Modellanpassungsgüten für die spezifizierten Messmodelle als gut bezeichnet werden (SWE-SV-P, SWE-SV-D, SWE-EL-P, SWE-EL-D). In zwei Fällen genügen die Kennwerte nicht den Grenzwerten (SWE-EX-P, SWE-EX-D). Nach einer Revision der Skalen verbessern sich alle Modellfits und liegen auch für die Experimentierskalen im akzeptablen bis guten Bereich (vgl. Tabelle 9.1). Auch die Faktorladungen können grundsätzlich als akzeptabel bezeichnet werden (vgl. Meinhardt, 2011, S. 85).

Tabelle 9.1.: Modellfit der spezifizierten und revidierten Messmodelle sowie Skalenreliabilitäten (Cronbach's Alpha α_C), Pilotstudie I.

Modell	$\boldsymbol{\chi^2}$	**df**	$\frac{\chi^2}{df}$	**p**	**CFI**	**TLI**	**RMSEA**	$\boldsymbol{\alpha_C}$
PS spez.	36.836	14	2.631	.001	.882	.823	.140	.83
PS rev.	7.139	9	.739	.623	1.000	1.021	.000	.81
A-SWE spez.	24.682	35	.705	.903	1.000	1.098	.000	.79
L-SWE spez.	25.664	35	.733	.875	1.000	1.174	.000	.67
L-SWE rev.	22.618	27	.838	.705	1.000	1.078	.000	.69
SWE-EX-P spez.	3.622	14	2.187	.006	.826	.739	.120	.73
SWE-EX-P rev.	5.563	5	1.113	.351	.990	.980	.037	.70
SWE-EX-D spez.	21.932	14	1.567	.080	.881	.821	.083	.68
SWE-EX-D rev.	11.458	9	1.273	.246	.954	.923	.057	.65
SWE-EL-P spez.	18.397	14	1.314	.189	.945	.918	.062	.70
SWE-EL-P rev.	11.053	9	1.228	.272	.971	.952	.052	.71
SWE-EL-D spez.	11.768	14	.841	.625	1.000	1.027	.000	.79
SWE-SV-P spez.	17.245	14	1.232	.243	.958	.937	.053	.71
SWE-SV-P rev.	11.237	9	1.249	.260	.969	.949	.055	.72
SWE-SV-D spez.	16.963	14	1.212	.258	.975	.963	.051	.77

9.5.3. Korrelationsanalyse

Da keine Normalverteilung der Daten vorliegt, wird im Folgenden Spearman's ρ berichtet. Zunächst ist festzustellen, dass die Skala zum physikalischen Selbstkonzept (PS) weder mit den Skalen zur allgemeinen Selbstwirksamkeitserwartung (A-SWE) noch mit der Skala zur Lehrer-Selbstwirksamkeitserwartung (L-SWE) korreliert (vgl. Tabelle 9.2). Allgemeine und Lehrer-Selbstwirksamkeitserwartung korrelieren dagegen auf einem mittleren Niveau ($\rho = .65^{**}$).

Tabelle 9.2.: Korrelationskoeffizienten zwischen den Skalen.

Spearman's ρ	PS	L-SWE	A-SWE
PS	1		
L-SWE	−.020	1	
A-SWE	.072	.653**	1
SWE-EX-P	.216*	.187	.322**
SWE-EX-D	.185	.286**	.334**
SWE-EL-P	.342**	.282**	.381**
SWE-EL-D	.285**	.185	.316**
SWE-SV-P	.090	.154	.295**
SWE-SV-D	.120	.449**	.524**

* signifikant ($\alpha = 0.05$); ** hoch signifikant ($\alpha = 0.01$)

Die Korrelationen der zusätzlich erhobenen Skalen zu den neu entwickelten Skalen zeigen ein differenziertes Ergebnis. Die Skala des physikalischen Selbstkonzepts korreliert lediglich schwach mit den Skalen SWE-EL-P und SWE-EL-D sowie mit der Skala SWE-EX-P. Die Skala der allgemeinen Selbstwirksamkeitserwartung korreliert mit allen neu entwickelten Skalen schwach (Korrelation im mittleren Bereich mit SWE-SV-D), während die Skala zur Lehrer-Selbstwirksamkeitserwartung lediglich mit den Skalen SWE-EX-D und SWE-EL-P in einem schwachen Zusammenhang steht (ebenfalls Korrelation im mittleren Bereich mit SWE-SV-D). Die Korrelationen der allgemeinen Selbstwirksamkeitserwartungsskala zu den neu entwickelten Skalen sind dabei leicht höher als die Korrelationen der Lehrer-Selbstwirksamkeitserwartungsskala zu den entsprechenden Skalen.

Die neu entwickelten Skalen korrelieren im mittleren Bereich ($.311 \leq \rho \leq .599$) untereinander (vgl. Tabelle 9.3).

Tabelle 9.3.: Korrelationskoeffizienten zwischen den neu entwickelten Skalen.

Spearman's ρ	SWE-EX-P	SWE-EX-D	SWE-EL-P	SWE-EL-D	SWE-SV-P	SWE-SV-D
SWE-EX-P	1					
SWE-EX-D	.521**	1				
SWE-EL-P	.311**	.452**	1			
SWE-EL-D	.425**	.466**	.565**	1		
SWE-SV-P	.543**	.459**	.344**	.471**	1	
SWE-SV-D	.439**	.599**	.524**	.545**	.566**	1

* signifikant ($\alpha = 0.05$); ** hoch signifikant ($\alpha = 0.01$)

9.5.4. Varianzanalyse

Zum Vergleich der Mittelwerte der Skalen in verschiedenen Subgruppen wurden t-Tests bzw. nicht parametrische Verfahren zur Absicherung der Ergebnisse aufgrund der fehlenden Normalverteilung der Skalenwerte durchgeführt. Geschlechtsspezifische Unterschiede konnten für zwei Skalen (SWE-EL-P, SWE-EX-D: $p < .05$, $d > .47$) identifiziert werden.

Für drei von sechs Skalen (SWE-EX-D, SWE-SV-P, SWE-SV-D: $p \leq .006$, $d > .58$) ergibt sich ein signifikanter Mittelwertunterschied, wenn die Mittelwerte vor und nach Praxiserfahrungen (Schulpraktische Übungen) in Betracht gezogen werden.

Des weiteren ergeben sich signifikante Mittelwertunterschiede bezüglich der Planungs- und Durchführungsskalen für zwei Handlungsfelder (SWE-EX, SWE-EL: $p < .05$, $d > .45$), wobei die Mittelwerte hinsichtlich der Planung geringer ausfallen.

9.6. Interpretation der Ergebnisse und Validitätsargumente

Die Ergebnisse können mit Blick auf einige der gestellten Forschungsfragen (vgl. Abschnitt 8.1) interpretiert werden. Wie in Tabelle 8.2 ersichtlich, geben die Ergebnisse der konfirmatorischen Faktorenanalyse insbesondere bezüglich FF6 (H6.1: Eindimensionalität) Auskunft, während die Korrelationsanalysen hauptsächlich Indizien zur Beantwortung von FF8 (H8.2: Korrelationen der Skalen untereinander) und FF9 (H9.1, H9.2, H9.3: Korrelationen zu Skalen anderer Konstrukte) liefern. Schließlich können die Ergebnisse der Varianzanalyse Beiträge zur Beantwortung von FF11 (H11.1: geschlechtsspezifische Unterschiede, H11.3 (3): Unterschiede aufgrund von Praxiserfahrung im Studium) beisteuern sowie die Ergebnisse der ebenfalls durchgeführten Reliabilitätsanalyse hinsichtlich FF7 (H7.1: Reliabilität) interpretiert werden. Es ist bei dieser rückwirkend vorgenommenen Zuordnung der Analyseergebnisse zu den Forschungsfragen zu beachten, dass diese letztlich lediglich im Hinblick auf die dieser Pilotstudie zugrunde liegende Frage nach der grundsätzlichen Möglichkeit der Skalenkonstruktion hilfreich ist, jedoch nur eingeschränkt für das abzuleitende Validitätsargument nutzbar ist.

9.6.1. Reliabilitätsanalyse

Auch wenn die Reliabilitätskoeffizienten für eine erste Erhebung akzeptabel sind (vgl. Unterabschnitt 12.6), so ist die Frage der Zuverlässigkeit der intendierten Testwertinterpretation damit natürlich nicht zufriedenstellend beantwortet (vgl. Abschnitt 8.1.2, FF7). Das Ergebnisse kann jedoch vorsichtig dahingehend interpretiert werden, dass die in den Skalen zusammengefassten Items (inhaltlich) scheinbar ausreichend gut zusammenpassen.

Problematisch ist, dass auch für scheinbar etablierte Skalen (z. B. L-SWE-Skala), an die strengere Maßstäbe hinsichtlich der Reliabilität angelegt werden sollten, trotz

Revision keine zuverlässige Testwertinterpretation möglich ist (spez. Skala: $\alpha_C = .67$ bzw. rev. Skala $\alpha_C = .69$).

9.6.2. Konfirmatorische Faktorenanalyse

Aufgrund der guten Modellanpassungsgüten können die Skalen für die befragte Studierendenkohorte als eindimensional eingestuft werden (vgl. Abschnitt 8.1.2, FF6). Ob die Skalen damit tatsächlich die *intendierten* Konstrukte eindimensional abbilden, muss jedoch erneut offen bleiben. Einschränkend zu berücksichtigen ist weiterhin die relativ geringe Stichprobengröße ($N = 84$).

9.6.3. Korrelationsanalyse

Die dargestellten Korrelationen der Skala des physikalischen Selbstkonzepts zu den neu entwickelten Skalen entsprechen der Hypothese 9.3 (vgl. Abschnitt 8.1.2, FF9). Auch die geringen bis mittleren Korrelationen der neu entwickelten Skalen zu den Skalen globaler operationalisierter Konstrukte (A-SWE, L-SWE) und die höheren Korrelationen dieser globaleren Konstrukte untereinander entsprechen den Erwartungen (vgl. Hypothese 9.1, Abschnitt 8.1.2, FF9).

Dass jedoch die neu entwickelten Skalen im Mittel höhere Korrelationen zur Skala der allgemeinen Selbstwirksamkeitserwartung als zur Skala der Lehrer-Selbstwirksamkeitserwartung aufweisen, steht im Widerspruch zu den Erwartungen (vgl. Hypothese 9.2, Abschnitt 8.1.2, FF9). Gegebenenfalls ist das Konstrukt der Lehrer-Selbstwirksamkeitserwartung, obwohl es spezifischer operationalisiert ist als das Konstrukt der allgemeinen Selbstwirksamkeitserwartungen, tatsächlich genauso wenig (oder sogar weniger) ausschlaggebend für die Einschätzungen, die bezüglich der neu entwickelten Items eine Rolle spielen.

Dass die neu entwickelten Skalen jeweils untereinander in einem mittleren Maß korrelieren, entspricht den Erwartungen (vgl. Hypothese 8.2 (1), Abschnitt 8.1.2, FF8). Insgesamt können deren mittlere Korrelationen untereinander sowie die geringen/mittleren Korrelationen zu Skalen anderer Domänen (PS) oder zu Skalen auf anderen Spezifitätsniveaus (A-SWE, L-SWE) als Indiz dahingehend gedeutet werden, dass die neu entwickelten Skalen tatsächlich jeweils eigenständige Konstrukte abbilden. Auch wenn die Aspekte (2) und (3) der Hypothese 8.2 nicht stringent in den Daten zu finden sind, so spricht dieses Ergebnis *nicht* gegen die Annahme der Trennbarkeit der intendierten Konstrukte, es erschwert lediglich die inhaltliche Interpretation. Das vorhergesagte Muster der Korrelationen (Hypothese 8.2) konnte jedoch für den Pearson-Korrelationskoeffizienten beobachtet werden (vgl. Meinhardt, 2011, S. 58 ff.).

9.6.4. Varianzanalyse

Dass in der Regel keine geschlechtsspezifischen Unterschiede hinsichtlich des Antwortverhaltens auftreten, stimmt mit Hypothese 11.1 (vgl. Abschnitt 8.1.2, FF11) überein. Dass für zwei Skalen Unterschiede aufgrund des Geschlechts dokumentiert werden müssen, kann nicht schlüssig erklärt werden. Unter Umständen ist die geringe und ungleichmäßige Verteilung der Gruppen ein Faktor, der berücksichtigt werden muss.

Die Mittelwertunterschiede bezüglich der Praxiserfahrung müssen ebenfalls vorsichtig interpretiert werden, da hier keine längsschnittlichen Daten miteinander verglichen werden können. Dass für drei Skalen Mittelwertunterschiede auftreten, kann ebenfalls nicht schlüssig erklärt werden.

Interessant ist, dass Mittelwertunterschiede bezüglich der Planungs- und Durchführungsskalen eines Handlungsfeldes existieren, auch wenn dieses nicht mit Blick auf die eigentlich interessierende Forschungsfrage interpretiert werden kann, sondern eher aus einer hypothesengenerierenden Perspektive relevant ist. Dass die Einschätzungen bezüglich der Durchführung höher ausfallen, erschient zunächst erklärungsbedürftig, da davon ausgegangen werden kann, dass Studierende mehr Erfahrungen im Planen als im Durchführen von Physikunterricht besitzen, sodass ein entgegengesetzter Trend erwartbar wäre (vgl. Abschnitt 3.3). Unter Umständen sind diese Planungserfahrungen aufgrund der Wahrnehmung eigener Defizite zunächst jedoch eher negativer Art. Gleichzeitig ist es denkbar, dass eine hohe Anzahl stellvertretender Erfahrungen (gesamte eigene Schulzeit) dazu führen, sich selbst die Durchführung von (ähnlichem) Physikunterricht eher zuzutrauen. Hinzu kommt, dass eigene Unterrichtserfahrungen im Rahmen der schulpraktischen Studien aufgrund des guten Betreuungsverhältnisses eher selten negativ ausfallen.

9.7. Itemrevision

Im Zuge der erste Pilotierung wurden erste Überarbeitungsempfehlungen gegeben (vgl. Meinhardt, 2011, S. 86 f.), die anschließend im Team diskutiert wurden (vgl. Abschnitt 7.3). Entsprechend wurden erste Veränderungen an den Items vorgenommen. Dabei wurden Items jeweils dann aufgrund der Analyseergebnisse entfernt, wenn zusätzlich triftige inhaltliche Gründe angegeben werden konnten. Die überarbeiteten Itemversionen können dem Anhang G.2 entnommen werden.

9.8. Zusammenfassung

Die erste Pilotierung der entwickelten Skalen sollte insbesondere darüber Auskunft geben, ob es sich langfristig lohnt, das interessierende Konstrukt der Lehrer-Selbstwirksamkeitserwartung in physikdidaktischen Handlungsfeldern durch Items/Skalen

zu konzeptualisieren bzw. die Operationalisierungen des Konstruktes zu überarbeiten und zu optimieren (vgl. Abschnitt 8.2). Hinsichtlich dieses Ziels kann mit Blick auf die intendierte Testwertinterpretation gesagt werden, dass die zu Skalen zusammengefassten Items scheinbar eindimensionale sowie voneinander abgrenzbare Konstrukte relativ zuverlässig abbilden, was aus den Ergebnissen der konfirmatorischen Faktorenanalyse, der Korrelations- und der Reliabilitätsanalyse gefolgert werden kann. Ob mit den Skalen tatsächlich Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern abgebildet werden, muss zunächst offen bleiben und mit anderen Methoden weiter untersucht werden. Höchstens *Indizien* dafür liefern die Korrelationsanalysen, die zumindest hinsichtlich des Spezifitätsniveaus erwartbare Resultate liefern sowie die Varianzanalyse, die keine systematischen geschlechtsspezifischen Unterschiede hervorbrachte, was für Lehrer-Selbstwirksamkeitserwartungen typisch wäre (vgl. Abschnitt 4.7/Geschlecht bzw. FF11). Es ist zu beachten, dass nicht alle Ergebnisse schlüssig interpretiert werden können, was jedoch aus der geringen Stichprobengröße bzw. ungleich besetzten Gruppen resultieren könnte.

10. Pilotstudien II und III (qualitativer Natur)

Im Rahmen der ersten Pilotstudie wurde grob geprüft, ob sich überhaupt (eindimensionale) Skalen auf dem anvisierten Spezifitätsniveau entwickeln lassen (vgl. Kapitel 9). Daher wurden schwerpunktmäßig solche Annahmen der Testwertinterpretation auf ihre Gültigkeit geprüft, die sich auf der Skalenebene ansiedeln lassen (FF6, FF8, F10). In den beiden nachfolgend näher beschriebenen Studien sollen dagegen hauptsächlich Annahmen auf Itemebene auf ihre Gültigkeit geprüft werden (FF1-FF5), sodass insgesamt eher die inhaltliche Frage im Fokus steht, ob durch die Skalen tatsächlich die intendierten Konstrukte abgebildet werden.

Die Ergebnisse beider Pilotstudien wurden bereits veröffentlicht und auf Tagungen vorgestellt (vgl. Meinhardt, Rabe & Krey, 2014a, 2014b). Nachfolgend werden die beiden Studien jeweils in ihrer Anlage kurz vorgestellt sowie anschließend für beide Studien zusammenfassend die Ergebnisse etwas detaillierter beschrieben und bezüglich der Forschungsfragen interpretiert.

10.1. Interviewstudie

Die Planung und Durchführung der Interviewstudie leitet sich aus den dargestellten Prüfmaßnahmen ab, die in Unterabschnitt 7.5.3 aufgelistet wurden. Wie bereits in Tabelle 8.2 auf Seite 211 dargestellt, können durch die Interviewstudie hauptsächlich die FF1 (Entsprechung Arbeitsdefinition), FF2 (kognitive Aktivierung), FF3 (Verständlichkeit), FF4 (Authentizität) und FF5 (Ratingskala) sowie in Ansätzen FF8 (Abgrenzbarkeit der Konstrukte untereinander) und FF10 (gruppenübergreifend ähnliches Verständnis) adressiert werden. In den nächsten Unterabschnitten werden u. a. die befragte Stichprobe beschrieben sowie die konkrete methodische Ausgestaltung der Studie begründet dargestellt.

10.1.1. Verfahren der Gewinnung von Interviewpartnern

Es sollten möglichst Personen aus jeder der drei relevanten Befragungsgruppen (Studierende, Referendare, Lehrpersonen) interviewt werden. Dazu wurden Studierende in Lehrveranstaltungen explizit über das Interviewanliegen informiert sowie Referendare angeschrieben, die vormals Studierende am Lehrstuhl waren. Außerdem wurden Lehrpersonen aus dem Umfeld der Mitarbeiterinnen und Mitarbeiter am Lehrstuhl direkt mit der Bitte um ein Interview angesprochen. Allen potentiellen Teilnehmern wurden

Kinogutscheine in Aussicht gestellt.

10.1.2. Stichprobe

Insgesamt wurden 20 Personen interviewt, wobei die Teilnehmerinnen und Teilnehmer allen potentiellen Befragungsgruppen angehörten und über unterschiedlich viel Praxiserfahrung verfügten. So wurden sowohl Studierende des Physiklehramts ohne Praxiserfahrung befragt ($N = 4$), als auch Studierende mit wenig (Schulpraktische Übungen, $N = 4$) und viel Praxiserfahrung (Praxissemester, $N = 4$). Darüber hinaus wurden Lehramtsanwärterinnen und -anwärter des Physiklehramts ($N = 4$) sowie Physiklehrpersonen ($N = 4$) für Interviews gewonnen.

10.1.3. Methodische Anmerkungen

Aus den gegebenen Antworten können in Interviews beispielsweise die Einstellungen zu Items aber auch das Itemverständnis von Interviewpartnern abgeleitet werden. Es ist dabei zu beachten, dass in Interviews Selbstauskünfte der Probanden gewonnen werden, die in jedem Fall interpretationsbedürftig sind (vgl. Niebert & Gropengießer, 2014, S. 121). Um die Interpretation nachvollziehbar zu machen, ist die lückenlose und ausführliche Dokumentation der Datenerhebung, -aufbereitung und -auswertung unabdingbar (vgl. Niebert & Gropengießer, 2014, S. 123). Es soll dargestellt werden, wie die konkret interessierenden Fragen, z. B. nach auftretenden kognitiven Prozessen, dem Itemverständnis oder der Einschätzung der Itemauthentizität (FF2-FF4), in Interviews adressiert und welche methodischen Aspekte aus welchen Gründen berücksichtigt wurden.

Eine Möglichkeit, die ablaufenden kognitiven Prozesse von Probanden zum Beispiel in einer Interviewsituation in Erfahrung zu bringen, besteht darin, die Methode des lauten Denkens in das Interview zu integrieren. Der Gesprächspartner wird dabei gebeten, alle Gedanken laut zu äußern, die ihr bei der Bearbeitung eines bestimmten Problems oder beim Durchführen einer bestimmten Handlung „in den Kopf kommen“ (vgl. u. a. Sandmann, 2014, S. 179). Interviewer sollten sich in solchen Phasen zurückhalten, um Probanden nicht abzulenken oder zu beeinflussen. Die zu führenden Interviews sollen daher möglichst lautes Denken zulassen, sowie dem Grundsatz folgen, möglichst selten in den Redefluss der Interviewten einzugreifen.

Um das laute Denken anzuregen, wurden die Interviewpartner gebeten, ein Item zunächst laut vorzulesen, anschließend zusammenzufassen bzw. zu paraphrasieren und sich dann bezüglich des Items begründet selbst einzuschätzen (vgl. Anhang H.1). Zusätzlich sollten die Interviewten alle Probleme oder Schwierigkeiten laut benennen. Es wird dabei davon ausgegangen, dass aus laut vorgenommenen Paraphrasierungen der Items sowie aus laut vorgenommenen, begründeten Selbsteinschätzungen auf das Itemverständnis

sowie auf ablaufende kognitive Prozesse der Interviewten geschlussfolgert werden kann, sodass diese mit dem intendierten Verständnis sowie intendierten kognitiven Prozessen abgeglichen werden können. Letztlich soll durch die Aufforderung zum lauten Denken der ansonsten still ablaufende gedankliche Prozess der Itembeantwortung verbalisiert werden (vgl. Unterabschnitt 2.2.1/Modell der Itembeantwortung), sodass kognitive Prozesse während der Handlung *Bearbeiten eines Fragebogens* sichtbar werden. Den Interviewten soll der dargestellte Vierschritt (lautes Vorlesen, Paraphrasierung/Zusammenfassung, Selbsteinschätzung, Begründung) zu Beginn des Interviews erläutert werden (vgl. Anhang H.2.1) und dann als Gedächtnisstütze in Form eines Merkzettels mit dem Ziel vorgelegt werden, ggf. nötige Impulse durch die Interviewerin zu reduzieren, sodass das laute Denken möglichst selten für die Items einer Skala unterbrochen wird (vgl. Anhang H.1).

Es ist jedoch davon auszugehen, dass Interviewpartner Phasen des lauten Denkens weniger gewöhnt sind, sodass trotz der ergriffenen Maßnahmen Impulse von Nöten sind, die zum Weitersprechen anregen und daher eher offen angelegt sein sollten (vgl. Niebert & Gropengießer, 2014, S. 128 ff.). Solche Impulse, die prinzipiell geeignet erschienen, die genannte Phase zu moderieren bzw. in Gang zu halten, wurden in einem Leitfaden zusammengefasst (vgl. Anhang H.2.2).

An die Phasen des lauten Denkens schließt sich jeweils eine stärker strukturierte Phase an, in der verbleibende interessierende Aspekte konkret angesprochen werden sollen, weil davon ausgegangen werden kann, dass auf diese weniger stringent anhand der getätigten Äußerungen rückgeschlossen werden kann. Dazu zählen Fragen nach besonders schwierigen Items (vergleichende Frage), der Itemauthentizität bzw. Itemrelevanz sowie der Eignung des verwendeten Antwortformates. Die abschließenden Fragen sind ebenfalls im Leitfaden angeführt (vgl. Anhang H.2.3).

Die Interviews sind aufgrund der Zielstellung klar strukturiert, wobei einige Phasen offener angelegt sind (Phase des lauten Denkens) als andere. Der Interviewleitfaden (vgl. Anhang H.2) dient dazu, zu Beginn keine der Instruktionen und zum Ende keine der Fragen zu vergessen und nötige Impulse bereit zu halten, um das laute Denken zu animieren.

Zu Trainingszwecken und der empfohlenen Erprobung und Überarbeitung des Interviewleitfadens (vgl. Niebert & Gropengießer, 2014, S. 130) wurde ein Probeinterview mit einer studentischen Hilfskraft durchgeführt. Zur Dokumentation wurden die Interviews zweifach aufgezeichnet (Video und Ton) und Notizen (vgl. exemplarisch Anhang H.3) angefertigt. Weitere Maßnahmen zur Gewährung der Transparenz vorgenommener Ableitungen und damit Interpretationen werden in Unterabschnitt 10.1.5 erläutert.

10.1.4. Datenerhebung – Interviewdurchführung

Die Interviews wurden durch die Autorin dieser Arbeit selbst durchgeführt und fanden von November 2012 bis März 2013 statt. Ein Interview dauerte im Schnitt 60 Minuten. Jedem Studierenden wurden die Items zweier Handlungsfelder, d. h. insgesamt vier Skalen vorgelegt, während mit den Lehrpersonen und Referendaren jeweils die Items eines Handlungsfeldes diskutiert wurden. Somit sind die Items eines jeden Handlungsfeldes Gegenstand von jeweils acht Interviews.

Vor dem Interview wurde im Büro der Interviewerin ein Tisch vorbereitet, auf dem das Audioaufnahmegerät sowie benötigte Materialien platziert wurden. Zu diesen Materialien gehörten für die Interviewerin die zu besprechenden Items (inkl. Notizspalten, vgl. Anhang H.3) und der Leitfaden sowie für die Interviewten die Items und ein Erinnerungskärtchen mit dem beschriebenen Vierschritt (vgl. exemplarisch Anhang H.1). Nach der Begrüßung des Interviewpartners sowie einer kurzen „Smalltalkrunde“, in der die Aufnahmegeräte präpariert wurden und die mit Blick auf die Herstellung einer offenen und gelösten Gesprächssituation nützlich erschien, begann in der Regel relativ zügig das Interview. Dieses gliedert sich, wie beschrieben, in drei Phasen, wobei die letzten beiden Phasen von den Studierenden insgesamt vier mal durchlaufen wurden:

Beginn Begrüßung und Schaffung einer angenehmen Gesprächsatmosphäre, Getränkeangebot

Phase 1: Einstieg (Impulse vgl. Anhang H.2.1)

- Zusicherung von Anonymität
- Erläuterungen des Ablaufs

Phase 2: lautes Denken (Impulse vgl. Anhang H.2.2)

- skalenweises Bearbeiten der Items durch Probanden
- lautes Vorlesen der Items durch Probanden
- lautes Denken der Probanden (Zusammenfassen der Items, ggf. Schwierigkeiten benennen, Selbsteinschätzung vornehmen, Begründungen angeben)
- ggf. Impulse oder Nachfragen durch Interviewerin

Phase 3: Ergänzungsfragen (Impulse vgl. Anhang H.2.3)

- Aufforderung zur Einschätzung bez. der gesamten Skala
- Identifizieren unauthentischer und sehr schwerer/leichter Items
- Frage zur Passung der Ratingskala
- Möglichkeit für beide Interviewpartner, Nachfragen zu stellen oder Anmerkungen zu machen

Ende Dank und Verabschiedung

Während der Interviews wurden Notizen bezüglich der interessierenden Fragen wie der Verständlichkeit, der Zuordnung der Items zu einer Dimension und der Authentizität/Relevanz der Items angefertigt (vgl. Anhang H.3).

10.1.5. Datenaufbereitung und Auswerteverfahren

Die Interviewdaten wurden nicht vollständig transkribiert. Unter Zuhilfenahme der während des Interviews angefertigten Notizen sowie der Audiomitschnitte wurden für jedes Item und jeden Befragten die folgenden Kriterien analysiert und ggf. durch Transkripte oder Kommentare dokumentiert:

1. Verständnis insgesamt, d.h. Item als Einheit von Handlung/Hürde entsprechend intendierter Domäne (Handlungsfeld/Dimension) verstanden?
    a) Handlung wie intendiert umschrieben?
    b) Barriere überhaupt relevant/wahrgenommen?
    c) Assoziationen entsprechend des intendierten Handlungsfeldes?
    d) Assoziationen entsprechend der intendierten Dimension (Planung/Durchführung)?
2. Kognitive Prozesse wie intendiert aktiviert?
    a) Selbstreferentialität: Selbsteinschätzungen grundsätzlich möglich und vorgenommen?
    b) Handlungsbezug: Handlungen als komplex wahrgenommen?
    c) Schwierigkeitsbezug: Hürden als schwierigkeitserzeugend wahrgenommen?
    d) Domänenspezifität: Einheit von Handlung/Hürde domänenkonform verstanden; domänenrelevante Erfahrungen als Grundlage der Selbsteinschätzung assoziiert?
3. Antwortformat geeignet?
4. Items authentisch/relevant?
5. leichte/schwierige Items?

Zusätzlich wird notiert, ob die Teilnehmenden Beantwortungsschwierigkeiten äußern – beispielsweise aufgrund fehlender Erfahrung oder weil sie davon ausgehen, dass die Beantwortung vom inhaltlichen Thema abhängig sei (Problematik des Spezifitätsniveaus). Außerdem werden Hinweise und diskussionswürdige Bemerkungen der

Probanden notiert (z. B. hinsichtlich der Itemschwierigkeit, schwieriger Begrifflichkeiten, missverständlicher Formulierungen etc.), die aus Sicht des Interviewerin bezüglich der Itemüberarbeitung im Team besprochen werden sollten. Für jedes Item ergibt sich somit ein Überblick bezüglich der genannten Kriterien für acht unterschiedliche Personen, sodass ggf. (kohortenübergreifende oder kohortenspezifische) Häufungen oder Singularitäten in Verständnisschwierigkeiten etc. sichtbar werden (vgl. exemplarisch Anhang H.4). Die gewonnenen Einsichten werden unter Berücksichtigung der Ergebnisse der Expertenbefragung zur Überarbeitung der Items herangezogen (vgl. Abschnitt 10.5).

10.2. Expertenbefragung

Auch die Planung und Durchführung einer Expertenbefragung ergibt sich aus den in Unterabschnitt 7.5.3 dargestellten Prüfmaßnahmen. Entsprechend ihrer Adressierung bestimmter auf ihre Gültigkeit zu prüfender Annahmen, resultieren einige Forschungsfragen, die im Zuge der Expertenbefragung zumindest in Teilen beantwortet werden können. Dazu zählen, wie bereits in Tabelle 8.2 dargestellt, im Wesentlichen die FF1 (Entsprechung Arbeitsdefinition), FF3 (Verständlichkeit), FF4 (Authentizität) sowie FF6 (umfassende und treffende Abbildung der Domäne) und FF8 (Abgrenzbarkeit der Konstrukte untereinander). In den nächsten Unterabschnitten werden u. a. die befragte Expertengruppe vorgestellt und der Aufbau des Expertenfragebogens erläutert.

10.2.1. Verfahren der Gewinnung von Experten

Die Expertenbefragung wurde auf dem GDCP-Doktorierendenkolloquium 2012 in Leuenberg/Schweiz angekündigt und von den anwesenden Betreuenden als sinnvoll erachtet. Insgesamt wurden im Dezember 2012 25 Physikdidaktikerinnen und Physikdidaktiker sowohl postalisch als auch per E-Mail angeschrieben und der Expertenfragebogen zugesandt. Darüber hinaus wurden vier intensiv zu Selbstwirksamkeitserwartungen forschende Professorinnen und Professoren der Psychologie gebeten, an der Expertenbefragung teilzunehmen.

10.2.2. Methodische Anmerkungen – Aufbau des Fragebogens

Es wurde ein Fragebogen entwickelt (vgl. Anhang I.1), der die genannten Forschungsfragen direkt aufgreift. Wie bereits erwähnt, wurde der Fragebogen den Befragten sowohl als digitale Version als auch in Papierform zur Verfügung gestellt. Er enthielt neben einer kurzen Ankündigung des Ablaufs, eine knappe Darlegung der inhaltlichen Grundlagen bezüglich des Konstruktes und der Itemkonstruktion. Dann wurden die Experten aufgefordert, die Items von insgesamt vier Skalen bezüglich ihrer Passung (Relevanz) zum Handlungsfeld, ihrer Schwierigkeit und der Authentizität (Schwerpunkt

Handlungsbarriere) zu bewerten. Für jedes Item konnten in einem Freitextfeld zusätzliche Kommentare zum Beispiel zur Verständlichkeit des Items hinterlassen werden. Außerdem sollten die Experten zu jedem Handlungsfeld angeben, ob die Konstrukte umfassend abgebildet sind und ob die Differenzierung des Handlungsfeldes in Planung und Durchführung gelungen ist. Wiederum konnten die Experten zu diesen Fragen Bemerkungen und Ergänzungen vornehmen. Im abschließenden Teil des Fragebogens wurden die Experten gebeten, die Auswahl der Handlungsfelder zu kommentieren und ggf. weitere Handlungsfelder anzugeben, bezüglich derer sie Selbstwirksamkeitserwartungsskalen für relevant erachten. Zudem wurde abgefragt, ob die Experten die Unterscheidung der Dimensionen Planung und Durchführung grundsätzlich für sinnvoll erachten oder nicht. Kommentare und Begründungen konnten angegeben werden. Zum Schluss wurden die Experten aufgefordert, zusätzliche Hinweise, Anmerkungen oder Ideen zu verbalisieren.

Zur Befragung der Psychologen wurde ein zusätzlicher Fragebogen entwickelt (vgl. Anhang I.2), der eher die methodisch-fachliche Seite der Instrumententwicklung beleuchten sollte und insbesondere zur Beantwortung von FF1 konzipiert war. Folglich wurden die Experten im Fragebogen aufgefordert, die Angemessenheit und die Umsetzung der abgeleiteten Konstruktionsvorschriften zu bewerten. Darüber hinaus sollten sie Einschätzungen bezüglich der Eignung des für die Operationalisierung gewählten Spezifitätsniveaus, der Plausibilität der Differenzierung Planung/Durchführung sowie bezüglich der Eignung des Antwortformates vornehmen. Auch Fragen zum konkreten Aufbau des Fragebogens, beispielsweise bezüglich der Anordnung der Items (durchmischte vs. skalenweise Anordnung) sowie Fragen zum Itemformat (einheitliche vs. variierende Itemstruktur) und zur Erhebung zusätzlicher Konstrukte wurden den Experten gestellt.

10.2.3. Stichprobe

Insgesamt haben 18 Experten den Fragebogen ausgefüllt und zurückgesandt. Unter den Experten waren drei Mitarbeiterinnen und Mitarbeiter eines Lehrstuhls für Physikdidaktik und ein Professor für Psychologie. Die restlichen Experten hatten eine (Junior-)Professur der Physikdidaktik inne.

10.2.4. Datenaufbereitung und Auswerteverfahren

Die Expertenfragebögen wurden bezüglich der Item- und Skalenebene ausgewertet. Im Schnitt beurteilten acht Experten die Items eines Handlungsfeldes. Für jedes Item wurden separat die Experteneinschätzungen hinsichtlich der Passung des Items zum Handlungsfeld, der Passung von Itemstamm zur Handlungsbarriere und der Itemschwierigkeit notiert. Zusätzlich wurden die Expertenkommentare zum Item notiert. Es ergibt sich eine detaillierte Auswertungsübersicht (vgl. exemplarisch Anhang I.3). In einem

zweiten Schritt wurden alle Einschätzungen und Kommentare bezüglich der angemessenen und vollständigen Operationalisierung des Handlungsfeldes sowie der Trennung von Planungs- und Durchführungsdimension tabellarisch systematisiert (vgl. exemplarisch Anhang I.4).

10.3. Exkurs: Akzeptanzbefragung bez. des Antwortformates

Um insbesondere Forschungsfrage 5 aufzugreifen, haben Olaf Krey und Thorid Rabe einen Fragebogen entwickelt (vgl. Anhang J), um mit dessen Hilfe und anhand einer Auswahl der entwickelten Items das verwendete Antwortformat dahingehend zu untersuchen, ob es von potentiellen Befragten als geeignet zur Beantwortung der Items eingestuft wird bzw. welches Antwortformat überhaupt als geeignet bewertet wird. An dieser Befragung nahmen insgesamt 33 Studierende des Physiklehramts teil, wobei 23 Studierende im ersten oder dritten Semester und 10 Studierende im fünften Semester immatrikuliert waren.

Für eine begrenzte Itemauswahl wurden den Studienteilnehmern jeweils vier Antwortskalen unterschiedlicher Stufigkeit (4-, 6-, 8- und 10-stufig) vorgelegt. Die Studierenden sollten dann jeweils die Beantwortung auf der Antwortskala vornehmen, die sie für die Selbsteinschätzung am geeignetsten hielten. Zusätzlich konnten die Befragten angeben, wie viele Stufen ihrer Meinung nach prinzipiell anvisiert werden sollten, um treffende Einschätzungen zu gewährleisten (vgl. Fragebogen im Anhang J).

Etwas mehr als die Hälfte aller Studierenden ($51,5\%$) gab an, für eine treffende Selbsteinschätzung idealerweise sechs Abstufungen zu benötigen. Eine vierstufige Antwortskala präferierten immerhin $21,2\%$ der Befragten. Eine vier- oder sechsstufige Skalierung war für jedes der dargebotenen Items die am häufigsten gewählte Beantwortungsmöglichkeit, wobei die Sechsstufigkeit jeweils präferiert wurde.

Interessant ist, dass 70% der Studierenden des fünften Semesters die Sechsstufigkeit bevorzugten, während in der Gruppe der Studierenden niedriger Semester etwas weniger als die Hälfte der Befragten (10 von 23) die Sechsstufigkeit präferierten. Eventuell deutet sich hier die Tendenz an, sich mit zunehmender Erfahrung bezüglich der in den Items beschriebenen Handlungen auch differenzierter einschätzen zu können.

10.4. Ergebnisse der Pilotstudien II und III

Im Folgenden sollen die Ergebnisse beider Studien dargestellt werden (vgl. zusammenfassend auch Meinhardt u. a., 2014a, 2014b).

10.4.1. Umsetzung der Arbeitsdefinition

Insgesamt konnten nicht ausreichend Experten gewonnen werden, die tatsächlich die Expertise aufweisen, einschätzen zu können, ob durch die Items tatsächlich Lehrer-Selbstwirksamkeitserwartungen entsprechend der definitorischen Vorgaben beschrieben werden (vgl. Frage 3 im Expertenfragebogen für die Psychologen, Anhang I.2). Der Experte wies darauf hin, sich bezüglich der Itemkonstruktion nicht nur auf Schmitz und Schwarzer zu beziehen, sondern auch auf Bandura. Er hielt mit Bezug auf Bandura Handlungsbarrieren nicht für essentielle Bestandteile eines jeden Items, sondern für mögliche Formulierungsoptionen.

Physikdidaktikerinnen und Physikdidaktiker können in der Regel nicht als Experten für bestimmte psychologische Konstrukte im engeren Sinne eingestuft werden, weshalb diese bezüglich der Umsetzung definitorischer Vorgaben nicht explizit befragt wurden. Allen befragten Experten der Physikdidaktik wurden die zugrunde liegenden Konstruktionsvorschriften jedoch transparent gemacht und theoretisch begründet (vgl. Anhang I.1). Keiner der Experten im Bereich Physikdidaktik kritisierte diese oder brachte sonstige Einwände diesbezüglich ein.

10.4.2. Itemverständnis

Insgesamt kann konstatiert werden, dass die Items verständlich waren, d. h. in der Regel wurden die Items wie intendiert paraphrasiert oder zusammengefasst. Dabei wurden insbesondere die Handlungen wie intendiert beschrieben und Assoziationen geäußert, die dem gewählten Handlungsfeld entsprechen. Nicht immer war jedoch die intendierte Dimension exakt in den Äußerungen der Probanden erkennbar. Gehäuft kam es vor, dass bezüglich intendierter Planungshandlungen eher Assoziationen bezüglich der Durchführungsdimension beschrieben wurden. Außerdem kam es zum Teil vor, dass die in den Items formulierten Handlungsbarrieren in den Paraphrasierungen nicht explizit genannt wurden.

Wenn Items missverstanden wurden, so konnten in der Regel konkrete Anlässe dafür in den Items identifiziert werden. Das Itemverständnis war häufig dann beeinträchtigt, wenn die Items Fachbegriffe („Lernaufgaben“, „Schülervorstellungen“, etc.), Zuspitzungen („Ich kann *immer*...“) oder Verneinungen („Es bereitet mir *keine* Probleme“) enthielten. Auch die Experten wiesen im Zuge ihrer Beurteilung der Items häufig auf potentielle Verständnisschwierigkeiten durch in den Items enthaltene Zuspitzungen („doppelte Hürden“) oder Verneinungen hin.

Zusammenfassend ergibt sich, dass das Spezifitätsniveau, welches durch die Items adressiert wird, in der Regel Akzeptanz fand, das heißt, dass geeignete Situationen oder Themen assoziiert wurden. Außerdem zeigten fortgeschrittene Studierende im Allgemeinen weniger Probleme beim grundsätzlichen Itemverständnis und es konnten

keine Hinweise auf ein kohortenspezifisches Itemverständnis identifiziert werden. So zeigten sich in den Itemparaphrasierungen keine kohortenspezifischen Unterschiede und auch die Itemschwierigkeiten wurden kohortenübergreifend ähnlich bestimmt. Aus der Analyse des Itemverständnisses können Schwerpunkte für die Itemüberarbeitung abgeleitet werden (vgl. nachfolgender Abschnitt 10.5).

10.4.3. Aktivierung kognitiver Prozesse bei der Itembeantwortung

Anhand der Äußerungen im Zuge der Itemparaphrasierung und der Durchführung der begründeten Selbsteinschätzung konnten Indizien bezüglich der ablaufenden kognitiven Prozesse beim Beantworten der Items gewonnen werden. Es kann insbesondere festgehalten werden, dass die Selbsteinschätzung von den Probanden in der Regel durchgeführt wurde, sodass der Anspruch, das Merkmal der Selbstreferentialität durch die Items zu adressieren, in der Regel eingelöst wird. Ausnahmen bildeten solche Items, die Zuspitzungen enthielten, sodass von einigen Probanden geäußert wurde, dass sie sich bezüglich dieser Items nicht einschätzen könnten bzw. dass durch solche Items die Antwortkategorie „stimmt nicht“ im Prinzip vorbestimmt sei.

In der Regel wurden entsprechend der Intention im Rahmen der Itemparaphrasierung domänenkonforme komplexe Handlungen beschrieben, sowie bezüglich der getätigten Selbsteinschätzung domänenspezifische Assoziationen berichtet wie zum Beispiel passende individuelle Erfahrungen (vgl. Unterabschnitt 10.4.2). Deshalb kann davon ausgegangen werden, dass erwartete kognitive Prozesse bezüglich der Merkmale „Handlungsbezug“ und „Domänenspezifität“ der intendierten Konstrukte hervorgerufen werden.

Wie bereits berichtet, wurden einzelne Hürden im Rahmen der Itemparaphrasierung überlesen (vgl. Unterabschnitt 10.4.2). Wenn Hürden überlesen wurden bzw. in den Paraphrasierungen nicht bedacht wurden, so wurden entsprechende Hürden zumeist ebenfalls nicht bezüglich der Selbsteinschätzung als Bewertungskriterium herangezogen. Es kann daher davon ausgegangen werden, dass die formulierten Hürden nicht immer als schwierigkeitserzeugend wahrgenommen wurden. Das Merkmal des Schwierigkeitsbezugs wird daher nicht konsequent durch die Items adressiert.

Einige Experten wiesen darauf hin, dass die relativ langen Satzkonstruktionen der Items unter Umständen kognitiv schwierig zu erschließen und damit auch fehleranfällig bezüglich des Itemverständnisses seien. Sie schlugen zur kognitiven Entlastung daher eine einheitliche Satzkonstruktion der Form „Ich kann..., auch wenn...“ vor.

10.4.4. Authentizität der Items

In der Regel wurden die in den Items beschriebenen Situationen (Einheit aus Handlung und Hürde) sowohl von den Interviewten als auch von den Experten als relevant mit

Bezug auf den Schulalltag bewertet. Es konnte festgestellt werden, dass häufig solche Items als weniger relevant für den Schulalltag und damit als unauthentisch eingestuft wurden, für die die Handlungsbarrieren als unpassend bewertet wurden bzw. in denen Zuspitzungen die Items als unrealistisch erscheinen ließen. Aus den Urteilen beider Befragungsgruppen (Experten und Interviewte) konnten ebenfalls Hinweise für die Itemüberarbeitung abgeleitet werden (vgl. nachfolgender Abschnitt 10.5).

10.4.5. Exkurs: Itemschwierigkeiten

Die wahrgenommene Itemschwierigkeit sollte in dieser Studie u. a. deshalb erhoben werden, weil nach Bandura nur bezüglich hinreichend schwieriger Handlungen die Erfassung von Selbstwirksamkeitserwartungen sinnvoll ist (vgl. Abschnitt 3.2.1). Gleichzeitig kann aus der Beurteilung der Schwierigkeit eines Items ein Indiz hinsichtlich des kohortenübergreifenden Verständnisses des Items abgeleitet werden. Einerseits konnten Hinweise hinsichtlich der Itemschwierigkeit sowohl implizit aus den begründeten Selbsteinschätzungen der Befragten als auch aus den konkreten Äußerungen bezüglich als besonders schwierig/leicht wahrgenommener Items im Rahmen der Interviews gewonnen werden. Andererseits wurden die Experten explizit dazu aufgefordert, die Itemschwierigkeit einzuschätzen.

Aus den Interviews kann gefolgert werden, dass die Items als hinreichend schwierig eingeschätzt wurden, da eine vollkommene Zustimmung seitens der Interviewten sehr selten gegeben wurde. Teils wurden die Items eher als zu schwierig wahrgenommen. Die Ergebnisse der Expertenbefragung hinsichtlich der Itemschwierigkeiten sind schwieriger zu interpretieren und nur unter Einschränkungen weiterzuverwenden. Die Experten waren aufgefordert worden, die Items durch Zuordnung der Zahlenwerte von 1 bis 7 nach der eingeschätzten Schwierigkeit zu sortieren. Die meisten Experten haben jedoch keine Reihenfolge erstellt, sondern Zahlenwerte mehrfach vergeben. Einige Experten machten auch transparent, dass für sie unklar war, was unter Itemschwierigkeit verstanden werden sollte (vgl. Experte 8,5 Anhang I.4)Die Aufgabenstellung für die Experten kann damit als wenig eindeutig bezeichnet werden. Um einen Eindruck von den zugewiesenen Itemschwierigkeiten zu erhalten, wurde der Durchschnitt der zugewiesenen Zahlenwerte (1-7 bzw. 1-8 je nach Anzahl der Items) berechnet. Eine sinnvolle Interpretation oder Weiterverwendung der durchschnittlichen Itemschwierigkeiten ist jedoch nicht möglich, weil unter anderem ...

- ... unklar ist, ob die Experten die Schwierigkeiten für potentielle Befragte eingeschätzt haben oder eine persönliche Bewertung unabhängig von Befragungsgruppen vorgenommen haben,
- ... die Beurteilung durch die in diesem Stadium in den Items enthaltenen Zuspitzungen beeinflusst ist,

- ... unklar ist, ob die Bewertung der Experten eine sinnvolle Schätzung für die wahrgenommene Schwierigkeit durch Probanden ist,

Die Experteneinschätzungen und die Informationen, die aus den Interviews (Welche Handlungen werden in der Regel als besonders schwierig eingestuft?) gewonnen wurden, können jedoch dafür verwendet werden, Hypothesen bezüglich der Itemschwierigkeiten für Skalenversion 3 zu formulieren. Dabei ist es argumentativ fragwürdig, eine konkrete Reihenfolge anzugeben. Es sollen aber wenigstens die jeweils potentiell schwierigsten bzw. leichtesten Items benannt werden, sodass sich eine ungefähre Reihenfolge der Items ergibt, die in weiteren Analysen geprüft werden kann (vgl. Tabelle 10.1). Die Hypothesen wurden im Team diskutiert.

Tabelle 10.1.: Hypothesen bez. der Itemschwierigkeiten (vgl. Skalenversion 3, Anhang G.3).

Skala	**eher leichte Items**	**eher schwere Items**
SWE-EX-P	exp2, exp6	exp3, exp5, exp8
SWE-EX-D	exd2	exd3, exd5, exd7
SWE-EL-P	elp2, elp4	elp3, elp9
SWE-EL-D	eld4	eld6
SWE-A-P	ap4, ap6	aap3
SWE-A-D	ad2, ad8	ad5, ad7
SWE-SV-P	svp6, svp7	svp4, svp8
SWE-SV-D	svd7	svd1, svd6

10.4.6. Eignung des Antwortformats

Die Reaktionen auf die Frage nach der Eignung der verwendeten Antwortskala waren gemischt. Einige Befragte schätzten an der vierstufigen Skala, dass diese intuitiv sei und nicht so viel Zeit für deren Interpretation aufgebracht werden müsse:

> „So bin ich halt gezwungen, mich selber schnell einzuschätzen. 'Stimmts', 'stimmts nicht' und nicht so ewig damit zu verbringen, wies nun passt."

Zudem wird öfter bezweifelt, dass höherstufige Skalen ähnlich einfach zu verstehen seien. Es wird außerdem in Zweifel gezogen, ob mehr Kategorien genauso trennscharf sein können und ob zwischen den einzelnen Kategorien tatsächlich sinnvoll unterschieden werden könne:

> „Wenns noch feiner ist, würde ich mal das eine und mal das andere ankreuzen, obwohl ich hier [im Falle der Vierstufigkeit] immer ‚stimmt eher' ankreuzen würde, weil so richtig unterscheiden kann ich dazwischen [im Falle einer höheren Stufigkeit] nicht mehr."

> „[...] Wenn's zu kleinschrittig wär', würde ich einfach irgendwann anfangen, Kreuze zu setzen, wo's mir grade passt, weil so 'stimmt nicht', 'stimmt ein bisschen', da gibt's zu viel Spielraum, sich zu entscheiden und da würd' ich eher an der Entscheidung, ob's nun passt oder nicht, viel zu lange arbeiten, und mir schon so viele Gedanken drüber machen, dass es meinen eigentlichen Eindruck von der Frage verfälscht [...]."

> „Das sollte auch nicht mehr werden, weil man dann noch länger überlegen muss, was die einzelnen Abstufungen sind."

Andererseits geben einige Studierende an, dass spätestens mit etwas Praxiserfahrung, die Kategorie „stimmt nicht' für sie zumeist nicht mehr in Frage kommt und die Kategorie 'stimmt genau' aufgrund der wenigen Praxiserfahrung eigentlich ebenfalls keine Option darstellt. Somit bleiben in der Regel nur zwei Kategorien für die Beantwortung der Frage übrig:

> „Also ich sag mal, ich hab mich bewusst immer für 'stimmt kaum' oder 'stimmt eher' entschieden, weil ich bei keinem Punkt der Meinung war, dass es überhaupt gar nicht stimmt. Aber ich meine, wenn ich jetzt noch gar keine Erfahrung hätte, dann hätt' ich vielleicht auch 'stimmt nicht' angekreuzt, aber wenigstens durch die SPÜ [Schulpraktische Übungen] hat man ja wenigstens so'n bisschen Anwendung bekommen und 'stimmt genau' würde ich auch nicht sagen, weil ich einfach nicht die Erfahrung habe [...]."

Einige Studierende geben direkt an, dass ihnen zusätzliche Kategorien bei der Beantwortung gefehlt haben bzw. dass die vorgegebene Abstufung im Antwortprozess als unbefriedigend wahrgenommen wurde:

> „Ne größere Differenzierung wär' mir trotzdem leicht gefallen, weil ich hatte ja die Situation: 'stimmt eher' oder 'stimmt kaum', es ist irgendwo dazwischen."

> „[...] wobei ich hier [vierstufige Skala] finde, dass der Unterschied zwischen 'stimmt eher' und 'stimmt kaum' schon ziemlich gravierend ist. Also das eine ist ziemlich positiv, das andere schon ziemlich negativ behaftet in meinen Augen."

Einige Befragte legen außerdem ihr Bestreben offen, möglichst konsistent zu antworten, wobei die Umsetzung dieses Bestrebens ihrer Meinung nach durch höherstufige Formate erschwert wird:

> „Für mich persönlich ist das schwierig, weil dann weiß ich nicht mehr was, dann hab ich da 'ein bisschen' angekreuzt und dann les' ich mir 'ne Frage durch, wo ich eigentlich 'ein bisschen' ankreuzen will und merk dann, hätte ich vorher was anderes ankreuzen müssen. Und wenn ich den jetzt ausgefüllt hätte, dann hätte ich hier bestimmt schon rumgekrakelt und hätte das verändert."

Dieses „Konsistenzbestreben“ kann im Allgemeinen als Tendenz bewertet werden, sozial erwünscht zu antworten, wobei ein intransparenter persönlicher Maßstab angelegt wird. Die Unabhängigkeit der Itembeantwortung – ein notwendige Voraussetzung der Fragebogenkonzeption – wird durch dieses Verhalten gefährdet.

Es ergibt sich damit ein differenziertes Stimmungsbild hinsichtlich der Praktikabilität der verwendeten Skala. Alle genannten Argumente fasst Tabelle 10.2 zusammen. Insgesamt gab kein Befragter an, dass eine Einschätzung bezüglich der vierstufigen Skala überhaupt nicht möglich wäre. Problematisch ist gegebenenfalls, dass den Studierenden keine konkrete Alternativskala vorgelegt wurde, bezüglich derer sie einen Vergleich ziehen konnten, so wie es in der Akzeptanzbefragung der Fall war (vgl. Abschnitt 10.3). Somit haben sich die meisten Studierenden willkürlich selbst mögliche zusätzliche Antwortkategorien ausgedacht („bisschen mehr“) und sind bezüglich derer zu dem Schluss gekommen, dass eine Erhöhung der Stufigkeit eher verwirrend wäre. Es ist also durchaus möglich, dass bei geeigneter Wahl der zusätzlichen Antwortkategorien, die Studierenden, die einer Erhöhung der Stufigkeit eher kritisch gegenüberstehen, ebenfalls zu dem Schluss kommen, dass es sich um eine geeignete Skala handelt.

Tabelle 10.2.: Pro und Kontra bez. der Stufigkeit der Antwortskala.

Stufen	**Pro**	**Kontra**
4	intuitiv	zu grob/nur zwei Kategorien tatsächlich nutzbar (stimmt eher/kaum)
	schnell zu erfassen	ggf. keine Äquidistanz zwischen den Abständen (vgl. Abstand stimmt kaum/stimmt eher)
>4	differenziertere Erfassung	ggf. höhere kognitive Beanspruchung durch Enkodierung der Kategorienbedeutung (Zeitfaktor)
	Äquidistanz ggf. eher simulierbar	
	„Konsistenzbestreben“ eher unterbindbar	

Nach Abwägung der Argumente und unter Berücksichtigung der Ergebnisse der Akzeptanzbefragung (vgl. Abschnitt 10.3) wird das Antwortformat geändert und die Einführung einer sechsstufigen Antwortskala präferiert. Es ist davon auszugehen, dass bei geeigneter Wahl der Antwortkategorien nach einer kurzen Eingewöhnung die Erfassung ähnlich schnell, intuitiv und konsistent erfolgen kann. Einen Hinweis darauf liefert die Akzeptanzbefragung, in der Studierende zwei vorgegebene Skalen direkt miteinander vergleichen konnten. Darüber hinaus bietet die Sechsstufigkeit die Möglichkeit, ein differenzierteres Bild von der Selbsteinschätzungen zu erhalten. Die Einführung der Nutzung einer höherstufigen Skala orientiert sich außerdem auch an der Empfehlung Banduras (vgl. Unterabschnitt 3.6.2).

10.4.7. Angemessenheit der Abbildung der Handlungsfelder

Die Items wurden in der Regel von den Experten als für das Handlungsfeld repräsentativ charakterisiert. Teils wurde durch die Experten oder die Interviewten auf einzelne Handlungen hingewiesen, die ebenfalls als Kern des Handlungsfeldes beurteilt wurden. Kamen diesbezüglich Häufungen vor, so wurden einzelne Items neu konstruiert und der jeweiligen Skala hinzugefügt.

10.4.8. Trennschärfe der Konstrukte

17 von 18 Experten halten die Unterteilung der Handlungsfelder nach Planung und Durchführung von Physikunterricht grundsätzlich für relevant. Die Anmerkungen zeichnen jedoch ein differenzierteres Bild. So wird zum Teil in Frage gestellt, ob die Unterscheidung auch in der Praxis relevant ist, oder ob die Differenzierung ggf. vom Handlungsfeld abhängig ist:

> „Da Lehrpersonen aus Zeitmangel i. d. R. nur grobe Planungen vornehmen und stark auf eigene Vorarbeiten und Erfahrungen setzen, bildet sich die Unterscheidung ggf. nicht gut empirisch ab."
>
> „Die Differenzierung liegt nahe, aber in der Beantwortung der Items hat der Lehrer den Fokus auf dem Geschehen im Unterricht und weniger seine Arbeit am Schreibtisch. Die Relevanzfrage mag man konzeptionell mit 'ja' beantworten, aber in der Itembeantwortung mit einem Gefühl der 'Vermischung'. Das gilt für die Schülervorstellungen. Für die Aufgabenstellungen ist das klarer, weil Aufgaben am Schreibtisch formuliert und notiert werden [...]."

Zusammenfassend wird die Unterteilung in zwei Dimensionen grundsätzlich befürwortet, eine empirische Überprüfung jedoch angeraten. Ein Experte regte an, eine dritte Dimension zu berücksichtigen (Reflexion von Physikunterricht). Dies ist grundsätzlich zu befürworten, kann in diesem Projekt jedoch nicht umgesetzt werden.

10.4.9. Exkurs: Relevanz der Handlungsfelder

Die Items eines jeden Handlungsfeldes wurden im Schnitt von acht Experten begutachtet. Lediglich das Handlungsfeld „Umgang mit Aufgaben" wurde von zehn Experten gewertet, da es zum Zeitpunkt von Pilotstudie I noch nicht operationalisiert war und deshalb im Gegensatz zu den anderen Handlungsfeldern noch keiner Prüfung unterzogen wurde (vgl. Kapitel 9). Insgesamt liegen demnach 34 Experteneinschätzungen zu den Handlungsfeldern vor. In 33 von 34 Fällen wurden die dargelegten Handlungsfelder von den Experten als relevant eingestuft. Lediglich ein Experte gab an, das Handlungsfeld „Umgang mit Aufgaben" für weniger relevant zu erachten. Er relativierte diese Einschätzung jedoch dahingehend, dass insbesondere in der Operationalisierung Aufgaben mit

experimentellem Charakter nicht berücksichtigt worden seien. Insgesamt ergibt sich damit ein klares Votum für die Relevanz der berücksichtigten Handlungsfelder.

Die Anmerkung des Experten wird aus messtheoretischen Erwägungen nicht weiter verfolgt, da zu befürchten ist, dass durch Items, die eine Verknüpfung von Aufgabenbearbeitungen und Experimentieren die Eindimensionalität der Skalen gefährdet ist, auch wenn aus inhaltlicher Perspektive ein Überlapp tatsächlich existiert.

Zusätzliche Handlungsfelder Nachfolgende Auflistung zeigt alle 25 Vorschläge der Experten für zusätzlich zu operationalisierende Handlungsfelder, wobei dreizehn der genanten Vorschläge zu bereits vorliegenden Skalen zugeordnet werden können. Zu berücksichtigen ist, dass den Experten nicht bekannt war, zu welchen Handlungsfeldern schon Skalen konstruiert wurden. Die mehrfache Nennung der bereits berücksichtigten Handlungsfelder kann damit als Bestätigung für die Relevanz dieser gewertet werden.

- **Experimentieren (5x)**
- **Aufgaben:**
  - Kompetenzorientierung: Aufgaben mit Passung zu den Bildungsstandards
  - experimentelle Aufgaben
- **Schülervorstellungen:**
  - Einbezug des Vorverständnisses (Präkonzepte) der Lernenden; Umgang mit Schülervorstellungen (3x)
  - Initiierung von Konzeptveränderungen
- **Elementarisierung**
  - Instruktionsstrategien
  - didaktische Rekonstruktion bzw. Elementarisierung
- Evtl. Erstellen und Korrigieren von schriftlichen Prüfungen
- Erklären von Physik (2x)
- Verbindung von Theorie und Experiment
- Umgang mit Mathematisierungen (2x)
- Medien: Umgang mit AV und computergestützten Medien
- unterrichtsmethodische Kompetenzüberzeugungen für Motivation und Lernfortschritte der Schüler
- Einbinden in für Schülerinnen und Schüler interessante Kontexte
- Fachwissen

- etwas Neues/selbstständig lernen
- Modelle im Physikunterricht

Einige der Vorschläge liegen gewissermaßen quer zu den bereits operationalisierten Handlungsfeldern, spielen also in mehreren Handlungsfeldern eine Rolle. Hierzu zählt beispielsweise das Erklären (von Physik), was beispielsweise sowohl im Kontext von Aufgabenstellungen oder Experimenten denkbar ist. Deshalb gibt es mehrere Items unterschiedlicher Skalen, die Erklärungen adressieren.[1]

Nach Abzug der Doppelnennungen verbleiben neun zusätzliche Vorschläge, wobei einige zum Teil in den Skalenversionen enthalten sind oder aufgenommen wurden.[2] Einige Vorschläge sind für die Zukunft zu berücksichtigen, wenn die Konstruktion weiterer Skalen zur Debatte steht (Umgang mit Mathematisierungen, Umgang mit Medien, Motivation und Interessenförderung). Hinsichtlich der Thematik Motivierung/Interessenförderung ist jedoch zu bedenken, dass entsprechende Items häufig schwierig als Selbstwirksamkeitserwartungen zu formulieren sind, weil zumeist direkt ein Ziel adressiert wird, nämlich beispielsweise eine Steigerung der Motivation zu erwirken. Demnach soll eingeschätzt werden, ob sich eine bestimmte Handlung mit dem konkreten Ziel der Motivierung zugetraut wird. Gerade dann liegt jedoch eine Verwechslung der Konstrukte Selbstwirksamkeitserwartung und Handlungsergebniserwartung nahe, da eher eine „means-ends“-Relation beschrieben wird (vgl. Abschnitt 3.2.1).

10.5. Itemrevision

10.5.1. Vorgehen und Ableitungen bez. der Itemüberarbeitung

Aus den Interviews und den Expertenkommentaren ergaben sich einige grundsätzliche Aspekte für die Itemrevision (vgl. Tabelle 10.3) hinsichtlich derer alle Items überprüft wurden.

Darüber hinaus wurden für den „inhaltlichen“ Überarbeitungsprozess zum einen die Übersichten, die für die Auswertung der Interviews (vgl. Unterabschnitt 10.1.5 bzw. Anhang H.4) und der Expertenbefragung (vgl. Unterabschnitt 10.2.4 bzw. Anhang I.3) angelegt wurden, genutzt. Zusätzlich wurde eine weitere zusammenfassende Darstellung

[1] vgl. Skalenversion 4 im Anhang G.4: „Ich kann im Physikunterricht eine physikalische Erklärung, die die Schülerinnen und Schüler nicht verstehen, weiter vereinfachen, ohne dass die Erklärung physikalisch falsch wird“ (eld3). „Ich kann meine Schülerinnen und Schüler bei der Bearbeitung von Physikaufgaben unterstützen, auch ohne den Lösungsweg einfach vorzugeben“ (ad2). „Ich kann auf unvorhergesehene Verständnisschwierigkeiten meiner Schülerinnen und Schüler beim Experimentieren so reagieren, dass sie selbstständig weiterarbeiten können, auch ohne einfach einen Lösungsweg vorzugeben“ (exd4).

[2] vgl. Skalenversion 4 im Anhang G.4: „Kontexte“ (elp5), „Erklären“ (eld3), „Verbindung von Theorie und Experiment“ (exd3), „selbstständiges Lernen (ap5), „Erstellen von Prüfungen“ (ap1).

erarbeitet, die sich für den Überarbeitungsprozess als nützlich erwies (vgl. exemplarisch Anhang K). Für jedes Item enthält diese Zusammenfassung neben dem ursprünglichen Itemtext, die durchschnittliche Itemschwierigkeit (Zuordnung der Experten) sowie die zentralen Problemstellen, die in den Interviews und den Expertenkommentaren zu Tage traten. Auf Grundlage dieser Daten wurde ein Verbesserungsvorschlag dokumentiert, der als Ausgangspunkt für die Itemüberarbeitung im Team diente.

In nachfolgendem Unterabschnitt 10.5.2 kann der Überarbeitungsprozess nicht erschöpfend dargestellt werden. Es sollen jedoch exemplarisch zwei Beispiele herausgegriffen werden, die die Art der Überarbeitung illustrieren und nachvollziehbar machen. Die revidierten Items bilden die dritte Skalenversion (vgl. Anhang G.3).

Tabelle 10.3.: Resultate der Pilotstudien (I: Interviewstudie, E: Expertenbefragung) sowie resultierende Festlegungen für die Itemüberarbeitung.

Resultate	**Ableitungen für die Itemkonstruktion**
Itemformulierungen/Handlungsbarrieren (Verständlichkeit):	
E: Empfehlung einer einheitlichen Itemstruktur zur besseren Lesbarkeit der Items und der kognitiven Entlastung.	(1) Einführung einer einheitlichen Satzstruktur: „Ich kann..., auch wenn...“.
I: Teils missverständliche, unpräzise, unauthentische Itemformulierungen. Diese werden gehäuft beobachtet, wenn Fachbegriffe verwendet werden („Schülervorstellungen“), Negationen auftreten (Ich habe *keine* Probleme...“) oder Zuspitzungen enthalten sind („Ich kann *immer*...“).	(2) Entfernung aller Negationen, Allaussagen und sonstigen Zuspitzungen. (3) Dem Handlungsfeld SWE-SV wird ein erläuternder Instruktionstext vorangestellt.
I/E: Teilweise werden mehrere Hürden in den Items wahrgenommen (z. T. aufgrund enthaltener Zuspitzungen), weshalb diese Items häufig als zu schwierig charakterisiert werden.	(4) Überprüfung der Items auf doppelte Hürden. Ggf. Schwierigkeit der Items durch inhaltliche Anpassungen erhöhen.
Unterscheidung der Dimensionen Planung und Durchführung von Physikunterricht:	
E: Teils nicht eindeutig entscheidbar, ob Item eher Planung oder Durchführung adressiert.	(5) Insbesondere Stärkung des Planungscharakters einzelner Items der Planungsdimension.
I: Tendenz zur Assoziation von Durchführungssituationen selbst dann, wenn Item Planungsdimension adressiert.	siehe Ableitung (5)
Relevanz und Vollständigkeit der Handlungsfelder:	
E: Teilweise werden Handlungsfelder nicht vollständig abgebildet.	(6) Ergänzung einzelner Items.

10.5.2. Exemplarische Illustration der Itemüberarbeitung

Beispiel 1

„Ich kann die Vorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie sehr vielfältig sind“ (Skala: SWE-SV-P, Version 2, Item 1, vgl. Anhang G.2.5).

Aus den Pilotstudien ergeben sich unter anderem folgende Hinweise (vgl. Anhang K):

- Die Experten betonen, dass Vielfältigkeit eine sehr unbestimmte Hürde darstelle.
- In den Interviews zeigt sich, dass die Hürde von den meisten Probanden überlesen wird oder nicht als Hürde wahrgenommen wird, da Vielfältigkeit als immanente Eigenschaft von Schülervorstellungen betrachtet wird.
- Zudem werden in den Interviews vermehrt Verständnisschwierigkeiten dahingehend geäußert, dass unklar ist, was genau unter Vorstellungen der Schülerinnen und Schüler verstanden werden soll. Teilweise werden sogar Interessen assoziiert.

Entsprechend der in Tabelle 10.3 dargestellten Ableitung (3) wird die Formulierung der „Alltagsvorstellung“, die im vorangestellten Instruktionstext der Skala erläutert wird, in das Item integriert. Darüber hinaus wird die Hürde angepasst. Es resultiert das folgende Item:

⇒ „Ich kann die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn die Vorstellungen einander widersprechen“ (Skala: SWE-SV-P, Version 3, Item svp6, vgl. Anhang G.3.5).

Beispiel 2

„Fehler und Schwierigkeiten, die bei Schülerinnen und Schülern auftreten, wenn sie Aufgaben bearbeiten, kann ich mit Sicherheit vorhersehen, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe“ (Skala: SWE-A-P, Version 2, Item ap6, vgl. Anhang G.2.7).

Aus den Pilotstudien ergeben sich unter anderem folgende Hinweise:

- Die Experten weisen darauf hin, dass das Item sehr extrem formuliert ist („mit Sicherheit vorhersagen“) und daher zwei Handlungsbarrieren enthält.
- Auch in den Interviews zeigt sich, dass die Formulierung als zu extrem wahrgenommen wird („es ist einfach zu vielseitig, dass man alles vorhersagen kann“).
- Außerdem werden zu dem Item häufig Durchführungssituationen assoziiert.

Für die Itemüberarbeitung werden insbesondere die Ableitungen (1), (2) und (4) angewandt (vgl. Tabelle 10.3). Zusätzlich wurde versucht, den Planungsaspekt stärker zu betonen (Ableitung (5)). Das überarbeitete Item lautet wie folgt:

⇒ „Ich kann bei der Unterrichtsvorbereitung Schwierigkeiten vorhersehen, die bei der Aufgabenbearbeitung auftreten könnten, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe“ (Skala: SWE-A-P, Version 3, Item ap6, vgl. Anhang G.3.7).

10.6. Interpretation der Ergebnisse und Validitätsargumente

Die dargestellten Ergebnisse können mit Blick auf die eingangs genannten Forschungsfragen (vgl. Abschnitte 10.1 und 10.2) interpretiert werden.

10.6.1. Aktivierung kognitiver Prozesse und Umsetzung der Arbeitsdefinition

Im Rahmen der Interviews konnten Eindrücke bezüglich der ablaufenden kognitiven Prozesse gesammelt werden. Insgesamt kann festgestellt werden, dass bis auf die Wahrnehmung der Handlungsbarrieren als schwierigkeitserzeugend, die intendierten kognitiven Prozesse weitgehend hervorgerufen wurden, sodass Hypothese 2.1 (vgl. Abschnitt 8.1.2, FF2) als weitgehend zutreffend eingestuft werden kann. Es ist davon auszugehen, dass durch die Itemrevision auch der Schwierigkeitsbezug durch die Items abgebildet wird. Auch dass die Befragten im Zuge der Selbsteinschätzung andeuten, dass die Beantwortung der Items mit zunehmender Erfahrung anders ausfallen würde, bzw. sehr häufig auf eigene Erfahrungen rekurrieren, die als Hauptquelle für Selbstwirksamkeitserwartungen gelten (vgl. Abschnitt 3.3), spricht dafür, dass entsprechend des intendierten Konstruktes passende kognitive Aktivitäten hervorgerufen werden.

Auch wenn FF1 (vgl. Abschnitt 8.1.2) anhand der Expertenbefragung nicht hinreichend vertiefend eruiert werden konnte, so scheint es zunächst keine Indizien zu geben, die gegen eine angemessene Umsetzung der Arbeitsdefinition sprechen. Insbesondere, dass die Experten keinen Anstoß an den dargelegten Konstruktionsvorschriften üben und der Fakt, dass weitgehend die intendierten kognitiven Prozesse durch die Items angesteuert werden, kann dahingehend gedeutet werden, dass Hypothese 1.1 zutrifft.

Bei der Interpretation der Ergebnisse sollte berücksichtigt werden, dass beim lauten Denken in der Regel trotz vorherigen Übens zumeist nicht alle kognitiven Prozesse verbalisiert werden können, da viele routinisiert und unbewusst ablaufen. Unter Umständen ist es möglich, dass das Verbalisieren eine zusätzliche kognitive Anstrengung darstellt, die die zu verbalisierenden gedanklichen Prozesse (negativ) beeinflusst. Gegebenenfalls ist die eigene Fähigkeit der Verbalisierung von Gedankenprozessen ein großer Einflussfaktor auf die Art der gewonnenen Daten. Schließlich sollte beachtet

werden, dass Tendenzen der sozialen Erwünschtheit unter Umständen dann besonders durchschlagen, wenn das laute Denken unter „Aufsicht“ stattfindet (vgl. Sandmann, 2014, S. 187 ff.). Dass trotz dieser potentiellen Einschränkungen in den Äußerungen der Befragten genügend Hinweise bezüglich des Ablaufens intendierter kognitiver Prozesse gefunden wurden, spricht für die dargestellte Deutung der Ergebnisse.

10.6.2. Itemverständnis

Es kann konstatiert werden, dass die Items grundsätzlich nicht unverständlich waren. In der Regel stimmten die Intention der Autoren und das Verständnis der Befragten weitgehend überein, sodass Hypothese 3.1 (vgl. Abschnitt 8.1.2, FF3) als zutreffend gewertet werden kann. Trotzdem konnten Hindernisse identifiziert werden, die das Verständnis der Items in der gewünschten Weise erschweren. Zu diesen zählen die Verwendung von Fachtermini, Negationen oder die Zuspitzung von Items. Als besonders ermutigend mit Blick auf FF10 erweist sich die Tatsche, dass keine Indizien auf ein kohortenspezifisches Verständnis hindeuten, auch wenn mit zunehmender Erfahrung der Studierenden die Verständnisschwierigkeiten tendenziell abnahmen. Es kann nach der akribisch vorgenommenen Revision der Items davon ausgegangen werden, dass die Items für potentielle Befragungsgruppen verständlich sind.

10.6.3. Relevanz und Authentizität

Bereits in der Itemkonstruktion gestaltete sich das Finden passender Hürden als Schwierigkeit, sodass damit gerechnet wurde, hier auf Probleme zu stoßen. Die offensive Thematisierung dieser Problematik (beispielsweise auch im Expertenfragebogen) führte dazu, dass Items, die aufgrund der dargestellten Handlungsbarriere übergreifend als weniger authentisch wahrgenommen wurden, identifiziert werden konnten und zugleich im Gespräch oder durch die Experten Alternativen diskutiert bzw. vorgeschlagen werden konnten. Aus den genannten Gründen ist davon auszugehen, dass die revidierten Items nun von potentiellen Befragten als weitgehend authentisch und relevant wahrgenommen werden. Hypothese 4.1 (vgl. Abschnitt 8.1.2, FF4) kann damit als zutreffend gelten.

Die Relevanz der formulierten Items für das betreffende Handlungsfeld wurde in der Regel durch die Experten bestätigt, was ein Indiz für das Zutreffen von Hypothese 6.1. darstellt. Teilweise wurden Ergänzungen vorgeschlagen, die in neu konstruierten Items aufgenommen wurden, sodass davon ausgegangen werden kann, dass die Handlungsfelder nun entsprechend der Hypothese 6.2 repräsentativ durch die entwickelten Items abgebildet werden (vgl. Abschnitt 8.1.2, FF6).

10.6.4. Eignung des Antwortformates

Es deutet sich an, dass eine sechsstufige Antwortskala zur differenzierten Erfassung der intendierten Konstrukte besser geeignet ist. Hypothese 5.1 (Abschnitt 8.1.2, FF5) muss daher verworfen werden.

10.6.5. Trennschärfe der Konstrukte

Ein interessantes Ergebnis ist die Tatsache, dass insbesondere zu Planungsitems häufig Durchführungssituationen assoziiert wurden. Dies könnte natürlich einerseits an der mangelnden sprachlichen Betonung des Planungsaspektes liegen. Es könnte aber auch sein, dass in der Realität häufig viel weniger scharf zwischen der Planung und der Durchführung einer Unterrichtsstunde unterschieden und beides eher als eine Symbiose empfunden wird, so wie es auch einzelne Experten beschrieben haben. Gerade bei der Planung und Durchführung von Physikunterricht ist diese Lesart naheliegend, da zur Planung häufig auch die Präparation und Testung von Experimenten etc. gehört und daher bereits in der Planungsphase Elemente mit „durchführendem“ Charakter integriert sind. Auch die Experten halten die theoretische Trennung der Konstrukte für nachvollziehbar, empfehlen aber dringend eine empirische Überprüfung. Das Zutreffen von Hypothese 8.1 (Abschnitt 8.1.2, FF8) kann bisher nicht abschließend beurteilt werden.

10.7. Zusammenfassung

Bezüglich der eingangs formulierten Fragen kann festgestellt werden, dass beide Studien wesentlich dazu beigetragen haben, die Items und deren Wirkung auf potentielle Befragte besser zu verstehen. Es konnten Schwachstellen und Probleme aufgezeigt und in die Überarbeitung der Items eingebracht werden. Es ist davon auszugehen, dass damit die Qualität der Items erheblich gesteigert werden konnte. Insgesamt wurden viele Indizien gesammelt, die nahelegen, dass die konstruierten Skalen tatsächlich die intendierten Konstrukte, das heißt Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern, abbilden.

Darüber hinaus zeigt sich, wie wichtig eine derartige Itemüberprüfung ist, denn obwohl die Items gewissenhaft konstruiert und bereits überarbeitet wurden, treten Verständnisprobleme auf, die im Konstruktionsprozess z. B. aufgrund der Involviertheit der Itementwickler nicht antizipiert werden können.

11. Weiterentwicklung des Instruments

Auf Grundlage der Informationen, die in den Pilotstudien gewonnen wurden, konnte das Instrument in wesentlichen Teilen überarbeitet werden (vgl. Abschnitte 9.7 und 10.5). Das nachfolgende Kapitel stellt eine Zusammenfassung der Überarbeitungen dar und beschreibt die Grundlagen und den Aufbau des Fragebogens, der in der Hauptstudie des Projektes eingesetzt wird. Teile dieses Kapitels wurden bereits in Meinhardt u. a. (2016c) veröffentlicht.

11.1. Überarbeitung der Itemkonstruktionsregeln

Aufgrund der Ergebnisse der Pilotstudien II und III (vgl. Kapitel 10 ab Seite 223) werden die folgenden Änderungen an den Itemkonstruktionsregeln in Einklang mit den theoretischen Notwendigkeiten vorgenommen (vgl. zusammenfassend Tabelle 11.1):

1. Jedes Item beginnt mit der Formulierung „Ich kann ...“.
2. Der Nebensatz beginnt in der Regel mit der Formulierung „..., auch wenn ...“.
3. Die Items enthalten keine zusätzlichen Zuspitzungen, wie sie durch Formulierungen wie „immer“ oder „in jedem Fall“ adressiert werden.

Die sehr wichtigen Hinweise zur Itemkonstruktion unter Berücksichtigung des Bezuges zu anderen Konstrukten gelten weiterhin (vgl. Abschnitt 7.1.3).

Tabelle 11.1.: Überarbeitete Itemkonstruktionsregeln.

Eigenschaft/Facette des Konstruktes	**abgeleitete Konstruktionsregel**
Subjektivität/Selbstreferentialität: Selbsteinschätzung eigener (aktueller) Fähigkeiten	Formulierung in der ersten Person Singular, Präsens
Handlungsorientierung: Selbsteinschätzung eigener Fähigkeiten bez. komplexer Handlungen	Verwendung der Phrase „Ich kann ...“ gefolgt von nicht trivialen bzw. komplexen Handlungen
Domänenspezifität: Selbsteinschätzung eigener Fähigkeiten bez. komplexer unterrichtsbezogener Handlungen	Handlungen müssen spezifisch für die Planung oder Durchführung von Physikunterricht sein
Schwierigkeitsbezug: Selbsteinschätzung eigener Fähigkeiten bez. komplexer Handlungen bei auftretenden Schwierigkeiten	Formulierung passender Handlungsbarrieren durch Verwendung der Phrase „auch wenn“

In der Regel ergibt sich durch die Festlegungen die nachfolgende spezifische Satzstruktur aus Haupt- und Nebensatz:

> „Ich kann ... [Handlung], auch wenn ... [Handlungsbarriere]."

11.2. Überarbeitung des Antwortformates

In Anbetracht der Ergebnisse der Pilotstudie II (vgl. Abschnitt 10.1) und der zusätzlichen Akzeptanzbefragung durch Thorid Rabe und Olaf Krey (vgl. Abschnitt 10.3) wird das vierstufige Antwortformat durch ein sechsstufiges Format ersetzt. Die Erhöhung der Stufigkeit und die Beibehaltung der Unipolarität stehen im Einklang mit den Empfehlungen Banduras (vgl. Abschnitt 3.6.2). Es wird auf eine gerade Anzahl an Wahlmöglichkeiten zurückgegriffen, auch wenn damit gewissermaßen eine Entscheidung erzwungen wird. Auch wird damit vermieden, eine „Mitte" für die Befragten zu suggerieren. Um für die Auswertung auf die Intervallskalierung zurückgreifen zu können, werden die einzelnen Kategorien der Skala nicht nur benannt, sondern auch mit Zahlenwerten belegt (stimmt nicht (1) bis stimmt genau (6)). Die Befragten werden im Fragebogen aufgefordert, ihre Zustimmung zu den dargebotenen Item-Aussagen entsprechend der vorgegebenen Skala anzugeben.

11.3. Zusammensetzung des Fragebogens

Eine Version des verwendeten Fragebogens kann dem Anhang L entnommen werden. Dieser enthält im Wesentlichen drei Komponenten: demografische Angaben, die eigens entwickelten Items/Skalen zu Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern sowie zusätzliche relevante Items/Skalen. Die genannten Komponenten werden in ihrer Auswahl und Zusammenstellung nachfolgend detaillierter vorgestellt.

11.3.1. Demografie

Es werden nur solche demografischen Angaben erfragt, die bezüglich der Beantwortung der Forschungsfragen relevant sind, beispielsweise solche, die Rückschlüsse auf die Praxiserfahrung bzw. Berufserfahrung zulassen (Dienstjahre). Die Fächerkombination wird zur Kontrolle abgefragt. Tabelle 11.2 gibt einen Überblick über die erhobenen demografischen Angaben.

11.3.2. Items

Die Items, die für den Einsatz in der Hauptstudie Verwendung finden, sind unter Berücksichtigung der angepassten Konstruktionsregeln (vgl. Abschnitt 11.1) mittlerweile

Tabelle 11.2.: Demografische Angaben nach Kohorte (erfragt: x, nicht erfragt: –).

Angaben	**Studierende**	**Referendare**	**Lehrkräfte**
allgemein			
Geschlecht	x	x	x
Alter	x	x	x
Dienstjahre	–	–	x
zum Studium			
Studiengang/Abschluss	x	x	x
Fachsemester	x	–	–
Ausbildungsart (Uni/PH)	x	x	x
Bundesland (Studium)	x	x	x
Lehramt	x	x	x
Fachkombination	x	x	x
∅-Leistung Physikklausuren/Abschluss	x	x	
Anzahl Lehrveranstaltungen Physikdidaktik	x	–	
Art schulpraktischer Erfahrungen	x	–	
Unterrichtserfahrung in Zeitstunden	x	–	
zur Unterrichtssituation			
Schulform	–	x	x
Schulstufe	–	x	x
Bundesland (Schule)	–	x	x
Fachkombination	–	x	x

dreifach überarbeitet worden und umfassen die in Tabelle 7.2 dargestellten Themenfelder. Im Vergleich zu Pilotstudie I werden zusätzlich die neu entwickelten Items der Skalen Umgang mit Aufgaben/Planung (SWE-A-P) und Umgang mit Aufgaben/Durchführung (SWE-A-D) in den Fragebogen aufgenommen.

11.3.3. Zusätzliche Konstrukte

Ebenfalls aufgrund der Relevanz bezüglich der Forschungsfragen (vgl. FF9, ab Seite 204) werden mit Hilfe des Fragebogens zusätzliche Skalen eingesetzt, die (potentiell) die folgenden Konstrukte abbilden sollen:

- Allgemeine Selbstwirksamkeitserwartung (A-SWE): Schwarzer und Jerusalem (1999),
- Lehrer-Selbstwirksamkeitserwartung (L-SWE): Schmitz und Schwarzer (2000),
- physikalisches Selbstkonzept (PS): Adaption nach Hoffmann u. a. (1998),
- Persönlichkeit (Big-Five-Inventory – BFI-K): Rammstedt und John (2005)
- Kontrollüberzeugungen, Internalität (K-I): Krampen (1991)

Die zusätzlich eingesetzten Skalen werden mit jeweils unterschiedlichen Antwortformaten dargeboten (vgl. Tabelle 11.4):

Anmerkungen

Einige der Skalen (A-SWE, L-SWE, PS) wurden bereits in Pilotstudie I verwendet (vgl. Kapitel 9), sodass vergleichende Analysen möglich werden. Die Skalen von Schwarzer und Jerusalem (1999) und Schmitz und Schwarzer (2000) können im deutschsprachigen Raum als Standardinstrumente zur Erfassung von (domänenspezifischen) Selbstwirksamkeitserwartungen angesehen werden (vgl. Kapitel 3 und 4). Die Skala von Hoffmann u. a. (1998) wurde für *Studierende* des Physik-Lehramtsstudiums angepasst und ist daher auch nicht in den Fragebögen für die Kohorten der Referendare und Lehrkräfte enthalten. Ursprünglich erfasst die Skala das physikalische Selbstkonzept von Schülerinnen und Schülern (vgl. Meinhardt, 2011, S. 32 bzw. S. 76).

Um Persönlichkeitsmerkmale abbilden zu können, wurde ein Instrument gesucht, welches die Testzeit nicht erheblich verlängert. Generell ist zu beachten, dass die Basistendenzen (Extraversion, Offenheit, etc.) nicht direkt erfasst werden, sondern lediglich das Selbstkonzept der betreffenden Person bezüglich der dargebotenen Indikatoren (vgl. z. B. Mayr, 2011, S. 132).

Tabelle 11.3.: Auswahl an Themenfeldern der Skalen und Itembeispiele nach den ersten drei Pilotierungsstudien.

SWE-EL-P	
Strukturierung physikal. Inhalte	elp2, elp4, elp5
Vereinfachung/Elementarisierung physikal. Inhalte für Lerngruppen	elp3, elp8
Planung von Unterrichtssequenzen, die Alltagsphänomene adressieren	elp5
Ausrichtung von Unterrichtssequenzen an Basiskonzepten/Kompetenzbereichen	elp6, elp9
Identifikation von Schlüsselstellen in Schulbuchdarstellungen	elp7
SWE-EL-D	
Anpassung von Erklärungen im Unterricht	eld3
inhaltl. Strukturierung und Akzentuierungen von Unterrichtsstunden	eld1, eld2
Herstellen thematischer Querverbindungen/Alltagsbezüge	eld4, eld5, eld6
SWE-A-P	
Entwicklung/Zusammenstellung von Prüfungs- & Übungsaufgaben	ap1, ap2, ap7
Entwicklung offener Aufgaben/Lernaufgaben	ap3, ap5
Auswahl und Beurteilung der Eignung/Qualität von Aufgaben	ap4
Antizipation potentieller Bearbeitungsschwierigkeiten/Erstellung von Lernhilfen	ap6, ap8
Planung von Auswertungsphasen	ap9
SWE-A-D	
Aufgabenbearbeitungen für Schülerinnen und Schüler einsichtig machen	ad7
Unterstützung bei der Aufgabenbearbeitung	ad2, ad3, ad4, ad6
Aufgaben spontan ausdenken	ad5
Auswertung von Aufgabenbearbeitungsphasen	ad1, ad8, ad9
SWE-SV-P	
Erkennen von Schülervorstellungen	svp2, svp8
Adressierung/Einbeziehung von Schülervorstellungen	svp1, svp3, svp6
Strategien zum Umgang mit Schülervorstellungen	svp4
Umgang mit Koexistenz von Schülervorstellungen und physikal Vorstellungen	svp5, svp7
SWE-SV-D	
Erkennen/Beurteilen von SV	svd1, svd2
Adressierung/Einbeziehung von Schülervorstellungen	svd3, svd4
Strategien zum Umgang mit Schülervorstellungen	svd5, svd6
Umgang mit Koexistenz von Schülervorstellungen und physikal. Vorstellungen	svd7
SWE-EX-P	
Auswahl von Experimenten	exp1, exp7, exp10
Variation von Experimenten	exp2, exp5
Zusammenstellung von Experimenten	exp3, exp9
Entwicklung von Experimenten	exp4, exp8,
Motivierung durch Experimente	exp6
SWE-EX-D	
Durchführung/Variation von Experimenten	exd1, exd2, exd5, exd7
Natur der Naturwissenschaften	exd3
Unterstützung beim Experimentieren	exd4, exd9
Umgang mit Messwerten	exd6
Inszenierung von Experimenten	exd8

Tabelle 11.4.: Antwortformate der Zusatzskalen.

Skala	Format
A-SWE	4-stufig (trifft nicht zu, trifft kaum zu, trifft eher zu, trifft genau zu)
L-SWE	4-stufig (stimmt nicht, stimmt kaum, stimmt eher, stimmt genau)
PS	5-stufig (sehr gut, gut, mittel, schlecht, sehr schlecht)
BFI-K	5-stufig (sehr unzutreffend, eher unzutreffend, weder noch, eher zutreffend, sehr zutreffend)
K-I	6-stufig (sehr falsch (---) bis sehr richtig (+++))

11.4. Instruktionen

Dem Fragebogen vorangestellt wurde der folgende einleitende Instruktionstext, dem sich die Beantwortung der demographischen Fragen anschloss:

> „Liebe Studierende/Referendare/Lehrkräfte,
>
> vielen Dank für die Teilnahme an der Befragung. Sie unterstützen damit mein Promotionsprojekt an der Universität Potsdam. Die Befragung ist natürlich freiwillig. Ihre Angaben werden anonymisiert und ausschließlich zu Forschungszwecken verwendet!
>
> Wenn Sie Fragen haben, wenden Sie sich gern an: Claudia.Meinhardt@uni-potsdam.de[1]

Darüber hinaus wurden folgende Hinweise zum Ausfüllen des Fragebogens gegeben:

> „Wir bitten Sie, sich auf den nachfolgenden Seiten **selbst einzuschätzen**. Es handelt sich um Situationen, die sich auf das Handeln von Physiklehrkräften beziehen. Es kann sein, dass Sie diese Situationen noch nicht selbst als Physiklehrkraft erlebt haben. Versuchen Sie bitte dennoch sich **in die Situationen hineinzuversetzen** und entscheiden Sie, ob Sie mit ihrem **heutigen** (!) Wissen und Können diese Situation bewältigen können oder nicht.
>
> Wir bitten Sie, sich möglichst **realistisch** einzuschätzen.
>
> Für unsere Auswertung ist es wichtig, dass Sie möglichst **alle** Fragen beantworten“ [Hervorhebungen im Original].

Anschließend wurden die Befragten aufgefordert, jeweils ihre Zustimmung zu den nachfolgenden Aussagen anzugeben. Entsprechend dem Vorschlag bei ggf. unklaren Begrifflichkeiten Definitionen anzugeben (vgl. Tabelle 2.1 auf Seite 38), um die Eindeutigkeit der Aussagen zu gewährleisten, wurde dem Handlungsfeld „Umgang mit

[1] Diese E-Mail-Adresse existiert mittlerweile nicht mehr.

Schülervorstellungen“ ein zusätzlicher Instruktionstext vorangestellt. Ausgangspunkt dieser Maßnahme war die Überlegung, dass unter Umständen Studierende der ersten Semester mit dem Begriff „Schülervorstellungen“ wenig vertraut sind – ein inhaltlich angemessenes Verständnis zur Beantwortung der Aussagen jedoch eine Voraussetzung darstellt. Die Instruktion lautete wie folgt:

> „In den folgenden Formulierungen ist häufig die Rede von **Schüler- bzw. Alltagsvorstellungen**. Darunter sollen Denkweisen und Vorstellungen zu physikalischen Begriffen und Phänomenen verstanden werden, die häufig in Widerspruch zu den entsprechenden physikalischen Modellen und Konzepten stehen. Schülervorstellungen sind häufig im Alltag nützlich und haben sich dort auch bewährt. Beispiele wären: *Elektronen sind winzige Kügelchen* oder *Glühlampen verbrauchen Strom*“ [Hervorhebungen im Original].

11.5. Anordnung der Skalen

Aus den Vorstudien ergibt sich ein Konsens darüber, dass die Handlungsfelder eindeutig sind und nach Ergänzungen vollständig in ihrem Kern durch die Items abgebildet werden. Inwiefern jedoch die Planung und Durchführung von Physikunterricht tatsächlich als zwei getrennte Dimensionen zu konzeptualisieren sind, kann nach den Vorstudien nicht erschöpfend beantwortet werden. Die genannte Frage ist daher von zentraler Bedeutung für die Hauptstudie. Um zu prüfen, ob sich zwei Dimensionen trennen lassen, werden die Items handlungsfeldspezifisch, jedoch nicht nach Planungs- und Durchführungsdimension separiert, sondern zufällig durchmischt dargeboten. Die konkrete Itemreihenfolge wird für jedes Handlungsfeld von einem Zufallsgenerator bestimmt (vgl. Tabelle 11.5).

Tabelle 11.5.: Zufällige Itemreihenfolge für die Handlungsfelder.

SWE-EL	elp8 elp1 eld1 elp6 elp5 eld5 elp7 eld6 eld3 elp4 eld2 elp3 eld4 elp2 elp9
SWE-A	ad3 ad8 ap9 ap6 ad2 ad9 ap8 ap2 ad4 ap7 ad1 ad6 ap3 ap4 ap5 ad7 ad5 ap1
SWE-SV	svd6 svd7 svp6 svp3 svp2 svd4 svp7 svd3 svp5 svp8 svd5 svd1 svp1 svp4 svd2
SWE-EX	exd2 exp2 exd3 exp4 exp10 exd1 exp1 exd6 exp7 exd7 exp9 exd4 exp3 exp8 exd5 exd9 exd8 exp6 exp5

Insgesamt werden vier verschiedene Fragebogenversionen erstellt, in denen die Reihenfolge der Darbietung der Items der einzelnen Handlungsfelder systematisch variiert. Dieses Vorgehen diente der Vorbeugung von Reihenfolgeeffekten. Folgende Versionen wurden gedruckt:

1. SWE-A, SWE-EX, SWE-EL, SWE-SV
2. SWE-SV, SWE-A, SWE-EX, SWE-EL

3. SWE-EL, SWE-SV, SWE-A, SWE-EX
4. SWE-EX, SWE-EL, SWE-SV, SWE-A

11.6. Überlegungen bez. der Tendenz zu sozialer Erwünschtheit

Dem in Abschnitt 2.2.2 beschriebenen Effekt der sozialen Erwünschtheit, der Verzerrungen des Antwortverhaltens beschreibt, wird dahingehend begegnet, dass den Befragten absolute Anonymität zugesichert und betont wird, dass die erhobenen Daten lediglich zu Forschungszwecken verwendet werden. Auch wenn in Folge der Organisation der Testdurchführung Dozentinnen und Dozenten bei der Beantwortung zugegen sind, so sollte dies keinen nennenswerten Einfluss auf die Beantwortung haben, da die Befragung einerseits anonym erfolgt und andererseits der eigentliche Urheber der Befragung gegenüber den Probanden transparent gemacht wird (Angabe der verantwortlichen Institution und der Kontaktdaten auf dem Deckblatt, Hinweis zu E-Mail-Kontakt im einleitenden Instruktionstext, vgl. Abschnitt 11.4 und Anhang L). Darüber hinaus sind für die Probanden keine Konsequenzen mit der Beantwortung verknüpft. Ihr Testergebnis hat folglich keinerlei Relevanz für ihre Zukunft, weshalb ebenfalls ein Antwortverhalten entsprechend sozialer Erwünschtheit eher unwahrscheinlich ist.

11.7. Papier-und-Bleistift versus Online-Befragung

Die Übersetzung des Fragebogens in ein Online-Format bringt natürlich eine zusätzliche Variable ins Spiel, die kontrolliert werden muss. Alternative Möglichkeiten der Befragung von Physiklehrpersonen konnten jedoch nicht realisiert werden. Um einen Effekt des Erhebungsformates beim Beantworten des Fragebogens möglichst zu vermeiden, wurde der Online-Fragebogen identisch zu dem Papier-Fragebogen aufgebaut und designt. Um Reihenfolgeeffekte bei der Beantwortung möglichst zu vermeiden, wurden die Handlungsfelder (inkl. der zugehörigen Items) in einer zufälligen Reihenfolge dargeboten. Die Itemreihenfolge entsprach der in Tabelle 11.5 dargestellten.

12. Hauptstudie

„Measurement is always uncertain, be it in a physics lab or in collecting data from humans.“

Boone, 2014, S. 201

Die Ergebnisse der bereits beschriebenen Pilotstudien tragen wesentlich zur Argumentationskette des Validitätsargumentes bei. Insbesondere inhaltliche Validierungsfragen standen dabei im Fokus. Vieles spricht dafür, dass die konstruierten Items wie intendiert verstanden und intendierte kognitive Prozesse generiert werden. Damit können Argumente angegeben werden, die für die Gültigkeit einiger der Testwertinterpretation zugrunde liegender Annahmen auf Itemebene sprechen (vgl. Unterabschnitt 7.5.2). Die nachfolgend dargestellte Hauptstudie des Validierungsprojektes untersucht im Wesentlichen die Annahmen auf Skalenebene. Dazu zählen beispielsweise die Abgrenzbarkeit der durch die Skalen repräsentierten Konstrukte zu anderen (benachbarten) Konstrukten sowie die statistische Prüfung der Eindimensionalität der Skalen und deren gleiche Funktionsweise für unterschiedliche Gruppen.

Zunächst werden kurz die Datenerhebung und -aufbereitung erläutert. Es schließen sich methodische Bemerkungen an, die jedoch nicht allzu stark ausgedehnt werden sollen, da wesentliche Vorgehensweisen im Rahmen der bereits publizierten Skalendokumentation ausgeführt wurden (vgl. Meinhardt u. a., 2016c, Vorbemerkungen). Sehr viele der nachfolgend dargestellten Ergebnisse können ebenfalls dieser Dokumentation entnommen werden. In den sich anschließenden Abschnitten werden bezüglich der Forschungsfragen relevante Ergebnisse zusammenfassend dargestellt, wobei jeweils auf entsprechende Abschnitte in der Skalendokumentation verwiesen wird, die ggf. ausführlichere Informationen enthält. Korrelationsanalysen und Mittelwertvergleiche sind nicht Teil des Skalenhandbuches und werden deshalb ausführlicher behandelt. Den Abschluss des Kapitels bildet die Interpretation der Ergebnisse hinsichtlich der Forschungsfragen, die für die Validierung der intendierten Testwertinterpretation relevant sind. Dabei soll der oben genannte Hinweis Boones berücksichtigt werden.

12.1. Datenerhebung

Um genügend Teilnehmerinnen und Teilnehmer – insbesondere aus der Gruppe der Lehrerinnen und Lehrer – für die Studie zu gewinnen, wurden nach Pötschke (2010, S. 44 f.) folgende Aspekte berücksichtigt: Appell an die Hilfsbereitschaft potentieller Teilnehmerinnen und Teilnehmer, Wecken des Interesses für die Studie sowie Adressierung persönlicher Vorteile. Konkret erfolgte die Umsetzung dieser Aspekte durch personalisierte Aufrufe zur Teilnahme in Internetforen bzw. durch Flyer und Aushänge (vgl. Anhänge M.1 und M.2). Vor der Befragung wurden Angebote unterbreitet, über die Inhalte des Fragebogens im Anschluss zu berichten oder zu diskutieren. Zusätzlich wurden die Fragebogeninhalte als Reflexionsgelegenheit vorgestellt, um sowohl mögliches Interesse zu wecken, als auch um auf einen potentiellen persönlichen Vorteil hinzuweisen. Letzteres sollte jedoch besonders durch eine initiierte Verlosung angesprochen werden, in der attraktive berufsgruppenspezifische Preise ausgelobt wurden.[1]

12.1.1. Rekrutierung von Teilnehmerinnen und Teilnehmern

Studierende Um möglichst viele Physik-Lehramtsstudierende befragen zu können, wurden alle Professuren der Physikdidaktik in Deutschland und Österreich mit der Bitte um die Bereitstellung von Ressourcen für die Befragung (Zeit in Lehrveranstaltungen) angeschrieben sowie auf Kolloquien und Tagungen persönlich angesprochen.

Referendare Um die Befragung von Lehramtsanwärterinnen und -anwärtern in Physik-Fachseminaren zu realisieren, mussten in mehreren Bundesländern Genehmigungen zuständiger Behörden beantragt werden. Insgesamt konnten Referendare der Bundesländer Brandenburg, Berlin, Niedersachsen und Hamburg sowie Baden-Württemberg und Rheinland-Pfalz befragt werden.

Lehrkräfte Um Physiklehrerinnen und Physiklehrer für die Befragung zu gewinnen, wurde eine Online-Version des Fragebogens erstellt. Genutzt wurde dazu die Software Unipark der Firma Questback.[2] Alle Schulen bzw. deren Physikfachkonferenzen der Bundesländer Berlin, Brandenburg, Rheinland-Pfalz und Hamburg wurden postalisch kontaktiert und auf den Online-Fragebogen inkl. der Verlosung hingewiesen. Gleichzeitig wurde ein Aufruf in gängigen Lehrerforen im Internet gestartet.[3]

[1] Es handelte sich bei den Preisen um zwei Experimentierkoffer zum Thema Energie im Wert von jeweils 500 Euro und insgesamt 10 Amazon-Gutscheine im Wert von jeweils 30 Euro.
[2] http://www.unipark.com/de/ (letzter Zugriff: 01.07.2016)
[3] Der Aufruf startete im November 2013 in den Foren 4teachers.de, lehrerforum.de, referendar.de, lehrerforen.de und lehreronline.de (vgl. Anhang M.2).

12.1.2. Durchführung der Datenerhebung

Die Befragung wurde an Universitäten (in physikdidaktischen Lehrveranstaltungen) oder in Studienseminaren (in Physikfachseminaren) in der Regel durch die Leiterinnen und Leiter der entsprechenden Veranstaltungen durchgeführt. Dazu wurden Papier-Fragebögen zu Beginn oder Ende einer Lehrveranstaltung/eines Seminars ausgeteilt und nach dem Ausfüllen wieder eingesammelt. Lediglich in Berlin und Brandenburg konnten einige Befragungen durch die Autorin selbst durchgeführt werden. Dann bestand für die Teilnehmerinnen und Teilnehmer im Anschluss an die Befragung die Möglichkeit zur Diskussion, weshalb ein kurzer Impulsvortrag vorbereitet wurde. Lehrpersonen wurden ausschließlich mittels der Online-Version des Fragebogens befragt (vgl. Abschnitt 12.4). Der Befragungszeitraum erstreckte sich von Oktober 2013 bis Mitte April 2014. Die mittlere Bearbeitungszeit sowohl des Online- als auch des Papier-Fragebogens betrug ca. 25 Minuten.

12.2. Datenaufbereitung

Die mittels Papier-Fragebogen erhobenen Daten wurden in einem Team aus drei Personen in die Statistik Software SPSS 22 (vgl. IBM Corp. Released, 2013) übertragen, wobei jeweils eine Person die Antworten laut vorlas und eine zweite Person diese Daten in die Eingabemaske übertrug. Aus organisatorischen Gründen und aufgrund der großen Datenmenge gab es zwei Personen, die jeweils im Wechsel die Daten eintrugen. Jeweils rund 10% der Fälle wurden sowohl für die Kohorte der Studierenden ($N = 57$) als auch für die Kohorte der Referendare ($N = 22$) zufällig ausgewählt und erneut eingegeben, um die Fehleranfälligkeit der Eingabe abzuschätzen. Insgesamt wurden für jeden Fall 116 bzw. 123 Variablen (demografische Angaben bleiben unberücksichtigt) auf Abweichungen überprüft. Für die Studierenden ergaben sich bei 7011 Eingaben 10 Abweichungen (.1%), für die Referendare bei 2552 Eingaben 8 Abweichungen (.3%). Insgesamt kann die Fehleranfälligkeit des Verfahrens damit als äußerst gering bezeichnet und das Verfahren als geeignet angesehen werden.

Unipark/Questback stellt nach Beendigung der Datenerhebung eine Datendatei zur Verfügung, die mit den Datendateien der Papier-und-Bleistift-Erhebung zusammengeführt wurde. Nicht alle Fälle konnten für die Datenauswertung herangezogen werden. Nachfolgend werden einige Kriterien genannt, die zum Ausschluss von Fällen aus dem Datensatz führten.

12.2.1. Datenselektion

Insgesamt haben 1002 Personen eine der beiden Fragebogenversionen bearbeitet. Nach Ausschluss einzelner Fälle verblieben insgesamt 931 auswertbare Fälle im Datensatz

(vgl. Tabelle 12.1). Folgende *inhaltliche* Kriterien führten zum Ausschluss einzelner Fälle aus dem Datensatz:

- Weder das Schul- noch das Studienfach ist Physik.
- Als Studiengang wurde Lehramt an Berufsschulen/Grundschulen/Sonderschulen angegeben oder es wurde gar kein Studiengang genannt (unbekannt).

Auch *formale* Kriterien führten zum Ausschluss einzelner Fälle aus dem Datensatz, und zwar:

- Mehrere komplette Skalen wurden nicht ausgefüllt.
- Mindestens zwei Skalen wurden identisch bewertet oder sonstige Muster waren erkennbar (Verdacht des „Durchkreuzens“).

Einzelne Fälle wurden vor allem immer dann ausgeschlossen, wenn die hauptsächlich zu validierenden Skalen (SWE-SV-P, SWE-SV-D, SWE-EX-P, ...) nicht komplett bearbeitet wurden. Wurden dagegen die zusätzlichen Skalen (L-SWE, A-SWE, PS, BFI-K, K-I) nicht vollständig bearbeitet, so verblieben die Fälle zunächst im Datensatz. Details zu den Ausschlussgründen können dem Anhang N entnommen werden.

Tabelle 12.1.: Übersicht der ausgeschlossenen Fälle (online/Papier-&-Bleistift).

Kohorte	**„finisher“**	**Ausschluss (inhaltlich)**	**Ausschluss (formal)**	**Stichprobe (resultierend)**
Lehrkräfte	189 (189/-)	14 (14/-)	7 (7/-)	168 (168/-)
Referendare	249 (42/207)	6 (0/6)	5 (3/2)	238 (39/199)
Studierende	564 (35/529)	23 (1/22)	16 (0/16)	525 (34/491)
gesamt	1002 (266/736)	43 (15/28)	28 (10/18)	931 (241/690)

12.3. Datenqualität

Nachfolgend werden die Anzahl fehlender Werte (vgl. Unterabschnitt 12.3.1) sowie die Verteilungsparameter für die Item- und Skalenwerte mitgeteilt (vgl. Unterabschnitt 12.3.2). Nach der Datenselektion verbleiben äußerst wenige fehlende Daten im Datensatz und die Abweichungen von der Normalverteilung fallen moderat aus.

12.3.1. Fehlende Werte

Die Anzahl fehlender Werte fasst Tabelle 12.2 zusammen, wobei in der Darstellung lediglich fehlende Werte der Items berücksichtigt wurden und nicht die fehlenden Angaben bezüglich demografischer Aspekte.

Auf den neu entwickelten Skalen liegt der Anteil fehlender Werte für die drei Kohorten zwischen .0% und .5% (vgl. Meinhardt u. a., 2016c, S. 20, 70, 150, 200, 260, 309, 388, 437). Das Auftreten fehlender Werte unterliegt keinem erkennbaren System und kann daher als zufällig angenommen werden.

Tabelle 12.2.: Übersicht über fehlende Werte.

		online-FB[1]			**Papier-FB**			**gesamt**		
Kohorte	# Items	N	# fW[2]	%	N	# fW	%	N	# fW	%
Studierende	123*	34	10	.2	491	168	.3	525	178	.3
Referendare	116	39	17	.4	199	133	.6	238	150	.5
Lehrkräfte	116	168	56	.3	–	–	–	168	56	.3
gesamt		241	83	.3	690	301	.4	931	384	.3

[1] FB: Fragebogen [2] # fW: Anzahl fehlender Werte
* Die zusätzlichen Items für die Kohorte der Studierenden resultieren aus dem Einsatz der Skala des physikalischen Selbstkonzeptes (vgl. Unterabschnitt 11.3.3).

12.3.2. Verteilung der Item- und Skalenwerte

Details zu den Verteilungen der Item- und Skalenwerte der neu konstruierten Skalen können der Skalendokumentation entnommen werden (Meinhardt u. a., 2016c), ebenso wie Details zu verwendeten Algorithmen etc. (vgl. Meinhardt u. a., 2016c, S. 9).

In der Regel sind die Item- und Skalenwerte nicht normalverteilt (vgl. Meinhardt u. a., 2016c, S. 24, 73, 153, 203, 263, 31, 391, 440), wobei die Schiefe der Itemwerte zwischen -1.161 und .005 schwankt und die Kurtosis[4] zwischen -.651 und 2.306. Die Schiefe der Skalenwerte liegt zwischen -.916 und -.237, während die Kurtosis der Skalenwerte zwischen -.159 und 1.870 liegt. Die Verteilung der Item- als auch Skalenwerte ist damit in der Regel linksschief. Bezüglich der Kurtosis liegt kein einheitlicher Trend vor – die Verteilungen sind im Vergleich zur Normalverteilung sowohl spitzer (Kurtosis größer als Null) als auch flacher (Kurtosis kleiner als Null). Insgesamt können die Abweichungen von der Normalverteilung als moderat bezeichnet werden (vgl. auch Histogramme bei Meinhardt u. a., 2016c, S. 21 ff., 71 ff., 151 ff., 201 f., 261 ff., 310 ff., 389 f., 438 f.).

Die Verteilungsparameter der Skalenwerte der zusätzlich erhobenen Skalen können Tabelle 12.3 entnommen werden. Sie liegen in ähnlichen Bereichen, wie die Werte für die

[4] Es wird die Kurtosis des SPSS-Outputs berichtet, der als Exzess interpretiert werden muss, da Werte größer, kleiner oder gleich Null auftreten können.

neu konstruierten Skalen. Auch die Normalverteilungsannahme muss in der Regel für die Skalenwerte der zusätzlich erhobenen Skalen verworfen werden. Lediglich die Skalenwerte der Skalen der Lehrer-Selbstwirksamkeitserwartung und der Kontrollüberzeugungen (Internalität) können für die Kohorte der Lehrkräfte als normalverteilt angesehen werden (vgl. Tabelle 12.4).

Tabelle 12.3.: Verteilungsparameter der Skalenwerte der zusätzlich erhobenen Skalen.

	Studierende		**Referendare**		**Lehrkräfte**	
Skala	Schie.[1](S.E.[2])	Kurt.[3](S.E.)	Schie. (S.E.)	Kurt. (S.E.)	Schie. (S.E.)	Kurt. (S.E.)
A-SWE	-.290 (.107)	.546 (.213)	-.634 (.158)	1.731 (.315)	.014 (.187)	-.137 (.373)
L-SWE	-.100 (.107)	.488 (.213)	-.650 (.158)	1.622 (.314)	-.222 (.187)	-.181 (.373)
PS	-.158 (.107)	.469 (.213)	–	–	–	–
K-I	.059 (.107)	.310 (.213)	-.600 (.158)	2.125 (.316)	-.329 (.187)	.461 (.373)
BFI-K (E)	-.272 (.107)	.082 (.213)	-.051 (.158)	-.663 (.315)	-.106 (.187)	.012 (.373)
BFI-K (V)	-.553 (.107)	.605 (.213)	-.485 (.158)	.894 (.315)	-.470 (.187)	.505 (.373)
BFI-K (G)	-.485 (.107)	.268 (.213)	-.556 (.158)	.230 (.315)	-1.120 (.187)	2.009 (.373)
BFI-K (N)	.050 (.107)	-.170 (.213)	-.026 (.158)	-.360 (.315)	.164 (.187)	-.199 (.373)

[1] Schiefe [2] Standardfehler [3] Kurtosis

Tabelle 12.4.: Test auf Normalverteilung der Skalenwerte der zusätzlich erhobenen Skalen.

	Studierende					**Referendare**					**Lehrkräfte**				
		K.-S.[1]		S.-W.[2]			K.-S.		S.-W.			K.-S.		S.-W.	
	df[3]	Stat.[4]	Sig.[5]	Stat.	Sig.	df	Stat.	Sig.	Stat.	Sig.	df	Stat.	Sig.	Stat.	Sig.
A-SWE	524	.064	.000	.987	.000	237	.098	.000	.969	.000	168	.095	.001	.983	.043
L-SWE	525	.078	.000	.987	.000	238	.088	.000	.968	.000	168	.066	.069	.984	.050
PS	525	.074	.000	.989	.001	–	–	–	–	–	–	–	–	–	–
K-I	522	.070	.000	.991	.003	236	.089	.000	.971	.000	168	.069	.052	.987	.120
BFI-K (E)	525	.072	.000	.974	.000	237	.080	.001	.972	.000	168	.092	.001	.975	.004
BFI-K (V)	525	.129	.000	.969	.000	237	.135	.000	.962	.000	168	.117	.000	.966	.000
BFI-K (G)	525	.135	.000	.969	.000	237	.120	.000	.965	.000	168	.149	.000	.920	.000
BFI-K (N)	525	.062	.000	.985	.000	237	.077	.002	.985	.013	168	.085	.005	.981	.023
BFI-K (O)	525	.093	.000	.966	.000	237	.076	.002	.974	.000	168	.117	.000	.971	.001

[1] Kolmogorov-Smirnov-Test [2] Shapiro-Wilk-Test [3] Freiheitsgrade [4] Statistik [5] Signifikanz

12.4. Stichprobe und deskriptive Analysen

Analysen der Daten können sich – wie bereits erwähnt – auf die ausgefüllten Fragebögen von insgesamt 931 Physik-Lehramtsstudierenden, Physik-Referendaren und Physik-Lehrkräften stützen. Insgesamt handelt es sich für die Kohorten nicht um repräsentative

Stichproben. Die Kohorten unterliegen aufgrund der Freiwilligkeit der Teilnahme einer gewissen Selbstselektion, die für die Gruppe der Lehrkräfte im besonderen Maße zutrifft. Die drei Kohorten und für die Auswertung relevante Subgruppen werden im Folgenden näher beschrieben.

12.4.1. Detaillierte Beschreibung der befragten Kohorten

Insgesamt können nach der Datenselektion die Fragebögen von 168 Physiklehrpersonen zur Auswertung herangezogen werden. Von diesen Personen gaben 38% ($N = 63$) an, weiblichen Geschlechts zu sein. Die befragten Lehrkräfte waren im Mittel 45 Jahre alt (26-68 Jahre) und konnten auf eine durchschnittliche Berufserfahrung von 16 Jahren (0-40 Jahre) zurückblicken. Je ca. 30% der befragten Lehrpersonen gaben an, in Brandenburg oder Rheinland-Pfalz zu unterrichten. In Berlin unterrichteten 20%, in Hamburg 10% der befragten Physik-Lehrkräfte. Die meisten Lehrpersonen gaben Unterricht an einem Gymnasium (ca. 56%), während die verbliebenen Lehrkräfte an anderen Schulformen (Realschulen, Gesamtschulen, Oberschulen, Gemeinschaftsschulen, Stadtteilschulen, ...) unterrichteten. Zwei Drittel der Befragten arbeitete vorrangig in der Sekundarstufe I. Am häufigsten vertraten die Physik-Lehrpersonen die Fachkombination Physik und Mathematik (87%), wobei 44% angaben, ein zusätzliches drittes Fach zu unterrichten. Informatik war dabei am beliebtesten (10%). 17% der befragten Physik-Lehrkräfte gab an, durch einen Quereinstieg zum Physiklehramt gekommen zu sein.

Von den 238 Physik-Lehramtsanwärterinnen und -anwärtern (Referendaren), deren Daten für die Auswertung genutzt werden können, gaben 71 Personen an, weiblichen Geschlechts zu sein (30 %). Im Durchschnitt waren die befragten Referendare 30 Jahre alt (23-58 Jahre) und absolvierten ihr Referendariat hauptsächlich in Berlin (32%), Niedersachsen (27%), Baden-Württemberg (17%) oder Hamburg (13%) an einem Gymnasium (53%). Insgesamt waren ca. 54% der Befragten zuvor in einem Gymnasiallehramtsstudiengang immatrikuliert, während 16% angaben, das Lehramt an Sekundarschulen (Realschullehramt) studiert zu haben. 71 Referendare (30%) haben keinen Lehramtsstudiengang absolviert (Quereinstieg). Am häufigsten wurde das Fach Physik mit Mathematik kombiniert (66%). 20% der befragten Referendare gab an, zusätzlich ein drittes Fach studiert zu haben.

Nach der Datenselektion stehen die Fragebögen von 525 Physik-Lehramtsstudierenden zur Auswertung zur Verfügung. 188 der befragten Studierenden gab ab, weiblichen Geschlechts zu sein (ca. 36%). Die Studierenden waren zwischen 18 und 46 Jahre alt (Durchschnitt: 24 Jahre). 39% der Lehramtsstudierenden waren in einem Bachelor-Studiengang immatrikuliert, 18% in einem Masterstudiengang und 42% studierten auf Staatsexamen. Die meisten Studierenden absolvierten das 3. (21%), 5. (31%) oder 7. Semester (21%). Bis auf vier Studierende, die an einer Pädagogischen Hochschule

studierten, waren alle an einer Universität eingeschrieben. Eine Mehrheit von 72% strebte den Abschluss eines Gymnasiallehramtstudiums an, wobei Physik am häufigsten mit dem Fach Mathematik kombiniert wurde (67%). Der auswertbare Datensatz enthält Antworten von Studierenden aller Bundesländer. Die größte Gruppe wurde am Standort Gießen befragt (ca. 9%). Insgesamt 20% der Befragten studierten an einer Berliner Universität.

12.4.2. Beschreibung relevanter Subgruppen

Die Verteilung der Probanden auf relevante Subgruppen (z. B. für durchzuführende Mehrgruppenvergleiche) ist Tabelle 12.5 zu entnehmen.

Tabelle 12.5.: Relevante Subgruppen der Gesamtstichprobe (vgl. Meinhardt, Rabe & Krey, 2016c, S. 6).

	Studierende	**Referendare**	**Lehrkräfte**	**Σ**
Format				
online	34	39	168	241
Papier-&-Bleistift	491	199	0	690
Σ	525	238	168	931
Geschlecht				
männlich	335	167	104	606
weiblich	188	71	63	322
ohne Angabe	2	0	1	3
Σ	525	238	168	931
Studiengang				
sekundar	140	39	–	179
gymnasial	385	128	–	513
Quereinstieg	0	71	–	71
Σ	525	238	–	763
Schulform				
Sekundarschule	–	–	68	
Gymnasium	–	–	100	
Σ			168	
Praxiserfahrung				
institutionelle	385	–	–	
anderweitige	20	–	–	
keine	120	–	–	
Σ	525			

– (unbesetzte Zelle): Merkmal für Kohorte nicht relevant

Unter der Schulform „Sekundarschule“ werden alle Schulformen zusammengefasst, die nicht explizit eine gymnasiale Oberstufe anbieten. Realschulen oder Hauptschulen werden daher als Sekundarschule klassifiziert. Gesamtschulen fallen nicht unter diese Klassifikation. Die Praxiserfahrung Physik-Lehramtsstudierender wird nach institutionellen (z. B. Schulpraktische Übungen/SPÜ, Schulpraktikum, Praxissemester) und nicht institutionellen, anderweitigen Erfahrungen (z. B. Nachhilfe) unterschieden (vgl. Meinhardt u. a., 2016c, S. 5 f.).

12.5. Analyseverfahren und benutzte Analysesoftware

Die Daten wurden systematisch verschiedenen Analyseverfahren unterzogen, wobei für jedes Verfahren eine andere Software benutzt wurde:

- Klassische Item- und Skalenanalyse: SPSS 22.0 (IBM Corp. Released, 2013),
- Konfirmatorische Faktorenanalyse: Mplus 7.11 (Muthén & Muthén, 2012),
- Raschanalyse: Winsteps 3.90.0 (Linacre, 2014).

12.6. Methodische Anmerkungen

Nachfolgend wird keine explizite Begründung für die Wahl der verwendeten Methoden dargelegt. Diese legitimieren sich implizit, da die Ergebnisse der Auswertungen zur Absicherung eines Validitätsargumentes beitragen. Auch die Funktionsweise der verwendeten Methoden wird nicht dargestellt, sondern diesbezüglich auf Spezialliteratur verwiesen. Wesentliche Analyseschritte und Kriterien zur Beurteilung der Analyseergebnisse einzelner Verfahren wurden bereits in der Skalendokumentation dargelegt (Meinhardt u. a., 2016c). Um die Interpretation der Ergebnisse (vgl. Abschnitt 12.8) zu erleichtern, werden die genannten Kriterien hier erneut aufgeführt sowie einige ergänzende Bemerkungen zu zusätzlich durchgeführten Analysen wie Korrelationsanalysen oder Mittelwertvergleichen dargelegt.

Insgesamt kann die Datenauswertung in fünf große Blöcke untergliedert werden. Zum einen wurden allgemeine Untersuchungen bezüglich fehlender Werte oder vorliegender Verteilungen vorgenommen, da diese häufig als Voraussetzungen für weitere Analysen abgeklärt werden müssen. Die Ergebnisse wurden bereits berichtet (vgl. Abschnitt 12.3). Außerdem wurden im Zuge einer klassischen Item- und Skalenanalyse Inter-Item-Korrelationen, Itemtrennschärfen sowie Skalenreliabilitäten bestimmt/geschätzt (vgl. u. a. Bühner, 2011, Kap. 5). Im Rahmen konfirmatorischer Faktorenanalysen wurden die Skalen auf ihre Eindimensionalität geprüft, eine Reliabilitätsprüfung vorgenommen sowie Mehrgruppen- und Dimensionsvergleiche durchgeführt (vgl. u. a. Brown, 2006; Weiber

& Mühlhaus, 2010; Geiser, 2011; Christ & Schlüter, 2012; Byrne, 2012). Ähnliche Untersuchungen werden im Rahmen einer Raschanalyse durchgeführt (vgl. u. a. Bond & Fox, 2007; Boone, Staver & Yale, 2014), wobei Mehrgruppenvergleiche durch DIF-Analysen[5] realisiert und Dimensionsvergleiche mittels Rasch-Hauptkomponentenanalysen der Residuen durchgeführt werden. Die Eindimensionalität kann jedoch mit dem Raschmodell nicht geprüft werden – diese stellt vielmehr eine Voraussetzung für die Durchführung einer Raschanalyse dar.

12.6.1. Anmerkungen zur klassischen Item- & Skalenanalyse

Es wird eine klassische Reliabilitätsprüfung mit SPSS durchgeführt, wobei Bühner (2011, Kapitel 5) als Orientierung dient.

Inter-Item-Korrelationen Es werden jeweils signifikante Pearson-Korrelationen angegeben, die die tatsächlichen Korrelationen wahrscheinlich überschätzen. Dies ist jedoch deshalb kein Problem, da geringe Korrelationen zwischen .20 und .40 für die Unabhängigkeit der Items sprechen (vgl. Bühner, 2011, S. 243). Liegen die angegebenen Korrelationen trotz potentieller Überschätzung in diesem Bereich, so spricht das für die zu testende Unabhängigkeitshypothese.

Itemtrennschärfen & Skalenreliabilitäten Es werden jeweils part-whole korrigierte Itemtrennschärfen angegeben. Als Reliabilitätsmaß dient Cronbach's Alpha α_C. Zur Interpretation der Ergebnisse können die Grenzwerte aus Tabelle 12.6 herangezogen werden, wobei je nach Grad der Homogenität/Heterogenität des zu messenden Konstruktes Trennschärfen/Reliabilitäten ggf. unterschiedlich interpretiert werden müssen. Schermelleh-Engel und Werner (2012, S. 135 f.) geben als zufriedenstellenden Richtwert für Persönlichkeitsfragebögen $\alpha_C \geq .70$ an.

Tabelle 12.6.: Grenzwerte für Reliabilität und Trennschärfe nach Bühner (2011, S. 81).

	niedrig	mittel	hoch
korrigierte Trennschärfe	$< .30$	$.30 - .50$	$> .50$
Reliabilität	$< .80$	$.80 - .90$	$> .90$

12.6.2. Konfirmatorische Faktorenanalyse

Alle angegebenen Beurteilungskriterien der jeweiligen Modellpassung beruhen auf einem robusten Maximum-Likelihood-Schätzverfahren, da der Datensatz fehlende Werte

[5] DIF: „differential item functioning“.

enthält und die Item- und Skalenwerte in der Regel nicht normalverteilt sind (vgl. Byrne, 2012, S. 234).

Beurteilung der Modellpassung Mit Bezug auf Hu und Bentler (1999), Schermelleh-Engel, Moosbrugger und Müller (2003), Brown (2006) sowie Weiber und Mühlhaus (2010) werden die in Tabelle 12.7 dargestellten Grenzwerte zur Beurteilung der Modellpassung herangezogen. Aufgrund der eher heterogenen Konstrukte werden etwas niedrigere Grenzwerte für die Faktorladungen akzeptiert.

Tabelle 12.7.: Grenzwerte zur Beurteilung des Modellfits.

	gut bis sehr gut	**akzeptabel**	**nicht akzeptabel**
globale Ebene			
χ^2/df	≤ 2.50	$2.50 < \chi^2/\text{df} \leq 3.00$	> 3.00
CFI[1]	$\geq .950$	$.900 \leq \text{CFI} < .950$	$< .900$
TLI[2]	$\geq .950$	$.900 \leq \text{TLI} < .950$	$< .900$
RMSEA[3]	$\leq .050$	$.050 < \text{RMSEA} \leq .080$	$> .080$
$[90\%KI]$[4]	Null enthalten		
SRMR[5]	$\leq .050$	$.050 < \text{SRMR} \leq .080$	$> .080$
lokale Ebene			
Faktorladung λ	$> .60$	$.50 - .60$	$< .50$

[1] Comparative-Fit-Index [2] Tucker-Lewis-Index [3] Root-Mean-Square-Error of Approximation
[4] 90%-Konfidenzintervall des RMSEA [5] Standardized-Root-Mean-Residual

Schätzung der Reliabilität Im Rahmen konfirmatorischer Faktorenanalysen können mehrere Reliabilitätskoeffizienten geschätzt werden. Einerseits die Indikatorreliabilität λ^2, die sich durch Quadrieren der jeweiligen Faktorladung ergibt, sowie die Faktorreliabilität (FR) und die durchschnittlich extrahierte Varianz (DEV). Grenzwerte zur Beurteilung der Reliabilitätsmaße können Tabelle 12.8 entnommen werden.

Tabelle 12.8.: Grenzwerte zur Beurteilung der Reliabilität (vgl. Weiber & Mühlhaus, 2010, S. 122 f.).

	gut bis sehr gut
Indikatorreliabilität	$\geq .40$
Faktorreliabilität	$\geq .60$
durchschnittlich extrahierte Varianz	$\geq .50$

Messinvarianz: Mehrgruppenvergleiche/KFA Im Rahmen des konfirmatorischen Ansatzes kann überprüft werden, inwiefern die spezifizierten Modelle für verschiedene interessierende Subgruppen (vgl. Unterabschnitt 12.4.2) in gleichem Maße auf die jeweiligen Daten passen, das heißt Ähnliches durch die Skalen abgebildet wird. Messinvarianz ist eine Voraussetzung dafür, dass die Testwerte in unterschiedlichen Gruppen gleich interpretiert und somit auch Mittelwerte unterschiedlicher Gruppen miteinander verglichen werden dürfen. Brown (2006, S. 290, 300) empfiehlt zur Durchführung von Mittelwertvergleichen sowohl das Vorliegen (partieller) metrischer also auch (partieller) skalarer Invarianz. Da die Prüfung schrittweise („step-up“) erfolgt, wobei die Modelle unter anderem mittels χ^2-Differenzentests (Satorra-Bentler-korrigiertes χ^2, vgl. https://www.statmodel.com/chidiff.shtml, letzter Zugriff: 01.01.2016) miteinander verglichen werden, wird als Ergebnis immer die höchste Stufe vorliegender Invarianz mitgeteilt.[6]

Dimensionsvergleich Mittels des konfirmatorischen Ansatzes können die konkurrierenden ein- bzw. zweidimensionalen Modellierungen des jeweiligen Handlungsfeldes miteinander vergleichen werden. Entsprechend der dargestellten Kriterien (vgl. Tabelle 12.7) kann die jeweilige Modellpassung beurteilt werden. Zusätzlich wird ein χ^2-Differenzentest durchgeführt, um die jeweils bessere Modellpassung zu prüfen.

12.6.3. Raschanalyse

Eine Raschanalyse kann nicht dazu verwendet werden, Argumente für die Eindimensionalität der Skalen zu sammeln, da diese eine Voraussetzung für die Anwendung der Methode ist: „Rasch measurement is based on the requirement that an instrument, as a set of items, must focus on a single trait“ (Boone u. a., 2014, S. 279). Vielmehr wird geprüft, ob das Rasch-Modell für die spezifischen Daten gilt. „[T]he concept of fit must be considered hand-in-hand with that of unidimensionality. [...] The benefits and properties of Rasch measurement exist only to the extent that the data fit the model's demanding requirements“ (Bond & Fox, 2007, S. 235). Zur Prüfung der Annahme, dass das Raschmodell gilt, werden bestimmte Parameter des Item- und Personenfit betrachtet und auf ihre Angemessenheit beurteilt.

Im vorliegenden Fall basieren alle Parameter auf einem Ratingskalen-Modell, weil allen Items dieselbe Antwortskala zugrunde gelegt wird und auch keine halben Punkte vergeben werden.

Beurteilung der Modellpassung Die Beurteilung, ob das Raschmodell auf die Daten passt, kann insbesondere anhand der Infit- und Outfit-MNSQ beurteilt werden, die

[6] Für eine Skala, für die (partielle) skalare Messinvarianz berichtet wird, liegt demnach mindestens auch partielle metrische Invarianz vor.

sowohl auf Item- als auch auf Personenebene angegeben werden. Boone u. a. (2014, S. 165 ff.) empfehlen, jeweils die Outfit-Statistiken als Kriterium zu verwenden, da diese die Ausreißer besser berücksichtigten. Ein Wert nahe Eins zeigt eine perfekte Modellpassung an. Werte zwischen .50 und 1.5 gelten als akzeptabel (vgl. Wright & Linacre, 1994), wobei Werte kleiner als .50 messtheoretisch eher unproblematisch sind, während größere Werte ($MNSQ > 2$) darauf hindeuten, dass das Modell nicht auf die Daten passt (vgl. Boone u. a., 2014, S. 167). Auf Personenebene werden in der Regel durchschnittliche MNSQ-Werte angegeben sowie der prozentuale Anteil der Personen, für die messtheoretisch problematische MNSQ-Werte ermittelt werden.

Reliabilität & Separation Im Rahmen der Prüfung auf Passung des Raschmodells werden verschiedene Reliabilitätsparameter geschätzt. Dazu zählen Personen- und Item-Separation sowie Personen- und Item-Reliabilität, wobei die Personen-Reliabilität mit dem klassischen Cronbach's Alpha vergleichbar ist (vgl. Tabelle 12.9). Alle Reliabilitätsmaße werden auf Grundlage einer Stichprobe geschätzt, die um extreme Antworten bereinigt wurde. Die entspricht einer konservativeren Schätzung und einer Empfehlung von Boone u. a. (2014, S. 221).

Tabelle 12.9.: Grenzwerte zur Beurteilung der Reliabilität (vgl. Linacre, 2014, S. 620-623).

	gut bis sehr gut
Personen-Separation	≥ 2.0
Item-Separation	≥ 3.0
Personen-Reliabilität	$\geq .80$
Item-Reliabilität	$\geq .80$

Qualität der Ratingskala Winsteps bietet mehrere Möglichkeiten, im Rahmen einer Raschanalyse die Qualität der verwendeten Ratingskala zu überprüfen und stellt diesbezüglich einige Tabellen zur Verfügung (z. B. Winsteps-Tabellen 2.2, 2.5, 3.2, 14.3). Tabelle 12.10 führt einige Kriterien an, die zur Beurteilung herangezogen werden können.

Eine zusätzliche, anschauliche Möglichkeit der Prüfung auf distinkte Antwortkategorien liefert das Diagramm zu Winsteps-Tabelle 3.2, das die Wahrscheinlichkeitsverteilung der Antwortkategorien für die jeweilige Skala darstellt. Jede Wahrscheinlichkeitskurve sollte ein unüberlagertes Maximum aufweisen, damit die Antwortkategorien als distinkt gelten (vgl. Bond & Fox, 2007, S. 224).

Messinvarianz: Mehrgruppenvergleiche/DIF-Analyse Im Zuge einer Raschanalyse kann ebenfalls eine Prüfung der Messinvarianz vorgenommen werden, wobei hier der

Tabelle 12.10.: Kriterien zur Beurteilung der Qualität der Ratingskala (Bond & Fox, 2007, S. 222 ff.).

Kriterium	**gut bis sehr gut**
Verteilung	möglichst regulär (Normal-/Gleichverteilung)
Kategorienbesetzung	hinreichend groß
durchschnittl. Kategorien-Raschwert	monoton steigend mit Anstieg der intendierten Kategorienschwierigkeit
Kategorien-Schwellwerte*	monoton steigend mit Anstieg der intendierten Kategorienschwierigkeit
Abstände der Kategorien-Schwellwerte	mindestens: 1.4 Logits, maximal: 5 Logits
O-MNSQ der Kategorien	≤ 2.0

* Rasch-Andrich Thresholds: „For rating-scale data, not only does each item have a difficulty estimate, but the scale also has a series of thresholds. This is the level at which the likelihood of being observed in a given response category (below the threshold) is exeeded by the likelihood of being observed in the next higher category (above the threshold)“ (Bond & Fox, 2007, S. 105 f.).

Ansatz der DIF-Analyse verfolgt wird. Letztlich wird geprüft, ob sich für verschiedene Gruppen die Reihenfolgen der nach ihrer Schwierigkeit geordneten Items signifikant voneinander unterscheiden (Boone u. a., 2014, S. 274). „DIF [...] models the invariance of item difficulty estimates across two or more samples“ (Bond & Fox, 2007, S. 231).

Zur Beurteilung können die Grenzwerte in Tabelle 12.11 herangezogen werden. Im Skalenhandbuch (Meinhardt u. a., 2016c) werden Scatterplots der Item-Schwierigkeiten kohortenspezifischer Raschanalysen dargestellt, die eine anschauliche Interpretation des DIF-Ansatzes liefern. Das Vorgehen orientiert sich an Empfehlungen von Boone u. a. (2014, S. 279-282).

Tabelle 12.11.: Grenzwerte zur Beurteilung von DIF (vgl. Linacre, 2014, S. 510).

gering bis moderat	**moderat bis groß**
$.64 > \Delta DIF \geq .43$	$\Delta DIF \geq .64$

Rasch-Hauptkomponentenanalyse Winsteps bietet nicht die Möglichkeit, mehrdimensionale Modelle miteinander zu vergleichen. Aber es kann eine Rasch-Hauptkomponentenanalyse der Residuen vorgenommen werden, um Indizien hinsichtlich möglicher zusätzlich im Datensatz enthaltener Dimensionen zu gewinnen. Auch wenn faktorenanalytische Verfahren verwendet werden, so ist die Interpretation im Vergleich zu klassischen Faktorenanalysen zu unterscheiden (vgl. Linacre, 2014, S. 516).

Dazu wird zunächst eine Raschanalyse durchgeführt, wobei alle Items der Planungs- und Durchführungsdimension im Modell enthalten sind. Die dabei „entstehenden“

Residuen werden dann einer Hauptkomponentenanalyse unterzogen. Laden einige Residuen substantiell auf einen weiteren Faktor (vgl. 1. Hauptkomponente, 1. Kontrast), so kann dies als Indiz für zusätzliche Dimensionen im Datensatz gewertet werden (vgl. Bond & Fox, 2007, S. 253). Es ist zu beachten, dass die Interpretation der Rasch-Hauptkomponentenanalyse nicht einfach wie bei klassischen faktorenanalytischen Verfahren vorgenommen werden kann (vgl. Linacre, 2014, S. 516).

Folgende Indizien sprechen für die Existenz zusätzlicher Strukturen im Datensatz: (1) Eigenwert der ersten Hauptkomponente (1. Kontrast) ist größer als Zwei, (2) Korrelation potentieller Itemcluster (Minderungskorrektur) ist kleiner als .57 (vgl. Linacre, 2014, S. 370 f.).

Zur grafischen Analyse stehen die Winsteps-Kontrastplots sowie die kohortenspezifischen Scatterplots der Personen-Raschwerte der Planungs- und Durchführungsdimension im Skalenhandbuch (vgl. Meinhardt u. a., 2016c) zur Verfügung.

12.6.4. Korrelationsanalysen

Um Zusammenhänge zu anderen Konstrukten, aber auch die Zusammenhänge der neu konstruierten Skalen untereinander zu analysieren, werden Korrelationen der Skalenwerte berechnet. Als Maß wird Spearman's ρ berichtet, um Verteilungsaspekte und das ggf. vorliegende Ordinalskalenniveau zu berücksichtigen (vgl. Bortz, 2005, S. 232 f.; Rasch, Friese, Hofmann & Naumann, 2006, S. 142 f.). In Anlehnung an Cohen nennen Rasch u. a. (2006) die in Tabelle 12.12 dargestellten Grenzwerte zur Beurteilung der Höhe von Korrelationskoeffizienten.

Tabelle 12.12.: Grenzwerte zur Beurteilung der Größe einer Korrelation (vgl. Rasch, Friese, Hofmann & Naumann, 2006, S. 133).

geringe Korrelation	**mittlere Korrelation**	**große Korrelation**
$r \leq .10$	$.10 < r \leq .30$	$r \geq .30$

12.6.5. Mittelwertvergleiche

Mittelwertvergleiche werden im Zuge des konfirmatorischen Ansatzes vorgenommen (vgl. Byrne, 2012, Kapitel 8). Brown (2006, S. 289) weist auf die Vorteile hin, die dieses Vorgehen gegenüber klassischen t-Tests oder der Berechnung einer ANOVA hat. Dazu zählt, dass (partielle) Messinvarianzen direkt bei der Schätzung der Mittelwerte berücksichtigt werden, während bei klassischen Analysen zumeist einfach von der Messinvarianz ausgegangen wird. Darüber hinaus sind die Mittelwerte messfehlerkorrigiert, sodass von einer größeren Präzision und statistischen Aussagekraft ausgegangen werden kann. Bei der Interpretation der Ergebnisse muss berücksichtigt werden, dass aufgrund

des Schätzverfahrens der Mittelwert einer Referenzgruppe auf Null fixiert wird und daher nur Mittelwertdifferenzen zweier Gruppen interpretiert werden können (vgl. Byrne, 2012, S. 251). Da keine Hypothesen bezüglich der konkreten Ausprägung der Mittelwerte vorliegen, die es zu prüfen gilt, ist dieses Vorgehen besonders geeignet.

12.7. Ergebnisse

Die im Fragebogen eingesetzten Skalen/Items wurden einer Vielzahl von Analysen unterzogen, deren Ergebnisse in der Regel in der bereits publizierten Skalendokumentation ausführlich dokumentiert wurden (vgl. Meinhardt u. a., 2016c). Einzelne, kohortenspezifische Ergebnisse der Korrelationsanalysen und Mittelwertvergleiche finden sich bei Meinhardt, Rabe und Krey (2015, 2016a, 2016b). Die Ergebnisse der Analysen bildeten die Basis eines weiteren Überarbeitungszyklus, in dessen Rahmen jedoch nur einzelne Items einer Revision unterzogen wurden. In wenigen Fällen wurden Items aus den Skalen entfernt, wobei es das Ziel war, *keine* kohortenspezifischen Skalen zu entwickeln, sondern *ein* Instrumentarium für alle drei Befragungsgruppen. In solchen Fällen wurden die derart revidierten Modelle erneut den bereits beschriebenen Analysen unterzogen. Die Überarbeitungen wurden detailliert bei Meinhardt u. a. (2016c, S. 37, 87, 167, 215, 277, 326, 404, 452) dargelegt und werden daher hier nicht noch einmal im Einzelnen aufgegriffen. Die zur Verwendung empfohlenen Skalen, die nun insgesamt eine dreifache Überarbeitung erfahren haben, sind dem Anhang G.4 zu entnehmen. Nachfolgend sollen exemplarisch solche Ergebnisse berichtet werden, die mit Blick auf die gestellten Forschungsfragen interpretiert werden können (vgl. Abschnitt 12.8).

12.7.1. Eindimensionalität der Skalen

Neu konstruierte Skalen Details zu den Ergebnissen der konfirmatorischen Faktorenanalysen (spezifizierte/revidierte Modelle) können der Skalendokumentation entnommen werden (vgl. Meinhardt u. a., 2016c, S. 27, 42, 77, 92, 157, 172, 206, 267, 281, 316, 330, 394, 409, 443, 456). Eine Zusammenfassung enthält Tabelle 12.13. Die Modellanpassungsgüten der spezifizierten Messmodelle sind durchweg akzeptabel; lediglich für die Skala SWE-SV-D (Referendare) ist das Modell zu verwerfen (vgl. Tabelle 12.13). Für die revidierten Modelle verbessern sich die Modellanpassungsgüten – diese können als akzeptabel bis sehr gut bezeichnet werden. Die Faktorladungen liegen in der Regel für jedes spezifizierte Modell und jede Kohorte im akzeptablen bis sehr guten Bereich. Nur in sehr wenigen Ausnahmefällen sind die Faktorladungen zu gering und werden daher nicht als misfit interpretiert. Gleiches gilt für die Faktorladungen nach der Revision (vgl. Meinhardt u. a., 2016c).

Tabelle 12.13.: Modellfit für die spezifizierten und revidierten Messmodelle.

	Modelle	entf.	χ^2	df	$\frac{\chi^2}{df}$	p	CFI	TLI	RMSEA [90% CI]	SRMR
Studierende	SWE-EX-P spez.		98.34	35	2.81	.000	.954	.941	.059 [.045; .073]	.038
	SWE-EX-P rev.	exp4/5	49.18	20	2.46	.000	.971	.959	.053 [.034; .072]	.032
	SWE-EX-D spez.		83.98	27	3.11	.000	.950	.933	.063 [.048; .079]	.036
	SWE-EX-D rev.	exd7	59.23	20	2.96	.000	.959	.942	.061 [.043; .079]	.035
	SWE-EL-P spez.		82.03	27	3.04	.000	.954	.939	.062 [.047; .078]	.035
	SWE-EL-P rev.	elp1	40.68	20	2.03	.004	.978	.969	.044 [.024; .064]	.030
	SWE-EL-D spez.		21.60	9	2.40	.010	.982	.970	.052 [.024; .080]	.026
	SWE-A-P spez.		67.08	27	2.48	.000	.957	.943	.053 [.037; .069]	.033
	SWE-A-P rev.	ap2	34.44	20	1.72	.023	.981	.973	.037 [.014; .058]	.027
	SWE-A-D spez.		102.78	27	3.81	.000	.910	.880	.073 [.058; .088]	.045
	SWE-A-D rev.	ad8	68.73	20	2.38	.000	.928	.900	.068 [.051; .086]	.040
	SWE-SV-P spez.		42.01	20	2.10	.003	.974	.963	.046 [.026; .065]	.031
	SWE-SV-P rev.	svp3	21.10	14	1.51	.099	.989	.984	.031 [.000; .057]	.025
	SWE-SV-D spez.		49.14	14	3.51	.000	.954	.931	.069 [.049; .091]	.036
	SWE-SV-D rev.	svd7	13.57	9	1.51	.139	.993	.988	.031 [.000; .063]	.020
Referendare	SWE-EX-P spez.		77.71	35	2.22	.000	.920	.897	.072 [.050; .093]	.052
	SWE-EX-P rev.	exp4/5	21.61	20	1.08	.362	.995	.993	.018 [.000; .060]	.034
	SWE-EX-D spez.		61.56	35	1.76	.000	.935	.913	.073 [.049; .098]	.047
	SWE-EX-D rev.	exd7	29.07	20	1.45	.086	.979	.970	.044 [.000; .076]	.036
	SWE-EL-P spez.		70.79	27	2.62	.000	.919	.892	.083 [.059; .106]	.050
	SWE-EL-P rev.	elp1	39.05	20	1.95	.007	.958	.941	.063 [.033; .093]	.040
	SWE-EL-D spez.		11.02	9	1.22	.275	.992	.987	.031 [.000; .083]	.029
	SWE-A-P spez.		55.14	27	1.58	.001	.946	.928	.066 [.041; .091]	.045
	SWE-A-P rev.	ap2	45.73	20	2.29	.001	.938	.914	.074 [.045; .102]	.046
	SWE-A-D spez.		57.91	27	2.15	.001	.921	.895	.069 [.045; .094]	.050
	SWE-A-D rev.	ad8	34.62	20	1.73	.022	.954	.935	.055 [.021; .086]	.042
	SWE-SV-P spez.		51.95	20	2.60	.000	.934	.908	.082 [.055; .110]	.047
	SWE-SV-P rev.	svp3	24.20	14	1.73	.043	.973	.959	.055 [.010; .092]	.036
	SWE-SV-D spez.		51.22	14	3.66	.000	.890	.835	.106 [.076; .137]	.054
	SWE-SV-D rev.	svd7	20.16	9	2.24	.017	.955	.925	.072 [.029; .115]	.039
Lehrkräfte	SWE-EX-P spez.		71.04	35	2.03	.000	.938	.920	.078 [.052; .104]	.047
	SWE-EX-P rev.	exp4/5	34.90	20	1.75	.009	.954	.935	.073 [.036; .108]	.042
	SWE-EX-D spez.		57.44	27	2.13	.007	.939	.919	.082 [.052; .111]	.045
	SWE-EX-D rev.	exd7	40.89	20	2.05	.004	.948	.927	.079 [.044; .113]	.044
	SWE-EL-P spez.		41.71	27	1.45	.035	.949	.932	.057 [.015; .089]	.045
	SWE-EL-P rev.	elp1	19.36	20	.97	.499	1.000	1.004	.000 [.000; .064]	.038
	SWE-EL-D spez.		19.68	9	2.19	.020	.951	.919	.084 [.032; .135]	.045
	SWE-A-P spez.		47.62	27	1.76	.009	.948	.931	.067 [.034; .098]	.048
	SWE-A-P rev.	ap2	26.23	20	1.01	.158	.979	.970	.043 [.000; .084]	.044
	SWE-A-D spez.		42.60	27	1.58	.028	.941	.921	.059 [.019; .091]	.050
	SWE-A-D rev.	ad8	29.68	20	1.46	.075	.958	.941	.054 [.000; .092]	.046
	SWE-SV-P spez.		21.35	20	1.07	.377	.995	.994	.020 [.000; .071]	.039
	SWE-SV-P rev.	svp3	11.29	14	.81	.663	1.000	1.017	.000 [.000; .061]	.032
	SWE-SV-D spez.		42.97	14	3.07	.000	.914	.871	.111 [.074; .150]	.052
	SWE-SV-D rev.	svd7	18.72	9	2.08	.028	.962	.937	.080 [.026; .132]	.038

Zusätzliche Skalen Die Modellanpassungsgüten für die zusätzlich erhobenen Skalen sind teils ungenügend (vgl. Tabelle 12.14), weshalb die Modelle teils größeren Revisionen unterzogen werden müssen. Nach der Revision sind die Modellanpassungsgüten akzeptabel bis sehr gut.

12.7.2. Exkurs: Raschanalyse

„We are not suggesting that Rasch modelling is the universal panacea for measurement in the human sciences. Nothing replaces thoughtful theory-driven research."

Bond & Fox, 2007, S. 138

Die vollständigen Parameter der Raschanalyse finden sich im Skalenreport (vgl. Meinhardt u. a., 2016c, S. 29, 44, 79, 94, 159, 174, 208, 269, 283, 318, 332, 396, 411, 445, 458). Insgesamt kann konstatiert werden, dass die Werte des Item-Outfit-MNSQ jeweils (vor und nach der Revision) im geforderten Bereich liegen ($.5 <$ Item-O-MNSQ < 1.5) bzw. um den Wert 1 streuen. Für den Personen-Outfit-MNSQ kann festgestellt werden, dass jeweils weniger als 10% aller Teilnehmenden einen Wert größer als Zwei zugewiesen bekommen. Insgesamt kann die Modellpassung damit für jede Kohorte (vor und nach der Revision) bestätigt werden.

12.7.3. Zuverlässigkeit der Skalen

„Reliability analysis must be a part of any assessment development or use."

Boone u. a., 2014, S. 218

Inter-Item-Korrelationen Die durchschnittlichen Inter-Item-Korrelationen (dIK) der Skalen liegen für die Kohorte der Studierenden zwischen .31 und .41 (nach der Revision: $.30 \leq dIK \leq .41$), für die Kohorte der Referendare zwischen .31 und .41 (nach der Revision: $.32 \leq dIK \leq .40$) und für die Kohorte der Lehrpersonen zwischen .31 und .46 (nach der Revision: $.33 \leq dIK \leq .46$) (vgl. Meinhardt u. a., 2016c, S. 25, 40, 75, 90, 155, 170, 204, 265, 279, 314, 328, 392, 407, 441, 454).

Itemtrennschärfen Die Item-Trennschärfen der Skalen (vgl. Tabelle 12.15) liegen für die Kohorten vor und nach der Revision überwiegend im mittleren Bereich (vgl. Meinhardt u. a., 2016c, S. 26, 41, 76, 91, 156, 171, 205, 266, 280, 315, 329, 393, 408, 442, 455).

Tabelle 12.14.: Modellfit für die spezifizierten und revidierten Messmodelle der zusätzlichen Skalen.

	Modelle	entf.	χ^2	df	$\frac{\chi^2}{df}$	p	CFI	TLI	RMSEA [90% CI]	SRMR
Studierende	A-SWE spez.		164.07	35	4.69	.000	.902	.874	.084 [.071; .097]	.049
	A-SWE rev.	asw4	92.01	27	3.41	.000	.941	.921	.068 [.053; .083]	.040
	L-SWE spez.		86.37	35	2.44	.000	.879	.845	.053 [.039; .067]	.043
	L-SWE rev.	lsw5/9	32	20	1.62	.039	.956	.938	.034 [.008; .055]	.032
	BFI-K (E) spez.		19.40	2	9.70	.000	.969	.907	.129 [.081; .184]	.035
	BFI-K (V) spez.		10.20	2	5.10	.006	.963	.889	.088 [.040; .145]	.023
	BFI-K (G) spez.		4.25	2	2.13	.120	.995	.984	.046 [.000; .109]	.013
	BFI-K (N) spez.		1.09	2	.55	.580	1.000	1.008	.000 [.000; .072]	.008
	BFI-K (O) spez.		44.29	2	22.15	.000	.909	.819	.122 [.091; .157]	.063
	BFI-K (O) rev.	bfi21	1.83	2	.92	.400	1.000	1.003	.000 [.000; .084]	.013
	K-I spez.		81.63	20	4.08	.000	.869	.816	.077 [.060; .095]	.048
	K-I rev.	ki2/4/8	7.37	5	1.47	.195	.988	.976	.030 [.000; .073]	.021
	PS spez.		108.07	14	7.72	.000	.917	.876	.113 [.094; .133]	.047
	PS rev	ps1	26.88	9	2.99	.002	.978	.963	.062 [.036; .089]	.028
Referendare	A-SWE spez.		74.85	35	2.14	.000	.915	.891	.069 [.048; .091]	.052
	A-SWE rev.	asw4	42.85	27	1.59	.027	.959	.945	.050 [.017; .077]	.043
	LSWE spez.		50.49	35	1.44	.044	.902	.874	.042 [.007; .068]	.051
	L-SWE rev.	lsw5/9	26.48	20	1.32	.151	.948	.927	.037 [.000; .071]	.046
	BFI-K (E) spez.		2.83	2	1.41	.243	.995	.984	.042 [.000; .142]	.021
	BFI-K (V) spez.		Modell problematisch							
	BFI-K (G) spez.		7.56	2	3.78	.023	.969	.906	.108 [.034; .195]	.031
	BFI-K (N) spez.		2.06	2	1.03	.358	1.000	.999	.011 [.000; .130]	.016
	BFI-K (O) spez.		17.67	2	8.84	.003	.925	.85	.103 [.054; .158]	.065
	BFI-K (O) rev.	bfi21	.675	2	.34	.713	1.000	1.070	.000 [.000; .093]	.012
	K-I spez.		50.51	20	2.53	.000	.834	.767	.080 [.053; .108]	.062
	K-I rev.	ki2/4	13.92	9	1.55	.125	.962	.937	.048 [.000; .095]	.036
Lehrkräfte	A-SWE spez.		49.02	35	1.40	.058	.957	.944	.049 [.000; .079]	.050
	L-SWE spez.		61.24	35	1.75	.004	.890	.859	.067 [.038; .094]	.061
	L-SWE rev.	lsw9	42.05	27	1.56	.033	.931	.908	.058 [.017; .090]	.053
	BFI-K (E) spez.		2.59	2	1.29	.275	.995	.985	.042 [.000; .165]	.020
	BFI-K (V) spez.		8.97	2	4.48	.011	.916	.748	.144 [.058; .246]	.028
	BFI-K (G) spez.		.15	2	.08	.928	1.000	1.063	.000 [.000; .048]	.006
	BFI-K (N) spez.		.05	2	.027	.974	1.000	1.054	.000 [.000; .000]	.003
	BFI-K (O) spez.		9.89	2	4.95	.078	.945	.890	.076 [.000; .146]	.060
	BFI-K (O) rev.	bfi21	.02	2	.01	.988	1.000	1.440	.000 [.000; .000]	.003
	K-I spez.		50.40	20	2.52	.000	.851	.791	.095 [.063; .128]	.065
	K-I rev.	ki2/4	10.30	9	1.14	.327	.991	.986	.029 [.000; .095]	.035

Tabelle 12.15.: Itemtrennschärfen (vor der Revision/nach der Revision).

Skala	Studierende	Referendare	Lehrkräfte
SWE-EX-P	.45-.69/.44-.65	.49-.62/.48-.60	.46-.70/.43-.67
SWE-EX-D	.50-.64/.50-.65	.42-.63/.42-.65	.49-.70/.49-.68
SWE-EL-P	.50-.64/.48-.61	.53-.63/.54-.63	.46-.64/.43-.66
SWE-EL-D	.49-.59/–	.53-.58/–	.40-.63/–
SWE-A-P	.43-.60/.42-.57	.44-.65/.42-.61	.46-.76/.47-.62
SWE-A-D	.42-.55/.42-.56	.40-.54/.40-.53	.40-.61/.41-.59
SWE-SV-P	.50-.61/.46-.60	.51-.64/.49-.62	.57-.67/.55-67
SWE-SV-D	.46-.60/.50-.59	.43-.59/.44-.60	.59-.65/.54-.64

Interne Konsistenz Die Schätzungen für die Skalenreliabilitäten können Tabelle 12.16 entnommen werden (vgl. Meinhardt u. a., 2016c, S. 26, 41, 76, 91, 156, 171, 205, 266, 280, 315, 329, 393, 408, 442, 455).

Tabelle 12.16.: Interne Konsistenz α_C (vor der Revision/nach der Revision).

Skala	# Items*	Studierende	Referendare	Lehrkräfte
SWE-EX-P	10/8	.87/.84	.85/.82	.89/.86
SWE-EX-D	9/8	.85/.83	.84/.83	.88/.86
SWE-EL-P	9/8	.85/.83	.85/.84	.82/.80
SWE-EL-D	6/6	.80/–	.79/–	.78/–
SWE-A-P	9/8	.83/.80	.85/.83	.86/.84
SWE-A-D	9/8	.80/.77	.81/.79	.80/.79
SWE-SV-P	8/7	.83/.81	.85/.82	.86/.84
SWE-SV-D	7/6	.81/.80	.80/.78	.85/.83

* #: Anzahl

Reliabilitätsmaße des konfirmatorischen Ansatzes 45% aller Faktorladungen vor und nach der Revision erreichen den Grenzwert von $\lambda^2 \geq .4$ für eine sehr gute Indikatorreliabilität (vgl. Meinhardt u. a., 2016c, S. 28, 43, 78, 93, 158, 173, 207, 282, 268, 317, 331, 395, 410, 444, 457). In der Regel stimmen die geschätzten Werte für die Faktorreliabilität (vor und nach der Revision) mit den Werten für Cronbach's Alpha überein (vgl. Tabelle 12.16). Die durchschnittlich extrahierte Varianz (DEV) liegt für die Skalen (vor und nach der Revision) im unteren Bereich ($.30 \leq DEV \leq .46$).

Reliabilitätsmaße des probabilistischen Ansatzes Alle Reliabilitätsberechnungen im Rahmen der Raschanalyse können bei Meinhardt u. a. (2016c, S. 32, 47, 82, 97,

162, 177, 210, 272, 286, 321, 335, 399, 415, 447, 461) eingesehen werden. Das Rasch-Pendant zu Cronbach's Alpha – die Personen-Reliabilität – fällt aufgrund der Wahl des konservativeren Schätzers (REAL-Wert als untere Grenze der Reliabilität) sowohl vor als auch nach der Revision etwas niedriger aus als α_C. Die Item-Reliabilität liegt vor und nach der Revision zwischen .84 und .99. Der Personen-Separationsindex liegt vor der Revision zwischen 1.70 und 2.50 und nach der Revision zwischen 1.72 und 2.17. Der Item-Separationsindex nimmt vor der Revision Werte zwischen 2.26 und 9.53 und nach der Revision Werte zwischen 2.31 und 9.57 an.

Reliabilität der zusätzlichen Skalen Die Reliabilitätsmaße für die zusätzlich erhobenen (revidierten) Skalen können Tabelle 12.17 entnommen werden.

Tabelle 12.17.: Interne Konsistenz α_C der zusätzlichen Skalen (vor der Revision/nach der Revision).

Skala	# Items*	Studierende	Referendare	Lehrkräfte
A-SWE	10/9	.85/.84	.82/.81	.82/–
L-SWE	10/8 bzw. 9	.65/.61 (8)	.67/.64 (8)	.76/.72 (9)
PS	7/6	.83/.80	–	–
K-I	8/6 bzw. 5	.71/.61 (5)	.67/.68 (6)	.70/.72 (6)
BFI-K (E)	5	.78	.76	.71
BFI-K (V)	5	.58	.52	.58
BFI-K (G)	5	.77	.78	.71
BFI-K (N)	5	.74	.73	.74
BFI-K (O)	6/5	.70/.60	.69/.58	.59/.44

* #: Anzahl

12.7.4. Eignung des Antwortformates

Detaillierte Informationen zur Qualität der Ratingskalen können im Skalenreport eingesehen werden (vgl. Meinhardt u. a., 2016c, S. 33-36, 48-51, 83-86, 98-101, 163-166, 178-181, 211-214, 273-276, 287-290, 322-325, 336-339, 400-403, 416-419, 448-451, 462-465). Zusammenfassen werden nachfolgend wesentliche Ergebnisse berichtet.

Verteilung der Antwortkategorien Die Verteilungen der Kategorienhäufigkeiten sind jeweils leicht linksschief, da die Schiefe i. d. R. kleiner als Null ist (vgl. Meinhardt u. a., 2016c, S. 20, 70, 150, 200, 260, 309, 388, 437), d. h. Kategorien mit dem Zahlenwert größer als „3“ wurden etwas häufiger gewählt. Nichtsdestotrotz können die Abweichungen von der Normalverteilung als moderat eingestuft werden, sodass die Verteilungen der Antworten keine großen Irregularitäten aufweisen.

Kategorienbesetzung Für einen optischen Eindruck der Besetzungsstärke der Antwortkategorien eigenen sich die Histogramme im Skalenreport (vgl. Meinhardt u. a., 2016c, S. 21 ff., 71 ff., 151 ff., 201 f., 261 ff., 310 ff., 389 f., 438 f.). Die exakten Häufigkeitsverteilungen können den Tabellen bezüglich der Charakteristika der Antwortkategorien entnommen werden (vgl. Meinhardt u. a., 2016c, S. 33 f., 48 f., 83 f., 98 f., 163 f., 178 f., 211 f., 273 f., 287 f., 322 f., 336 f., 400 f., 416 f., 448 f., 462 f.). Diese bestätigen zunächst den Eindruck, dass für jedes Item in der Regel alle Kategorien besetzt sind. In einigen Fällen ist Antwortkategorie „1“ (stimmt nicht) nicht besetzt, was insbesondere für die Gruppe der Lehrpersonen zutrifft, da diese einerseits die wenigsten Teilnehmerinnen und Teilnehmer aufweist und andererseits in dieser Kohorte auch eher hohe spezifische Lehrer-Selbstwirksamkeitserwartungen erwartet werden (vgl. Hypothese 11.3 (1), Unterabschnitt 8.1.2).[7] Die Häufigkeitsverteilungen bezüglich der Kategorien für die jeweiligen (revidierten) Gesamtskalen sind ebenfalls dem Skalenreport zu entnehmen (vgl. Meinhardt u. a., 2016c, S. 36, 51, 86, 101, 166, 181, 214, 276, 290, 325, 339, 403, 419, 451, 465).

Kategorien: Durchschnittliche Raschwerte und Outfit-MNSQ Es kann festgestellt werden, dass der durchschnittliche Raschwert pro Antwortkategorie sowohl auf Item- als auch auf Skalenebene mit zunehmendem Zahlenwert der Antwortkategorie ebenfalls zunimmt (d. h. die durchschnittliche Personenfähigkeit/Lehrer-Selbstwirksamkeitserwartung steigt). Der Outfit-MNSQ für die Kategorien ist sowohl auf Item- als auch auf Skalenebene (vor und nach der Revision) jeweils kleiner als 2.0 und schwankt um den Wert 1.0.

Schwellwerte Auf Skalenebene steigt mit zunehmendem Zahlenwert der Antwortkategorie ebenfalls der Schwellenwert (Andrich-Threshold), der die Personenfähigkeit/Lehrer-Selbstwirksamkeitserwartung angibt, bei der eine 50/50-Chance für die jeweilige Kategorie vorliegt. Die Abstände zwischen den Andrich-Thresholds sind für die äußeren Kategorien jeweils größer als 1.4 Logits, zwischen den Kategorien „3“ und „4“ sowie zwischen den Kategorien „4“ und „5“ liegen die Abstände in der Regel bei 1.0 Logit. Aus den Darstellungen der Wahrscheinlichkeitskurven für die einzelnen Kategorien der Skalen lassen sich jeweils distinkte Peaks für jede Kategorie ablesen (vgl. Meinhardt u. a., 2016c, S. 36, 51, 86, 101, 166, 181, 214, 276, 290, 325, 339, 403, 419, 451, 465).

12.7.5. Messinvarianz

Itemschwierigkeiten Aus den qualitativen Studien ergab sich kein Hinweis darauf, dass Items kohortenspezifisch verstanden bzw. dass die Itemschwierigkeiten kohor-

[7] vgl. z. B. Skala SWE-EX-P: Für die Items exp2 und exp6 ist in der Gruppe der Lehrpersonen Kategorie „1“ unbesetzt.

tenspezifisch eingeschätzt wurden. Auch statistisch kann geprüft werden, ob Items kohortenspezifisch als unterschiedlich schwer wahrgenommen werden. Zunächst ist ein einfacher Vergleich der Item-Mittelwerte/Raschwerte (vgl. Meinhardt u. a., 2016c) dahingehend möglich, ob einerseits die Itemschwierigkeiten mit den Vorhersagen (vgl. Abschnitt 10.4.5, Tabelle 10.1) übereinstimmen und ob diese kohortenspezifisch abweichen. Das Ergebnis kann Tabelle 12.18 entnommen werden.

Tabelle 12.18.: Vergleich der Hypothesen bez. der Itemschwierigkeiten (Skalenversion 3). Hervorhebungen: Item kohortenübergreifend unter den beiden leichtesten/schwersten Items.

	Hypothesen		**Vergleich mit emp. MW**	
Skala	**eher leicht**	**eher schwer**	**eher leicht**	**eher schwer**
SWE-EX-P	exp2, exp6	exp3, exp5, exp8	exp2, exp6, **exp7**	**exp3, exp5**, exp9
SWE-EX-D	exd2	exd3, exd5, exd7	**exd2**	**exd5**, exd7
SWE-EL-P	elp2, elp4	elp3, elp9	**elp1**, elp4	**elp9**
SWE-EL-D	eld4	eld6	**eld4**	**eld6**
SWE-A-P	ap4, ap6	ap3	ap4, ap6, ap8	ap1, **ap2, ap3**
SWE-A-D	ad2, ad8	ad5, ad7	**ad2, ad8**	**ad5, ad7**
SWE-SV-P	svp6, svp7	svp4, svp8	**svp6, svp7**	**svp2**
SWE-SV-D	svd7	svd1, svd6	**svd7**	**svd1**

Mehrgruppenvergleiche/DIF-Analysen Die Prüfungen der Messinvarianz sind ausführlich im Skalenreport dokumentiert (vgl. Meinhardt u. a., 2016c, S. 52-67, 102-117, 182-197, 216-229, 291-306, 340-355, 420-434, 466-479), wobei zu beachten ist, dass diese lediglich für die revidierten Modelle dokumentiert und mehrere methodische Herangehensweisen genutzt wurden (Mehrgruppenvergleiche/konfirmatorische Faktorenanalysen und Mehrgruppenvergleiche/DIF-Analysen). Einen Überblick über die Ergebnisse für den Personengruppenvergleich beider Verfahren enthält Tabelle 12.19. Aus dieser geht hervor, dass beide Verfahren in der Regel ähnliche Ergebnisse liefern. Für jede Skala liegt mindestens partielle skalare Messinvarianz vor.

Zusätzlich wurden Mehrgruppenvergleiche für verschiedene Subgruppen vorgenommen. Diese sind ebenfalls ausführlich in der Skalendokumentation dargestellt (vgl. Meinhardt u. a., 2016c). Zu beachten sind jedoch die teilweise gering besetzten Gruppen ($N \approx 70$). Beide Analyseverfahren kommen wiederum häufig zu sehr ähnlichen Ergebnissen. Für das Unterscheidungskriterium „Studiengang“ kann jeweils (partielle) skalare Messinvarianz konstatiert werden. Gleiches gilt für das Kriterium „Geschlecht“. Wird für die Gruppe der Studierenden das Instrument auf Messinvarianz hinsichtlich des Kriteriums „Praxiserfahrung“ geprüft, so liegt in der Regel (partielle) skalare Invarianz vor. In einem Fall konnte keine Messinvarianz nachgewiesen werden (SWE-EX-D). Für die

Tabelle 12.19.: Übersicht Mehrgruppenvergleiche/DIF-Analysen: Personengruppen.

		MGV			**DIF-Analyse**
Modell	**Gruppen**[1]	Typ[2]	λ[3]	[Intercept][4]	Gruppen-Vergleich[5]
SWE-EX-P	S, R, L	mMI, p_sMI	–	[exp10], [exp2], [exp7], [exp6]	SL(exp2/exp1/exp6*), RL(exp6)
SWE-EX-D	S, R, L	mMI, p_sMI	–	[exd8], [exd9];	SL(exd3/exd8*), RL(exd8)
SWE-EL-P	S, R, L	mMI, p_sMI	–	[elp9], [elp2], [elp5], [elp4]	SR (elp9), SL (elp2/elp5), RL (elp4*/elp5/elp8)
SWE-EL-D	S, R, L	mMI, sMI	–	–	–
SWE-A-P	S, R, L	p_mMI, p_sMI	ap6	[ap4], [ap6], [ap8]	SL(ap8/ap4*), RL(ap6/ap4)
SWE-A-D	S, R, L	mMI, p_sMI	–	[ad6], [ad3], [ad5], [ad4]	SL(ad6/ad7), RL(ad7/ad5)
SWE-SV-P	S, R, L	mMI, p_sMI	–	[svp1]	SL(svp7)
SWE-SV-D	S, R, L	mMI, p_sMI	–	[svd2]	SL(svd2)

[1] S: Studierende, R: Referendare, L: Lehrkräfte
[2] p_mMI: partielle metrische Messinvarianz, mMI: metrische Messinvarianz, p_sMI: partielle skalare Messinvarianz, sMI: skalare Messinvarianz
[3] frei zu schätzende Faktorladung von Item ...
[4] frei zu schätzendes Intercept von Item ...
[5] SR: Vergleich Studierende/Referendare, SL: Vergleich Studierende/Lehrkräfte, RL: Vergleich Referendare/Referendare, *$\Delta DIF \geq .64$

Gruppe der Lehrkräfte kann für jede Skala (partielle) skalare Messinvarianz bezüglich des Kriteriums „Schulform“ festgestellt werden.

Für die zusätzlich erhobenen, revidierten Skalen kann ebenfalls jeweils (partielle) skalare Messinvarianz nachgewiesen werden (vgl. Tabellen im Anhang O.1). Teils müssen jedoch sehr viele Intercepts frei geschätzt werden (je fünf Intercepts für die Skalen A-SWE und L-SWE). Um Mittelwertvergleiche durchführen zu können, wurde für einen Teil der zusätzlichen Skalen (A-SWE- und L-SWE-Skala) die Messinvarianz auch für die genannten Subgruppen überprüft. Wiederum ergibt sich jeweils (partielle) skalare Messinvarianz (vgl. Anhang O.2 bis O.4), wobei eine Ausnahme für die Kohorte der Referendare und das Kriterien „Schulform“ bezüglich der L-SWE-Skala vorliegt (vgl. Anhang O.3).

12.7.6. Dimensionsanalyse

Konfirmatorische Faktorenanalyse (Dimensionsvergleich) Für jede der neu konstruierten Skalen konnten gute bis sehr gute Modellanpassungsgüten für die revidierten Modelle festgestellt werden (vgl. Tabelle 12.13). In einem zweiten Schritt soll geprüft werden, ob die jeweiligen zwei Skalen eines Handlungsfeldes tatsächlich unterschiedliche Konstrukte abbilden. Dazu werden eindimensionale Modelle (Planungs- und Durchführungsdimension zusammen bilden jeweils die Konstrukte SWE-EX, SWE-EL, SWE-A, SWE-SV) mit zweidimensionalen Modellen (Planungs- und Durchführungsdimension korrelieren miteinander) verglichen (vgl. Meinhardt u. a., 2016c, S. 119-129, 231-240, 357-368, 481-488). Zunächst können die Vergleiche in einigen Fällen nicht ohne vorherige Anpassungen der Modelle für einzelne Kohorten durchgeführt werden. χ^2-Differenzentests sprechen in der Regel jeweils für die zweidimensionalen Modelle. Lediglich für das Handlungsfeld SWE-A (Referendare) und das Handlungsfeld SWE-EL (Lehrpersonen) passen beide Modelle gleich gut auf die Daten. Hinzu kommen für die zweidimensionalen Modelle in jeder Kohorte sehr hohe latente Korrelationen der Dimensionen (SWE-EX: $.89 \leq r \leq .96$, SWE-EL: $.88 \leq r \leq .97$, SWE-A: $.92 \leq r \leq .97$, SWE-SV: $.82 \leq r \leq .87$). Die angegebenen Korrelationen sind jedoch jeweils von eins verschieden.

Rasch-Hauptkomponentenanalyse Auch die Ergebnisse der Rasch-Hauptkomponentenanalyse sind detailliert im Skalenhandbuch einsehbar (vgl. Meinhardt u. a., 2016c, S. 119-129, 231-240, 357-368, 481-488). Zusammenfassend ergibt sich, dass die Eigenwerte der Hauptkomponentenanalyse der Rasch-Residuen für die Skalen zwischen 1.54 und 2.14 liegen. Die jeweiligen Minderungskorrekturen potentiell korrelierender Komponenten liegen für die Skalen in einem Bereich von .62 bis .78. Darüber hinaus ordnen sich die Faktorladungen der Residuen nicht stringent getrennt nach Planungs- und Durchführungsitems an.

Korrelationen untereinander Die Korrelationen der neu entwickelten (revidierten) Skalen können Tabelle 12.20 entnommen werden. Die Korrelationen zwischen den Skalen unterschiedlicher Dimensionen eines Handlungsfeldes sind geringer als die soeben berichteten latenten Korrelationen ($.640 \leq \rho \leq .808$). Für die Gruppe der Studierenden korrelieren die Skalen der Planungs- und Durchführungsdimension eines Handlungsfeldes jeweils am höchsten miteinander. Zusätzlich korrelieren Planungsskalen unterschiedlicher Handlungsfelder jeweils in einem höheren Maße untereinander als mit Durchführungsskalen anderer Handlungsfelder. Für die anderen beiden Kohorten trifft dieses beschriebene Korrelationsmuster in der Regel ebenfalls zu; es sind lediglich einzelne Abweichungen zu beobachten. Es wird darauf verzichtet, zwei Korrelationen auf signifikante Unterschiede zu prüfen.

Tabelle 12.20.: Kohortenspezifische Korrelationen (Spearman's ρ) zwischen den neu entwickelten (revidierten) Skalen.

					SWE-			
Studierende	EX-P	EX-D	EL-P	EL-D	A-P	A-D	SV-P	SV-D
SWE-EX-P	1.000							
SWE-EX-D	.745**	1.000						
SWE-EL-P	.667**	.608**	1.000					
SWE-EL-D	.594**	.646**	.732**	1.000				
SWE-A-P	.617**	.601**	.652**	.578**	1.000			
SWE-A-D	.561**	.627**	.608**	.641**	.735**	1.000		
SWE-SV-P	.623**	.534**	.657**	.567**	.631**	.568**	1.000	
SWE-SV-D	.578**	.631**	.595**	.695**	.571**	.624**	.663**	1.000
Referendare	EX-P	EX-D2	EL-P	EL-D	A-P	A-D	SV-P	SV-D
SWE-EX-P	1.000							
SWE-EX-D	.669**	1.000						
SWE-EL-P	.682**	.604**	1.000					
SWE-EL-D	.511**	.610**	.681**	1.000				
SWE-A-P	.703**	.498**	.711**	.582**	1.000			
SWE-A-D	.632**	.636**	.685**	.651**	.749**	1.000		
SWE-SV-P	.635**	.531**	.661**	.563**	.647**	.613**	1.000	
SWE-SV-D	.565**	.653**	.571**	.712**	.520**	.634**	.651**	1.000
Lehrkräfte	EX-P	EX-D	EL-P	EL-D	A-P	A-D	SV-P	SV-D
SWE-EX-P	1.000							
SWE-EX-D	.809**	1.000						
SWE-EL-P	.609**	.599**	1.000					
SWE-EL-D	.494**	.581**	.758**	1.000				
SWE-A-P	.609**	.542**	.644**	.483**	1.000			
SWE-A-D	.584**	.641**	.630**	.658**	.770**	1.000		
SWE-SV-P	.508**	.541**	.672**	.647**	.589**	.626**	1.000	
SWE-SV-D	.424**	.546**	.603**	.734**	.444**	.585**	.640**	1.000

12.7.7. Abgrenzbarkeit zu anderen Konstrukten

Da auch die zusätzlich erhobenen Skalen einer Prüfung auf Eindimensionalität unterzogen wurden, die teils zu einer Revision der Skalen führte (vgl. Tabelle 12.14), können für diese revidierten (als eindimensional angenommenen) Skalen nachfolgend auch Korrelationen mit den neu entwickelten Skalen berichtet werden (vgl. Tabelle 12.21).

Es ergibt sich das folgende Bild: Die neu entwickelten Skalen korrelieren mit der Skala zu allgemeinen Selbstwirksamkeitserwartungen für jede Kohorte im schwachen bis mittleren Bereich (Studierende: $.230^{**} \leq r \leq .375^{**}$, Referendare: $.409^{**} \leq \rho \leq .512^{**}$, Lehrkräfte: $.211^{**} \leq \rho \leq .348^{**}$). Mit der Skala zur Lehrer-Selbstwirksamkeitserwartung korrelieren die neu entwickelten Skalen jeweils im mittleren Bereich, aber jeweils etwas höher als mit der A-SWE-Skala (Studierende: $.439^{**} \leq \rho \leq .544^{**}$, Referendare: $.419^{**} \leq \rho \leq .516^{**}$, Lehrkräfte: $.371^{**} \leq \rho \leq .529^{**}$). Lediglich für die Kohorte der Referendare ist dieser Befund leicht abweichend – es kann keine einheitliche Tendenz der Korrelationen der neu entwickelten Skalen mit den Skalen A-SWE und L-SWE angegeben werden. Die A-SWE-Skala und die L-SWE-Skala korrelieren jeweils im mittleren Bereich miteinander (Studierende: $\rho = .479^{**}$, Referendare: $\rho = .484^{**}$, Lehrkräfte: $\rho = .552^{**}$).

Die revidierte Skala des physikalischen Selbstkonzeptes korreliert für die Studierenden sehr schwach mit den neu entwickelten Skalen (Studierende: $.104^{**} \leq \rho \leq .185^{**}$). Die Skala zu internalen Kontrollüberzeugungen korreliert schwach bis mittel mit den neu entwickelten Skalen (Studierende: $.168^{**} \leq \rho \leq .272^{**}$, Referendare: $.227^{**} \leq \rho \leq .317^{**}$, Lehrkräfte: $.060 \leq \rho \leq .255^{**}$).

Bezüglich der erhobenen Persönlichkeitsvariablen (Extraversion, Verträglichkeit, Gewissenhaftigkeit, Neurotizismus und Offenheit) ergibt sich folgendes Bild: Kohortenübergreifend können keine signifikanten Korrelationen der neu entwickelten Skalen (und der Skalen A-SWE und L-SWE) zur Verträglichkeitsskala gefunden werden. Für die Kohorte der Studierenden korrelieren die Skalen A-SWE und L-SWE jeweils in höherem Maße mit den Skalen zu den genannten übrigen Persönlichkeitsvariablen ($-.473^{**} \leq \rho \leq .473^{**}$) als die Skalen der neu entwickelten Skalen mit diesen Persönlichkeitsskalen ($-.271^{**} \leq \rho \leq .202^{**}$). Dies trifft grob auch für die Kohorten der Referendare und Lehrkräfte zu, allerdings korrelieren nicht immer beide der zusätzlichen Selbstwirksamkeitsskalen höher mit den Persönlichkeitsvariablen (vgl. z. B. Kohorte der Lehrpersonen: SWE-A-D korreliert höher mit der Skala Gewissenhaftigkeit (BFI-K-(G)) als z. B. die A-SWE-Skala zu BFI-K-(G)). Für die Kohorte der Lehrkräfte wird eine Ausnahme ersichtlich: Die Skala Offenheit (BFI-K-(O)) korreliert jeweils in höherem Maße mit den neu entwickelten Skalen als mit der A-SWE- oder der L-SWE-Skala. Insgesamt korrelieren die neu entwickelten Skalen mit den Skalen der Persönlichkeitsvariablen im schwachen bis mittleren Bereich.

Tabelle 12.21.: Kohortenspezifische Korrelationen (Spearman's ρ) zwischen den neu entwickelten (revidierten) und den zusätzlich erhobenen (revidierten) Skalen.

Studierende	A-SWE	L-SWE	K-I	E	V	G	N	O	PS
A-SWE	1.000	.479**	.416**	.309**	.040	.251**	−.473**	.255**	.219**
L-SWE	.479**	1.000	.310**	.310**	.114**	.283**	−.281**	.191**	.104**
K-I	.416**	.310**	1.000	.280**	.027	.292**	−.222**	.133**	.090*
BFI-K (E)	.309**	.310**	.280**	1.000	.238**	.203**	−.218**	.309**	.102**
BFI-K (V)	.040	.114**	.027	.238**	1.000	.175**	−.092*	.057	.047
BFI-K (G)	.251**	.283**	.292**	.203**	.175**	1.000	.023	.145**	.189**
BFI-K (N)	−.473**	−.281**	−.222**	−.218**	−.092*	.023	1.000	.004	−.111**
BFI-K (O)	.255**	.191**	.133**	.309**	.057	.145**	.004	1.000	.107**
PS	.219**	.104**	.090*	.102**	.047	.189**	−.111**	.107**	1.000
SWE-EX-P	.316**	.465**	.208**	.164**	.034	.193**	−.205**	.167**	.158**
SWE-EX-D	.370**	.486**	.256**	.124**	−.058	.067	−.271**	.109**	.180**
SWE-EL-P	.332**	.486**	.212**	.188**	.021	.202**	−.184**	.152**	.160**
SWE-EL-D	.304**	.473**	.179**	.107**	−.027	.141**	−.195**	.112**	.185**
SWE-A-P	.338**	.483**	.255**	.157**	−.035	.201**	−.202**	.147**	.104**
SWE-A-D	.375**	.544**	.272**	.167**	−.012	.156**	−.261**	.177**	.129**
SWE-SV-P	.230**	.439**	.176**	.173**	.044	.195**	−.121**	.143**	.112**
SWE-SV-D	.343**	.479**	.168**	.160**	−.039	.064	−.252**	.150**	.181**

Referendare	A-SWE	L-SWE	K-I	E	V	G	N	O
A-SWE	1.000	.484**	.416**	.250**	−.065	.209**	−.455**	.217**
L-SWE	.484**	1.000	.332**	.328**	.071	.333**	−.297**	.107
K-I	.416**	.332**	1.000	.191**	.114*	.385**	−.255**	.162**
BFI-K (E)	.250**	.328**	.191**	1.000	.142*	.160**	−.258**	.269**
BFI-K (V)	−.065	.071	.114*	.142*	1.000	.140*	−.073	.147*
BFI-K (G)	.209**	.333**	.385**	.160**	.140*	1.000	−.046	.075
BFI-K (N)	−.455**	−.297**	−.255**	−.258**	−.073	−.046	1.000	−.035
BFI-K (O)	.217**	.107	.162**	.269**	.147*	.075	−.035	1.000
SWE-EX-P	.409**	.500**	.305**	.301**	.023	.177**	−.203**	.200**
SWE-EX-D2	.492**	.463**	.298**	.180**	−.040	.176**	−.309**	.200**
SWE-EL-P	.512**	.516**	.317**	.233**	.002	.350**	−.269**	.182**
SWE-EL-D	.414**	.440**	.229**	.136*	−.027	.165**	−.216**	.142*
SWE-A-P	.433**	.501**	.294**	.215**	−.032	.252**	−.228**	.185**
SWE-A-D	.495**	.455**	.279**	.213**	−.059	.198**	−.266**	.141*
SWE-SV-P	.434**	.475**	.302**	.125*	.000	.285**	−.195**	.129*
SWE-SV-D	.411**	.419**	.227**	.162**	−.052	.047	−.196**	.107*

Lehrkräfte	A-SWE	L-SWE	K-I	E	V	G	N	O
A-SWE	1.000	.552**	.473**	.392**	−.024	.234**	−.435**	.115
L-SWE	.552**	1.000	.311**	.330**	.030	.326**	−.348**	.192**
K-I	.473**	.311**	1.000	.369**	−.114	.344**	−.159*	.224**
BFI-K (E)	.392**	.330**	.369**	1.000	.172*	.229**	−.236**	.134*
BFI-K (V)	−.024	.030	−.114	.172*	1.000	.002	−.184**	.051
BFI-K (G)	.234**	.326**	.344**	.229**	.002	1.000	−.155*	.128*
BFI-K (N)	−.435**	−.348**	−.159*	−.236**	−.184**	−.155*	1.000	.027
BFI-K (O)	.115	.192**	.224**	.134*	.051	.128*	.027	1.000
SWE-EX-P	.297**	.446**	.255**	.101	.045	.227**	−.249**	.360**
SWE-EX-D2	.348**	.495**	.173*	.126	.028	.139*	−.300**	.323**
SWE-EL-P	.300**	.529**	.214**	.181**	.128*	.283**	−.304**	.261**
SWE-EL-D	.237**	.419**	.079	.188**	.048	.146*	−.277**	.183**
SWE-A-P	.229**	.394**	.195**	.090	.080	.312**	−.263**	.326**
SWE-A-D	.282**	.460**	.204**	.170*	.048	.325**	−.303**	.268**
SWE-SV-P	.239**	.401**	.246**	.152*	.003	.236**	−.221**	.171*
SWE-SV-D	.211**	.371**	.060	.106	−.024	.168*	−.250**	.227**

12.7.8. Abbildung theoretischer Erwartungen

Mittelwertunterschiede: Personengruppe (neu konstruierte Skalen) Die nachfolgend berichteten latenten Mittelwertdifferenzen wurden im Zuge der Mehrgruppenvergleiche mittels konfirmatorischer Faktorenanalysen gewonnen (vgl. Tabelle 12.22). Es kann konstatiert werden, dass für vier von acht Skalen die latenten Mittelwerte für die Gruppe der Referendare signifikant niedriger geschätzt werden, als die der Studierenden und dass die latenten Mittelwerte der Skalenwerte der Lehrkräfte jeweils signifikant höher geschätzt werden als die der Studierenden und Referendare (vgl. auch Abbildungen zu klassischen Mittelwertvergleichen im Anhang P.2). Für die Skalen SWE-EX-D, SWE-EL-P, SWE-A-P und SWE-SV-D werden die latenten Skalenmittelwerte für die Gruppen der Studierenden und Referendare gleich groß geschätzt (vgl. Tabelle 12.22). Für die Vergleiche der Personengruppen „Studierende vs. Lehrpersonen“ bzw. „Referendare vs. Lehrpersonen“ liegen jeweils mittlere bis große Effekte vor ($.53 \leq d \leq 1.15$), für die Mittelwertdifferenzen der Personengruppe „Studierende vs. Referendare“ jeweils kleine Effekte ($.07 \leq d \leq .22$).[8]

Mittelwertunterschiede: Personengruppe (zusätzliche Skalen) Für die L-SWE-Skala ist ein ähnlicher Trend zu verzeichnen, wie er für die neu konstruierten Skalen beschrieben wurde: Der latente Mittelwert wird für die Referendare im Vergleich zum latenten Mittelwert der Studierenden signifikant niedriger geschätzt, während der latente Mittelwert der Lehrpersonen signifikant größer geschätzt wird als der latente Mittelwert der Studierenden/Lehrkräfte. Für die A-SWE-Skala hingegen ist der Schätzwert des latenten Mittelwerts für die Referendare höher als für die Studierenden und für die Lehrpersonen ebenfalls höher als für die Referendare (vgl. Anhang P.1). Für die Verträglichkeitsskala konnte kein Mehrgruppenvergleich und folglich auch keine Analyse der Mittelwertdifferenzen vorgenommen werden, da für die Kohorte der Referendare kein Modellfit möglich war (vgl. Tabelle 12.14).

Mittelwertunterschiede: Geschlecht (neu konstruierte Skalen) Für die Gruppe der Studierenden können für drei Skalen signifikante Mittelwertdifferenzen bezüglich des Geschlechts festgestellt werden (SWE-EX-D, SWE-A-D, SWE-SV-D, vgl. Tabelle 12.23). Für die Kohorte der Referendare können keine signifikanten Mittelwertunterschiede bezüglich des Geschlechts festgestellt werden und für die Gruppe der Lehrpersonen können lediglich für zwei Skalen signifikante Mittelwertdifferenzen bezüglich des Geschlechts identifiziert werden (SWE-EX-D, SWE-SV-D). Wenn Mittelwertunterschiede

[8] Im Zuge eines systematischen Methodenvergleichs, der hier nicht dargestellt werden kann und soll, wurde über die Skalenmittelwerte eine ANOVA mit Bonferroni-Korrektur gerechnet, sodass sehr einfach die Effektstärke Cohen's d bei Varianz- und Stichprobenungleichheit berechnet werden konnte.

Tabelle 12.22.: Mittelwertdifferenzen nach Personengruppe für die neu konstruierten und die zusätzlich erhobenen Skalen (Referenzgruppe: jeweils erstgenannte Personengruppe).

	Neu entwickelte Skalen						**Zusätzliche Skalen**				
Skala	GV[1]	ΔMW[2]	S.E.[3]	$\frac{\Delta MW}{S.E.}$	p	Skala	GV	ΔMW	S.E.	p $\frac{\Delta MW}{S.E.}$	p
SWE-EX-P	SR	-.242	.092	-2.632	.01	A-SWE	SR	.114	.094	1.207	.23
	SL	.449	.107	4.189	.00		SL	.665	.104	6.395	.00
	RL	.690	.123	5.618	.00		RL	.581	.118	4.092	.00
SWE-EX-D	SR	-.021	.090	-.229	.82	L-SWE	SR	-.431	.127	-3.391	.00
	SL	.841	.105	8.020	.00		SL	.660	.145	4.544	.00
	RL	.862	.121	7.138	.00		RL	1.091	.173	6.315	.00
SWE-EL-P	SR	.017	.102	.165	.87	K-I	SR	.018	.100	.180	.86
	SL	.936	.099	9.413	.00		SL	.092	.102	.901	.37
	RL	.919	.122	7.553	.00		RL	.074	.119	.622	.53
SWE-EL-D	SR	-.228	.092	-2.471	.01	E	SR	-.009	.095	-.093	.97
	SL	1.019	.094	10.824	.00		SL	-.300	.094	-3.192	.00
	RL	1.247	.115	10.847	.00		RL	-.291	.111	-2.627	.01
SWE-A-P	SR	-.181	.094	-1.920	.06	G	SR	.002	.095	.021	.98
	SL	.607	.109	5.556	.00		SL	.532	.101	5.289	.00
	RL	.788	.128	6.176	.00		RL	.530	.117	4.528	.00
SWE-A-D	SR	-.288	.101	-2.853	.00	N	SR	-.131	.093	-1.402	.16
	SL	.564	.104	5.424	.00		SL	-.297	.105	2.823	.01
	RL	.852	.124	6.887	.00		RL	-.167	.117	-1.421	.16
SWE-SV-P	SR	-.285	.092	-3.104	.00	O	SR	.024	.109	.223	.82
	SL	.569	.102	5.562	.00		SL	.285	.112	2.55	.01
	RL	.854	.122	6.990	.00		RL	.261	.126	2.065	.04
SWE-SV-D	SR	-.158	.09	-1.745	.08						
	SL	1.082	.105	10.330	.00						
	RL	1.24	.122	10.200	.00						

[1] Gruppenvergleich (SR: Studierende vs. Referendare, SL: Studierende vs. Lehrkräfte, RL: Referendare vs. Lehrkräfte)
[2] Schätzwert für die Mittelwertdifferenz [3] Standardfehler des Schätzwertes

auftreten, so werden die latenten Mittelwerte jeweils für die Gruppe der weiblichen Studierenden/Lehrkräfte geringer geschätzt als die latenten Mittelwerte der männlichen Studierenden. Es handelt sich jeweils lediglich um schwache Effekte ($.29 \leq d \leq .39$). Außerdem treten Mittelwertunterschiede lediglich für Skalen der Durchführungsdimension auf (vgl. Tabelle 12.23).

Mittelwertunterschiede: Geschlecht (zusätzliche Skalen, Auswahl) Sowohl für die L-SWE- als auch die A-SWE-Skala sind die geschätzten Skalenmittelwertdifferenzen in der Regel nicht signifikant (Ausnahme: A-SWE-Skala/Referendare).

Tabelle 12.23.: Mittelwertdifferenzen nach Geschlecht für die neu konstruierten und einige der zusätzlichen Skalen (Referenzgruppe: Männer).

	Studierende				**Referendare**				**Lehrkräfte**			
	ΔMW[1]	S.E.[2]	$\frac{\Delta MW}{S.E.}$	p	ΔMW	S.E.[3]	$\frac{\Delta MW}{S.E.}$	p	ΔMW	S.E.	$\frac{\Delta MW}{S.E.}$	p
SWE-EX-P	-.163	.103	-1.321	.19	.011	.168	.065	.95	-.091	.171	-.531	.60
SWE-EX-D	-.429	.103	-4.146	.00	-.218	.166	-1.311	.19	-.382	.179	-2.130	.03
SWE-EL-P	-.011	.101	-.111	.91	-.012	.157	-.078	.94	.043	.163	.264	.79
SWE-EL-D	-.195	.108	-1.801	.07	-.230	.158	-1.453	.15	-.189	.184	-1.025	.31
SWE-A-P	-.106	.101	-1.055	.29	-.107	.159	-.663	.51	.252	.174	1.446	.15
SWE-A-D	-.297	.109	-2.733	.01	-.229	.170	-1.350	.18	.082	.18	.457	.65
SWE-SV-P	-.005	.102	-.045	.96	.258	.159	1.624	.10	.176	.172	1.027	.30
SWE-SV-D	-.425	.105	-4.043	.00	-.235	.171	-1.480	.14	-.384	.179	-2.141	.03
L-SWE	.039	.124	.319	.75	-.110	.185	-.597	.55	-.085	.202	-.421	.67
A-SWE	-.025	.105	-.240	.81	-.366	.160	-2.289	.02	.125	.185	.680	.50

[1] Schätzwert für die Mittelwertdifferenz [2] Standardfehler des Schätzwertes

Mittelwertunterschiede: Schulform/Studiengang (neu konstruierte Skalen) Mittelwertdifferenzen bezüglich der Schulform/des Studiengangs sind in Tabelle 12.24 dargestellt. Für die Gruppe der Studierenden werden für vier Skalen signifikante Mittelwertunterschiede bezüglich des Studiengangs angezeigt (SWE-EX-D, SWE-EL-D, SWE-A-P, SWE-SV-P). Dabei werden die latenten Mittelwerte für Studierende des gymnasialen Lehramtsstudiengangs geringer (!) als die Mittelwerte für Studierende eines Sekundarstufenlehramts geschätzt. Die Effektstärken sind jeweils gering ($.20 \leq d \leq .25$).

Für die Kohorten der Referendare und Lehrkräfte wird jeweils bezüglich einer Skala eine signifikante Mittelwertdifferenz angezeigt (SWE-SV-P bzw. SWE-EL-P). Zwischen den Mittelwerten von Referendaren, die ein Gymnasiallehramt studiert haben und solchen, die als Quereinsteiger das Referendariat absolvieren, gibt als also in der Regel keine Unterschiede der latenten Skalenmittelwerte. Gleiches kann für die latenten Mittelwerte der Lehrkräfte konstatiert werden, die an einem Gymnasium arbeiten und solchen Lehrpersonen, die an einer Schule im Sekundarschulbereich arbeiten.

Tabelle 12.24.: Mittelwertdifferenzen nach Schulform/Studiengang für die neu konstruierten und einige der zusätzlichen Skalen (Referenzgruppe jeweils erstgenanntes Kriterium; Studierende: Studiengang Sekundarstufe I/Studiengang Gymnasium, Referendare: gymnasialer Studiengang/Quereinstieg, Lehrkräfte: Gymnasium/kein Gymnasium).

	Studierende SG SekI/Gym				**Referendare** SG Gym/Quer				**Lehrkräfte** Gym/nicht Gym			
	ΔMW[1]	S.E.[2]	$\frac{\Delta MW}{S.E.}$	p	ΔMW	S.E.	$\frac{\Delta MW}{S.E.}$	p	ΔMW	S.E.	$\frac{\Delta MW}{S.E.}$	p
SWE-EX-P	-.212	.111	-1.911	.06	-.166	.172	-.966	.33	.127	.164	.772	.44
SWE-EX-D	-.216	.106	-2.029	.04	-.044	.177	-.251	.80	.167	.166	1.006	.31
SWE-EL-P	-.114	.105	-1.091	.28	.026	.161	.161	.87	.382	.174	2.197	.03
SWE-EL-D	-.230	.111	-2.077	.04	.214	.163	1.314	.19	.130	.176	.743	.46
SWE-A-P	-.267	.108	-2.465	.01	-.237	.18	-1.319	.19	-.034	.169	-.200	.84
SWE-A-D	-.098	.110	-.893	.37	-.087	.164	-.530	.60	.005	.182	.027	.98
SWE-SV-P	-.263	.108	-2.430	.02	-.402	.167	-2.409	.02	-.008	.164	-.050	.96
SWE-SV-D	-.100	.113	-.882	.38	.323	.177	1.829	.07	.289	.184	1.573	.12
L-SWE	-.383	.134	2.846	.00	–	–	–	–	.028	.190	.147	.88
A-SWE	-.095	.107	-.885	.38	.153	.155	.985	.32	.073	.172	.422	.67

[1] Schätzwert für die Mittelwertdifferenz [2] Standardfehler des Schätzwertes

Mittelwertunterschiede: Schulform/Studiengang (zusätzliche Skalen, Auswahl)
Bezüglich der Skala zur allgemeinen Selbstwirksamkeitserwartung können für keine der Kohorten signifikante Mittelwertdifferenzen berichtet werden. Für die L-SWE-Skala ist die Skalenmittelwertdifferenz nur für die Studierendenkohorte signifikant, wobei die Effektstärke gering ausfällt ($d = .27$). Für die Gruppe der Referendare kann aufgrund der Modellierungsproblematik keine Berechnung vorgenommen werden (vgl. Abschnitt 12.7.5 bzw. Anhang O.3).

Mittelwertunterschiede: Praxiserfahrung (neu konstruierte Skalen) Für fünf Skalen zeigen sich signifikante Mittelwertunterschiede bezüglich der Praxiserfahrung der Studierenden (vgl. Tabelle 12.25). Die latenten Mittelwerte der Studierenden, die institutionelle Praxiserfahrungen aufweisen, werden dabei höher geschätzt als die latenten Mittelwerte der Studierenden, die keine institutionelle Praxiserfahrung sammeln konnten. Es handelt sich um kleine bis mittlere Effekte ($.21 \leq d \leq .52$).

Mittelwertunterschiede: Praxiserfahrung (zusätzliche Skalen, Auswahl) Die geschätzten Skalenmittelwertdifferenzen werden erneut für die Skala zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen signifikant ($d = .23$), während dies für die Skala zur Erhebung allgemeiner Selbstwirksamkeitserwartungen nicht der Fall ist (vgl. Tabelle 12.25).

Tabelle 12.25.: Mittelwertdifferenzen nach Praxiserfahrung der Studierenden für die neu konstruierten und einige der zusätzlichen Skalen (Referenzgruppe: Studierende ohne Praxiserfahrung).

	ΔMW[1]	**S.E.**[2]	$\frac{\Delta MW}{S.E.}$	**p**
SWE-EX-P	.561	.130	4.332	.00
SWE-EX-D*	–	–	–	–
SWE-EL-P	.266	.131	2.037	.04
SWE-EL-D	.234	.129	1.815	.07
SWE-A-P	.301	.139	2.165	.03
SWE-A-D	.335	.128	2.630	.01
SWE-SV-P	.233	.129	1.803	.07
SWE-SV-D	.331	.135	2.454	.01
L-SWE	.315	.141	2.231	.026
A-SWE	.147	.122	1.205	.228

[1] Schätzwert für die Mittelwertdifferenz
[2] Standardfehler des Schätzwertes
* Es wurden keine Mittelwertdifferenzen berechnet, da keine Messinvarianz vorliegt (vgl. Unterabschnitt 12.7.5).

12.8. Interpretation der Ergebnisse und Validitätsargumente

12.8.1. Eindimensionalität der Skalen

Aufgrund der sehr guten Modellanpassungsgüten (insbesondere nach der Revision) können die Skalen als eindimensional betrachtet werden. Mit Blick auf die Forschungsfrage bzw. Hypothese 6.1 (vgl. Abschnitt 8.1.2, FF6) und unter Berücksichtigung der Ergebnisse der qualitativen Studien kann davon ausgegangen werden, dass die Skalen die intendierten Konstrukte jeweils eindimensional abbilden.

Auch die *revidierten*, zusätzlich eingesetzten Skalen können als eindimensional eingestuft werden. Es ist jedoch zu berücksichtigen, dass durch die Revision die Skalen ggf. inhaltlich differieren und somit schwieriger zu interpretieren sind bzw. Ergebnisse schwieriger mit Literaturangaben verglichen werden können.

12.8.2. Zuverlässigkeit der Skalen

Insgesamt kann bezüglich der siebten Forschungsfrage (FF7) konstatiert werden, dass mit Blick auf die ausgewerteten Reliabilitätsmaße davon ausgegangen werden kann, dass die konstruierten Skalen die intendierten Konstrukte zuverlässig abbilden. Dieses Urteil wird durch die nachfolgenden Ausführungen untermauert.

Klassische Reliabilitätsmaße Die niedrigen mittleren Inter-Item-Korrelationen sprechen einerseits für die Unabhängigkeit der Items (schwache Korrelationen) und andererseits für die Zusammengehörigkeit der Items (positive Korrelationen). Die Itemtrennschärfen weisen keine großen Streuungen auf und liegen für jede Skala in jeder Kohorte im mittleren Bereich. Dies kann ebenfalls dahingehend gedeutet werden, dass die Items zu den Skalen kompatibel sind, da sie jeweils mit dem Gesamtwert der Skala einen positiven Zusammenhang bilden. Zu beachten ist, dass Inter-Item-Korrelationen und auch Cronbach's Alpha im eigentlichen Sinne keine Homogenitätsindizes darstellen, da sie letztlich lediglich auf Korrelationen der Items untereinander beruhen. Aus deren Ausprägung kann daher nicht direkt auf die Eindimensionalität des gemessenen Konstruktes geschlossen werden (vgl. Bühner, 2011, S. 166 ff.). Diese Kennzahlen können höchstens als zusätzliche Indizien herangezogen werden. Das Standardverfahren der Eindimensionalitätsprüfung ist die Faktorenanalyse, deren Ergebnisse bereits berichtet wurden.

Die Skalenreliabilitäten, die hier durch Cronbach's Alpha α_C repräsentiert werden, können als sehr gut charakterisiert werden, insbesondere wenn man bedenkt, dass es sich um einen Persönlichkeitsfragebogen handelt und keinen Leistungstest. Die ermittelten Werte für Cronbachs Alpha können dahingehend gedeutet werden, dass für die konstruierten Skalen im Schnitt gilt, dass weniger als 20% der Gesamtvarianz auf Messfehler zurückzuführen sind (vgl. Bühner, 2011, S. 146).

Reliabilitätsmaße des konfirmatorischen Ansatzes Dass immerhin 45% aller Faktorladungen eine Indikatorreliabilität von $\lambda^2 \geq .40$ erreichen, ist für das Entwicklungsstadium des Fragebogens als sehr gut zu werten. Das bedeutet, dass für fast die Hälfte aller Items mindestens 40% des Varianzanteils der Indikatoren durch das jeweilige Konstrukt erklärt werden kann. Dass die Faktorreliabilität (FR) jeweils in der Größenordnung von Cronbach's Alpha liegt, zeugt von der Qualität der entwickelten Skalen, da damit der Grenzwert von $FR \geq .60$ jeweils deutlich überschritten wird. Auch inhaltlich können Faktorreliabilität und Cronbach's Alpha miteinander verglichen werden (FR: Anteil der Gesamtvarianz, die durch das Konstrukt erklärt wird.). Die Werte für die durchschnittlich extrahierte Varianz (DEV) liegen dagegen jeweils unter den geforderten Grenzwerten von $DEV \geq .50$, was daran liegt, dass die Faktorladungen zwar im angemessenen Bereich, jedoch nicht sehr hoch ausfallen. Insofern muss davon ausgegangen werden, dass im Schnitt lediglich 30% bis 46% der Faktorvarianz pro Indikator erklärt wird (vgl. Weiber & Mühlhaus, 2010, S. 122 f.).

Reliabilitätsmaße des probabilistischen Ansatzes Auch im probabilistischen Paradigma genügen die Reliabilitätsmaße den gesetzten Kriterien und können als gut bis sehr gut bezeichnet werden. Da die Personen-Reliabilitäten Werte um .80 annehmen,

könnten nach Linacre (2014, S. 230) anhand des Datensatzes maximal drei Gruppen unterschiedlicher Lehrer-Selbstwirksamkeitserwartungen unterschieden werden. Die hohen Item-Reliabilitäten weisen darauf hin, dass die Stichprobe jeweils groß genug ist, um die Items entlang des latenten Konstruktes zu lokalisieren (vgl. Linacre, 2014, S. 230).

Der Personen-Separationsindex nimmt akzeptable bis gute Werte an (Person-Separation > 1.5), während die Item-Separation sehr gute Werte aufweist (Item-Separation > 2.5) (Boone u. a., 2014, S. 231). Insgesamt können die Indizes dahingehend gedeutet werden, dass das Instrument eine ausreichende Sensitivität aufweist, um auf Personenebene zwischen Personen unterschiedlicher Lehrer-Selbstwirksamkeitserwartungen und auf Itemebene Items unterschiedlicher Schwierigkeit unterscheiden zu können (vgl. Boone u. a., 2014, S. 226 f.).

Reliabilität der zusätzlichen Skalen Die Reliabilität der zusätzlich erhobenen, sogar revidierten Skalen ist erschreckend gering, auch wenn teils weniger Items eine Skala bilden und es sich um Persönlichkeitsvariablen handelt. Der Grenzwert von .70 (vgl. Schermelleh-Engel & Werner, 2012, S. 135 f.) wird für die Subskalen Verträglichkeit und Offenheit für alle Kohorten, und für die L-SWE-Skala für zwei von drei Kohorten nicht erreicht. Für die Skala A-SWE kann davon ausgegangen werden, dass eine zuverlässige Testwertinterpretation möglich ist.

12.8.3. Eignung des Antwortformates

Alle formalen Kriterien sprechen für die Verwendung der sechsstufigen Antwortskala und damit für die Eignung des verwendeten Antwortformates. Auch wenn die geforderten Logit-Abstände für die mittleren Kategorien teils etwas geringer als vorgeschlagene Grenzwerte ausfallen, so können die Darstellungen der Wahrscheinlichkeitsverteilungen für die Antwortkategorien bezogen auf die Differenz von Personenfähigkeit und Itemschwierigkeit dahingehend interpretiert werden, dass die Sechsstufigkeit der Skala zur Erfassung des Konstruktes geeignet ist, da jeweils distinkte Peaks in den Kurven für alle Kategorien sichtbar sind.

Hinsichtlich der fünften Forschungsfrage bzw. Hypothese 5.1 kann damit konstatiert werden, dass durch die Verwendung der sechsstufigen Antwortskala ein differenzierteres Urteil bezüglich der Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern von den Befragten erfasst werden kann als bei Verwendung eines vierstufigen Formates.

12.8.4. Messinvarianz

Bezüglich der vorhergesagten und der empirisch ermittelten Itemschwierigkeiten gibt es einen sehr großen Überlapp, auch wenn nicht immer kohortenübergreifende Überein-

stimmungen bezüglich der zwei leichtesten/schwersten Items vorliegen. Dies spiegelt sich auch in den Mehrgruppenvergleichen, in denen einzelne Items identifiziert werden können, deren Faktorladungen und oder Intercepts frei geschätzt werden müssen bzw. die hohe DIF-Werte aufweisen. Da in der Regel (partielle) skalare Messinvarianz vorliegt, können Mittelwertvergleiche vorgenommen werden.

Die Interpretation wird ein wenig dadurch erschwert, dass die Mehrgruppenanalysen für zum Teil erheblich unterschiedliche Gruppengrößen vorgenommen wurden, auch wenn dieses Verfahren gängige Praxis und durchaus möglich ist (Brown, 2006, S. 279). Um die Ergebnisse abzusichern, wurden alle Analysen für gleichverteilte Zufallsstichproben wiederholt. Die Ergebnisse dieser Analysen bestätigen die referierten Ergebnisse, sodass diese insgesamt dahingehend interpretiert werden können, dass die Items in den meisten Fällen gruppenübergreifend sehr ähnlich verstanden werden. Die Operationalisierung der intendierten Konstrukte kann damit für unterschiedliche Gruppen auch aus statistischer Perspektive als gelungen angesehen werden (vgl. Forschungsfrage 10, Unterabschnitt 8.1.2).

12.8.5. Dimensionsanalyse

Die Ergebnisse der Dimensionsanalysen sind nicht leicht zu interpretieren. Die χ^2-Differenzentests sprechen eher für die Unterscheidung der Konstrukte entsprechend der gewählten Dimensionen. Allerdings liegen die latenten Korrelationen der Skalen derart hoch, dass der Unterschied der Skalen minimal erscheint. Aus den Ergebnissen der Rasch-Hauptkomponentenanalyse lassen sich ebenfalls keine eindeutigen Hinweise auf eine zusätzliche Dimension im Datensatz ableiten.

Die klassische Korrelationsanalyse spricht dagegen für die Differenzierung der Handlungsfelder in Planungs- und Durchführungsdimension. Das Korrelationsmuster entspricht für die Kohorte der Studierenden vollständig und für die Kohorten der Referendare und Lehrkräfte im Wesentlichen den Erwartungen (vgl. Hypothese 8.2 (1-3), Unterabschnitt 8.1.2). Insgesamt scheint es vertretbar zu sein, davon auszugehen, dass die Planungs- und Durchführungsskalen eines Handlungsfeldes tatsächlich unterschiedliche Konstrukte abbilden.

Exkurs: Kurzskalen Aufgrund der teils uneindeutigen Ergebnisse der Dimensionsanalysen wurden Kurzskalen für jedes Handlungsfeld entwickelt, die sich aus Items beider Dimensionen zusammensetzen. Die Ergebnisse der Analysen können der Skalendokumentation entnommen werden (vgl. Meinhardt u. a., 2016c, S. 130-147, 241-257, 369-385, 489-505), müssen aufgrund der post-hoc gebildeten Skalen jedoch vorsichtig interpretiert werden. Nichtsdestotrotz stimmen die Ergebnisse optimistisch, dass sinnvolle Skalen gebildet wurden, die mit hoher Wahrscheinlichkeit Lehrer-Selbstwirksamkeitserwartungen entsprechend der jeweiligen Handlungsfelder abbilden. Welche Skalen in der Zukunft

von Forschenden eingesetzt werden sollten, erscheint insbesondere von der jeweiligen konkreten Forschungsfrage abhängig zu sein, zum Beispiel in dem Sinne, ob diese auf die Unterscheidung von Planung und Durchführung von Physikunterricht fokussiert.

12.8.6. Abgrenzbarkeit zu anderen Konstrukten

Insgesamt entsprechen die berichteten Korrelationen im Wesentlichen den in Unterabschnitt 8.1.2 formulierten Hypothesen 9.1 bis 9.5, auch wenn die Korrelationen der Skalen A-SWE und L-SWE im Vergleich zu anderen Studien vergleichsweise niedriger ausfallen. Schmitz und Schwarzer (2000, S. 19) berichten beispielsweise $r = .70^{**}/73^{**}$.[9] Mit Bezug zur Forschungsfrage (FF9) kann konstatiert werden, dass aufgrund der Ergebnisse der Korrelationsanalysen davon ausgegangen werden kann, dass die neu entwickelten Skalen zu anderen, ähnlichen Konstrukten hinreichend abgrenzbar sind.

12.8.7. Abbildung theoretischer Erwartungen

„Entwicklung“ von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern Auch wenn es sich nicht um längsschnittliche Daten handelt, so scheinen Referendare im Schnitt niedrigere spezifische Lehrer-Selbstwirksamkeitserwartungen aufzuweisen als Studierende oder Lehrpersonen, was ggf. mit dem Phänomen des Praxisschocks erklärt werden kann. Dass die latenten Mittelwerte der Gruppe der Lehrkräfte derart stark von den geschätzten Mittelwerten der Studierenden bzw. Referendare abweichen, kann unter Umständen mit der Positivauswahl der Kohorte zu tun haben. Schließlich fand bei der Teilnahme eine Art Selbstselektion statt und es kann davon ausgegangen werden, dass eher selbstwirksame Physik-Lehrpersonen an der Befragung teilgenommen haben. Insofern werden die latenten Mittelwerte der Kohorte der Lehrpersonen ggf. überschätzt. Der beobachtete Trend entspricht im Wesentlichen den eingangs formulierten Hypothesen 11.3 (1/3) (vgl. FF11, Unterabschnitt 8.1.2). Auch dass im Studium kein Praxisschock beobachtet werden kann (unter der Berücksichtigung, dass keine längsschnittlichen Daten vorliegen), entspricht den Erwartungen und der These, dass Praxisphasen im Studium in der Regel gut betreut werden und daher nicht mit einem Abfall domänenspezifischer Selbstwirksamkeitserwartungen zu rechnen ist (vgl. FF11, Hypothese 11.3 (2), Unterabschnitt 8.1.2). Der beschriebene Mittelwertunterschied zwischen den Kohorten der Studierenden und Referendare, der ggf. auf den Praxisschock zurückzuführen ist, kann nur für Skalen nachgewiesen werden, die Selbstwirksamkeitserwartungen auf einem domänenspezifischen Niveau erheben (L-SWE-Skala, neu konstruierte Skalen), während für die A-SWE-Skala kein signifikanter Mittelwertunterschied zwischen den beiden Kohorten beobachtet werden

[9] Zu beachten ist, dass beide Skalen revidiert wurden und sich somit von den Originalskalen unterscheiden.

kann.

Unterschiede aufgrund des Geschlechts Es können keine systematischen Mittelwertunterschiede aufgrund des Geschlechts konstatiert werden, was mit den Erwartungen in Einklang steht (vgl. FF11, Hypothese 11.1, Unterabschnitt 8.1.2). Wenn Mittelwertunterschiede auf einzelnen Skalen (für die Kohorte der Studierenden) vorliegen, dann werden für Frauen geringere spezifische Lehrer-Selbstwirksamkeitserwartungen geschätzt. Es muss weiter erforscht werden, ob tatsächlich die Dimension ein ausschlaggebender Faktor für geschlechtsspezifische Unterschiede darstellt oder ob es sich hierbei um ein Artefakt handelt, denn tatsächlich können lediglich signifikante Mittelwertdifferenzen für die Durchführungsdimension dokumentiert werden.

Unterschiede aufgrund der Schulform Die Ergebnisse der latenten Mittelwertvergleiche korrespondieren nicht immer mit den Erwartungen (vgl. FF11, Hypothese 11.2, Unterabschnitt 8.1.2). Entgegen der Erwartung werden für vier von acht Skalen niedrigere Mittelwerte für Lehramtsstudierende eines Gymnasiallehramts angezeigt. Für Lehrpersonen konnten dagegen keine signifikanten Mittelwertdifferenzen aufgrund der Schulform gefunden werden, sodass ebenfalls offen bleiben muss, ob es sich hierbei um ein Spezifikum oder ein Artefakt handelt.

12.9. Externe Validierung

Das Instrument wurde zeitgleich zu seiner hier beschriebenen Validierung im Rahmen mehrerer Projekte an anderen Einrichtungen[10] eingesetzt, sodass zusätzlich eine Art externe Validierung stattfand. Für die Schweiz wurde eine Version entwickelt, die speziell auf den naturwissenschaftlichen Unterricht fokussierte. Bisher sind keine Ergebnisse dieser Studien bekannt, die gegen die intendierte Testwertinterpretation sprechen.

12.10. Zusammenfassung

Es lässt sich urteilen, dass zur Beantwortung der anvisierten Forschungsfragen ein relativ hoher Aufwand betrieben wurde. Insgesamt wurden knapp 1000 Personen befragt, die Daten wurden akribisch unter Berücksichtigung unterschiedlicher methodischer Herangehensweisen ausgewertet und alle Ergebnisse transparent dokumentiert (vgl. auch Meinhardt u. a., 2016c). Die methodische Kreuzvalidierung kann insofern als Erfolg

[10] Zum Beispiel in den Bereichen Physikdidaktik der Universität Frankfurt, der Freien Universität Berlin und des IPN Kiel, dem Bereich der Chemiedidaktik der Universität Regensburg sowie der Didaktik der Naturwissenschaften an der PH Zürich.

gewertet werden, als Ergebnisse zumeist nicht von den angewandten Methoden abzuhängen scheinen, sondern lediglich in Nuancen differieren. Die Bedeutung des mehrfach erwähnten Skalenreports muss hier erneut hervorgehoben werden. Dieser ist nicht nur eine systematisch angelegte Dokumentation (fast) aller methodischen Entscheidungen und Ergebnisse, sondern ein wichtiger Bestandteil dieser Arbeit ohne den einige der Interpretationen nicht vollständig nachvollzogen werden können. Insgesamt konnten eine Vielzahl von Indizien für ein übergeordnetes Validitätsargument gesammelt werden, da sich die aufgestellten Hypothesen zumeist als korrekt erwiesen. Die Untersuchung der Angemessenheit der Unterscheidung der Handlungsfelder in die Dimensionen „Planung“ und „Durchführung“ muss Gegenstand weiterer Validierungsmaßnahmen sein.

Teil IV.

Schluss

13. Zusammenfassung, Diskussion und Ausblick

„Unterkomplexe Antworten haben ihre eigene Suggestion."

Münkler, 2016, Die Zeit

Das nachfolgende Kapitel dient einem systematischen Resümee und einer kritischen Reflexion der Ergebnisse der vorliegenden Arbeit. Zunächst werden jedoch die Ergebnisse der Arbeit auf theoretischer wie empirischer Ebene zusammenfassend dargelegt, wobei letztere in der Reihenfolge der Forschungsfragen aufgelistet werden. Aus der kompakten Darstellung der Ergebnisse wird im sich anschließenden Abschnitt ein vorläufiges Validitätsargument abgeleitet und unter Berücksichtigung der konkurrierenden Testwertinterpretationen diskutiert. Mögliche Testnutzungen wurden im Rahmen dieser Arbeit antizipiert, wobei eine spezifische Nutzung nicht favorisiert wurde und daher auch nicht im Fokus der Validierungsbemühungen stand. Ein Blick in die Zukunft konkretisiert potentielle Einsatzmöglichkeiten, indem Fragestellungen aufgeworfen werden, die es aus Sicht der Autorin Wert sind, in Folgeprojekten untersucht zu werden.

13.1. Zusammenfassung der Ergebnisse

13.1.1. Ergebnisse auf theoretischer Ebene

Theoriebasierte Entwicklung von Itemkonstruktionsregeln Aus dem Bestreben heraus das Konstrukt der Selbstwirksamkeitserwartung nach Bandura möglichst exakt zu fassen (vgl. Abschnitt 3.2), um daraus nachvollziehbare und theoretisch fundierte Regeln für die Itementwicklung eines domänenspezifischen Instrumentes abzuleiten (vgl. Abschnitte 7.1.2, 7.1.3 und 11.1), resultiert ein Katalog von Eigenschaften des Konstruktes:

- Subjektivität/Selbstreferentialität,
- (konkreter) Handlungsbezug,
- Schwierigkeitsbezug,
- Situationsspezifität/Kontextabhängigkeit,
- Domänenspezifität,
- „agent-means"-Relation.

Insbesondere die Auffassung von Selbstwirksamkeitserwartungen als Kontrollkonstrukt und die damit verbundene Charakterisierung als „agent-means“-Relation nach einem Klassifikationsschema von Skinner (1996) erweisen sich als fruchtbar, um Selbstwirksamkeitserwartungen von anderen Konstrukten wie Handlungsergebniserwartungen abzugrenzen bzw. neue Items zu entwerfen, die Selbstwirksamkeitserwartungen abbilden. Im Wesentlichen muss bei der Formulierung von Items darauf geachtet werden, auf konkrete (komplexe) Handlungen zu fokussieren und nicht auf Konsequenzen, die aus diesen Handlungen womöglich resultieren oder die angestrebt werden sollen. Es wurde herausgearbeitet, dass im englischen Sprachraum unter „teacher efficacy“ in der Regel keine Lehrer-Selbstwirksamkeitserwartungen im Sinne der sozial-kognitiven Theorie Banduras verstanden werden dürfen, weil neben anderen Problematiken gerade die benannte Trennung von Handlungen und Handlungskonsequenzen nicht oder selten beachtet wird, woraus Konfundierungen mit anderen, benachbarten Konstrukten resultieren. Mit dem Ableiten von Itemkonstruktionsregeln aus den dargelegten Eigenschaften wurden die notwendigen Grundlagen für eine theoriekonforme Operationalisierung des Konstruktes der Selbstwirksamkeitserwartungen auf einem fachspezifischen Niveau gelegt.

Literaturrecherche, Evaluation existierender Instrumente und Aufbereitung des Forschungsstandes Um Einblicke in domänenspezifische Auslegungen des Konstruktes und daraus resultierend Anregungen für die eigene Instrumentneuentwicklung zu erhalten, wurden existierende Instrumente gesichtet, die beanspruchen, Selbstwirksamkeitserwartungen von Lehrpersonen („teacher efficacy“) zu erfassen. Die ausgewählten Instrumente wurden einer Evaluation unterzogen, wobei die herausgearbeiteten Eigenschaften und Merkmale des Konstruktes der Selbstwirksamkeitserwartungen nach Bandura als Kriterien herangezogen wurden, um die domänenspezifischen Operationalisierungen auf ihre Passung zu den theoretischen Anforderungen zu prüfen. Als Ergebnis der Evaluation wurden bei einer Reihe der gesichteten Instrumente teils erhebliche Mängel in der theoretischen Fundierung oder der methodischen Herangehensweise der Validierung identifiziert.

Auch wenn das Konstrukt der Selbstwirksamkeitserwartungen schnell und intuitiv zugänglich erscheint, so zeigt sich anhand der Publikationen eindrücklich, dass die Adaption des Konstruktes auf die interessierende Domäne der Planung und Durchführung von Physikunterricht keineswegs trivial ist, da bereits bei einem allgemeineren Spezifitätsniveau (Operationalisierung von *Lehrer*-Selbstwirksamkeitserwartungen) Schwierigkeiten auftreten. Insbesondere konnte aufgezeigt werden, dass derzeit kein zu empfehlendes Erhebungsinstrument existiert, welches Lehrer-Selbstwirksamkeitserwartungen auf dem gewünschten Spezifitätsniveau, also in physikdidaktischen Handlungsfeldern, erfasst und in Einklang mit theoretischen Annahmen der sozial-kognitiven Theorie Banduras steht.

Diese Einschätzung basiert auf der Analyse von mehr als 110 wissenschaftlichen Publikationen zum Themenkomplex „teacher (self) efficacy (beliefs)“ bzw. Lehrer-Selbstwirksamkeitserwartungen, die nach dem Jahr 2000 veröffentlicht wurden. Zuerst wurden diese nach genutztem Erhebungsinstrument, nach Fragestellungen und befragter Kohorte klassifiziert. Eine sehr große Anzahl von Forschenden nutzten die „Teacher Efficacy Scale“ (TES) von Gibson und Dembo (1984), da dieses Instrument lange Zeit als Standard galt. Viele Studien weisen mittlerweile auf die Unzulänglichkeiten der TES hin, sodass Forschungsbefunde, die mit diesem Instrument generiert wurden, mindestens in Bezug auf die zugrunde gelegten Testwertinterpretationen zu hinterfragen sind. Insgesamt muss erstaunlicherweise festgestellt werden, dass die empirische Basis in Bezug auf die Erforschung von Lehrer-Selbstwirksamkeitserwartungen aufgrund der Schwächen existierender Instrumente trotz jahrzehntelanger Forschungstraditionen eher dünn ausfällt bzw. Forschungsergebnisse nur sehr eingeschränkt interpretiert werden können.

Weiterentwicklung eines Modells zur Abgrenzung von „teacher efficacy“ und „teacher self-efficacy beliefs“ Es stellte sich im Rahmen der theoretischen Fundierung heraus, dass es nicht *das* „teacher efficacy“-Konstrukt gibt, sondern dass Autoren wahlweise unterschiedliche Definitionen anlegen oder voneinander abweichende theoretische Einbettungen vornehmen. In der Regel konfundieren die Operationalisierungen mit Handlungsergebniserwartungen oder Kontrollüberzeugungen. Dellinger u. a. (2008) schlagen daher vor, „teacher efficacy“ von „teacher self-efficacy beliefs“ zu unterscheiden, wobei diffus bleibt, was ersteres genau sein soll und letzteres Lehrer-Selbstwirksamkeitserwartungen im Sinne Banduras darstellen. Die Autoren schlagen außerdem ein Modell vor, welches die Unterschiede und Gemeinsamkeiten beider Konstrukte veranschaulichen soll. Dieses Modell wurde als Ausgangspunkt für eine grundlegende Überarbeitung und Erweiterung genutzt (vgl. Abbildung 4.5 auf Seite 100). Ergänzungen wurden insbesondere hinsichtlich relevanter Elemente zur Struktur und Genese des Konstrukts vorgenommen, wobei versucht wurde, die wesentlichen Aussagen der sozial-kognitiven Theorie Banduras möglichst vollständig grafisch umzusetzen.

Selbstwirksamkeitserwartungen als Kompetenzfacette Die vorliegende Forschungsarbeit ist in mehrerer Hinsicht anschlussfähig. In Kapitel 6 wurde schwerpunktmäßig herausgearbeitet, dass Lehrer-Selbstwirksamkeitserwartungen im Rahmen der Kompetenzdiskussion um Lehrpersonen eine Rolle spielen, da sie sowohl als Kompetenz determinierender Einflussfaktor als auch als kompetenzkonstituierende Dimension aufgefasst werden können (vgl. Abschnitt 6.1.1), wobei letztgenannter Ansatz in besonderem Maße kompatibel mit den Postulaten der sozial-kognitiven Theorie erscheint. In jedem Fall konnte plausibel gemacht werden, dass Selbstwirksamkeitserwartungen berücksich-

tigt werden sollten, wenn Kompetenzmodellierungen vorgenommen werden. Inwiefern die Kompetenzorientierung in der Lehrerbildung insgesamt als tragfähig anzusehen ist, bleibt abzuwarten. Unabhängig vom „Kompetenz-Paradigma“ können Selbstwirksamkeitserwartungen als relevantes Konstrukt begriffen und in ihren Wirkungen bzw. Wechselbeziehungen zu anderen Konstrukten und/oder Kriterien erforscht werden.

13.1.2. Ergebnisse auf empirischer Ebene in Bezug auf die Forschungsfragen

Die Bedeutung einer umfassenden theoretischen Auseinandersetzung mit dem interessierenden Konstrukt als Grundlage der Itementwicklung ist hoffentlich aus den Darstellungen deutlich geworden. Dass Verkürzungen häufig zu inadäquaten Adaptionen interessierender Konstrukte führen, wurde versuchsweise und beispielhaft an einer Vielzahl von konkreten Operationalisierungen aufgezeigt. Für die Durchführung der konzipierten Studien war die Umsetzung des dargelegten Validierungskonzepts (vgl. Kapitel 2) handlungsleitend. Zu diesem gehörte unter anderem die transparente Ableitung von Annahmen, die der intendierten Testwertinterpretation zugrunde liegen sowie das Aufstellen von Maßnahmen, um die Gültigkeit selbiger zu prüfen und damit den Forschungsfragen (vgl. Unterabschnitt 8.1.2) auf den Grund zu gehen. Nachfolgend werden diese schrittweise beantwortet, wobei auf die übergeordnete Forschungsfrage abschließend eingegangen wird, da diese das übergeordnete Validitätsargument berührt.

Beantwortung der Forschungsfragen Tabelle 8.2 auf Seite 211 listet die einzelnen durchgeführten Studien auf, sowie deren Beiträge zu einzelnen Forschungsfragen. Für jede einzelne Studie wurden ausführlich alle Ergebnisse im Hinblick auf die Forschungsfragen interpretiert (vgl. Abschnitte 9.6, 10.6 und 12.8). Aufgrund der mehrfachen Überarbeitung der Items sowie des zugrunde liegenden Validierungskonzepts ist eine abschließende Beantwortung der Forschungsfragen nicht möglich. Im Sinne einer fortdauernden Validierung sollten die überarbeiteten Items erneut Probanden vorgelegt werden, um weitere Hinweise bezogen auf interessierende Fragestellungen zu erhalten. Zusammenfassend kann Folgendes konstatiert werden.

FF1 Hinsichtlich der ersten Forschungsfrage konnten wenige Indizien gesammelt werden, da hauptsächlich Experten mit dem nötigen theoretischen Hintergrund geeignet erschienen, um die Frage fundiert zu beantworten. Für die entsprechende Befragung konnte jedoch lediglich ein Experte gewonnen werden, der bezüglich der abgeleiteten Konstruktionsvorschriften und der dargelegten Arbeitsdefinition allerdings nichts einzuwenden hatte.

Die Items wurden in der Zwischenzeit von einer großen Zahl von Forschenden zum Zweck der Erhebung von domänenspezifischen Selbstwirksamkeitserwartungen eingesetzt. Bisher wurde die inhaltliche Validität nicht offen in Zweifel gezogen,

was natürlich nicht heißt, dass tatsächlich Items entsprechend der Arbeitsdefinition und der theoretischen Vorarbeiten formuliert wurden.[1] Entsprechend der formulierten Hypothese ist aufgrund der sorgfältigen Theoriearbeit jedoch davon auszugehen, dass die Items den Charakteristika von Selbstwirksamkeitserwartungen in der Tradition der sozial-kognitiven Theorie Banduras entsprechen und damit eine domänenspezifische Adaption des Konstruktes darstellen. Die theoretischen Überlegungen müssen weiteren Experten zugänglich gemacht werden, um eine Diskussion anzuregen.

FF2 Forschungsfrage 2 wurde ausschließlich im Zuge der Interviewstudie untersucht. Dabei zeigt sich erwartungskonformm dass bei den Probanden beim Durcharbeiten der Items vorwiegend intendierte kognitive Prozesse ausgelöst wurden. Dies zeigte sich darin, dass die Gesprächspartner, die meisten Items (aufgrund der formulierten Hürden) als schwierig wahrnahmen (Schwierigkeitsbezug), die in den Items beschriebenen komplexen Handlungen inhaltlich passend paraphrasierten (Domänenspezifität/Handlungsbezug) oder passende eigene Erfahrungen assoziierten, sowie Selbsteinschätzungen vornehmen konnten (Selbstreferentialität). Immer dann, wenn es Anhaltspunkte dafür gab, dass erwartete kognitive Aktivitäten nicht ausgelöst wurden, was die Befragten zum Teil selbst zum Ausdruck brachten, wurde dies als Anlass für eine Itemüberarbeitung genutzt. Häufiger mussten Handlungsbarrieren ersetzt oder angepasst werden, oder Items wurden als zu schwierig eingeschätzt, weil die Schwierigkeit sowohl durch die Handlung als auch die Hürde oder eine zusätzliche Zuspitzung realisiert wurde. Im Sinne einer kognitiven Entlastung wurde die Satzstruktur der Items vereinheitlicht. Insgesamt hat die Itemüberarbeitung wesentlich dazu beigetragen, das Maß der anvisierten kognitiven Prozesse weiter zu erhöhen, sodass nun davon ausgegangen werden kann, dass die formulierte Hypothese weitgehend zutreffend ist.

FF3 Das Itemverständnis wurde einerseits im Rahmen der Interviewstudie untersucht und andererseits wurden Experten gebeten, die Items auf ihre Verständlichkeit hin zu beurteilen. In der Regel wurden die Items wie intendiert verstanden, was beispielsweise an passenden Paraphrasierungen ersichtlich wurde. Hinweise seitens der Experten oder gehäufte Verständnisprobleme wurden wiederum in der Itemrevision aufgegriffen. Insbesondere Fachbegriffe wurden – wann immer möglich – ersetzt oder wie im Falle des Begriffs der Schülervorstellungen durch einen einleitenden Text erläutert. Aus den Interviews konnten Indizien dahingehend gewonnen werden, dass zu Planungsitems nicht immer Planungssituationen assoziiert wurden. Umgekehrt wurden beim Paraphrasieren von Items der Durchführungsdimension

[1] Jahrzehntelang wurde die TES in dem Glauben der Forschenden eingesetzt, damit Lehrer-Selbstwirksamkeitserwartungen in der Tradition Banduras zu erheben.

zum Teil Planungssituationen erwähnt. Als Konsequenz wurde in der Itemrevision darauf geachtet, die jeweilige Dimension ggf. stärker hervorzuheben. Durch die vielen Überarbeitungszyklen der Items und die begründet vorgenommenen Revisionen ist davon auszugehen, dass die Items für die Probandengruppen der Physik-Lehramtsstudierenden, der Physik-Referendare und der Physik-Lehrkräfte verständlich sind.

FF4 Die Authentizität der formulierten Items wurde sowohl in der Interviewstudie als auch der Expertenbefragung untersucht, wobei diese Frage ihre Bedeutung dadurch erhält, dass davon ausgegangen wird, dass lediglich Selbsteinschätzungen bezüglich solcher Handlungen die Handlungsinitiative beeinflussen, die als relevant erachtet werden. In dem Sinne ist die Frage wichtig, ob Probanden und Experten den beschriebenen Tätigkeiten bez. der Planung und Durchführung von Unterricht eine gewisse Wichtigkeit beimessen bzw. die beschriebenen Tätigkeiten als schulalltagsrelevant einstufen. In der Regel wurden die Items als authentisch bewertet. In Einzelfällen wurden Anpassungen vorgenommen, insbesondere dann, wenn ein Item mehrfach als unrealistisch (zu schwer bzw. Hürde unpassend) eingeschätzt wurde. Insgesamt kann nach der mehrfachen Überarbeitung davon ausgegangen werden, dass es sich bei den formulierten Items um solche handelt, die von (angehenden) Physik-Lehrpersonen als relevant beurteilt werden.

FF5 Die Frage nach einem geeigneten Antwortformat wurde sowohl aus einem qualitativen, als auch quantitativen Blickwinkel beleuchtet und darüber hinaus durch Thorid Rabe und Olaf Krey in einer zusätzlichen Akzeptanzbefragung. Anhand der Ergebnisse aus der Zusatzbefragung sowie den Kommentaren und Einschätzungen der Befragten in den Interviews wurde die Entscheidung getroffen, von einem vierstufigen auf ein sechsstufiges Antwortformat zu wechseln, welches dann in der Hauptstudie erprobt wurde. Im Rahmen der Raschanalyse konnten keine Hinweise hinsichtlich eines Nicht-Funktionierens der Antwortskala abgeleitet werden. Insofern wird die Entscheidung der Nutzung eines sechsstufigen Antwortformates weiterhin präferiert, da davon ausgegangen wird, dass mit dieser Skala differenziertere Selbsteinschätzungen von Befragten vorgenommen und damit erfasst werden können.

FF6 Die Eindimensionalität der Skalen wurde zum einen mittels konfirmatorischer Faktorenanalysen in der ersten Pilot- und der Hauptstudie untersucht; über die Repräsentativität und Angemessenheit der Skalen gaben die Experten Auskunft. Tatsächlich konnten anhand der Expertenurteile fehlende inhaltliche Facetten durch Itemneukonstruktionen ergänzt, ansonsten aber sichergestellt werden, dass die jeweiligen Domänen treffend durch die jeweils vorgelegte Itemauswahl beschrieben werden. Die Ergebnisse der faktorenanalytischen Verfahren sprechen für die

Eindimensionalität der entwickelten Skalen. Somit können die beiden formulierten Hypothesen vorläufig als zutreffend angenommen werden. Aufgrund des ständigen Wandels einer als zeitgemäß zu wertenden Unterrichtspraxis sind in der Zukunft Anpassungen zu erwarten, die dann erneuter Validierungsmaßnahmen bedürfen.

FF7 Die statistischen Analysen sowohl im Zuge des klassischen als auch des probabilistischen Ansatzes sprechen für eine zuverlässige Abbildung der intendierten Konstrukte durch die konstruierten Skalen, sodass Hypothese 7.1 als zutreffend angesehen werden kann. Der Reliabilitätskoeffizient, der im klassischen Paradigma die häufigste Verwendung findet, ist Cronbach's Alpha α_C, wobei dieser zufriedenstellende Werte zwischen .79 und .89 für die neu konstruierten Skalen annimmt.

FF8 Forschungsfrage 8 kann kontrovers diskutiert werden. Im Rahmen der qualitativen Studien äußerten die Experten, dass sie die konzeptionelle Trennung der Dimensionen „Planung“ und „Durchführung“ grundsätzlich für sinnvoll erachteten. Gleichzeitig betonten einige ihre Unsicherheit dahingehend, ob sich die Unterscheidung auch empirisch nachweisen lassen würde. In den Interviews wurden zu Planungsitems gelegentlich Durchführungssituationen assoziiert und umgekehrt, was entweder daran gelegen haben könnte, dass die Items uneindeutig formuliert waren, oder aber die Befragten nicht stringent zwischen Planung und Durchführung von Physikuntersicht unterscheiden. Aus statistischen Untersuchungen geht einerseits hervor, dass sowohl Planungs- als auch Durchführungsskalen zunächst scheinbar eindimensionale Konstrukte abbilden, die nach Meinung der Experten auch inhaltlich zu differenzieren sind (vgl. FF6). Auch die Korrelationsmuster entsprechen weitgehend der Vorhersage (vgl. Hypothese 8.2). Werden eindimensionale Modelle, die die Unterscheidung von Planungs- und Durchführungsdimension aufheben, mit zweidimensionalen Modellen, in denen die Planungsdimension mit der Durchführungsdimension korreliert ist, verglichen, so sprechen statistische Tests in der Regel für die zweidimensionalen Modelle. Die latenten Korrelationen sind von eins verschieden, liegen aber derart hoch, dass deshalb die zweidimensionalen Modellierungen in Zweifel gezogen werden können. Die Raschanalyse legt keine zusätzliche Dimension im Datensatz, der aus Planungs- *und* Durchführungsitems gebildete wird, nahe. Insgesamt ist letztere Analyse im Rahmen der Berechnung des Raschmodells ggf. nicht die geeignetste Methode zur Untersuchung der Fragestellung. Angesichts dieser Ergebnisse wurden post-hoc Kurzskalen zu jedem Handlungsfeld entwickelt, die sowohl Items auf Planungs- als auch auf Durchführungsebene enthalten. Die umfangreiche Validierung der Kurzskalen steht noch aus. Zusammenfassend erscheint die grundsätzliche Aufhebung der Unterscheidung der Dimensionen nicht gerechtfertigt und/oder sinnvoll zu sein. Gegebenenfalls muss

je nach Forschungsfrage zwischen den zur Verfügung gestellten Skalen gewählt werden. Die gegenseitige Abgrenzbarkeit der neu konstruierten Skalen für die intendierten Konstrukte sollte weiterhin Gegenstand von Validierungsbemühungen sein.

FF9 Die Korrelationsanalysen, in denen Zusammenhänge der neu konstruierten Skalen zu Skalen anderer Konstrukte untersucht wurden, entsprechen im Wesentlichen den Erwartungen und damit den Hypothesen 9.1 bis 9.5. Die Skalen zu Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern korrelieren zu den eingesetzten Skalen der allgemeinen Selbstwirksamkeitserwartung und zur Skala der Lehrer-Selbstwirksamkeitserwartung im schwachen bis mittleren Bereich, wobei die Korrelationen erwartungsgemäß zum letztgenannten Konstrukt etwas höher ausfallen.[2] Ähnliches gilt für die Korrelationen zu den Skalen, die internale Kontrollüberzeugungen und Persönlichkeitsvariablen abbilden, wobei keine Korrelationen zur Persönlichkeitsdimension „Verträglichkeit“ gefunden wurden. Die Korrelationen der neu entwickelten Skalen zur Skala des physikalischen Selbstkonzeptes fallen erwartungsgemäß äußerst gering aus. Insgesamt sprechen diese Ergebnisse dafür, dass sich die neu entwickelten Skalen, die Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern operationalisieren, jeweils empirisch von verwandten Konstrukten abgrenzen lassen. Die Neukonstruktion erweist sich damit als gerechtfertigt.

FF10 Bereits die Auswertung der Interviewdaten lieferte keine Anhaltspunkte für ein gruppenspezifisches Verständnis der Items. Auch die durchgeführten Mehrgruppenvergleiche im Rahmen klassischer wie probabilistischer Ansätze erbrachten diesbezüglich keine Hinweise. In der Regel konnte für die Skalen (partielle) skalare Messinvarianz konstatiert werden, wobei „verursachende“ Items jeweils benannt wurden. Obwohl die Gruppengrößen teils niedrig ausfielen, wurden auch innerhalb der drei Befragungsgruppen Mehrgruppenvergleiche für verschiedene Subgruppen (nach Geschlecht, Praxiserfahrung, Schulform/Studiengang) durchgeführt. Auch bei diesen Analysen ließ sich zumeist als Ergebnis (partielle) skalare Messinvarianz der Skalen nachweisen. Die Durchführung von Mittelwertvergleichen ist damit sowohl zwischen den Befragungsgruppen als auch zwischen den genannten Subgruppen gerechtfertigt. Insgesamt können die Operationalisierungen damit für unterschiedliche Befragungsgruppen als gelungen angesehen werden.

FF11 Aus den erhobenen Daten scheinen sich Indizien für den sogenannten Praxisschock im Referendariat ableiten zu lassen (vgl. Hypothese 11.3 (3)), da die Skalenmittelwerte der Physik-Referendare in der Regel (signifikant) niedriger

[2] Die Ergebnisse der Korrelationsanalyse aus der ersten Pilotstudie sind damit höchstwahrscheinlich als Artefakte zu charakterisieren.

ausfallen als die Skalenmittelwerte der Physik-Lehramtsstudierenden und die Skalenmittelwerte der Physiklehrkräfte signifikant höher sind als die der beiden anderen Kohorten. Für die Skala der Lehrer-Selbstwirksamkeitserwartung konnte ein ähnlicher „Trend“ belegt werden, während der Vergleich der Skalenmittelwerte der Skala zur allgemeinen Selbstwirksamkeitserwartung bezogen auf die drei Befragungsgruppen keinen Praxisschock nahelegt. Beachtet werden muss, dass keine längsschnittlichen Daten vorliegen, also keine Entwicklung untersucht wurde. Dass für Physiklehrkräfte die höchsten Lehrer-Selbstwirksamkeitserwartungen dokumentiert werden konnten, entspricht Hypothese 11.3 (1). Für die Praxisphasen im Studium wurde aufgrund der zumeist guten Betreuungsverhältnisse (Mentoreneffekt) nicht mit einem Praxisschock gerechnet (Hypothese 11.3 (2)). Tatsächlich zeigten sich für Physiklehramtsstudierende mit Praxiserfahrungen im Studium sogar (signifikant) höhere Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern als für Studierende ohne solche Erfahrungen. Insofern ist Hypothese 11.3 (2) nicht aufrecht zu erhalten.

In der Literatur wurden in der Regel für Lehrkräfte an Gymnasien höhere Lehrer-Selbstwirksamkeitserwartungen angegeben als für Lehrkräfte, die an Schulen anderer Schulformen arbeiten und es wurde davon ausgegangen, dass diese Tendenz mit höheren Schülerleistungen an den Gymnasien zusammenhängt. Für die in dieser Arbeit befragten (angehenden) Lehrpersonen kann ein differenziertes Bild beschrieben werden, das teils konträr zu den Erwartungen ist: Zunächst einmal konnten für Physik-Lehrpersonen je nach Schulform, an der diese unterrichteten, in der Regel keine signifikanten Mittelwertunterschiede auf den Skalen beobachtet werden. Das gleiche gilt in fast allen Fällen für Physik-Referendare, deren Skalenmittelwerte mit denen von Quereinsteigern verglichen wurden. Lediglich für Physik-Lehramtsstudierende zeigt sich das erwartete Muster, allerdings in der umgekehrten (!) Richtung. Die Skalenmittelwerte von Physik-Lehramtsstudierenden mit einem anvisierten gymnasialen Abschluss fielen zumeist (signifikant) geringer aus, als die Skalenmittelwerte der Vergleichskohorte des Sekundarstufenlehramts. Ob es sich hierbei um ein Fachspezifikum oder ein Artefakt handelt, kann an dieser Stelle nicht beantwortet werden.

Außerdem wurden Hypothese 11.1 entsprechend keine geschlechtsspezifischen Unterschiede bezüglich der Skalenmittelwerte erwartet. Sowohl für die Kohorte der Physik-Referendare als auch die Kohorte der Physik-Lehrpersonen war diese Vermutung auch zutreffend. Für die Kohorte der Studierenden konnte die Tendenz beobachtet werden, dass Studentinnen des Physiklehramts scheinbar geringere Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern aufweisen als ihre männlichen Kommilitonen, wobei sich die Skalenmittelwerte insbesondere in der Durchführungsdimension signifikant unterscheiden (SWE-EX-

D, SWE-A-D, SWE-SV-D). Die Gründe für diese Ergebnisse können im Rahmen dieser Arbeit leider nicht beleuchtet werden.

üFF Nach der systematischen Sichtung sowie Abwägung der Interpretationen aller Ergebnisse, wird davon ausgegangen, dass die intendierten Konstrukte tatsächlich sowie zuverlässig durch die neu konstruierten Skalen abgebildet werden. Die Eindimensionalität der Konstrukte konnte für die durchgeführten Befragungen sowohl qualitativ (durch die Auswertung der Expertenbefragung) als auch quantitativ (durch statistische Analysen) nachgewiesen werden. Darüber hinaus spricht vieles für messinvariante Skalen bezüglich der drei Befragungsgruppen sowie ein funktionierendes Antwortformat. Es ergaben sich sowohl interpretierbare Korrelations- als auch Mittelwertstrukturen. Außerdem wurde eine Reihe von Teilergebnissen durch eine „methodische Kreuzvalidierung“ abgesichert. Damit wurden alle Validitätsfacetten nach Messick in mehreren Studien in den Blick genommen. Lediglich bezüglich der Zusammenhangsstruktur von Planungs- und Durchführungsdimension besteht nach wie vor größere Unsicherheit, was eine von mehreren strukturellen Validitätsfacetten betrifft. Die überwiegende Zahl der Ergebnisse spricht jedoch dafür, dass Testwerte, die mit Hilfe der neu entwickelten Skalen gewonnen werden, tatsächlich jeweils als Lehrer-Selbstwirksamkeitserwartungen in spezifischen physikdidaktischen Handlungsfeldern interpretiert werden dürfen. Damit wurde ein übergeordnetes Validitätsargument formuliert, welches im folgenden Abschnitt 13.2 neben anderen Aspekten Gegenstand der Diskussion ist.

13.1.3. Das wichtigste Ergebnis: Das Instrument

Neben der Vielzahl bereits genannter Ergebnisse, darf das eigentliche Ziel dieser Arbeit nicht vergessen werden. Nach dem umfangreichen Entwicklungs- und Validierungsprozess können im Endergebnis acht Skalen zur Verfügung gestellt werden, die mit einer hohen Wahrscheinlichkeit Lehrer-Selbstwirksamkeitserwartungen in den vier physikdidaktischen Handlungsfeldern Umgang mit Schülervorstellungen, Experimentieren, Elementarisieren und Umgang mit Aufgaben jeweils in den Dimensionen der Unterrichtsplanung und -durchführung abbilden. Die nach einer dreimaligen Itemüberarbeitung zum Einsatz empfohlenen Skalen können der Skalendokumentation oder dem Anhang G.4 (Skalenversion 4) entnommen werden. Außerdem wurden im Zuge der Diskussion um die Dimensionen post-hoc Kurzskalen für jedes Handlungsfeld erarbeitet und ersten Analysen unterzogen (vgl. Meinhardt u. a., 2016c). Die Kurzskalen sollten daher weitere Validierungszyklen durchlaufen.

13.2. Diskussion

13.2.1. Zur Reflexion der Itementwicklung und -überarbeitung

Rückblickend stand für die Instrumententwicklung jeweils das Bemühen im Vordergrund, zum herausgearbeiteten Theorierahmen passende Items bzw. Skalen zu entwickeln. Aus diesem Grunde musste zunächst ein theoretisches Grundgerüst erarbeitet werden, welches als Konsens vom Entwicklerteam getragen und dann schrittweise ergänzt oder überarbeitet wurde. Der theoretischen Aufarbeitung inhaltlich relevanter Themen wurde daher viel Zeit (teils auch parallel zur Itementwicklung) eingeräumt. Durch die mehrfache Überarbeitungen der Items konnten neuere Erkenntnisse – zum Beispiel immer „agent-means"-Relationen zu verbalisieren – in den Entwicklungsprozess integriert werden. Alle Entscheidungen oder Setzungen wie die Ableitung der Itemkonstruktionsregeln, Vermutungen über kognitive Prozesse oder die Wahl eines Antwortformates wurden begründet und mit der theoretischen Basis in Beziehung gesetzt.

Darüber hinaus wurden alle (Um)Formulierungen der Items im Team (kontrovers) diskutiert, sodass im Überarbeitungsprozess stets das Sechs-Augen-Prinzip angewandt wurde. Dieser sehr aufwendige und dadurch langwierige Prozess kann in seiner Dauer durchaus kritisiert werden. Letztlich standen für das Entwicklerteam aber immer die Qualität der Skalen im Vordergrund, was unter anderem aus den ernüchternden Analysen bereits existierender Instrumente resultierte. Tiefgründigkeit und Details benötigen leider sehr viel Zeit, was in der schnelllebigen Forschungscommunity nicht immer auf Verständnis trifft.

Auch die methodische Vielfältigkeit trug zur Verlängerung des Entwicklungsprojektes bei. Im Rückblick hatte jede Methodik jedoch ihren sinnvollen Beitrag, sodass auch an dieser Stelle ein Verzicht beispielsweise auf die Durchführung der Raschanalyse oder eine der qualitativ angelegten Studien nicht von Vorteil gewesen wäre. Gerade der systematische Vergleich des klassischen und des probabilistischen Forschungsparadigmas kann aus einer methodischen Perspektive als weiteres Ergebnis dieser Arbeit gewertet werden, auch wenn diesem im Rahmen dieser Arbeit keine kontinuierliche Aufmerksamkeit geschenkt wurde.

13.2.2. Zum übergeordneten Validitätsargument

Die Gesamtschau der erzielten Ergebnisse führte zu der Einschätzung, dass Daten, die mit dem neuen Instrument erhoben werden, valide, das heißt im Sinne der intendierten Testwertinterpretation, gedeutet werden können. Insbesondere konnten keine Hinweise gefunden werden, die für die konkurrierenden Testwertinterpretationen sprechen, was insbesondere damit zu tun hat, dass darauf geachtet wurde, „agent-means"-Relationen zu operationalisieren, sowie keine affektiven Parameter in die Items zu integrieren.

Einige Ergebnisse lassen jedoch keine eindeutigen Schlüsse zu – gerade die Dimensionsanalyse regt zu weiterer Forschung an. Einige Ergebnisse sprechen eher für die These der Unterscheidung der Dimensionen von Planung und Durchführung (klassische Korrelationsanalyse, Prüfung der Eindimensionalität der Skalen). Die hohen latenten Korrelationen (die sich jedoch von Eins unterscheiden!) können einerseits dahingehend gedeutet werden, dass eine sinnvolle Unterscheidung der Dimensionen ggf. nicht tragfähig ist oder aber einfach auf die praktische und nicht zu leugnende Verknüpfung beider (zu unterscheidenden) Dimensionen hinweisen. Die Ergebnisse der Raschanalyse diesbezüglich sollten nur vorsichtig interpretiert werden, da der Vergleich mehrdimensionaler Raschmodelle hier unter Umständen eher geeignet ist, um die Frage der Dimensionalität zu beantworten.

Weitere, die angestrebte Testwertinterpretation *möglicherweise* einschränkende Hinweise können aus den Varianzanalysen abgeleitet werden, da nicht alle Mittelwertunterschiede im Einklang mit den formulierten Hypothesen stehen. Hierzu ist zu sagen, dass die dokumentierten Mittelwertstrukturen zunächst einmal grundsätzlich interpretierbar sind. Weiterhin ist anzumerken, dass die Analyse der Mittelwertunterschiede keine notwendige Validierungsmaßnahme darstellt, sondern im besten Falle zusätzliche Plausibilitätsargumente liefert. Dies ist unter anderem damit zu begründen, dass Hypothesen beispielsweise lediglich für Selbstwirksamkeitsskalen auf weniger spezifischen Niveaus aus der Literatur abgeleitet werden können, weil – wie in diesem Fall – keine spezifischeren Skalen existieren. Aus diesem Grunde sind die Ergebnisse zum jetzigen Zeitpunkt in keinem Fall als Kontraargumente zu werten. Unter Umständen weisen sie sogar auf Domänenspezifika hin, die bisher noch nicht im Fokus standen. Natürlich ist es auch möglich, dass es sich einfach um Artefakte handelt. Interessant ist, dass die geschlechtsspezifischen Unterschiede in der Studierendenkohorte nur auf den neu entwickelten Skalen (und nur auf Skalen der Durchführungsdimension) beobachtet werden konnten und beispielsweise nicht durch die L-SWE-Skala detektiert wurden. Weitere Fragen ergeben sich auch daraus, dass entgegen der Vermutung angehende Gymnasiallehrkräfte eher niedrigere physikdidaktische Selbstwirksamkeitserwartungen zu haben scheinen.

Insgesamt muss einschränkend angegeben werden, dass die zusätzlich eingesetzten Skalen zum Teil den Erwartungen an ihre Qualität nicht vollumfänglich entsprachen und teils größere Adaptionen vorgenommen werden mussten, um beispielsweise die Eindimensionalität oder die Messinvarianz zu gewährleisten. Die Testwertinterpretationen der zusätzlich verwendeten Skalen weichen daher in der Regel leicht von denen in anderen Studien ab, in denen keine Items aus den Skalen entfernt werden mussten.

Methodisch problematisch könnten ggf. die ungleichen Gruppengrößen im Rahmen der Messinvarianzprüfung und der Mittelwertvergleiche sein, auch wenn diese Vorgehensweise üblich ist. Wann immer möglich und sinnvoll wurden Zufallsstichproben gezogen,

um die Ergebnisse zusätzlich abzusichern. Eine weitere Schwierigkeit bestand in der Rekrutierung erfahrener Physiklehrpersonen. Probanden dieser Gruppe wurden ausschließlich auf freiwilliger Basis mit dem entwickelten Online-Fragebogen befragt, wobei die Initiative für die Befragung von den Befragten ausging und nicht wie in den anderen Kohorten durch Dozentinnen und Dozenten oder Seminarleitende. Die Ergebnisse dieser Kohorte unterliegen daher gegebenenfalls erhebungstechnisch bedingten Verzerrungen. Insbesondere dass in der Gruppe der Physiklehrpersonen eine sehr hohe Ausprägung der Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern vorliegt, kann möglicherweise mit dieser Positivauswahl zusammenhängen. Deshalb sollte die Entwicklung von physikdidaktischen Selbstwirksamkeitserwartungen längsschnittlich für verschiedene Gruppen mit unterschiedlicher Praxiserfahrung weiter untersucht werden.

Abschließend soll betont werden, dass im Rahmen der argumentbasierten Validierung inhaltliche, substantielle, generalisierende, externale und strukturelle Validitätsaspekte untersucht wurden. Damit wurden bis auf die konsequenzenbezogenen alle von Messick (1995) vorgeschlagenen und in den Standards (vgl. AERA, APA, NCME, 2014) aufgegriffenen Validierungsfacetten berücksichtigt. Darüber hinaus wurde das vorgeschlagene Verfahren von Kane (2001) schrittweise durchgeführt (vgl. Unterabschnitt 2.1.4), wobei eine Revision der Testwertinterpretation nicht nötig war. Da spezifische Testnutzungen und die damit verbundenen Konsequenzen im Validierungsprozess ausgespart wurden, könnten sich diesbezüglich Folgestudien anschließen bzw. Testanwender sollten sich dieser Tatsache bewusst sein. Das Validitätsargument ist damit als rein deskriptiv zu charakterisieren. Außerdem resultiert aus dieser Art des Validitätsarguments sowie der in dieser Arbeit im Zentrum stehenden Testwertinterpretation, dass der Einsatz des Instruments und die Interpretation der Daten entsprechend der Intention bisher wenige Schlussfolgerungen zulässt. Insbesondere extrapolierende, bewertende und entscheidungsrelevante Schlüsse bleiben unzulässig. Bewertende sowie generalisierende Schlüsse sind durch das gewählte Validierungsverfahren mit großer Sicherheit sinnvoll möglich.

13.2.3. Zur Subjektivität und einem mittleren Spezifitätsniveau

An den in den Items integrierten Hürden wird eine Besonderheit von Selbstwirksamkeitserwartungsitems besonders sichtbar: Sie unterliegen *immer* einem gewissen Deutungsspielraum seitens der Probanden. Je spezifischer die Items formuliert sind, desto kleiner wird dieser. Der Interpretationsspielraum kann daher minimiert, aber nie gänzlich eliminiert werden.

Unter Umständen kann die Problematik als Messproblematik verstanden und kritisiert werden. Je allgemeiner ein Item formuliert ist, desto größer und vor allem individueller kann der persönliche Referenzrahmen sein, der zur Beantwortung des Items herangezogen wird und desto weniger messen die Items dasselbe für verschiedene Personen. Dieser Problematik kann aus mehreren Perspektiven heraus begegnet werden.

1. **Minimierung durch „Vignettentests“:** Will man Selbstwirksamkeitserwartungen so vergleichbar wie möglich erheben, so sollten bei gleichzeitiger Kontrolle zusätzlicher Variablen wie den tatsächlichen Fähigkeiten möglichst konkrete (physikdidaktische) Aufgaben in konkreten Situationen vorgegeben werden, bezüglich derer sich die Probanden einschätzen müssen. Damit wird der Referenzrahmen zur Beurteilung – sozusagen eine Art objektive Schwierigkeit – vorgegeben. Ein Beispiel für diese Art der Erfassung wäre das Folgende: Einem Probanden wird ein Schülertext zu einer physikalischen Fragestellung vorgelegt. Der Proband soll einschätzen, inwiefern er sich selbst in der Lage sieht, in einer festgelegten Zeitspanne aus diesem konkreten Text die enthaltenen Schülervorstellungen zu enkodieren und aus diesen Ideen für die weitere Unterrichtsgestaltung abzuleiten.

   Es muss berücksichtigt werden, dass die Problematik der unterschiedlichen Schwierigkeitseinschätzung auch bei diesem Messverfahren bestehen bleibt. Es kann jedoch argumentiert werden, dass sich gerade in der individuell unterschiedlich wahrgenommenen Schwierigkeit äußert, ob eine Person sich als wirksamer oder weniger wirksam einschätzt. Das Messverfahren rekurriert auf das Spezifitätsniveau, welches am spezifischsten Selbstwirksamkeitserwartungen erfassen kann. Es ist davon auszugehen, dass ein sehr hohes Maß an Prädiktivität gegeben ist, wenn die Tätigkeiten im Anschluss an die Erhebung der Selbstwirksamkeitserwartung tatsächlich durchgeführt werden müssen. Nicht immer macht die Erfassung solch spezifischer Selbstwirksamkeitserwartungen Sinn; insbesondere dann nicht, wenn allgemeinere Forschungsfragen von Interesse sind.

2. **Wahl eines mittleren Spezifitätsniveaus; Bewusstmachen des Interpretationsspielraums:** Werden Selbstwirksamkeitserwartungen also auf einem weniger spezifischen Niveau gemessen, beispielsweise auf einem mittleren Spezifitätsniveau, so kommt die Problematik des subjektiven Interpretationsspielraumes deutlicher zum Tragen. Dies wäre der Fall, würde man Probanden bei gleichzeitiger Kontrolle zusätzlicher Variablen keine ganz konkrete physikdidaktische Aufgabe zur Selbsteinschätzung vorlegen, sondern statt dessen die subjektive Gewissheit erfragen, eine Aufgabe eines bestimmten Typs lösen zu können (z. B. Ableiten von Schülervorstellungen aus Schülertexten). An dieser Stelle könnte argumentiert werden, dass die Selbsteinschätzungen nicht mehr vergleichbar bzw. interpretierbar wären, weil Probanden auf einer Skala zum selben Urteil über ihre Fähigkeiten gelangen könnten, jedoch auf Grundlage gänzlich unterschiedlicher Referenzrahmen, also unterschiedlich schwieriger Aufgabenassoziationen. Man könnte solche Messungen als unzulässig abweisen, da auf dem Papier beide Probanden die gleiche Selbstwirksamkeitserwartung hätten, in der Praxis eine Person jedoch als selbstwirksamer gelten würde, die ihrem Urteil schwierigere Aufgaben zugrunde legt.

Dieses Dilemma trifft auf das Spezifitätsniveau der in dieser Arbeit entwickelten Skalen zu, da diese ein mittleres Spezifitätsniveau adressieren. Dass trotzdem sinnvolle Messungen möglich sind, kann anhand nachfolgender Argumente gestützt werden. Zunächst ist es unwahrscheinlich, dass ein Proband lediglich eine singuläre Situation als Referenzrahmen vornimmt. Wahrscheinlicher ist es, dass die Summe eigener Erfahrungen, etc. als eine Art mittlerer Erfahrungshorizont zur Beantwortung des Items herangezogen wird. Dies konnte auch in den Interviewsituationen beobachtet werden.

Die formulierten Hürden stehen ggf. im Widerspruch zu dieser Argumentation, da beispielsweise unter wenig Planungszeit tatsächlich jeder Proband etwas anderes verstehen kann und auch unterschiedliche Erfahrungen dieses Maß nicht „mitteln". Hierzu ist zu sagen, dass die Hürden jedoch den Zweck erfüllen, in jedem Fall schwierigkeitserzeugend zu wirken, auch wenn das Maß individuell festgelegt wird. Damit wird jedoch verhindert, dass ein Item als überhaupt nicht schwierig wahrgenommen werden kann, wie es auch bei ganz konkreten Hürden („ein Tag Planungszeit") der Fall sein könnte.

3. **Wahl eines globaleren Spezifitätsniveaus:** Es ist davon auszugehen, dass es mit globaler erfassten Selbstwirksamkeitserwartungen eher nicht möglich ist, einen als ähnlich schwierig wahrgenommenen Erfahrungshorizont über Probanden hinweg abzubilden. Die Unsicherheit der Items würde zunehmen; ihre Prädiktivität sinken. Diese These wäre im Einklang zu Banduras Hinweis, dass globalere Messungen weniger aussagekräftig sind (vgl. Kapitel 3).

13.3. Ausblick

> „In conclusion, I believe we have learned a great deal in our efforts to accurately and reliably measure the construct we label „teacher efficacy," but we still have much to learn. Lessons from the past have taught us that we must be more careful, more thoughtful, and more sophisticated in our measurement efforts, and most especially in our interpretations of those measures."
>
> *Guskey, 1998, S. 4*

Resümierend ergeben sich für mich aus der vorliegenden Forschungsarbeit drei große Themenkomplexe für gewinnbringende Anschlussforschung, die ich kurz skizzieren möchte.

Zum einen kann aufgrund der zusammenfassenden Darstellung der Literatur konstatiert werden, dass die Entwicklung von (domänenspezifischen) Lehrer-Selbstwirk-

samkeitserwartungen weitgehend auf Hypothesen und weniger auf empirischer Evidenz beruht, sodass ein Desiderat dahingehend formuliert werden kann, diese Lücke zu schließen. Das nun neu vorliegende Forschungsinstrument ermöglicht bezüglich dieser Fragestellung längsschnittliche Untersuchungen. Auch hinsichtlich der Quellen und Determinanten physikdidaktischer Lehrer-Selbstwirksamkeitserwartungen besteht weiterer Forschungsbedarf, gerade dann, wenn man aufgrund der Überzeugung, dass hohe bzw. realistisch ausgeprägte Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern für die Planung und Durchführung von gutem Physikunterricht eine wichtige Rolle spielen, diese gezielt in den verschiedenen Ausbildungsphasen entwickeln möchte. Darüber hinaus können aus der vorliegenden Arbeit bezüglich der Entwicklung domänenspezifischer Lehrer-Selbstwirksamkeitserwartungen konkrete Hypothesen abgeleitet und in Folgeuntersuchungen überprüft werden. So scheint sich der sogenannte Praxisschock insbesondere mit Eintritt in den Vorbereitungsdienst bemerkbar zu machen, während universitäre Praxisphasen scheinbar weniger an den eigenen Selbstwirksamkeitserwartungen rütteln und im Gegenteil dieses sogar stärken. Unter Umständen muss eine genauere Differenzierung der Praxisphasen vorgenommen werden. Beispielsweise ist es denkbar, dass das Praxissemester, welches in einigen Bundesländern im Rahmen des Lehramtsstudiums absolviert wird, eher eine praxisschockähnliche Wirkung auf die Ausprägung physikdidaktischer Lehrer-Selbstwirksamkeitserwartungen hervorruft.

Gerade wenn es um das Verständnis des Konstruktes der physikdidaktischen Lehrer-Selbstwirksamkeitserwartungen, sowie dessen Quellen und Determinanten geht, kann unter Umständen eine zusätzliche qualitative Forschungsperspektive erhellend sein. Denn Hypothesen über Einflussfaktoren etc. können durch eine rekonstruierende Herangehensweise im Rahmen von Gesprächs- bzw. Interviewsituationen gegebenenfalls überhaupt erst gewonnen werden.

Ein weiterer großer Themenkomplex eröffnet sich durch das Interesse an einem besseren Verständnis des Zusammenspiels domänenrelevanter Lehrer-Selbstwirksamkeitserwartungen und entsprechender Wissenskomponenten in Hinblick auf das Lehrerhandeln bzw. die Kompetenz einer Lehrperson. Durch das neue Instrument können nun Wissensstrukturen und Selbstwirksamkeitserwartungen auf einem ähnlichen Spezifitätsniveau erfasst werden, sodass davon auszugehen ist, das auch die Vorhersage einer bestimmten Performanz zunehmen müsste. Um das komplexe Verhalten von Lehrpersonen tatsächlich annähernd vorhersagen oder verstehen zu können, müssen sehr wahrscheinlich eine Vielzahl weiterer Variablen wie Zielvorstellungen oder Handlungsergebniserwartungen in das Modell integriert werden:

> „[E]xpectation is [not] the sole determinant of behavior. Expectation alone will not produce desired performance if the component capabilities are lacking. Moreover, there are many things that people can do with certainty of success that they do not perform because they have no incentives to do so. Given appropriate skills

> and adequate incentives, however, efficacy expectations are a major determinant of people's choice of activities, how much effort they will expend, and of how long they will sustain effort in dealing with stressful situations“ (Bandura, 1977a, S. 194).

Es bleibt ein Forschungsdesiderat, die Rolle von Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern beim Planen, Durchführen und Reflektieren von Physikunterricht zu erkunden.

Teil V.

Anhang

A. Auswahl an Instrumenten zur Erfassung des Selbstkonzeptes

A.1. Instrumente zur Erfassung des Selbstkonzeptes von Schülerinnen und Schülern.

Nachfolgend wird eine Auswahl von Messinstrumenten zur Erfassung des (globalen, akademischen, ...) Selbstkonzeptes aufgeführt (vgl. Tabelle A.1). Aufgrund des Schwerpunktes der Arbeit werden hauptsächlich Fragebogen-Instrumente referiert und damit andere Arten der Erfassung des Selbstkonzeptes vernachlässigt (vgl. u. a. Greenwald, McGhee & Schwartz, 1998; Greenwald & Farnham, 2000). Die Auflistung erhebt nicht den Anspruch der Vollständigkeit. Um einen Vergleichsmaßstab zur Erfassung von Selbstwirksamkeitserwartungen anzubieten, werden für einzelne Faktoren, die domänenspezifische Selbstkonzepte adressieren, einzelne Itembeispiele angeführt.

Tabelle A.1.: Instrumente zur Erfassung des Selbstkonzeptes von Schülerinnen und Schülern

allgemeines Selbstkonzept:
Perceived Competence Scale for Children: vgl., Harter (1982), dt. Version: Asendorpf und van Aken (1993)
Itembeispiele (Faktor: kognitive/schulische Kompetenz, Auszug dt. Version): – Bist Du: gar nicht gut oder einigermaßen gut im Rechnen? Bist Du: ziemlich gut oder sehr gut im Rechnen? – Kannst Du: sehr gut oder ziemlich gut addieren? Kannst Du: einigermaßen gut oder gar nicht gut addieren?
SKI (Selbstkonzept Inventar für Kinder im Vorschul- und Grundschulalter): Eggert, Reichenbach und Bode (2003)
SDQI (Self Description Questionnaire I): vgl. u. a., Marsh (1990a), dt. Version: Arens, Trautwein und Hasselhorn (2011)
Itembeispiele (Faktor: mathematisches Selbstkonzept, Auszug dt. Version): –“Mathe fällt mir leicht.“ –“Ich mag Mathe.“ –“Mathe interessiert mich.“

Fortsetzung auf der nächsten Seite...

Tabelle A.1.: *Fortsetzung*

SDQ II-S (Self Description Questionnaire II-Short): Marsh, Ellis, Parada, Richards und Heubeck (2005)
SDQ III (Self Description Questionnaire III): Marsh und O'Neill (1984) Itembeispiele (Faktor: Mathematics): – „At school, my friends always came to me for help in mathematics." – „I have always done well in mathematics classes."
FSKN (Frankfurter Selbstkonzeptskalen): Deusinger (1986)
akademisches/schulisches Selbstkonzept:
ASDQ (Academic Self Description Questionnaire I & II): vgl. u. a., Marsh (1990c) Itembeispiele (Faktoren: Mathematik, Schule im Allg.): – „I learn things quickly in [a specific subject area, e.g., mathematics]." (Ma-Skala) –"„Compared to others my age I am good at math." (Ma-Skala) – „I learn things quickly in most school subjects." (allg. Skala)
SKSLF (Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten): Rost und Lamsfuss (1992)
SASK (Skalen zum akademischen Selbstkonzept): Dickhäuser, Schöne, Spinath und Stiensmeier-Pelster (2002) Itembeispiele (Faktor: Fähigkeiten, Schülerversion): –"Wenn ich mir angucke, was wir in der Schule können müssen, finde ich, dass ich . . . wenig kann/ viel kann." –"Ich kann in der Schule . . . weniger als früher/mehr als früher." –"Ich kann in der Schule . . . weniger als meine Mitschüler(inne)n/ mehr als meine Mitschüler(inne)n." –"Ich kann in der Schule . . . wenig/viel."
DISK-Gitter (Differentielles Selbstkonzept schulischer Leistungen): Rost und Sparfeldt (2002) aufbauend auf Rost und Lamsfuss (1992) Itembeispiele (Faktor: ... entspricht Mathe/Deutsch/Physik/Geschichte): – „Ich weiß in ... die Antwort auf eine Frage schneller als die Anderen." – „Manchmal fühle ich mich in ... anderen überlegen und glaube, da[ss] sie noch manches von mir lernen können." – „Ich kann in ... Sachen selbst 'rauskriegen." – „Ich bin in ... zufrieden mit meiner Fähigkeit, vor der Klasse zu sprechen." – „Es fällt mir in ... leicht, Probleme zu lösen." – „In ... fallen mir gute Noten zu." – „Ich gehöre in ... zu den Guten." – „Ich habe ein gutes Gefühl, was meine Arbeit in ... angeht."

Fortsetzung auf der nächsten Seite...

Tabelle A.1.: *Fortsetzung*

Selbstkonzept junger Erwachsener: Schwanzer, Trautwein, Lüdtke und Sydow (2005) nach Marsh und O'Neill (1984)
Itembeispiele (Faktor: mathematische Fähigkeiten): – „Ich bin gut in Mathematik." – „Fähigkeiten Ich habe Verständnisschwierigkeiten bei allem, für das man Mathematik braucht." – „Ich war in Mathematik immer gut." – „Ich bin bei Aufgaben, die mathematisches Denken erfordern, nie gut."
fachspezifisches Selbstkonzept:
Physikalisches Selbstkonzept: (Hoffmann u. a., 1998) – „Mir fällt Physik leicht." – „Ich bin für Physik begabt." – „Ich bringe in Physik gute Leistungen.

A.2. Instrumente zur Erfassung des Selbstkonzeptes von (angehenden) Lehrerinnen und Lehrern.

Tabelle A.2.: Instrumente zur Erfassung des Selbstkonzeptes von (angehenden) Lehrerinnen und Lehrern

Guskey (1984), Guskey (1988); Neukonstruktion Itembeispiele: – „I am very proud of my performance as a teacher.“ – „I sometimes have doubts about the effectiveness of my teaching.“
Villa und Calvete (2001); Neukonstruktion Itembeispiele: – „I think I have good teaching abilities/skills.“ – „My collegues regard me as a competent teacher.“
Yeung, Craven und Kaur (2014, S. 320); Neukonstruktion, Orientierung an SDQ III von Marsh und O'Neill (1984) Itembeispiele: – „I am good at teaching most subjects that I teach.“ – „I am confident that if students misbehave I can manage the situation.“ – „I can teach numeracy skills well.“ – „I am confident that I can maintain a calm learning environment through engaging students in worthwhile learning.“ – „I can teach literacy skills well.“
Panel zum Lehramtsstudium (PaLea): Kauper u. a. (2012, S. 27-33), Retelsdorf, Bauer, Gebauer, Kauper und Möller (2014); Neukonstruktion beruflicher Selbstkonzepte (SK) orientiert an KMK (2004): SK Fach, SK Erziehen, SK Diagnostizieren, SK Innovieren, SK Medien, SK Beraten, Allgemeine Begabung Itembeispiele: – SK Fach: „In meinen Studienfächern bin ich sehr gut.“ – SK Fach: „Die Inhalte meiner Studienfächer bereiten mir keine Schwierigkeiten.“ – SK Fach: „Ich denke, meine Studienfächer liegen mir besonders.“ – SK Fach: „Was meine Studienfächer angeht, bin ich ziemlich fit.“
Panel zum Lehramtsstudium (PaLea): Kauper u. a. (2012, S. 36), Lehrerbezogenes Selbstkonzept als Facette der Berufswahlmotivation nach Pohlmann und Möller (2010): Itembeispiele: „Ich habe das Lehramtsstudium gewählt, weil...“ – „...ich gut erklären kann.“ – „...ich fachliche Inhalte interessant vermitteln kann.“ – „...ich denke, dass ich eine gute Lehrerin/ein guter Lehrer sein werde.“ – „...mir schon häufiger rückgemeldet wurde, Dinge verständlich erklären zu können.“ – „...ich bereits erfolgreich Nachhilfe gegeben habe oder Mitschülern Dinge erfolgreich erklärt habe.“ (Item **nicht** bei Pohlmann und Möller (2010) enthalten) – „...ich schon häufiger für meine Geduld beim Erklären von Sachverhalten gelobt wurde.“

B. Itemsammlung selten verwendeter Instrumente zur Erfassung von „Lehrer-Selbstwirksamkeitserwartungen“

Nachfolgend sind einige Studien aufgeführt, die Instrumente mit Bezug zur Theorie Banduras nutzen. Diese Instrumente wurden in der Regel für den Zweck der jeweiligen Studie entwickelt und sind selten erneut von anderen Forschenden verwendet bzw. validiert worden. In den meisten Fällen kann bereits durch die angegebenen Beispielitems ein Eindruck davon gewonnen werden, dass die Instrumente den Kriterien, die bei einer inhaltlichen Prüfung angelegt werden würden, nicht oder nur in Ansätzen genügen.

Tabelle B.1.: Itemsammlung selten verwendeter Instrumente zur Erfassung von „Lehrer-Selbstwirksamkeitserwartungen“.

Studie	**Itembeispiel**
Somech und Drach-Zahavy (2000, S. 653)	„Few teachers in my school can do a better job than I can.“
Midgley u. a. (2000, S. 38)	„I am certain that I am making a difference in the lives of my students.“
Friedman und Kass (2002, S. 682)	„I believe I enjoy a good rapport with the administrators at my school.“
Ho und Hau (2004, S. 318)	„Students would come to me when they have problems in their daily life because they know I can help.“
Rimm-Kaufman und Sawyer (2004, S. 328 f.)	„How much can you do to help other teachers with their teaching skills.“ (vgl. auch Bandura, 2006)
Evers, Tomic und Brouwers (2005, S. 40)	„I suffer from self-doubts.“ (vgl. Sherer u. a., 1982)
Betoret (2006, S. 523)	„I have sometimes had the feeling of not being cut out for this profession.“
Caprara, Barbaranelli, Steca und Malone (2006, S. 480)	„I am able to earn the trust and appreciation of all my colleagues.“
Senemoglu, Demirel, Yagci und Üstündag (2009, S. 167)	„I can lead my students to research and discovery.“

Fortsetzung auf der nächsten Seite...

Tabelle B.1.: *Fortsetzung*

Kastens (2009, S. 203)*	„Trotz des Zeitdrucks, unter dem ich als Lehrperson oftmals stehe und des Umfangs des Unterrichtsstoffes, gelingt es mir in meinem Unterricht ein Thema aus verschiedenen Perspektiven zu behandeln.“
Whitley (2010, S. 91)	„I have high expectations for the academic success of my students“
Devos, Dupriez und Paquay (2012, S. 211)	„I am capable of making students learn new material.“

* Das Item erscheint auf den ersten Blick gelungen zu sein. Tatsächlich fokussieren die meisten Items des Instrumentes auf Lehrerhandlungen. Allerdings werden alle 28 Items mit derselben Hürde dargeboten, sodass in Frage steht, ob die Items überhaupt authentisch sein können. Zudem ist die Hürde zweigeteilt, sodass die Items u. U. nicht eindimensional sind.

C. Exemplarische Darstellung der Bedeutung des Konstruktes ohne Berücksichtigung der Erhebungsinstrumente

Bei nachfolgendem längeren Zitat handelt es sich um einen Textausschnitt aus dem Artikel von van Dinther u. a. (2013, S. 170), in welchem die Bedeutung des Konstruktes der Lehrer-Selbstwirksamkeitserwartungen empirisch untermauert werden soll. Die als Beleg angeführten Quellen sind grün hinterlegt und werden in Tabelle C.1 darauf hin geprüft, welche Instrumente in der jeweiligen Studie genutzt werden.

> „The meaning of teacher self-efficacy as a type of self-efficacy regarding student achievement and motivation has been investigated in several studies (Woolfolk Hoy & Davis, 2006). Several researchers found significant relations between teacher sense of efficacy and student achievement. [...] Ashton und Webb (1986) demonstrated that students generally learn more from teachers with a high sense of efficacy than from teachers with a low sense of efficacy. Other researchers showed that students guided by high self-efficacious teachers achieved higher in subjects such as mathematics (Muijs & Reynolds, 2002; Ross, 1992, 1998) and reading (Ross, 1992, 1998) than did students guided by low self-efficacious teachers. Others connected teacher self-efficacy with student motivation (Midgley u. a., 1989) and students' interest in and attitude towards school (Woolfolk, Rosoff & Hoy, 1990). Referencing teacher behaviour, research has pointed out that teachers with a high sense of self-efficacy differ from those with low sense of self- efficacy in their teaching behaviour regarding issues such as classroom management, instruction, teacher feedback. Researchers as Chacón (2005), Woolfolk und Hoy (1990) and Woolfolk u. a. (1990) suggest that teacher efficacy is related to teacher classroom management. High efficacy teachers incline to less controlling and more humanistic behaviour in handling their students. High efficacious teachers apt to divide the class for small group instruction and direct teaching (Gibson & Dembo, 1984; Muijs & Reynolds, 2002) spend more time in interactive instruction (Smylie, 1988), demonstrate higher levels of planning and organisation (Allinder, 1994), and demonstrate more enthusiasm in their teaching (Allinder, 1994; Guskey, 1984) than do their low efficacious colleagues. Ashton, Webb und Doda (1983) found significant relations between teacher self-efficacy and interactions between teacher and students, and student accomplishments. High efficacy teachers focused more on high standards, instruction, student task behaviour and a supportive climate, than do low efficacy teachers. Gibson und Dembo (1984) and Dembo und Gibson

(1985) investigated the influence of teacher efficacy on academic focus and teacher feedback. Their results revealed that high efficacy teachers were more effective in leading students to correct responses by means of questioning than were low efficacy teachers. High efficacious teachers are less critical to and spent more time in working with and monitoring students who exhibited learning difficulties (Ashton & Webb, 1986; Gibson & Dembo, 1984), those teachers perceive all students as teachable (Soodak & Podell, 1993, 1996).“

Tabelle C.1.: Instrumente zitierter Autoren.

Studie	Beleg für ... laut van Dinther, Dochy, Segers und Braeken (2013)*	Instrument
Woolfolk Hoy und Davis (2006)	allg. Bedeutsamkeit	–
Ashton und Webb (1986)	Schülerleistung	RAND-Items
Muijs und Reynolds (2002)	Schülerleistung Instruktionsstrategien	Items nicht transparent
Ross (1992)	Schülerleistung	TES
Ross (1998)	Schülerleistung	–
Midgley, Feldlaufer und Eccles (1989)	Schülermotivation	5-Items, inkl. RAND-Item II
Woolfolk, Rosoff und Hoy (1990)	Schülerinteresse classroom management	TES-Short I
Chacón (2005)	classroom management	OSTES
Woolfolk und Hoy (1990)	classroom management	TES-Short I
Gibson und Dembo (1984)	Instruktionsstrategien	TES
Smylie (1988)	Instruktionsstrategien	RAND-Item II
Allinder (1994)	Unterrichtsplanung Enthusiasmus	TES
Guskey (1984)	Enthusiasmus	RSA
Ashton, Webb und Doda (1983)	...	RAND-Items
Dembo und Gibson (1985)	allg. Bedeutsamkeit	–
Soodak und Podell (1993)	Umgang mit schwierigen SuS	TES
Soodak und Podell (1996)	Umgang mit schwierigen SuS	TES

Anmerkung: Muijs und Reynolds im Original 2001, Titel jedoch 2002 erschienen.
* Es muss berücksichtigt werden, dass die Variablen, die van Dinther, Dochy, Segers und Braeken (2013) angeben, teils missverständlich sein können, sodass eine separate Prüfung dieser notwendig erscheint. Beispielsweise messen Woolfolk, Rosoff und Hoy (1990) keine Schülervariablen und „classroom management“ wird teils indirekt über Einstellungsfragebögen erfasst!

D. Itemsammlung selten verwendeter Instrumente zur Erfassung von „science teaching self-efficacy beliefs“

Nachfolgend sind einige Studien aufgeführt, bezüglich derer Instrumente mit Bezug zur Bandura'schen Theorie entwickelt und genutzt wurden. Die Instrumente wurden in der Regel für den Zweck der jeweiligen Studie entwickelt und sind selten erneut von anderen Forschenden verwendet bzw. validiert worden. Teils genügen die Items den Kriterien von Bandura (vgl. Abschnitte 3.2.1 und 3.6.2) nicht.

Tabelle D.1.: Itemsammlung selten verwendeter Instrumente zur Erfassung von „science teaching self-efficacy beliefs“.

Studie	**Itembeispiel**
Smolleck und Mongan (2011, S. 136)*	„I possess the skills necessary for guiding my students toward explanations that are consistent with experimental and observational evidence.“
Buss (2010, S. 292)	„Generally, I will teach science effectively.“
Avery und Meyer (2012, S. 407 f.)	„I feel confident to teach science in my elementary classroom.“
Walan und Chang Rundgren (2014, S. 59)	„I can undertake assessment in science.“

* Das Item scheint zunächst eher adäquat zu sein, auch wenn ein Schwierigkeitsbezug fehlt. Leider kann das Instrument nicht in seiner Gesamtheit analysiert werden, da nicht alle Items eingesehen werden können (herangezogene Literatur: Smolleck u. a., 2006; Smolleck & Yoder, 2008; Smolleck & Mongan, 2011).

E. Vergleich der „Teacher Efficacy Scale“ und des „Science Teaching Efficacy Beliefs Instrument“

Tabelle E.1.: Vergleich der Faktoren der „Teacher Efficacy Scale“ (TES) nach Gibson und Dembo (1984) und des „Science Teaching Efficacy Beliefs Instrument“ (STEBI) nach Riggs und Enochs (1990).

Items TES	**Items STEBI**
„When a student does better than usual, many times it is because I exerted a little extra effort.“ (PTE)	„When a student does better than usual in science, it is often because the teacher exerted a little extra effort.“ (STOE)
„Even a teacher with good teaching abilities may not reach many students.“ (TE).“	„Even teachers with good science teaching abilities cannot help some kids learn science.“ (STOE)
„When the grades of my students improve it is usually because I found more effective teaching approaches.“ (PTE)	„When the science grades of students improve, it is most often due to their teacher having found a more effective teaching approach.“ (STOE)
„When I really try, I can get through to most difficult students.“ (PTE)	„3. Even when I try very hard, I don’t teach science as well as I do most subjects.“ (PSTEB)

F. Physikdidaktische Themenfelder

Nachfolgend werden physikdidaktsiche Themenfelder unterschiedlicher Autoren aufgelistet, die von diesen als relevant für die Lehramtsausbildung Physik angesehen werden.

Tabelle F.1.: Für die Ausbildung von Physiklehrpersonen relevante physikdidaktische Themenfelder entsprechend der DPG-Quereinsteigerstudie (vgl. Korneck, Lamprecht, Wodzinski & Schecker, 2010) bzw. der DPG-Studie zum Lehramtsstudium (DPG, 2014).

DPG-Quereinsteigerstudie (vgl. Korneck, Lamprecht, Wodzinski & Schecker, 2010)
– Gestaltung des Physikunterrichts durch Lern- und Leistungsaufgaben
– Unterrichtsbezogenes Experimentieren
– Bildungsstandards und Kompetenzen
– Differenzierung, Motivation und Interesse
– Nature of Science
– Schülervorstellungen und Lernprozesse
– Unterrichtsmethoden
– Digitale Medien
– physikspezifische und übergreifende Unterrichtskonzeptionen
DPG-Studie zum Lehramtsstudium (vgl. DPG, 2014)
– Schülervorstellungen und Lernprozesse im Physikunterricht
– Motivation und Interesse an Physik
– Bildungsstandards und physikbezogene Kompetenzen
– Konzeptionen für den Physikunterricht
– Unterrichtsbezogenes Experimentieren
– Medien im Physikunterricht
– Physik erklären – Sprache und Repräsentationsformen im Physikunterricht (*)
– Aufgabenkultur: Lern- und Leistungsaufgaben im Physikunterricht
– Planung und Analyse von Physikunterricht
– Natur der Naturwissenschaften – Wissenschafts- und Erkenntnistheorie zur Physik
– forschungsbasierte physikdidaktische Vertiefung (*)

* Vertiefung Gymnasiallehramt

Tabelle F.2.: Physikdidaktische Handlungsfelder und zugehörige Wissensbereiche nach Gramzow u. a. (2013, S. 21-25).

Instruktionsstrategien: Wissen über/Kenntnis von ...
... Schwierigkeiten bei verschiedenen Formen der Inhaltsdarstellungen (z. B. Schulbuch)
... (Alltags)Beispielen
... Strukturierung von Inhalten
... Strukturierung von Aktivitäten (Arbeitsformen)
... Ermöglichung kumulativen und vernetzten Lernens
... Gestaltungsmöglichkeiten kognitiv anregender, motivierender, verständlicher Lernumgebungen
... Reflexion der Lernangebote
Schülervorstellungen: Wissen über/Kenntnis von ...
... themenspezifische/übergreifende Schülervorstellungen
... typische Argumentationen/Lernprozesse der Schülerinnen und Schüler (typische Rolle von Schülervorstellungen im Lernprozess)
... Prognose typischer Schülerantworten
... Diagnose von Fehlvorstellungen anhand von Schüleräußerungen
... Möglichkeiten der Berücksichtigung von Schülervorstellungen bei Planung und Durchführung von Physikunterricht
Experimente/Wissenschaftsverständnis: Wissen über/Kenntnis von ...
... Arten, Funktionen, Bedeutung von Experimenten im Erkenntnisprozess Lernender
... Vorgehen beim Experimentieren im Physikunterricht
... Mess- und Auswerteverfahren, Arbeitsweisen
... Strukturierung von Inhalten
... verschiedene experimentelle Zugänge/Anordnung von Experimenten
... Auswahl von Experimenten (kog. Potentiale, lernwirksamer Einsatz)
... Einbettung von Experimenten in Unterrichtsgang
... Schülerfehler, Lernschwierigkeiten beim Experimentieren (Prognose/Diagnose)
... Umgang mit unerwünschten Ergebnissen
... Differenzierungsmöglichkeiten anhand von Experimenten
... Rolle des Experiments für ein angemessenes Wissenschaftsverständnisses
Kontext/Interesse: Wissen über/Kenntnis von ...
... Interessensbereiche von Schülerinnen und Schülern
... interessenfördernde, fachspezifische Unterrichtsgestaltung
... Kontexte, Anwendungsbereiche
... kriteriengeleitete Wahl bestimmter Kontexte, Anwendungen, etc.
... Strukturierung eines Unterrichtsgangs anhand eines Kontextes
... Interessenunterschiede von Jungen und Mädchen
... Selbstkonzeptförderung

Fortsetzung auf der nächsten Seite...

Tabelle F.2.: *Fortsetzung*

Curriculum/Bildungsstandards/Ziele: Wissen über/Kenntnis von ...
... Auswahlkriterien für Inhalte (vor Curriculumshintergrund)
... Relevanz bestimmter Fragestellungen und Sachverhalte
... Qualitätssicherung, Kompetenz(anforderungs)diagnose
... allg. Ziele von Physikunterricht (Reflexion, Legitimation, Abgrenzung, Umsetzung)
(Digitale) Medien: Wissen über/Kenntnis von ...
... Möglichkeiten und Anforderungen bei Mediennutzung
... Einsatzszenarien für e-learning
... Einsatz digitaler Messwerterfassung, & -auswertung, Lernsoftware, Simulationen, etc.
Fachdidaktische Konzepte: Wissen über/Kenntnis von ...
... Methodenkonzeptionen (exemplarisch, genetisch, ...)
... Grundsätze, Anwendung, Reflexion von Lerntheorien (Konzeptwechsel, Elementarisieurng),
... Grundsätze, Anwendung, Reflexion didaktischer Rekonstruktion
Aufgaben: Wissen über/Kenntnis von ...
... lernpsychologische, didaktische Bedeutung von Aufgaben
... Bewertungskriterien
... Entwicklung, Einbindung, kriterienorientierte Beurteilung von Aufgaben
... Differenzierungsmöglichkeiten

G. Skalenversionen

G.1. Skalenversion 1 (V1): Pilotstudie I

Die Items wurden auch bei Meinhardt (2011) abgedruckt. Die Items der Skalen Umgang mit Aufgaben/Planung sowie Umgang mit Aufgaben/Durchführung wurden erst nach Pilotstudie I entwickelt.

G.1.1. Skala zum Experimentieren/Planung

1. Es bereitet mir keine Probleme, zu meinem Unterricht passende Experimente vorzubereiten, auch wenn die Physiksammlung schlecht ausgestattet ist.
2. Ich bin stets in der Lage, in der Unterrichtsvorbereitung lernförderliche Experimente auszuwählen, auch wenn ich das Themengebiet, das erste Mal unterrichte.
3. Ich kann jedes Experiment auch für andere Phasen des Unterrichts einplanen, obwohl ich es bisher nur zur Erarbeitung verwendet habe.
4. Auch für eine sehr heterogene Klasse kann ich Schülerexperimente so planen, dass ich sie zur Differenzierung nutzen kann.
5. Auch Zeitdruck während der Unterrichtsvorbereitung hindert mich nicht daran, ein zu meinen Unterrichtszielen passendes Experiment zu entwickeln.
6. Es gelingt mir immer, zu einem Thema Schülerexperimente leistungsdifferenziert vorzubereiten, auch wenn es in der Literatur dazu keine Vorschläge gibt.
7. Ich kann jedes Experiment als Demonstrationsexperiment planen, auch wenn ich das betreffende Experiment bisher ausschließlich als Schülerexperiment eingesetzt habe.

G.1.2. Skala zum Experimentieren/Durchführung

1. Immer wenn der Unterrichtsverlauf es sinnvoll erscheinen lässt, gelingt es mir, ein Experiment angemessen zu variieren, auch wenn ich das vorher nicht geplant habe.
2. Ich kann ein Demonstrationsexperiment für Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine sehr komplexe Versuchsanordnung handelt.
3. Auch unvorbereitet schaffe ich es stets, anhand eines Experiments über naturwissenschaftliches Arbeiten mit den Schülerinnen und Schülern zu reflektieren, auch wenn ich mich nicht explizit darauf vorbereitet habe.
4. Während Schülerexperimentierphasen kann ich die Schülerinnen und Schüler individuell zielführend unterstützen, auch wenn ich nicht alle Schwierigkeiten vorhergesehen habe.
5. Ich bin in jedem Fall in der Lage, auch längere Demonstrationsexperimente so zu gestalten, dass die Schülerinnen und Schüler bei der Sache bleiben.
6. Ich weiß, dass ich beim Experimentieren in einer Vertretungsstunde auf Unvorhergesehenes reagieren kann, auch wenn ich das Experiment längere Zeit nicht eingesetzt habe.

7. Auch unter Zeitdruck bin ich fast immer in der Lage, ein Experiment zum Laufen zu bringen, wenn es im Unterricht nicht auf Anhieb funktioniert.

G.1.3. Skala zum Elementarisieren/Planung

1. Ich bin stets in der Lage, die Kernideen eines physikalischen Inhalts zu identifizieren, auch wenn diese nicht zum Kanon des Physikunterrichts gehören.
2. Es gelingt mir immer, physikalische Inhalte sachlogisch anzuordnen, obwohl ich nicht auf ein Physikbuch zurückgreifen kann.
3. Ich schaffe es, auch komplexe Themen der modernen Physik für eine Unterrichtseinheit so zu vereinfachen, dass meine Schülerinnen und Schüler sie verstehen.
4. Auch ohne Vorlage ist es kein Problem für mich, ein Thema für den Physikunterricht zu strukturieren, das sich auf mehrere Teilgebiete der Physik (Mechanik, Thermodynamik, ...) bezieht.
5. Ich weiß, dass ich auch bei der Aufbereitung komplexer physikalischer Inhalte lernförderliche Analogien und Modelle finde.
6. Im Prinzip kann ich für jedes abstrakte Thema des Physikunterrichts Aufgaben mit Alltagsbezügen konstruieren.
7. Es gelingt mir, Physikunterricht so zu planen, dass auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen deutlich werden.

G.1.4. Skala zum Elementarisieren/Durchführung

1. Es gelingt mir, die zentralen physikalischen Inhalte einer Physikstunde herauszustellen, auch wenn ich spontan für einen Kollegen einspringen muss.
2. Auch wenn ich von meiner Planung abweichen muss, schaffe ich es in meinen Physikstunden immer, eine sinnvolle inhaltliche Reihenfolge beizubehalten.
3. Im Physikunterricht bin ich stets in der Lage, eine Erklärung weiter zu vereinfachen, die die Schülerinnen und Schüler nicht verstehen, auch wenn ich darauf nicht gefasst war.
4. Wenn sich im Unterricht die Gelegenheit bietet, kann ich spontan auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen aufzeigen.
5. Wenn Schülerinnen und Schüler im Unterricht etwas nicht verstehen, ist es für mich kein Problem, ihnen durch geeignete Modelle und Analogien zu helfen, auch wenn ich das Problem nicht vorhergesehen habe.
6. Ich kann Alltagsbezüge zu einem physikalischen Inhalt herstellen, auch wenn ich mir vor dem Unterricht keine Beispiele überlegt hatte.
7. Wenn mich Schülerinnen und Schüler in der Stunde danach fragen, kann ich auch unvorbereitet Bezüge zu vorangegangenen Unterrichtssequenzen aufzeigen.

G.1.5. Skala zum Umgang mit Schülervorstellungen/Planung

1. Ich kann die Vorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie sehr vielfältig sind.

2. Es bereitet mir keine Schwierigkeiten Physikunterricht so zu planen, dass kognitive Konflikte bei den Schülerinnen und Schülern entstehen, auch wenn ich deren Vorstellungen nicht extra erhoben habe.
3. Ich kann für den Physikunterricht relevante Schülervorstellungen in den schriftlichen Äußerungen von Schülerinnen und Schülern erkennen, auch wenn die Äußerungen nur stichwortartig sind.
4. Ich bin immer in der Lage, geeignete Methoden auszuwählen, um Schülervorstellungen zur Physik zu erfassen, obwohl es keine fertigen Testinstrumente gibt.
5. Ich kann in meiner Planung für den Physikunterricht die Vorstellungen von Schülerinnen und Schülern als Ausgangspunkt von Lernprozessen vorsehen, auch wenn die Schülervorstellungen von physikalischen Vorstellungen weit entfernt sind.
6. Ich kann Lernumgebungen kreieren, die mit Sicherheit geeignet sind, Schülervorstellungen in Frage zu stellen, auch wenn es dazu noch keine Unterrichtsvorschläge gibt.
7. Ich kann mir Methoden überlegen, die geeignet sind, physikalische Vorstellungen anzuregen, ohne Schülervorstellungen und physikalische Sichtweise einfach zu kontrastieren.

G.1.6. Skala zum Umgang mit Schülervorstellungen/Durchführung

1. Auch auf ungewöhnliche Schülervorstellungen kann ich im Unterrichtsgespräch lernförderlich reagieren.
2. Ich kann geäußerte Schülervorstellungen noch in derselben Unterrichtsstunde für Lernprozesse nutzbar machen, auch wenn ich darauf nicht vorbereitet war.
3. Im Unterricht schaffe ich es immer, kognitive Konflikte zu erzeugen, auch wenn ich die geäußerten Schülervorstellungen nicht vorhergesehen habe.
4. Ich habe keine Schwierigkeiten damit, Schülervorstellungen im Unterrichtsgespräch zu identifizieren, auch wenn sie nicht direkt geäußert werden.
5. Ich weiß, dass ich im Physikunterricht physikalisch angemessene von unangemessenen Schülervorstellungen unterscheiden kann, auch wenn ich das Themengebiet selbst schwierig finde.
6. Es gelingt mir, Schülerinnen und Schüler im Physikunterricht dazu anzuleiten, ihre eigenen Vorstellungen zu physikalischen Sachverhalten zu hinterfragen, auch wenn sie sehr überzeugt von ihnen sind.
7. Es ist kein Problem für mich, auf Schülervorstellungen im Unterricht lernförderlich zu reagieren, ohne einfach die physikalisch angemessene Vorstellung dagegen zu setzen.

G.2. Skalenversion 2 (V2): Pilotstudie II und III

G.2.1. Skala zum Experimentieren/Planung

1. Es bereitet mir keine Probleme, zu meinem Unterricht passende Experimente vorzubereiten, auch wenn die Physiksammlung schlecht ausgestattet ist.
2. Ich bin stets in der Lage, für eine Unterrichtssequenz ein motivierendes Einstiegsexperiment zu planen, auch wenn ich die Sequenz das erste Mal unterrichte.
3. Ich kann jedes Experiment auch für andere Phasen des Unterrichts einplanen, obwohl ich es bisher nur zur Erarbeitung verwendet habe.
4. Auch für eine sehr heterogene Klasse kann ich Schülerexperimente so planen, dass ich sie zur Differenzierung nutzen kann.
5. Auch Zeitdruck während der Unterrichtsvorbereitung hindert mich nicht daran, ein zu meinen Unterrichtszielen passendes Experiment zu entwickeln.
6. Es gelingt mir immer, zu einem Thema Schülerexperimente leistungsdifferenziert vorzubereiten, auch wenn es in der Literatur dazu keine Vorschläge gibt.
7. Ich kann mich für jedes Experiment begründet entscheiden, ob ich es als Schüler- oder Demonstrationsexperiment plane, auch wenn ich noch keine Erfahrung mit dem Experiment habe.

G.2.2. Skala zum Experimentieren/Durchführung

1. Immer, wenn der Unterrichtsverlauf es sinnvoll erscheinen lässt, gelingt es mir, ein Experiment angemessen zu variieren, auch wenn ich das vorher nicht geplant habe.
2. Ich kann ein Demonstrationsexperiment für Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine sehr komplexe Versuchsanordnung handelt.
3. Auch unvorbereitet schaffe ich es stets, anhand eines Experiments über naturwissenschaftliches Arbeiten mit den Schülerinnen und Schülern zu reflektieren, auch wenn ich mich nicht explizit darauf vorbereitet habe.
4. Auch auf unvorhergesehene Schwierigkeiten von Schülerinnen und Schülern beim Experimentieren kann ich so reagieren, dass sie selbständig weiterarbeiten können.
5. Ich bin in jedem Fall in der Lage, ein Demonstrationsexperiment so durchzuführen, dass die Schülerinnen und Schüler bei der Sache bleiben, auch wenn die Durchführung viel Zeit in Anspruch nimmt.
6. Ich weiß, dass ich beim Experimentieren in einer Vertretungsstunde auf Unvorhergesehenes reagieren kann, auch wenn ich das Experiment längere Zeit nicht eingesetzt habe.
7. Auch unter Zeitdruck bin ich fast immer in der Lage, ein Experiment zum Laufen zu bringen, wenn es im Unterricht nicht auf Anhieb funktioniert.

G.2.3. Skala zum Elementarisieren/Planung

1. Ich bin stets in der Lage, die Kernideen eines physikalischen Inhalts zu identifizieren, auch wenn diese nicht zum klassischen Kanon des Physikunterrichts gehören.
2. Es gelingt mir immer, Inhalte physikalisch sinnvoll anzuordnen, obwohl ich nicht auf ein Physikbuch zurückgreifen kann.

3. Ich schaffe es, auch komplexe Themen der modernen Physik für eine Unterrichtseinheit so zu vereinfachen, dass meine Schülerinnen und Schüler sie verstehen.
4. Auch ohne Vorlage ist es kein Problem für mich, ein Thema für den Physikunterricht zu strukturieren, das sich auf mehrere Teilgebiete der Physik (Mechanik, Thermodynamik, ...) bezieht.
5. Ich weiß, dass ich auch bei der Aufbereitung komplexer physikalischer Inhalte lernförderliche Analogien und Modelle finde.
6. Ich kann Alltagsphänomene zum Ausgangspunkt einer Sequenz im Physikunterricht machen, auch wenn ich das Thema das erste Mal unterrichte.
7. Es gelingt mir, Physikunterricht so zu planen, dass auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen deutlich werden.
8. Ich bin in der Lage einen physikalischen Inhalt für verschiedene didaktische Schwerpunktsetzungen aufzubereiten, auch wenn ich ihn das erste Mal unterrichte.

G.2.4. Skala zum Elementarisieren/Durchführung

1. Es gelingt mir, die zentralen physikalischen Inhalte einer Physikstunde herauszustellen, auch wenn ich spontan für einen Kollegen einspringen muss.
2. Auch wenn ich meinen Physikunterricht spontan an Interessen der Schülerinnen und Schüler anpasse, gelingt es mir immer, eine sinnvolle inhaltliche Reihenfolge beizubehalten.
3. Im Physikunterricht bin ich stets in der Lage, eine Erklärung weiter zu vereinfachen, die die Schülerinnen und Schüler nicht verstehen, auch wenn ich darauf nicht gefasst war.
4. Wenn sich im Unterricht die Gelegenheit bietet, kann ich spontan auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen aufzeigen.
5. Wenn Schülerinnen und Schüler im Unterricht etwas nicht verstehen, ist es für mich kein Problem, ihnen durch geeignete Modelle und Analogien zu helfen, auch wenn ich das Problem nicht vorhergesehen habe.
6. Ich kann Alltagsbezüge zu einem physikalischen Inhalt herstellen, auch wenn ich mir vor dem Unterricht keine Beispiele überlegt hatte.
7. Wenn mich Schülerinnen und Schüler in der Stunde danach fragen, kann ich auch unvorbereitet Bezüge zu vorangegangenen Unterrichtssequenzen aufzeigen.

G.2.5. Skala zum Umgang mit Schülervorstellungen/Planung

1. Ich kann die Vorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie sehr vielfältig sind.
2. Es bereitet mir keine Schwierigkeiten, Physikunterricht so zu planen, dass Schülerinnen und Schülern die Grenzen ihrer Alltagsvorstellungen bewusst werden, auch wenn ich ihre Vorstellungen nicht extra erhoben habe.
3. Ich kann für den Physikunterricht relevante Schülervorstellungen in den schriftlichen Äußerungen von Schülerinnen und Schülern erkennen, auch wenn die Äußerungen nur stichwortartig sind.
4. Ich bin immer in der Lage, geeignete Methoden auszuwählen, um Schülervorstellungen zur Physik zu erfassen, obwohl es keine fertigen Testinstrumente gibt.
5. Auch wenn es mir schwer fällt, Schülervorstellungen inhaltlich zu verstehen, kann ich diese zum Ausgangspunkt von Lernprozessen machen.

6. Ich kann Lernumgebungen kreieren, die mit Sicherheit geeignet sind, Schülervorstellungen in Frage zu stellen, auch wenn es dazu noch keine Unterrichtsvorschläge gibt.
7. Ich kann mir Methoden überlegen, die geeignet sind, physikalische Vorstellungen anzuregen, ohne Schülervorstellungen und physikalische Sichtweise einfach zu kontrastieren.

G.2.6. Skala zum Umgang mit Schülervorstellungen/Durchführung

1. Auch auf ungewöhnliche Schülervorstellungen kann ich im Unterrichtsgespräch lernförderlich reagieren.
2. Ich kann geäußerte Schülervorstellungen noch in derselben Unterrichtsstunde für Lernprozesse nutzbar machen, auch wenn ich darauf nicht vorbereitet war.
3. Auch auf unvorhergesehene Schülervorstellungen kann ich im Physikunterricht so reagieren, dass die Schülerinnen und Schüler diese hinterfragen.
4. Ich kann Schülervorstellungen während des Unterrichtsgesprächs rekonstruieren, auch wenn sie nur indirekt erkennbar sind.
5. Ich weiß, dass ich in jeder Unterrichtssituation physikalisch angemessene von unangemessenen Schülervorstellungen unterscheiden kann, auch wenn ich das Themengebiet selbst schwierig finde.
6. Es gelingt mir, Schülerinnen und Schüler im Physikunterricht zu motivieren, ihre eigenen Vorstellungen zu physikalischen Sachverhalten zu hinterfragen, auch wenn sie sehr überzeugt von ihnen sind.
7. Ich kann Schülervorstellungen im Physikunterricht in Frage stellen, auch ohne einfach die physikalisch angemessene Vorstellung dagegen zu setzen.

G.2.7. Skala zum Umgang mit Aufgaben/Planung

1. Ich bin in der Lage, eigene Aufgaben für ein physikalisches Thema zu entwickeln, obwohl ich das Thema noch nicht unterrichtet habe.
2. Wenn ich einen Physiktest erstelle, gelingt es mir immer, passende Aufgaben für alle Anforderungsbereiche auszuwählen, auch wenn ich unter Zeitdruck stehe.
3. Ich kann eine typische Physik-Rechenaufgabe zu einer offenen Aufgabe mit mehreren Lösungswegen umgestalten, auch wenn ich nur wenig Planungszeit habe.
4. Bei der Planung meines Physikunterrichts gelingt es mir immer zu beurteilen, ob eine Aufgabe zu meinen Zielen passt, auch wenn ich die Aufgabe noch nicht im Unterricht eingesetzt habe.
5. Ich schaffe es, eine strukturierte Lernaufgabe zu entwickeln, mit der sich Schülerinnen und Schüler ein Themengebiet selbstständig erschließen können, auch wenn ich dieses Thema bisher anders unterrichtet habe.
6. Fehler und Schwierigkeiten, die bei Schülerinnen und Schülern auftreten, wenn sie eine Aufgabe bearbeiten, kann ich mit Sicherheit vorhersehen, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.
7. Aufgaben binnendifferenziert zu gestalten ist für mich kein Problem, auch wenn ich für das betreffende physikalische Thema noch kein Beispiel kenne.
8. Auch für sehr komplexe Physikaufgaben kann ich gestufte Lernhilfen entwickeln, die den Schülerinnen und Schülern ein selbstständiges Weiterarbeiten ermöglichen.

G.2.8. Skala zum Umgang mit Aufgaben/Durchführung

1. Es gelingt mir immer, Physikaufgaben so auszuwerten, dass alle auch in dieser Phase noch dazulernen, ohne dass sich einzelne Schülerinnen und Schüler dabei langweilen.
2. Wenn ich für einen Kollegen spontan einspringen muss, kann ich den Schülerinnen und Schülern bei der Bearbeitung von Physikaufgaben helfen, auch ohne den Lösungsweg einfach vorzugeben.
3. Wenn ich im Physikunterricht feststelle, dass meine Schülerinnen und Schüler die Aufgabenstellung nicht verstehen, kann ich sie spontan angemessen variieren, auch wenn ich darauf nicht gefasst war.
4. Wenn Schülerinnen und Schüler Schwierigkeiten bei der Bearbeitung einer Physikaufgabe haben, gelingt es mir immer herauszufinden, worin genau die Schwierigkeit besteht, auch wenn ich die Aufgabe das erste Mal einsetze.
5. Wenn der Verlauf einer Stunde von meiner Planung abweicht, kann ich auch spontan angemessene Aufgabenstellungen formulieren, selbst wenn ich das physikalische Thema das erste Mal unterrichte.
6. Eine von mir vorgesehene Physikaufgabe kann ich spontan im Unterricht an verschiedene Leistungsniveaus anpassen, obwohl ich vorher nicht mit dieser Situation gerechnet habe.
7. Bei der Bearbeitung einer Physikaufgabe kann ich meine Schülerinnen und Schüler auch dann sinnvoll unterstützen, wenn sie Lösungswege einschlagen, die ich nicht vorhergesehen habe.
8. Es gelingt mir immer, den Schülerinnen und Schülern Aufgaben so vorzustellen, dass sie erkennen, weshalb eine Bearbeitung sinnvoll ist, auch wenn das physikalische Thema wenig Anwendungsbezüge aufweist.

G.3. Skalenversion 3 (V3): Hauptstudie

Die Items wurden auch bei Meinhardt u. a. (2016c) abgedruckt und entsprechen den Indikatoren vor der Revision.

G.3.1. Skala zum Experimentieren/Planung

Tabelle G.1.: Indikatorvariablen, V3 (SWE-EX-P)

Abk.	Indikatoren
exp1	Ich kann in meiner Unterrichtsplanung zu den Lernzielen passende Experimente aufbauen, auch wenn die Physiksammlung schlecht ausgestattet ist.
exp2	Ich kann bei meiner Unterrichtsplanung ein Experiment gegebenenfalls so variieren, dass ich es in einer Übungsphase einsetzen kann, auch wenn ich es bisher nur als Einstiegsexperiment genutzt habe.
exp3	Ich kann Schülerexperimente so zusammenstellen, dass die praktischen Fähigkeiten meiner Schülerinnen und Schüler auf verschiedenen Niveaus gefördert werden, auch wenn ich bei der Planung unter Zeitdruck stehe.
exp4	Ich kann zu dem Kontext einer Unterrichtsreihe ein Experiment entwickeln, auch wenn es zu diesem Kontext keine fertigen Experimentieranweisungen gibt.
exp5	Ich kann verschiedene Varianten eines Experimentes planen, mit denen sich Physikunterricht leistungsdifferenziert gestalten lässt, auch wenn es dazu keine Unterrichtsvorschläge gibt.
exp6	Ich kann ein Experiment planen, das meine Schülerinnen und Schüler begeistert, auch wenn sie sich sonst wenig für Physik interessieren.
exp7	Ich kann für ein physikalisches Experiment begründet entscheiden, ob es didaktisch sinnvoller ist, es als Demonstrations- oder Schülerexperiment einzuplanen, auch wenn ich das Experiment noch nicht eingesetzt habe.
exp8	Ich kann in meiner Unterrichtsvorbereitung ein Experiment planen, welches meine Schülerinnen und Schüler dazu anregt, physikalische Fragestellungen zu entwickeln, auch wenn ich dieses Experiment neu entwickeln muss.
exp9	Ich kann mehrere Experimente so zusammenstellen, dass bei der Auswertung unterschiedliche Möglichkeiten des Umgangs mit Messdaten deutlich werden, auch wenn ich diesbezüglich keine Unterrichtsvorschläge kenne.
exp10	Ich kann bei der Unterrichtsplanung didaktisch begründet entscheiden, ob ein Experiment mit Hilfe von schultypischen Experimentiersets oder mit Alltagsgegenständen durchgeführt werden soll, auch wenn ich die Lerngruppe noch nicht lange kenne.

G.3.2. Skala zum Experimentieren/Durchführung

Tabelle G.2.: Indikatorvariablen, V3 (SWE-EX-D)

Abk.	Indikatoren
exd1	Ich kann physikalische Experimente an interessante Impulse meiner Schülerinnen und Schülern anpassen, auch wenn ich das vorher nicht geplant hatte.
exd2	Ich kann ein Demonstrationsexperiment für meine Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine komplexe Versuchsanordnung handelt.
exd3	Ich kann beim Experimentieren spontan mit den Schülerinnen und Schülern über das Wechselspiel von Theorie und Experiment reflektieren, auch wenn ich den Anlass nicht vorgesehen hatte.
exd4	Ich kann auf unvorhergesehene Verständnisschwierigkeiten meiner Schülerinnen und Schüler beim Experimentieren so reagieren, dass sie selbstständig weiterarbeiten können, auch ohne einfach einen Lösungsweg vorzugeben.
exd5	Ich kann ein Experiment, das im Physikunterricht nicht auf Anhieb funktioniert, zum Laufen bringen, auch wenn ich unter Zeitdruck stehe.
exd6	Ich kann unerwartete Messwerte aus einem Demonstrationsexperiment spontan als Lernanlass für meine Schülerinnen und Schüler nutzen, auch ohne „unpassende“ Werte einfach zu übergehen.
exd7	Ich kann spontan ein passendes Experiment einsetzen, um auf weiterführende physikalische Fragestellungen meiner Schülerinnen und Schüler zu reagieren, auch wenn ich das nicht vorgesehen hatte.
exd8	Ich kann Experiment so inszenieren, dass meine Schülerinnen und Schüler motiviert sind, eigene physikalische Fragestellungen zu entwickeln, auch wenn es sich um eine unbeliebte Randstunde handelt.
exd9	Ich kann meine Schülerinnen und Schüler bei der Planung ihres experimentellen Vorgehens unterstützen, auch wenn sie im Physikunterricht Ihren eigenen Fragestellungen nachgehen.

G.3.3. Skala zum Elementarisieren/Planung

Tabelle G.3.: Indikatorvariablen, V3 (SWE-EL-P)

Abk.	Indikatoren
elp1	Ich kann bei der Planung meines Physikunterrichts die Kernideen eines physikalischen Themas identifizieren, auch wenn es sich nicht um ein klassisches Thema der Schulphysik handelt.
elp2	Ich kann die im Schulbuch vorgeschlagene inhaltliche Strukturierung eines physikalischen Themas für meine Lerngruppe abwandeln, auch wenn ich wenig Planungszeit habe.
elp3	Ich kann ein physikalisches Thema beim Planen einer Unterrichtseinheit so vereinfachen, dass meine Schülerinnen und Schüler es verstehen können, auch wenn es sich dabei um Inhalte der modernen Physik handelt.
elp4	Ich kann eine gut strukturierte Unterrichtsstunde planen, auch wenn sie sich auf mehrere Teilgebiete der Physik (z.B. Energie in der Mechanik, Thermodynamik, ...) bezieht.
elp5	Ich kann eine Sequenz für den Physikunterricht so planen, dass ein Alltagsphänomen den weiteren Unterrichtsverlauf strukturiert, auch wenn dadurch die gängige Fachsystematik nicht eingehalten wird.
elp6	Ich kann eine Unterrichtssequenz so aufbereiten, dass neben dem physikalischen Fachwissen ein weiterer Kompetenzbereich (Kommunikation, Erkenntnisgewinnung, Bewertung) besonders gefördert wird, auch wenn ich diese Sequenz das erste Mal unterrichte.
elp7	Ich kann bei der Planung meines Physikunterrichts die inhaltlichen Problemstellen der Schulbuchdarstellung eines physikalischen Themas erkennen, auch wenn ich nach diesem Vorschlag noch nicht unterrichtet habe.
elp8	Ich kann in meiner Unterrichtsplanung ein physikalisches Thema in sinnvolle Lernschritte zerlegen, auch wenn dieses Thema bisher noch nicht didaktisch aufbereitet wurde.
elp9	Ich kann eine Unterrichtssequenz planen, die auf ein bestimmtes Basiskonzept (System, Wechselwirkung, Energie, Materie) so fokussiert, dass dieses in späteren Lernprozessen sinnvoll erweitert werden kann, auch wenn der Rahmenlehrplan dazu keine Anregungen bietet.

G.3.4. Skala zum Elementarisieren/Durchführung

Tabelle G.4.: Indikatorvariablen, V3 (SWE-EL-D)

Abk.	Indikatoren
eld1	Ich kann eine Physikstunde so durchführen, dass die Schülerinnen und Schüler die physikalischen Kernideen erkennen, auch wenn ich von meiner ursprünglichen Stundenplanung stark abweichen muss.
eld2	Ich kann eine Physikstunde in einer sinnvollen inhaltlichen Reihenfolge durchführen, auch wenn ich mein Vorgehen spontan an Fragen der Schülerinnen und Schüler anpasse.
eld3	Ich kann im Physikunterricht eine physikalische Erklärung, die die Schülerinnen und Schüler nicht verstehen, weiter vereinfachen, ohne das die Erklärung physikalisch falsch wird.
eld4	Ich kann in einer Physikstunde mit meinen Schülerinnen und Schülern die Verbindung zu vorangegangenen Themen erarbeiten, auch wenn sich die Gelegenheit dazu unerwartet ergibt.
eld5	Ich kann einen Alltagsbezug, den Schülerinnen und Schüler während des Physikunterrichts aufwerfen, spontan für den Lernprozess nutzbar machen, ohne meine Unterrichtsziele aus den Augen zu verlieren.
eld6	Ich kann in einer Physikstunde die Basiskonzepte (System, Energie, Wechselwirkung und Materien) nutzen, mit meiner Lerngruppe Querverbindungen zwischen unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik zu erarbeiten, auch wenn sich der Anlass überraschend ergibt.

G.3.5. Skala zum Umgang mit Schülervorstellungen/Planung

Tabelle G.5.: Indikatorvariablen, V3 (SWE-SV-P)

Abk.	Indikatoren
svp1	Ich kann für meine Planung inhaltlich begründet entscheiden, ob ich Schülervorstellungen zum Ausgangspunkt einer Lernsequenz mache oder dies bewusst vermeide, auch wenn ich wenig Unterrichtserfahrung mit diesem physikalischen Inhalt habe.
svp2	Ich kann bei der Planung meines Physikunterrichts einen Unterrichtsvorschlag daraufhin beurteilen, welche Schülervorstellungen vermutlich aktiviert werden, auch wenn ich ihn noch nicht ausprobiert habe.
svp3	Ich kann eine Physikstunde planen, die Schülervorstellungen zum Ausgangspunkt von Lernprozessen macht, auch wenn ich zu dem entsprechenden Thema keinen solchen Unterrichtsvorschlag kenne.
svp4	Ich kann meinen Physikunterricht so planen, dass er schrittweise von Schülervorstellungen zu physikalischen Vorstellungen führt, ohne diese Vorstellungen einfach einander gegenüberzustellen.
svp5	Ich kann Unterrichtssituationen planen, in denen meine Schülerinnen und Schüler eine neu erworbene physikalische Vorstellung als hilfreich erleben, aber Alltagsvorstellungen nicht als sinnlos erscheinen.
svp6	Ich kann die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn die Vorstellungen einander widersprechen.
svp7	Ich kann Unterrichtssituationen planen, in denen Schülerinnen und Schüler ihre Alltagsvorstellungen hinterfragen, auch wenn sich diese im Alltag bewährt haben.
svp8	Ich kann bei meiner Unterrichtsvorbereitung Schülervorstellungen aus Aufgabenbearbeitungen erschließen, auch wenn es sich nur um Stichpunkte handelt.

G.3.6. Skala zum Umgang mit Schülervorstellungen/Durchführung

Tabelle G.6.: Indikatorvariablen, V3 (SWE-SV-D)

Abk.	Indikatoren
svd1	Ich kann Schülervorstellungen während eines Unterrichtsgesprächs zu einem physikalischen Thema rekonstruieren, auch wenn sie nur indirekt erkennbar sind.
svd2	Ich kann im Physikunterricht geäußerte Vorstellungen meiner Schülerinnen und Schüler hinsichtlich ihrer fachlichen Angemessenheit beurteilen, auch wenn ich sofort reagieren muss.
svd3	Ich kann im Physikunterricht entscheiden, ob es sinnvoll ist, den Unterrichtsverlauf an eine geäußerte Schülervorstellung anzupassen, auch wenn mich diese überrascht.
svd4	Ich kann mit ergiebigen Schülervorstellungen im Physikunterricht inhaltlich weiterarbeiten, auch wenn sie nicht in meine ursprüngliche Planung passen.
svd5	Ich kann auch spontan geeignete Beispiele heranziehen, die die Schülerinnen und Schüler dazu veranlassen, ihre Alltagsvorstellungen zu hinterfragen, auch wenn sie sehr überzeugt von ihnen sind.
svd6	Ich kann im Physikunterricht die Grenzen von geäußerten Alltagsvorstellungen mit meinen Schülerinnen und Schülern erarbeiten, ohne bloß zu sagen, wie es physikalisch „richtig“ ist.
svd7	Ich kann im Physikunterricht an Beispielen verdeutlichen, dass die physikalische Sichtweise hilfreich ist, ohne die Alltagsvorstellungen der Schülerinnen und Schüler als nutzlos darzustellen.

G.3.7. Skala zum Umgang mit Aufgaben/Planung

Tabelle G.7.: Indikatorvariablen, V3 (SWE-A-P)

Abk.	Indikatoren
ap1	Ich kann Prüfungsaufgaben entwickeln, sodass Fähigkeiten der Schülerinnen und Schüler im Kompetenzbereich Kommunikation erfasst werden, auch wenn es keine passenden Materialien gibt.
ap2	Ich kann eine Prüfungsaufgabe so verändern, dass unterschiedliche Versionen mit verschiedenen Schwierigkeitsstufen entstehen, auch wenn ich unter Zeitdruck bin.
ap3	Ich kann bei Bedarf eine offene Physikaufgabe mit mehreren Lösungswegen für meinen Physikunterricht entwickeln, auch wenn ich unter Zeitdruck bin.
ap4	Ich kann bei der Planung meines Physikunterrichts beurteilen, ob eine komplexe Aufgabe zu meinen Lernzielen passt, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.
ap5	Ich kann in meiner Unterrichtsplanung eine Aufgabe entwickeln, mit der sich Schülerinnen und Schüler einen physikalischen Inhalt selbstständig erschließen können, auch wenn das Thema für die Lernenden schwierig ist.
ap6	Ich kann bei der Unterrichtsvorbereitung Schwierigkeiten vorhersehen, die bei der Aufgabenbearbeitung auftreten könnten, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.
ap7	Ich kann bei der Planung einer längeren Übungsphase Aufgaben so zusammenstellen, dass die unterschiedlichen Fähigkeiten meiner Lerngruppe berücksichtigt werden, auch wenn die Fähigkeiten stark variieren.
ap8	Ich kann bei der Unterrichtsvorbereitung für Physikaufgaben Lernhilfen erstellen, die den Schülerinnen und Schülern ein selbstständiges Weiterarbeiten ermöglichen, auch wenn es sich um komplexe Aufgaben handelt.
ap9	Ich kann die Auswertung von Aufgaben aus einem Physiktest so planen, dass alle Schülerinnen und Schüler inhaltlich etwas dazu lernen können, ohne dass einfach der Erwartungshorizont vorgestellt wird.

G.3.8. Skala zum Umgang mit Aufgaben/Durchführung

ad1 Ich kann Physikaufgaben im Unterricht so besprechen, dass die Schülerinnen und Schüler auch in der Auswertungsphase noch etwas dazulernen, ohne sich zu langweilen.

ad2 Ich kann meine Schülerinnen und Schülern bei der Bearbeitung von Physikaufgaben unterstützen, auch ohne den Lösungsweg einfach vorzugeben.

ad3 Ich kann bei Verständnisschwierigkeiten eine Aufgabestellung im Physikunterricht spontan umformulieren, ohne die Aufgabe dabei zu vereinfachen.

ad4 Ich kann erkennen, weshalb meine Schülerinnen und Schüler Schwierigkeiten bei der Bearbeitung einer Physikaufgabe haben, auch wenn sie ihr Problem noch nicht selbst in Worte fassen können.

ad5 Ich kann mir in unvorhergesehenen Situationen spontan eine herausfordernde Aufgabe ausdenken, ohne meine Schülerinnen und Schüler zu überfordern.

ad6 Ich kann meine Schülerinnen und Schüler bei der Bearbeitung einer Physikaufgabe auch dann sinnvoll unterstützen, wenn sie Lösungswege einschlagen, die ich nicht vorhergesehen habe.

ad7 Ich kann meinen Schülerinnen und Schülern einsichtig machen, weshalb die Bearbeitung einer Aufgabe für ihren eigenen Lernprozess sinnvoll ist, auch wenn ein Anwendungsbezug nicht direkt ersichtlich ist.

ad8 Ich kann bei der Auswertung von Physikaufgaben Fehler einzelner Schülerinnen und Schüler für den Lernprozess der gesamten Lerngruppe nutzen, ohne jemanden bloß zu stellen.

ad9 Ich kann in meinem Physikunterricht die Auswertung einer Aufgabe strukturiert gestalten, auch wenn die Lösungswege der Schülerinnen und Schüler vielfältiger sind, als ich es erwartet hatte.

G.4. Skalenversion 4 (V4): Zur Verwendung empfohlen

Die Items wurden auch bei Meinhardt u. a. (2016c) abgedruckt und entsprechen den Indikatoren nach der Revision.

G.4.1. Skala zum Experimentieren/Planung

Tabelle G.8.: Indikatorvariablen, V4 (SWE-EX-P)

Abk.	Indikatoren
exp1	Ich kann in meiner Unterrichtsplanung zu den Lernzielen passende Experimente aufbauen, auch wenn die Physiksammlung schlecht ausgestattet ist.
exp2	Ich kann bei meiner Unterrichtsplanung ein Experiment gegebenenfalls so variieren, dass ich es in einer Übungsphase einsetzen kann, auch wenn ich es bisher nur als Einstiegsexperiment genutzt habe.
exp3	Ich kann Schülerexperimente so zusammenstellen, dass die Fähigkeiten meiner Schülerinnen und Schüler auf verschiedenen Niveaus gefördert werden, auch wenn ich bei der Planung unter Zeitdruck stehe.
exp6	Ich kann ein Experiment planen, das meine Schülerinnen und Schüler begeistert, auch wenn sie sich sonst wenig für Physik interessieren.
exp7	Ich kann für ein physikalisches Experiment begründet entscheiden, ob es didaktisch sinnvoller ist, es als Demonstrations- oder Schülerexperiment einzuplanen, auch wenn ich das Experiment noch nicht eingesetzt habe.
exp8	Ich kann in meiner Unterrichtsvorbereitung ein Experiment planen, welches meine Schülerinnen und Schüler dazu anregt, physikalische Fragestellungen zu entwickeln, auch wenn ich dieses Experiment neu entwickeln muss.
exp9	Ich kann mehrere Experimente so zusammenstellen, dass bei der Auswertung unterschiedliche Möglichkeiten des Umgangs mit Messdaten deutlich werden, auch wenn ich diesbezüglich keine Unterrichtsvorschläge kenne.
exp10	Ich kann bei der Unterrichtsplanung didaktisch begründet entscheiden, ob ein Experiment mit Hilfe von schultypischen Experimentiersets oder mit Alltagsgegenständen durchgeführt werden soll, auch wenn ich die Lerngruppe noch nicht lange kenne.

G.4.2. Skala zum Experimentieren/Durchführung

Tabelle G.9.: Indikatorvariablen, V4 (SWE-EX-D)

Abk.	Indikatoren
exd1	Ich kann physikalische Experimente an interessante Impulse meiner Schülerinnen und Schüler anpassen, auch wenn ich das vorher nicht geplant hatte.
exd2	Ich kann ein Demonstrationsexperiment für meine Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine komplexe Versuchsanordnung handelt.
exd3	Ich kann beim Experimentieren spontan mit den Schülerinnen und Schülern über das Wechselspiel von Theorie und Experiment reflektieren, auch wenn ich den Anlass nicht vorhergesehen hatte.
exd4	Ich kann auf unvorhergesehene Verständnisschwierigkeiten meiner Schülerinnen und Schüler beim Experimentieren so reagieren, dass sie selbstständig weiterarbeiten können, auch ohne einfach einen Lösungsweg vorzugeben.
exd5	Ich kann ein Experiment, das im Physikunterricht nicht auf Anhieb funktioniert, zum Laufen bringen, auch wenn ich unter Zeitdruck stehe.
exd6	Ich kann unerwartete Messwerte aus einem Demonstrationsexperiment spontan als Lernanlass für meine Schülerinnen und Schüler nutzen, auch ohne „unpassende“ Werte einfach zu übergehen.
exd8	Ich kann ein Experiment so inszenieren, dass meine Schülerinnen und Schüler motiviert sind mitzuarbeiten, auch wenn es sich um eine unbeliebte Randstunde handelt.
exd9	Ich kann meine Schülerinnen und Schüler bei der Planung ihres experimentellen Vorgehens unterstützen, auch wenn sie im Physikunterricht ihren eigenen Fragestellungen nachgehen.

G.4.3. Skala zum Elementarisieren/Planung

Tabelle G.10.: Indikatorvariablen, V4 (SWE-EL-P)

Abk.	Indikatoren
elp2	Ich kann die im Schulbuch vorgeschlagene inhaltliche Strukturierung eines physikalischen Themas für meine Lerngruppe abwandeln, auch wenn ich wenig Planungszeit habe.
elp3	Ich kann ein physikalisches Thema beim Planen einer Unterrichtseinheit so vereinfachen, dass meine Schülerinnen und Schüler es verstehen können, auch wenn es sich dabei um Inhalte der modernen Physik handelt.
elp4	Ich kann eine gut strukturierte Unterrichtsstunde planen, auch wenn sie sich auf mehrere Teilgebiete der Physik (z.B. Energie in der Mechanik, Thermodynamik, ...) bezieht.
elp5	Ich kann eine Sequenz für den Physikunterricht so planen, dass ein Alltagsphänomen den weiteren Unterrichtsverlauf strukturiert, auch wenn dadurch die gängige Fachsystematik nicht eingehalten wird.
elp6	Ich kann eine Unterrichtssequenz so aufbereiten, dass neben dem physikalischen Fachwissen ein weiterer Kompetenzbereich (Kommunikation, Erkenntnisgewinnung, Bewertung) besonders gefördert wird, auch wenn ich diese Sequenz das erste Mal plane.
elp7	Ich kann bei der Planung meines Physikunterrichts die inhaltlichen Problemstellen der Schulbuchdarstellung eines physikalischen Themas erkennen, auch wenn ich dieses das erste Mal für den Unterricht vorbereite.
elp8	Ich kann in meiner Unterrichtsplanung ein physikalisches Thema in sinnvolle Lernschritte zerlegen, auch wenn dieses Thema bisher noch nicht didaktisch aufbereitet wurde.
elp9	Ich kann eine Unterrichtssequenz planen, die auf ein bestimmtes Basiskonzept (System, Wechselwirkung, Energie, Materie) so fokussiert, dass dieses in späteren Lernprozessen sinnvoll erweitert werden kann, auch wenn der Rahmenlehrplan dazu keine Anregungen bietet.

G.4.4. Skala zum Elementarisieren/Durchführung

Tabelle G.11.: Indikatorvariablen, V4 (SWE-EL-D)

Abk.	Indikatoren
eld1	Ich kann eine Physikstunde so durchführen, dass die Schülerinnen und Schüler die physikalischen Kernideen erkennen, auch wenn ich von meiner ursprünglichen Stundenplanung stark abweichen muss.
eld2	Ich kann eine Physikstunde in einer sinnvollen inhaltlichen Reihenfolge durchführen, auch wenn ich mein Vorgehen spontan an Fragen der Schülerinnen und Schüler anpasse.
eld3	Ich kann im Physikunterricht eine physikalische Erklärung, die die Schülerinnen und Schüler nicht verstehen, weiter vereinfachen, ohne dass die Erklärung physikalisch falsch wird.
eld4	Ich kann in einer Physikstunde mit meinen Schülerinnen und Schülern die Verbindung zu vorangegangenen Themen erarbeiten, auch wenn sich die Gelegenheit dazu unerwartet ergibt.
eld5	Ich kann einen Alltagsbezug, den Schülerinnen und Schüler während des Physikunterrichts aufwerfen, spontan für den Lernprozess nutzbar machen, ohne meine Unterrichtsziele aus den Augen zu verlieren.
eld6	Ich kann in einer Physikstunde mit meiner Lerngruppe Querverbindungen zwischen unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik erarbeiten, auch wenn sich der Anlass überraschend ergibt.

G.4.5. Skala zum Umgang mit Schülervorstellungen/Planung

Tabelle G.12.: Indikatorvariablen, V4 (SWE-SV-P)

Abk.	Indikatoren
svp1	Ich kann für meine Planung inhaltlich begründet entscheiden, ob ich Schülervorstellungen zum Ausgangspunkt einer Lernsequenz mache oder dies bewusst vermeide, auch wenn ich wenig Unterrichtserfahrung mit diesem physikalischen Inhalt habe.
svp2	Ich kann bei der Planung meines Physikunterrichts einen Unterrichtsvorschlag daraufhin beurteilen, welche Schülervorstellungen vermutlich aktiviert werden, auch wenn ich ihn noch nicht ausprobiert habe.
svp4	Ich kann meinen Physikunterricht so planen, dass er schrittweise von Schülervorstellungen zu physikalischen Vorstellungen führt, ohne diese Vorstellungen einfach einander gegenüberzustellen.
svp5	Ich kann Unterrichtssituationen planen, in denen meine Schülerinnen und Schüler eine neu erworbene physikalische Vorstellung als hilfreich erleben, aber Alltagsvorstellungen nicht als sinnlos erscheinen.
svp6	Ich kann eine Physikstunde planen, die die Alltagsvorstellungen meiner Schülerinnen und Schüler zum Ausgangspunkt von Lernprozessen macht, auch wenn die Vorstellungen einander widersprechen.
svp7	Ich kann Unterrichtssituationen planen, in denen Schülerinnen und Schüler ihre Alltagsvorstellungen hinterfragen, auch wenn sich diese im Alltag bewährt haben.
svp8	Ich kann bei meiner Unterrichtsvorbereitung Schülervorstellungen aus Aufgabenbearbeitungen erschließen, auch wenn es sich nur um Stichpunkte handelt.

G.4.6. Skala zum Umgang mit Schülervorstellungen/Durchführung

Tabelle G.13.: Indikatorvariablen, V4 (SWE-SV-D)

Abk.	Indikatoren
svd1	Ich kann Schülervorstellungen während eines Unterrichtsgesprächs zu einem physikalischen Thema rekonstruieren, auch wenn sie nur indirekt erkennbar sind.
svd2	Ich kann im Physikunterricht geäußerte Vorstellungen meiner Schülerinnen und Schüler hinsichtlich ihrer fachlichen Angemessenheit beurteilen, auch wenn ich sofort reagieren muss.
svd3	Ich kann im Physikunterricht entscheiden, ob es sinnvoll ist, den Unterrichtsverlauf an eine geäußerte Schülervorstellung anzupassen, auch wenn mich diese überrascht.
svd4	Ich kann mit ergiebigen Schülervorstellungen im Physikunterricht inhaltlich weiterarbeiten, auch wenn sie nicht in meine ursprüngliche Planung passen.
svd5	Ich kann auch spontan geeignete Beispiele heranziehen, die die Schülerinnen und Schüler dazu veranlassen, ihre Alltagsvorstellungen zu hinterfragen, auch wenn sie sehr überzeugt von ihnen sind.
svd6	Ich kann im Physikunterricht die Grenzen von geäußerten Alltagsvorstellungen mit meinen Schülerinnen und Schülern erarbeiten, ohne bloß zu sagen, wie es physikalisch „richtig“ ist.

G.4.7. Skala zum Umgang mit Aufgaben/Planung

Tabelle G.14.: Indikatorvariablen, V4 (SWE-A-P)

Abk.	Indikatoren
ap1	Ich kann Prüfungsaufgaben entwickeln, sodass Fähigkeiten der Schülerinnen und Schüler im Kompetenzbereich Kommunikation erfasst werden, auch wenn es keine passenden Materialien gibt.
ap3	Ich kann bei Bedarf eine offene Physikaufgabe mit mehreren Lösungswegen für meinen Physikunterricht entwickeln, auch wenn ich unter Zeitdruck bin.
ap4	Ich kann bei der Planung meines Physikunterrichts beurteilen, ob eine komplexe Aufgabe zu meinen Lernzielen passt, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.
ap5	Ich kann in meiner Unterrichtsplanung eine Aufgabe entwickeln, mit der sich Schülerinnen und Schüler einen physikalischen Inhalt selbstständig erschließen können, auch wenn das Thema für die Lernenden schwierig ist.
ap6	Ich kann bei der Unterrichtsvorbereitung Schwierigkeiten vorhersehen, die bei der Aufgabenbearbeitung auftreten könnten, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.
ap7	Ich kann bei der Planung einer längeren Übungsphase Aufgaben so zusammenstellen, dass die unterschiedlichen Fähigkeiten meiner Lerngruppe berücksichtigt werden, auch wenn die Fähigkeiten stark variieren.
ap8	Ich kann bei der Unterrichtsvorbereitung für Physikaufgaben Lernhilfen erstellen, die den Schülerinnen und Schülern ein selbstständiges Weiterarbeiten ermöglichen, auch wenn es sich um komplexe Aufgaben handelt.
ap9	Ich kann die Auswertung von Aufgaben aus einem Physiktest so planen, dass alle Schülerinnen und Schüler inhaltlich etwas dazulernen können, ohne dass einfach der Erwartungshorizont vorgestellt wird.

G.4.8. Skala zum Umgang mit Aufgaben/Durchführung

Tabelle G.15.: Indikatorvariablen, V4 (SWE-A-D)

Abk.	Indikatoren
ad1	Ich kann Physikaufgaben im Unterricht so besprechen, dass die Schülerinnen und Schüler auch in der Auswertungsphase noch etwas dazulernen, ohne sich zu langweilen.
ad2	Ich kann meine Schülerinnen und Schüler bei der Bearbeitung von Physikaufgaben unterstützen, auch ohne den Lösungsweg einfach vorzugeben.
ad3	Ich kann bei Verständnisschwierigkeiten eine Aufgabenstellung im Physikunterricht spontan umformulieren, ohne die Aufgabe dabei zu vereinfachen.
ad4	Ich kann erkennen, weshalb meine Schülerinnen und Schüler Schwierigkeiten bei der Bearbeitung einer Physikaufgabe haben, auch wenn sie ihr Problem noch nicht selbst in Worte fassen können.
ad5	Ich kann mir in unvorhergesehenen Situationen spontan eine herausfordernde Aufgabe ausdenken, ohne meine Schülerinnen und Schüler zu überfordern.
ad6	Ich kann meine Schülerinnen und Schüler bei der Bearbeitung einer Physikaufgabe auch dann sinnvoll unterstützen, wenn sie Lösungswege einschlagen, die ich nicht vorhergesehen habe.
ad7	Ich kann meinen Schülerinnen und Schülern einsichtig machen, weshalb die Bearbeitung einer Aufgabe für ihren eigenen Lernprozess sinnvoll ist, auch wenn ein Anwendungsbezug nicht direkt ersichtlich ist.
ad9	Ich kann in meinem Physikunterricht die Auswertung einer Aufgabe strukturiert gestalten, auch wenn die Lösungswege der Schülerinnen und Schüler vielfältiger sind, als ich es erwartet hatte.

H. Zum Interview

H.1. Skalen – Interviewversion

	stimmt nicht	stimmt kaum	stimmt eher	stimmt genau
1. Ich bin stets in der Lage, die Kernideen eines physikalischen Inhalts zu identifizieren, auch wenn diese nicht zum klassischen Kanon des Physikunterrichts gehören.				
2. Es gelingt mir immer, Inhalte physikalisch sinnvoll anzuordnen, obwohl ich nicht auf ein Physikbuch zurückgreifen kann.				
3. Ich schaffe es, auch komplexe Themen der modernen Physik für eine Unterrichtseinheit so zu vereinfachen, dass meine Schülerinnen und Schüler sie verstehen.				
4. Auch ohne Vorlage ist es kein Problem für mich, ein Thema für den Physikunterricht zu strukturieren, das sich auf mehrere Teilgebiete der Physik (Mechanik, Thermodynamik, …) bezieht.				
5. Ich weiß, dass ich auch bei der Aufbereitung komplexer physikalischer Inhalte lernförderliche Analogien und Modelle finde.				
6. Ich kann Alltagsphänomene zum Ausgangspunkt einer Sequenz im Physikunterricht machen, auch wenn ich das Thema das erste Mal unterrichte.				
7. Es gelingt mir, Physikunterricht so zu planen, dass auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen deutlich werden.				
8. Ich bin in der Lage einen physikalischen Inhalt für verschiedene didaktische Schwerpunktsetzungen aufzubereiten, auch wenn ich ihn das erste Mal unterrichte.				

H.2. Interviewleitfaden

H.2.1. Phase I – Anmerkungen zur Einführung

- Vielen Dank, dass Sie sich bereit erklärt haben, am Interview teilzunehmen und mich damit in meinem Promotionsprojekt unterstützen.
- Alle Informationen, die ich heute von Ihnen bekomme, dienen ausschließlich Forschungszwecken und werde nur ich zu Gehör bekommen bzw. anonymisiert zu Forschungszwecken verwendet.
- Ich werde Ihnen gleich eine Reihe von Fragebogen-Items vorlegen. Ich möchte gern etwas darüber erfahren, was Sie unter den Items verstehen und ob sie für Sie überhaupt verständlich sind.
- Deshalb möchte ich Sie bitten, zunächst jedes Item laut vorzulesen und anschließend laut zu sagen, wie Sie das Item zusammenfassen würden bzw. was Sie unter dem Item verstehen. Wenn Sie über schwierige Begriffe stolpern, dann teilen Sie mir diese bitte mit. Wir wollen die Items nämlich so einfach wie möglich gestalten.
- Das laute Sprechen/Denken ist vielleicht zunächst ungewohnt. Kein Problem – Ich werde Sie einfach immer wieder auffordern, Ihre Gedanken zu äußern und Sie so unterstützen.
- Außerdem würde mich interessieren, wie Sie sich selbst in Bezug auf das Item einschätzen würden und wie Sie dies begründen würden. Also, warum schätzen Sie sich so ein?
- Am Ende werde ich Sie jeweils fragen, wie Sie sich bez. aller Items einschätzen würden und welche Items sie als unauthentisch bzw. sehr scher/leicht wahrgenommen haben.
- Diesen Vierschritt *lautes Vorlesen – Worum gehts da? – Selbsteinschätzung – Begründung* lege ich als Gedächtnisstütze neben die Items.
- Und ich bin ja auch noch da, möchte aber versuchen, Sie so wenig wie möglich zu stören oder zu unterbrechen.

H.2.2. Phase II – Mögliche Impulse bez. des Vierschritts

Lautes Vorlesen:

- Lesen Sie Schritt für Schritt jedes Item laut vor!
- Lesen Sie bitte zuerst das Item laut vor.

Lautes Zusammenfassen:

- Formulieren Sie bitte für jedes Item den Inhalt, so wie Sie ihn verstanden haben.
- Worum geht es für Sie in dem Item?
- Welches sind für Sie die 1-2 zentralen Wörter in dem Item?
- Können Sie die für Sie wichtigen Kernbestandteile des Items nennen?
- Wie würden Sie dieses Item stichpunktartig zusammenfassen?

Selbsteinschätzung:

- Wie würden Sie sich selbst einschätzen?
- Wo würden Sie im Fragebogen Ihr Kreuz machen?

Begründung:

- Weshalb würden Sie sich so einschätzen?
- Woran machen Sie Ihre Einschätzung fest?
- Warum würden Sie sich eher schlecht/gut einschätzen?
- Denken Sie an eine bestimmte Erfahrung oder an eine bestimmte Situation?

H.2.3. Phase III – Anmerkungen zur Gesamtskala

Einschätzung Gesamtskala

Mit den Items wollen wir messen, was Sie sich selbst in einem bestimmten Bereich zutrauen, in diesem Falle:

- Inwiefern würden Sie sich zutrauen, Schülervorstellungen beim Planen/Durchführen von Physikunterricht zu berücksichtigen?
- Inwiefern würden Sie sich zutrauen, Experimente für den Physikunterricht zu planen bzw. im Physikunterricht durchzuführen.
- Inwiefern würden Sie sich zutrauen, physikalische Inhalte für den Physikunterricht aufzubereiten bzw. mit physikalischen Inhalten im Physikunterricht angemessen umzugehen.
- Inwiefern würden Sie sich zutrauen, Aufgaben für den Physikunterricht zu planen bzw. im Physikunterricht sinnvoll einzusetzen.

Authentizität

- Inwiefern würden Sie sagen, passen die Items zu dem genannten Themenfeld?
- Inwiefern würden Sie die beschriebenen Items als authentisch oder für den Schulalltag relevant einschätzen?
- Welche Items halten Sie für unrealistisch?
- Handelt es sich bei den in den Items beschriebenen Handlungen um für die Schule relevante Handlungen? Welche Handlungen halten sie für unnötig/unrealistisch?

Schwierigkeit

- Welche der Items haben Sie als besonders leicht/schwer wahrgenommen?
- Wenn Sie die Items in eine Reihenfolge bringen müssen und mit dem leichtesten/schwersten anfangen sollten, welche Reihenfolge würden Sie wählen?

Vollständigkeit

- Falls Ihnen noch zusätzliche sehr wichtige Handlungen ein, die in dem beschrieben Themenfeld unbedingt abgefragt werden müssten?
- Welche zusätzlichen Handlungen halten Sie für so wichtig, dass sie in dem Fragebogen berücksichtigt werden sollten?

Antwortformat

- Inwiefern waren die vorgegebenen Antwortkategorien für Sie hilfreich, ihre Selbsteinschätzung vorzunehmen?
- Wären mehr Kategorien hilfreich gewesen?

H.3. Beispiel Interviewnotizen

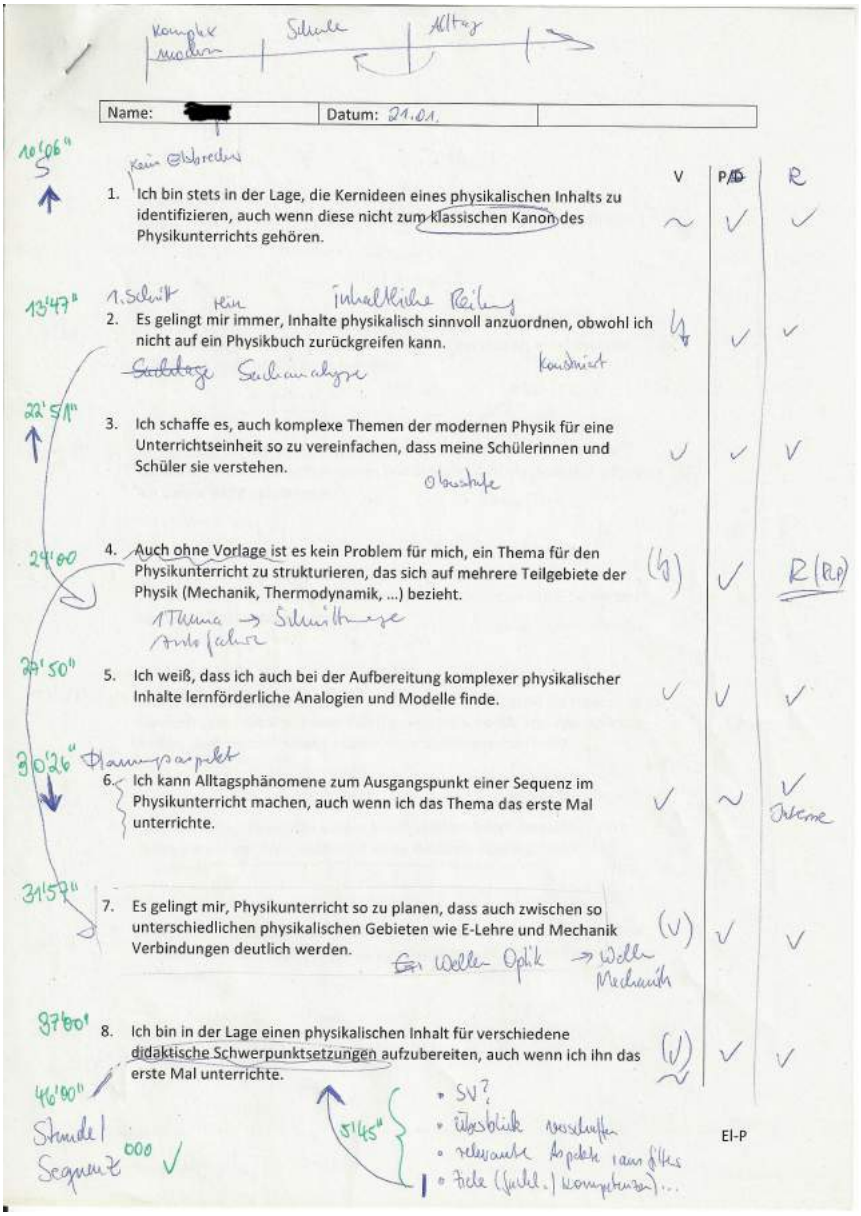


Abbildung H.1.: Beispiel für angefertigte Interviewnotizen sowie später hinzugefügte Zeitangaben des Audiomitschnitts.

H.4. Beispielübersicht Itemauswertung (Interviews)

Nachfolgende Tabelle gibt exemplarisch einen Einblick in die Verfahrensweise der kriteriengeleiteten Auswertung der Interviews. Beim erneuten Hören des Interviews (Audiomitschnitt) wurden direkt Urteile hinsichtlich der in Abschnitt 10.1.5 dargestellten Kriterien gebildet und diese ggf. durch Kommentare oder Transkripte belegt. Beachtet werden sollte, dass die Tabelle hauptsächlich Aspektes des Kriteriums „Verständnis“ enthält sowie Hinweise zur Authentizität/Relevanz bzw. zur Schwierigkeit des betreffenden Items. Auf Aspekte des Kriteriums „kognitive Aktivierung“ konnte in der Regel geschlussfolgert werden. Indem beispielsweise eine Selbsteinschätzung vorgenommen wurde, konnte die Selbstreferentialität des Items belegt werden.

Die Urteile, Kommentare und Transkripte in der Tabelle beziehen sich dabei auf das folgende Item:

„Ich kann die Vorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie sehr vielfältig sind.“ (Skala: SWE-SV-P, Version 2, Item 1, vgl. Anhang G.2.5)

Tabelle H.1.: Ausschnitt aus der Interview-Auswertungsübersicht zum Item 1 der Skala SWE-SV-P, Version 2.

IP[1]	V[2]	Barriere relevant?		Assoziation Dimension		R/A[4]	K-IP[5]	K-I[6]
		E[3]	Beleg	E[3]	Beleg			
1	teils	nein	–	e D	„Themengebiete abdecken, die SuS behandelt haben wollen“	ja	„was ist mit Vorstellungen gemeint? Schülervorstellungen vs. sollen SuS Unterricht mit gestalten?“	Vorstellung als Interesse, Themenvielfalt
2	ja	e nicht	eher im Sinne: viele SV im Voraus bedenken! Ich kann vielfältige Vorstellungen in meine Unterrichtsplanung...	P	–	ja	„sind damit auch „richtige“ Vorstellungen gemeint?“	Vorstellung als SV identifiziert, aber unklar ob neben Fehlvorstellungen auch „richtige“ Vorstellungen gemeint sind
3	teils	nein	„weil es mir Spaß macht, mit SuS zusammenzuarbeiten ...“	e D	„weil es mir Spaß macht, mit SuS zusammenzuarbeiten ...“	ja	–	Vorstellungen als Interessen der SuS; Vorstellungen als einbeziehen von Interessen
4	–	e nicht	Bezug SPÜ Stunde und Bezug auf Reaktion auf eine SuS-Äußerung	D	„äuf SV eingehen, die SuS mit in den Unterricht tragen", Bezug zu SPÜ Stunde: ës ist schwer im U zu reagieren, wenn SuS etwas einwerfen’	ja	–	–

[1] Interviewpartner [2] Einschätzung Itemverständnis [3] Einschätzung (e: eher, P: Planung, D: Durchführung) [4] Relevanz/Authentizität
[5] Kommentare Interviewpartner [6] Kommentare Interviewerin

Fortsetzung auf der nächsten Seite...

Tabelle H.1.: *Fortsetzung*

IP[1]	V[2]	Barriere relevant?		Assoziation Dimension		R/A[4]	K-IP[5]	K-I[6]
		E[3]	Beleg	E[3]	Beleg			
5	teils	ja	explizit in Einschätzung	P	–	ja	„was ist mit Vorstellungen gemeint? Schülervorstellungen vs. sollen SuS Unterricht mit gestalten?	Vorstellung als Interesse, Themenvielfalt
6	ja	e nein	Bezug auf eigene Erfahrungen, SV überhaupt kennenzulernen	e P	–	ja	–	Vorstellungen der SuS als das „was SuS von der Physik verstehen“
7	ja	e nein	Bezug: „mir fiel es schwer die SV zu erfassen“; „Was ich dachte, was leicht ist, fiel ihnen besonders schwer“	P/D	–	ja	–	Schwerpunkt auf dem Erfassen von SV, um sie dann in die Planung integrieren zu können?
8	ja	e nein	–	P	–	ja	„sehr umfangreich“	einbeziehen, insofern als dass Gelegenheiten im Unterricht einräumen, sodass SuS ihre Vorstellungen bewusst werden können; Unterschiede zu physikalischer Theorie deutlich machen

[1] Interviewpartner [2] Einschätzung Itemverständnis [3] Einschätzung (e: eher, P: Planung, D: Durchführung) [4] Relevanz/Authentizität
[5] Kommentare Interviewpartner [6] Kommentare Interviewerin

I. Zur Expertenbefragung

I.1. Expertenfragebogen (Physikdidaktik)

Beispiel für einen Expertenfragebogen (Fokus auf Handlungsfeldern SWE-EL und SWE-A):

Expertenbefragung im Rahmen einer Validierungsstudie bzgl. des Konstrukts der Selbstwirksamkeitserwartung (SWE) in physikdidaktischen Handlungsfeldern – EIA

Was erwartet Sie auf den folgenden Seiten?

Bisher haben wir für verschiedene physikdidaktische Handlungsfelder SWE-Skalen in den Dimensionen Planung und Durchführung entwickelt. Wir möchten Ihnen zunächst kurz darstellen, an welchen Regeln wir uns bzgl. der Itemformulierung orientiert haben und diese an einem Beispiel verdeutlichen.

Anschließend möchten wir Sie bitten, die entwickelten Items/Skalen für zwei Handlungsfelder nach bestimmten Kriterien, die wir Ihnen vorstellen werden, zu prüfen und ihre Qualität einzuschätzen. Die Bearbeitung dauert **ca. 40 Minuten**.

Über Kommentare und Anmerkungen jeglicher Art freuen wir uns sehr! Falls Ihnen der Platz auf den vorgesehenen Seiten dazu nicht ausreicht, nutzen Sie für Ihre Ergänzungen einfach die freie letzte Seite.

Grundlagen zur Itemformulierung

Wir beziehen uns auf Schmitz & Schwarzer (2000, S. 13), die Selbstwirksamkeitserwartungen wie folgt definieren:

> „Unter Selbstwirksamkeitserwartungen versteht man die subjektive Gewi[ss]heit, eine neue oder schwierige Aufgabe auch dann erfolgreich bearbeiten zu können, wenn sich Widerstände in den Weg stellen."

Für die Itemformulierung ergeben sich daraus ebenfalls nach Schmitz & Schwarzer (2000, S. 14) die folgenden drei Regeln:

1. Formulierung in der ersten Person singular: *Ich…*
2. Bezug zu eigenen Fähigkeiten durch Verwendung von Verbphrasen wie *Ich kann …* oder *Ich bin in der Lage …*,
3. Verdeutlichung von Schwierigkeiten (Barrieren), bei denen es sich nicht um Personeneigenschaften handelt (Möglichkeit der Intersubjektivierbarkeit), beispielsweise: Zeitdruck, fehlende Vorbereitung, fehlende Materialien, komplizierte Fachinhalte, unvorhergesehene Situationen. **Nicht** passend wäre: *… auch wenn ich das Thema nicht mag.*

Beispielitem

Regeln 1 & 2 **Regel 3**

Ich kann Alltagsbezüge zu einem physikalischen Inhalt herstellen, **auch wenn** ich mir vor dem Unterricht keine Beispiele überlegt hatte.

Bitte beurteilen und kommentieren Sie im Folgenden die Items der zwei Skalen zum Handlungsfeld **schülerorientierte Elementarisierung von physikalischen Inhalten** nach den folgenden Kriterien. Bitte geben Sie ggf. Verbesserungsvorschläge an.

a. Passung des Items zum **Handlungsfeld** (HF) → Handelt es sich um eine wesentliche Komponente des Handlungsfeldes? Ja: +, Nein: –
b. Passung der **Handlungsbarriere** (HB) → Ist die Barriere authentisch und ausreichend intersubjektivierbar? Ja: +, Nein: –
c. **Itemschwierigkeit** (IS) → Bitte ordnen Sie die Items danach, wie schwierig es ist, ihnen zuzustimmen! 1: am leichtesten, 7 bzw. 8: am schwersten

	Items: schülerorientierte Elementarisierung von physikalischen Inhalten/Planung	**HF (+/–)**	**HB (+/–)**	**IS (1-8)**	**Kommentar, Verbesserung**
P1	Ich bin stets in der Lage, die Kernideen eines physikalischen Inhalts zu identifizieren, auch wenn diese nicht zum klassischen Kanon des Physikunterrichts gehören.				
P2	Es gelingt mir immer, Inhalte physikalisch sinnvoll anzuordnen, obwohl ich nicht auf ein Physikbuch zurückgreifen kann.				
P3	Ich schaffe es, auch komplexe Themen der modernen Physik für eine Unterrichtseinheit so zu vereinfachen, dass meine Schülerinnen und Schüler sie verstehen.				
P4	Auch ohne Vorlage ist es kein Problem für mich, ein Thema für den Physikunterricht zu strukturieren, das sich auf mehrere Teilgebiete der Physik (Mechanik, Thermodynamik, …) bezieht.				
P5	Ich weiß, dass ich auch bei der Aufbereitung komplexer physikalischer Inhalte lernförderliche Analogien und Modelle finde.				
P6	Ich kann Alltagsphänomene zum Ausgangspunkt einer Sequenz im Physikunterricht machen, auch wenn ich das Thema das erste Mal unterrichte.				
P7	Es gelingt mir, Physikunterricht so zu planen, dass auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen deutlich werden.				
P8	Ich bin in der Lage einen physikalischen Inhalt für verschiedene didaktische Schwerpunktsetzungen aufzubereiten, auch wenn ich ihn das erste Mal unterrichte.				

	Items: schülerorientierte Elementarisierung von physikalischen Inhalten /Durchführung	**HF (+/–)**	**HB (+/–)**	**IS (1-7)**	**Kommentar, Verbesserung**
D1	Es gelingt mir, die zentralen physikalischen Inhalte einer Physikstunde herauszustellen, auch wenn ich spontan für einen Kollegen einspringen muss.				
D2	Auch wenn ich meinen Physikunterricht spontan an Interessen der Schülerinnen und Schüler anpasse, gelingt es mir immer, eine sinnvolle inhaltliche Reihenfolge beizubehalten.				
D3	Im Physikunterricht bin ich stets in der Lage, eine Erklärung weiter zu vereinfachen, die die Schülerinnen und Schüler nicht verstehen, auch wenn ich darauf nicht gefasst war.				
D4	Wenn sich im Unterricht die Gelegenheit bietet, kann ich spontan auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen aufzeigen.				
D5	Wenn Schülerinnen und Schüler im Unterricht etwas nicht verstehen, ist es für mich kein Problem, ihnen durch geeignete Modelle und Analogien zu helfen, auch wenn ich das Problem nicht vorhergesehen habe.				
D6	Ich kann Alltagsbezüge zu einem physikalischen Inhalt herstellen, auch wenn ich mir vor dem Unterricht keine Beispiele überlegt hatte.				
D7	Wenn mich Schülerinnen und Schüler in der Stunde danach fragen, kann ich auch unvorbereitet Bezüge zu vorangegangenen Unterrichtssequenzen aufzeigen.				

Bitte beantworten Sie zusätzlich die folgenden Fragen zu **beiden** Skalen:

- Wurden die Konstrukte insgesamt angemessen und vollständig abgebildet? ja ❑ nein ❑
- Sind die Dimensionen Planung/Durchführung aus Ihrer Sicht trennscharf? ja ❑ nein ❑
- Falls Sie eine Frage mit „Nein“ beantwortet haben: Bitte begründen Sie Ihre Wahl und nutzen Sie den Platz für Bemerkungen. →

Platz für Ihre Bemerkungen:

Bitte beurteilen und kommentieren Sie im Folgenden die Items der zwei Skalen zum Handlungsfeld **Umgang mit Aufgaben** nach den folgenden Kriterien. Bitte geben Sie ggf. Verbesserungsvorschläge an.

a. Passung des Items zum **Handlungsfeld** (HF) → Handelt es sich um eine wesentliche Komponente des Handlungsfeldes? Ja: +, Nein: –
b. Passung der **Handlungsbarriere** (HB) → Ist die Barriere authentisch und ausreichend intersubjektivierbar? Ja: +, Nein: –
c. **Itemschwierigkeit** (IS) → Bitte ordnen Sie die Items danach, wie schwierig es ist, ihnen zuzustimmen! 1: am leichtesten, 8: am schwersten

Items: Umgang mit Aufgaben /Planung		**HF** (+/–)	**HB** (+/–)	**IS** (1-8)	**Kommentar, Verbesserung**
P1	Ich bin in der Lage, eigene Aufgaben für ein physikalisches Thema zu entwickeln, obwohl ich das Thema noch nicht unterrichtet habe.				
P2	Wenn ich einen Physiktest erstelle, gelingt es mir immer, passende Aufgaben für alle Anforderungsbereiche auszuwählen, auch wenn ich unter Zeitdruck stehe.				
P3	Ich kann eine typische Physik-Rechenaufgabe zu einer offenen Aufgabe mit mehreren Lösungswegen umgestalten, auch wenn ich nur wenig Planungszeit habe.				
P4	Bei der Planung meines Physikunterrichts gelingt es mir immer zu beurteilen, ob eine Aufgabe zu meinen Zielen passt, auch wenn ich die Aufgabe noch nicht im Unterricht eingesetzt habe.				
P5	Ich schaffe es, eine strukturierte Lernaufgabe zu entwickeln, mit der sich Schülerinnen und Schüler ein Themengebiet selbstständig erschließen können, auch wenn ich dieses Thema bisher anders unterrichtet habe.				
P6	Fehler und Schwierigkeiten, die bei Schülerinnen und Schülern auftreten, wenn sie eine Aufgabe bearbeiten, kann ich mit Sicherheit vorhersehen, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.				
P7	Aufgaben binnendifferenziert zu gestalten ist für mich kein Problem, auch wenn ich für das betreffende physikalische Thema noch kein Beispiel kenne.				
P8	Auch für sehr komplexe Physikaufgaben kann ich gestufte Lernhilfen entwickeln, die den Schülerinnen und Schülern ein selbstständiges Weiterarbeiten ermöglichen.				

Items: Umgang mit Aufgaben/Durchführung		**HF** (+/–)	**HB** (+/–)	**IS** (1-8)	**Kommentar, Verbesserung**
D1	Es gelingt mir immer, Physikaufgaben so auszuwerten, dass alle auch in dieser Phase noch dazulernen, ohne dass sich einzelne Schülerinnen und Schüler dabei langweilen.				
D2	Wenn ich für einen Kollegen spontan einspringen muss, kann ich den Schülerinnen und Schülern bei der Bearbeitung von Physikaufgaben helfen, auch ohne den Lösungsweg einfach vorzugeben.				
D3	Wenn ich im Physikunterricht feststelle, dass meine Schülerinnen und Schüler die Aufgabenstellung nicht verstehen, kann ich sie spontan angemessen variieren, auch wenn ich darauf nicht gefasst war.				
D4	Wenn Schülerinnen und Schüler Schwierigkeiten bei der Bearbeitung einer Physikaufgabe haben, gelingt es mir immer herauszufinden, worin genau die Schwierigkeit besteht, auch wenn ich die Aufgabe das erste Mal einsetze.				
D5	Wenn der Verlauf einer Stunde von meiner Planung abweicht, kann ich auch spontan angemessene Aufgabenstellungen formulieren, selbst wenn ich das physikalische Thema das erste Mal unterrichte.				
D6	Eine von mir vorgesehene Physikaufgabe kann ich spontan im Unterricht an verschiedene Leistungsniveaus anpassen, obwohl ich vorher nicht mit dieser Situation gerechnet habe.				
D7	Bei der Bearbeitung einer Physikaufgabe kann ich meine Schülerinnen und Schüler auch dann sinnvoll unterstützen, wenn sie Lösungswege einschlagen, die ich nicht vorhergesehen habe.				
D8	Es gelingt mir immer, den Schülerinnen und Schülern Aufgaben so vorzustellen, dass sie erkennen, weshalb eine Bearbeitung sinnvoll ist, auch wenn das physikalische Thema wenig Anwendungsbezüge aufweist.				

Bitte beantworten Sie zusätzlich die folgenden Fragen zu **beiden** Skalen:

- Wurden die Konstrukte insgesamt angemessen und vollständig abgebildet? ja ❑ nein ❑
- Sind die Dimensionen Planung/Durchführung aus Ihrer Sicht trennscharf? ja ❑ nein ❑
- Falls Sie eine Frage mit „Nein" beantwortet haben: Bitte begründen Sie Ihre Wahl und nutzen Sie den Platz für Bemerkungen. →

Platz für Ihre Bemerkungen:

In der nachstehenden Tabelle sind die physikdidaktischen Handlungsfelder angegeben, zu denen Sie auf den letzten Seiten die entwickelten Skalen begutachtet haben. Bitte beurteilen Sie nun die Relevanz der genannten Handlungsfelder und begründen Sie (im vorgesehenen Kommentarfeld), falls Sie ein Handlungsfeld als weniger/nicht relevant einschätzen:

Handlungsfeld	**Relevant? (+/–)**	**Kommentar: Warum nicht relevant?**
schülerorientierte Elementarisierung von physikalischen Inhalten		
Umgang mit Aufgaben		

Für welche physikdidaktischen Handlungsfelder sollten Ihrer Meinung nach weitere SWE-Skalen konstruiert werden?

Halten Sie die Differenzierung in Planungs- und Durchführungsebene für relevant? Falls „Nein“, warum nicht?

ja ❑ nein ❑

Haben Sie noch zusätzliche Hinweise, Kritik, Anmerkungen oder Ideen?

Gern würden wir Sie bei auftretenden Fragen erneut kontaktieren. Würden Sie ggf. für ein Gespräch zur Verfügung stehen?

ja ❑ nein ❑

Wir bedanken uns sehr herzlich für Ihre Beteiligung!

I.2. Expertenfragebogen (Psychologie)

Expertenbefragung im Rahmen einer Validierungsstudie bzgl. des Konstrukts der Selbstwirksamkeitserwartung (SWE) in physikdidaktischen Handlungsfeldern

Was erwartet Sie auf den folgenden Seiten?

Wir stellen Ihnen unser Projekt und insbesondere den von uns verfolgten Ansatz zur Entwicklung und Validierung der Skalen kurz vor. Bezüglich einiger Facetten und Fragestellungen möchten wir Sie bitten, eine Bewertung vorzunehmen. Die Bearbeitung dauert **ca. 30 Minuten**.

Über Kommentare und Anmerkungen jeglicher Art freuen wir uns sehr! Falls Ihnen der Platz auf den vorgesehenen Seiten dazu nicht ausreicht, nutzen Sie für Ihre Ergänzungen einfach die freien letzten Seiten.

Vielen Dank!

1. Ziel der Skalenentwicklung/Spezifitätsniveau

Selbstwirksamkeitserwartungen angehender Physiklehrkräfte stellen ein bisher vernachlässigtes Konstrukt in der Forschung zur Lehrerprofessionalisierung dar. Insbesondere valide Erhebungsinstrumente zu Handlungsfeldern, die Physiklehrkräfte betreffen, sind bisher wenig vorhanden, obwohl von der Annahme ausgegangen werden kann, dass domänenspezifische, also „physikdidaktische" SWE-Skalen prädiktiv überlegen sind, wenn Handeln von Physiklehrkräften untersucht werden soll.

Bisher haben wir Items zu vier physikdidaktischen Handlungsfeldern entwickelt, die auch von Experten (Professoren, Lehrkräften) als relevant eingeschätzt wurden. Es handelt sich um die Handlungsfelder *Umgang mit Experimenten*, *Umgang mit Aufgaben*, *Umgang mit Schülervorstellungen* und *Elementarisierung von physikalischen Inhalten*.

- Bitte schätzen Sie ein, wie sinnvoll die Entwicklung von Items/Skalen auf dem beschriebenen Spezifitätsniveau ist.

Platz für Ihre Einschätzung:

2. Dimensionen Planung/Durchführung

Pro Handlungsfeld haben wir Skalen in den Dimensionen Planung und Durchführung von Physikunterricht entwickelt, weil diese die Kerntätigkeiten von Lehrpersonen darstellen.[1]

Itembeispiel Planung: Ich kann *bei meiner Unterrichtsvorbereitung* auch aus stichpunktartigen Aufgabenbearbeitungen Schülervorstellungen erschließen, die für die weitere Planung relevant sind.

Itembeispiel Durchführung: Ich kann Schülervorstellungen *während des Unterrichtsgesprächs* rekonstruieren, auch wenn sie nur indirekt erkennbar sind.

- Bitte beurteilen Sie, inwiefern die Erfassung von Selbstwirksamkeitserwartungen in zwei getrennten Dimensionen (Planung und Durchführung von Physikunterricht) sinnvoll und erfolgversprechend ist.

Platz für Ihre Einschätzung:

[1] Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. Zeitschrift für Erziehungswissenschaft, 9(4), 469–520.

3. Grundlagen der Itemformulierung

Für die Itemkonstruktion orientieren wir uns an der von Ihnen formulierten Definition von Selbstwirksamkeitserwartungen ...

> „Unter Selbstwirksamkeitserwartungen versteht man die subjektive Gewi[ss]heit, eine neue oder schwierige Aufgabe auch dann erfolgreich bearbeiten zu können, wenn sich Widerstände in den Weg stellen.“[2]

… und den sich daraus ergebenen semantischen Implikationen bzw. Regeln, die wir leicht modifiziert haben:

1. Formulierung in der ersten Person singular: *Ich…*
2. Bezug zu eigenen Fähigkeiten durch Verwendung von Verbphrasen wie *Ich kann …* oder *Ich bin in der Lage …*,
3. Verdeutlichung von Schwierigkeiten (Barrieren), bei denen es sich nicht um Personeneigenschaften, sondern um externe Handlungsbedingungen handelt, beispielsweise, fehlende Materialien.[3]

Beispielitem (Elementarisierung_Durchführung)

Regeln 1 & 2 **Regel 3**

Ich kann Alltagsbezüge zu einem physikalischen Inhalt herstellen, **auch wenn** ich mir vor dem Unterricht keine Beispiele überlegt hatte.

- Bitte nehmen Sie dazu Stellung, inwiefern die Items angemessen formuliert sind, um SWE im Sinne der angegebenen Definition zu erfassen. Im Anhang finden Sie zusätzlich eine Auswahl von Items.

Platz für Ihre Einschätzung:

[2] Schmitz, G. S. & Schwarzer, R. (2000). Selbstwirksamkeitserwartung von Lehrern: Längsschnittbefunde mit einem neuen Instrument. Zeitschrift für Pädagogische Psychologie, 14(1), S. 13.
[3] Ebd., S.14

4. Abstufung der Likertskalen

In einer ersten Pilotierungsstudie haben wir uns an der von Ihnen verwendeten Abstufung orientiert:

stimmt nicht – stimmt kaum – stimmt eher – stimmt genau

In einem Abstufungspräferenztest und Interviews mit Studierenden kristallisierte sich die folgende Abstufung heraus:

stimmt nicht – stimmt sehr wenig – stimmt wenig – stimmt eher – stimmt überwiegend – stimmt genau

- Welche Bedenken sprechen aus Ihrer Sicht gegen ein 6-stufiges Antwortformat?

Platz für Ihre Einschätzung:

5. Anordnung der Items

Bezüglich der Vollständigkeit und Trennschärfe der Items in den einzelnen Handlungsfeldern liegen uns Experteneinschätzungen vor. Allerdings stellt sich uns im Zusammenhang mit der Darbietung der Items die Frage, ob die Items den Befragten skalenweise und damit in inhaltlichen Blöcken oder durchmischt präsentiert werden sollen. Denkbar sind verschiedene Durchmischungsstufen, und zwar

1. keine Durchmischung (skalenweise Darbietung der Items: je Handlungsfeld getrennt nach Planungs- und Durchführungsdimension),
2. Durchmischung der Planungs- und Durchführungsitems innerhalb eines Handlungsfeldes,
3. Durchmischung aller Items unabhängig von Dimension und Handlungsfeld.

➢ Bitte erläutern Sie kurz die Vorzüge der von Ihnen favorisierten Vorgehensweise im Rahmen einer Validierungsstudie.

Platz für Ihre Einschätzung:

6. Itemformat

Bezüglich der Itemdarbietung diskutieren wir, ob die Items mit einem einheitlichen Satzanfang beginnen sollten.

Beispiel einheitlicher Satzanfang	**Beispiel uneinheitlicher Satzanfang**
Ich kann …	
… die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn diese widersprüchlich sind.	Ich bin in der Lage, die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einzubeziehen, auch wenn diese widersprüchlich sind.
… bei der Planung meines Physikunterrichts verschiedene Methoden gegeneinander abwägen, um die Alltagsvorstellungen meiner Schülerinnen und Schüler zu erfassen, auch wenn auf den ersten Blick mehrere Methoden geeignet erscheinen.	Bei der Planung meines Physikunterrichts kann ich verschiedene Methoden gegeneinander abwägen, um die Alltagsvorstellungen meiner Schülerinnen und Schüler zu erfassen, auch wenn auf den ersten Blick mehrere Methoden geeignet erscheinen.

➢ Bitte erläutern Sie kurz die Vorzüge des von Ihnen favorisierten Itemformats.

Platz für Ihre Einschätzung:

7. Benachbarte Konstrukte

- Welche psychologischen Konstrukte sollten Ihrer Meinung nach im Rahmen der weiteren Validierung zusätzlich erhoben werden?

Platz für Ihre Vorschläge:

Haben Sie noch zusätzliche Hinweise, Kritik, Anmerkungen oder Ideen?

Gern würden wir Sie bei auftretenden Fragen erneut kontaktieren. Würden Sie ggf. für ein Gespräch zur Verfügung stehen?

ja ❑ nein ❑

Wir bedanken uns sehr herzlich für Ihre Beteiligung!

Universität Potsdam, AG Didaktik der Physik
Expertenbefragung

ANHANG – Itemauswahl

Umgang mit Experimenten – Planung

- Auch für eine sehr heterogene Klasse kann ich Schülerexperimente so planen, dass ich sie zur Differenzierung nutzen kann.
- Ich bin in der Lage, zu meinem Unterricht passende Experimente vorzubereiten, auch wenn die Physiksammlung schlecht ausgestattet ist.

Umgang mit Experimenten – Durchführung

- Ich kann ein Demonstrationsexperiment für Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine sehr komplexe Versuchsanordnung handelt.
- Auch unter Zeitdruck bin ich fast immer in der Lage, ein Experiment zum Laufen zu bringen, wenn es im Physikunterricht nicht auf Anhieb funktioniert.

Umgang mit Aufgaben – Planung

- Bei der Planung meines Physikunterrichts kann ich angemessen beurteilen, ob eine komplexe Aufgabe zu meinen Lernzielen passt, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.
- Ich kann in meiner Unterrichtsplanung eine Aufgabe entwickeln, mit der sich Schülerinnen und Schüler einen physikalischen Inhalt selbstständig erschließen können, auch wenn das Thema für die Lernenden schwierig ist.

Umgang mit Aufgaben – Durchführung

- Bei der Bearbeitung einer Physikaufgabe kann ich meine Schülerinnen und Schüler auch dann sinnvoll unterstützen, wenn sie Lösungswege einschlagen, die ich nicht vorhergesehen habe.
- Ich kann Physikaufgaben im Unterricht so besprechen, dass auch in der Auswertungsphase die Schülerinnen und Schüler noch etwas dazulernen, ohne sich zu langweilen.

Umgang mit Schülervorstellungen – Planung

- Ich kann Unterrichtssituationen planen, in denen meine Schülerinnen und Schüler eine neu erworbene physikalische Vorstellung als hilfreich erleben, aber deren Alltagsvorstellungen nicht pauschal abgewertet werden.
- Ich kann die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn diese widersprüchlich sind.

Umgang mit Schülervorstellungen – Durchführung

- Wenn es sinnvoll ist, kann ich mit unerwarteten Schülervorstellungen im Physikunterricht inhaltlich weiterarbeiten, auch wenn sie nicht in meine Planung passen.
- Ich kann auch spontan geeignete Beispiele heranziehen, die die Schülerinnen und Schüler dazu veranlassen, ihre Alltagsvorstellungen zu hinterfragen, auch wenn sie sehr überzeugt von ihnen sind.

Elementarisierung von physikalischen Inhalten – Planung

- Es gelingt mir immer, Inhalte physikalisch sinnvoll anzuordnen, obwohl ich nicht auf ein Physikbuch zurückgreifen kann.
- Ich schaffe es, auch komplexe Themen der modernen Physik für eine Unterrichtseinheit so zu vereinfachen, dass meine Schülerinnen und Schüler sie verstehen.

Elementarisierung von physikalischen Inhalten – Durchführung

- Wenn sich im Unterricht die Gelegenheit bietet, kann ich spontan auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen aufzeigen.
- Auch wenn ich meinen Physikunterricht spontan an Interessen der Schülerinnen und Schüler anpasse, gelingt es mir immer, eine sinnvolle inhaltliche Reihenfolge beizubehalten.

I.3. Beispielübersicht Itemauswertung (Expertenbefragung)

Nachfolgende Tabelle enthält exemplarisch die Expertenmeinungen zum Item:

„Ich kann die Vorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie sehr vielfältig sind.“ (Skala: SWE-SV-P, Version 2, Item 1, vgl. Anhang G.2.5)

Tabelle I.1.: Experten-Auswertungsübersicht zum Item 1 der Skala SWE-SV-P, Version 2.

Experte	**HF**[1]	**HB**[2]	**IS**[3]	**Kommentare**
1	ja	nein	3	„Vielfältig“ ist für mich zu unbestimmt (zahlreich, umfassend, tief verwurzelt?).
2	ja	ja	5	–
3	ja	ja	1	–
4	ja	ja	5	–
5	ja	ja	1	–
6	ja	ja	5	–
7	ja	ja	–	–
8	ja	ja	–	–

[1] Passung Handlungsfeld [2] Passung der Barriere zum Item
[3] eingeschätzte Itemschwierigkeit

I.4. Beispielübersicht Skalenauswertung (Expertenbefragung)

Nachfolgende Tabelle enthält exemplarisch die Expertenmeinungen zu den Skalen SWE-SV-P und SWE-SV-D:

Tabelle I.2.: Experten-Auswertungsübersicht zum Handlungsfeld SWE-SV, Version 2.

Experte	A/V[1]	D[2]	Kommentare
1	ja	ja	„Ich halte die extremen Formulierungen für problematisch (jede Unterrichtssituation, immer in der Lage ...). Möglichkeiten zur Differenzierung bzw. Abstufung ließen sich mit einer Likert-Skala anbieten, dazu würden aber auch diese Formulierungen stören.“
2	o. A.	größtenteils	„Mir scheinen die Barrieren (und damit die Schwierigkeiten 1-7) im Schnitt zu hoch → Deckeneffekte; relativ häufig doppelte Barrieren, m. E. wenig sinnvoll“
3	ja	ja	„D1 und D3 sind sehr nah beieinander. D6 und D7 ebenfalls“
4	nein	ja	„evtl. explizit die Bekanntheit von Testinstrumenten wie z. B. FCI erheben; SuS-Vorstellungen und "cultural diversity"werden hier (wie auch sonst) m. E. unterschätzt!“
5	nein	nein	„Es werden beim Umgang mit SV nur das Aktivieren (falscher) und eine entsprechende Konfliktstrategie angesprochen. Der Aspekt, solche SV zu aktivieren, die ohne Konflikt (und explizite) Thematisierung zu fachlichen Vorstellungen führen (z. B. der Ansatz der Gruppe um Wiesner, der z. B. bei der Einführung der Mechanik von Phenomenological Primitives ausgeht, wird nicht mit genannt. Damit reduziert sich der Umgang mit SV auf Methodisches und der inhaltliche Aspekt, geeignete Lernwege bzw. das Umlernen vermeidende Elementarisierungen zu wählen, wird vernachlässigt. Bei der Durchführung fehlt mir die Fähigkeit, flexibel bei unvorhergesehenen SV zu reagieren und einen alternativen Zugang zu den fachlichen Konzepten zu wählen.“
6	ja	ja	„Z. T. sind die Items schwierig formuliert, z. B. P2; Die Handlungsbarrieren finde ich z. T. sehr konstruiert.“
7	o. A.	o. A.	„Im Gegensatz zum Umgang mit Aufgaben, fiel mir die Beurteilung hier schwer. Deshalb kann ich die Frage links nicht mit ja/nein beantworten. Die Trennung von Planung und Durchführung ist nicht trennscharf eingehalten. Eine Frage, die sich ein Lehrer stellen könnte: Sind mit Schülervorstellungen die gemeint, die die Fachdidkaktik in Untersuchungen zutage gefördert hat (z. B. Kraft-Haben-Vorstellung) oder sind es die Vorstellungen, die Schüler im laufenden Lernprozess zu dem anstehenden Problem haben.“
8	ja	ja	„IS habe ich nicht verstanden. Da es sehr subjektiv ist, wusste ich nicht, was hier mit Schwierigkeit gemeint war.“

[1] Angemessenheit/Vollständigkeit: Wurden die Konstrukte insgesamt angemessen und vollständig abgebildet?
[2] Dimension: Sind die Dimensionen (Planung/Durchführung) aus Ihrer Sicht trennscharf?
o. A.: ohne Angabe

J. Fragebogen zur Akzeptanzbefragung bez. des Antwortformates

Der Fragebogen wurde von Olaf Krey und Thorid Rabe erstellt und im Rahmen einer zusätzlichen Studie eingesetzt.

Aufgabenstellung

Mit dieser Befragung versuchen wir herauszufinden, wie viele Abstufungen des Kontinuums von „stimmt gar nicht“ bis „stimmt voll und ganz“ Sie benötigen, um sich bezüglich der folgenden acht Aussagen sinnvoll einzuschätzen.

1. Lesen die Aussagen in der folgenden Tabelle Item für Item und verorten Sie sich gedanklich auf dem Spektrum von „stimmt gar nicht“ bis „stimmt voll und ganz“.

2. Entscheiden Sie, wie viele Abstufungen Sie benötigen, um sich möglichst treffend einschätzen zu können und setzen Sie ihr Kreuz auf einer der vorgegebenen Skalen.

Nach Bearbeitung der ersten zwei Aufgaben:

3. Nachdem Sie alle Aussagen durchgegangen sind, versuchen Sie bitte für die feinste von Ihnen gewählte Abstufung eine Beschriftung vorzuschlagen. Die folgende Tabelle soll Ihnen dabei behilflich sein.

Ich benötige maximal _____ Abstufungen, um mich treffend einzuschätzen.

Mein (ggf. auch unvollständiger) Beschriftungsvorschlag wäre der folgende:

Stufe	☐	☐
Beschriftung	stimmt gar nicht	stimmt voll und ganz

Nr.	Aussage	Ausmaß der Zustimmung stimmt gar nicht ← → stimmt voll und ganz
1	Ich bin in der Lage, eigene Aufgaben für ein physikalisches Thema zu entwickeln, obwohl ich das Thema noch nicht unterrichtet habe.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
2	Wenn ich für einen Kollegen spontan einspringen muss, kann ich den Schülerinnen und Schülern bei der Bearbeitung von Physikaufgaben helfen, auch ohne den Lösungsweg einfach vorzugeben.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
3	Ich schaffe es, auch komplexe Themen der modernen Physik für eine Unterrichtseinheit so zu vereinfachen, dass meine Schülerinnen und Schüler sie verstehen.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
4	Wenn sich im Unterricht die Gelegenheit bietet, kann ich spontan auch zwischen so unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik Verbindungen aufzeigen.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
5	Auch Zeitdruck während der Unterrichtsvorbereitung hindert mich nicht daran, ein zu meinen Unterrichtszielen passendes Experiment zu entwickeln.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
6	Ich weiß, dass ich beim Experimentieren in einer Vertretungsstunde auf Unvorhergesehenes reagieren kann, auch wenn ich das Experiment längere Zeit nicht eingesetzt habe.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
7	Ich kann mir Methoden überlegen, die geeignet sind, physikalische Vorstellungen anzuregen, ohne Schülervorstellungen und physikalische Sichtweise einfach zu kontrastieren.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung
8	Auch auf ungewöhnliche Schülervorstellungen kann ich im Unterrichtsgespräch lernförderlich reagieren.	☐ ☐ ☐ ☐ 4-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ 6-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 8-stufige Einteilung ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ 10-stufige Einteilung

K. Beispiel einer zusammenfassenden Übersicht zur Itemüberarbeitung nach Pilotstudien II & III

Sv-p1
Ich kann die Vorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie sehr vielfältig sind

IS = 3,333

Probleme (Experten)

- Vielfältig zu unbestimmt

Probleme/Interviewte

- Was heißt in die Unterrichtsplanung einbeziehen???
- Planungsaspekt eher selten ausformuliert, eher werden U-Situationen assoziiert
- geht es nur um in Literatur beschriebene Vorstellungen oder muss ich die vorher noch erfassen --> Schüleräußerungen/U-Erfahrung wird quasi vorausgesetzt
- Schülervorstellungen erst mal nur heraussuchen/ kennen?
- Vielfältigkeit nicht wirklich eine Handlungsbarriere, weil das quasi den SV immanent ist und gerade SV sehr abstrus sein können
- Vorstellungen als Interessen

Vorschläge:

Bei der Planung einer Physikstunde berücksichtige ich die vielfältigen Schülervorstellungen, indem ich mir geeignete Szenarien überlege, um die SV im Unterricht zu thematisieren.

Ich kann die vielfältigen, in der Literatur beschriebenen Schülervorstellungen in die Unterrichtsplanung einer Physikstunde einbeziehen und mir Unterrichtssituationen überlegen, in denen den SuS die Grenzen ihrer Vorstellungen bewusst werden.

Sv-p1 (neu)
Ich kann die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn sie einander widersprechen.

L. Fragebogen – Hauptstudie

Nachfolgend ist exemplarisch der Fragebogen mit der Handlungsfeldreihenfolge SWE-SV, SWE-A, SWE-EX, SWE-EL für die Kohorte der Studierenden abgebildet.

FRAGEBOGEN

Selbsteinschätzungen bzgl. der Planung und Durchführung von Physikunterricht

Claudia Meinhardt, Thorid Rabe & Olaf Krey
Universität Potsdam
Institut für Physik und Astronomie
Lehrstuhl Didaktik der Physik
Karl-Liebknecht-Str. 24/25
14476 Potsdam-Golm

Liebe Studierende,

vielen Dank für die Teilnahme an der Befragung. Sie unterstützen damit mein Promotionsprojekt an der Universität Potsdam. Die Befragung ist natürlich **freiwillig**. Ihre Angaben werden **anonymisiert** und ausschließlich zu Forschungszwecken verwendet!

Wenn Sie Fragen haben, wenden Sie sich gern an: Claudia.Meinhardt@uni-potsdam.de

---

Zunächst bitten wir Sie um einige allgemeine Angaben zu Ihrer Person.

1. **Geschlecht:** ❑ männlich ❑ weiblich **2. Alter:**_____

3. **derzeitiger Studiengang:** ❑ Bachelor ❑ Master ❑ Staatsexamen

4. **Fachsemester** (ggf. Summe aus Bachelor und Master):_____
   Falls sich die Anzahl der der Fachsemester für Ihre Fächer unterscheiden, geben Sie bitte die Fachsemester für das Fach Physik an.

5. Ich studiere an einer ❑ **Uni/** ❑ **PH** im **Bundesland:**_____

6. Ich studiere das **Lehramt:**_____(z.B. an Gymnasien/Primarstufe/...)

7. **Fächerkombination:**

   Fach 1:_____ Fach 2:_____ ggf. Fach 3:_____

8. Die **Physikklausuren** habe ich bisher im Mittel wie folgt bestanden (z.B. 2,3 oder 3,0):

   _____

9. Schätzen Sie bitte den **Umfang** der bisher von Ihnen belegten **Lehrveranstaltungen im Bereich Physikdidaktik** in SWS und/oder credit points (cp)?

   SWS:_____ cp:_____

10. Bitte geben Sie an, welche **Art von schulpraktischen Erfahrungen (Durchführen von Physikunterricht)** Sie bereits im Rahmen eines Praktikums oder einer anderen Lehrveranstaltung sammeln konnten:

    ❑ keine
    ❑ Tagespraktikum/Schulpraktische Studien
    ❑ mehrwöchiges Schulpraktikum
    ❑ Praxissemester
    ❑ Nachhilfe
    ❑ Aushilfs-/Vertretungslehrkraft
    ❑ sonstiges:_____

11. Bitte schätzen Sie ein, über wie viel **Unterrichtserfahrung** als Lehrkraft im Fach Physik Sie insgesamt verfügen:

    in Zeitstunden:_____

Hinweise zum Ausfüllen des Fragebogens

Wir bitten Sie, sich auf den nachfolgenden Seiten **selbst einzuschätzen**. Es handelt sich um Situationen, die sich auf das Handeln von Physiklehrkräften beziehen. Es kann sein, dass Sie diese Situationen noch nicht selbst als Physik-Lehrkraft erlebt haben. Versuchen Sie bitte dennoch sich **in die Situationen hineinzuversetzen** und entscheiden Sie, ob Sie mit ihrem **heutigen** (!) Wissen und Können diese Situation bewältigen können oder nicht.

Wir bitten Sie, sich möglichst **realistisch einzuschätzen**.
Für unsere Auswertung ist es wichtig, dass Sie möglichst **alle** Fragen beantworten.

Los geht's…

Bitte geben Sie an, wie sehr sie den folgenden Aussagen zustimmen.

Aussagen	**stimmt nicht**	**stimmt kaum**	**stimmt eher**	**stimmt genau**
Ich weiß, dass ich es schaffe, selbst den problematischsten Schülerinnen und Schülern den Stoff zu vermitteln.	❑	❑	❑	❑
Ich traue mir zu, die Schülerinnen und Schüler für neue Projekte zu begeistern.	❑	❑	❑	❑
Ich kann Innovationen auch gegenüber skeptischen Kolleginnen und Kollegen durchsetzen.	❑	❑	❑	❑
Ich weiß, dass ich zu den Eltern guten Kontakt halten kann, selbst in schwierigen Situationen.	❑	❑	❑	❑
Ich bin mir sicher, dass ich auch mit den problematischen Schülerinnen und Schülern in guten Kontakt kommen kann, wenn ich mich darum bemühe.	❑	❑	❑	❑
Ich bin mir sicher, dass ich mich in Zukunft auf individuelle Probleme der Schülerinnen und Schüler noch besser einstellen kann.	❑	❑	❑	❑
Selbst wenn mein Unterricht gestört wird, bin ich mir sicher, die notwendige Gelassenheit bewahren zu können.	❑	❑	❑	❑
Selbst wenn es mir mal nicht so gut geht, kann ich doch im Unterricht immer noch gut auf die Schülerinnen und Schüler eingehen.	❑	❑	❑	❑
Auch wenn ich mich noch so sehr für die Entwicklung meiner Schülerinnen und Schüler engagiere, weiß ich, dass ich nicht viel ausrichten kann.	❑	❑	❑	❑
Ich bin mir sicher, dass ich kreative Ideen entwickeln kann, mit denen ich ungünstige Unterrichtsstrukturen verändere.	❑	❑	❑	❑

Kurze Begriffsklärung

Im den folgenden Formulierungen ist häufig die Rede von **Schüler- bzw. Alltagsvorstellungen**. Darunter sollen Denkweisen und Vorstellungen zu physikalischen Begriffen und Phänomenen verstanden werden, die häufig in Widerspruch zu den entsprechenden physikalischen Modellen und Konzepten stehen. Schülervorstellungen sind häufig im Alltag nützlich und haben sich dort auch bewährt. Beispiele wären: *Elektronen sind winzige Kügelchen* oder *Glühlampen verbrauchen Strom*.

Ich kann …	**stimmt nicht** 1					**stimmt genau** 6
…im Physikunterricht die Grenzen von geäußerten Alltagsvorstellungen mit meinen Schülerinnen und Schülern erarbeiten, ohne bloß zu sagen, wie es physikalisch "richtig" ist.	❑	❑	❑	❑	❑	❑
…im Physikunterricht an Beispielen verdeutlichen, dass die physikalische Sichtweise hilfreich ist, ohne die Alltagsvorstellungen der Schülerinnen und Schüler als nutzlos darzustellen.	❑	❑	❑	❑	❑	❑
…die Alltagsvorstellungen meiner Schülerinnen und Schüler in meine Unterrichtsplanung einbeziehen, auch wenn die Vorstellungen einander widersprechen.	❑	❑	❑	❑	❑	❑
…eine Physikstunde planen, die Schülervorstellungen zum Ausgangspunkt von Lernprozessen macht, auch wenn ich zu dem entsprechenden Thema keinen solchen Unterrichtsvorschlag kenne.	❑	❑	❑	❑	❑	❑
…bei der Planung meines Physikunterrichts einen Unterrichtsvorschlag daraufhin beurteilen, welche Schülervorstellungen vermutlich aktiviert werden, auch wenn ich ihn noch nicht ausprobiert habe.	❑	❑	❑	❑	❑	❑
…mit ergiebigen Schülervorstellungen im Physikunterricht inhaltlich weiterarbeiten, auch wenn sie nicht in meine ursprüngliche Planung passen.	❑	❑	❑	❑	❑	❑
…Unterrichtssituationen planen, in denen Schülerinnen und Schüler ihre Alltagsvorstellungen hinterfragen, auch wenn sich diese im Alltag bewährt haben.	❑	❑	❑	❑	❑	❑
…im Physikunterricht entscheiden, ob es sinnvoll ist, den Unterrichtsverlauf an eine geäußerte Schülervorstellung anzupassen, auch wenn mich diese überrascht.	❑	❑	❑	❑	❑	❑
…Unterrichtssituationen planen, in denen meine Schülerinnen und Schüler eine neu erworbene physikalische Vorstellung als hilfreich erleben, aber Alltagsvorstellungen nicht als sinnlos erscheinen.	❑	❑	❑	❑	❑	❑
…bei meiner Unterrichtsvorbereitung Schülervorstellungen aus Aufgabenbearbeitungen erschließen, auch wenn es sich nur um Stichpunkte handelt.	❑	❑	❑	❑	❑	❑
…auch spontan geeignete Beispiele heranziehen, die die Schülerinnen und Schüler dazu veranlassen, ihre Alltagsvorstellungen zu hinterfragen, auch wenn sie sehr überzeugt von ihnen sind.	❑	❑	❑	❑	❑	❑
…Schülervorstellungen während eines Unterrichtsgesprächs zu einem physikalischen Thema rekonstruieren, auch wenn sie nur indirekt erkennbar sind.	❑	❑	❑	❑	❑	❑
…für meine Planung inhaltlich begründet entscheiden, ob ich Schülervorstellungen zum Ausgangspunkt einer Lernsequenz mache oder dies bewusst vermeide, auch wenn ich wenig Unterrichtserfahrung mit diesem physikalischen Inhalt habe.	❑	❑	❑	❑	❑	❑
…meinen Physikunterricht so planen, dass er schrittweise von Schülervorstellungen zu physikalischen Vorstellungen führt, ohne diese Vorstellungen einfach einander gegenüberzustellen.	❑	❑	❑	❑	❑	❑
…im Physikunterricht geäußerte Vorstellungen meiner Schülerinnen und Schüler hinsichtlich ihrer fachlichen Angemessenheit beurteilen, auch wenn ich sofort reagieren muss.	❑	❑	❑	❑	❑	❑

Ich kann …	**stimmt nicht** 1					**stimmt genau** 6
…bei Verständnisschwierigkeiten eine Aufgabestellung im Physikunterricht spontan umformulieren, ohne die Aufgabe dabei zu vereinfachen.	❑	❑	❑	❑	❑	❑
…bei der Auswertung von Physikaufgaben Fehler einzelner Schülerinnen und Schüler für den Lernprozess der gesamten Lerngruppe nutzen, ohne jemanden bloß zu stellen.	❑	❑	❑	❑	❑	❑
…die Auswertung von Aufgaben aus einem Physiktest so planen, dass alle Schülerinnen und Schüler inhaltlich etwas dazu lernen können, ohne dass einfach der Erwartungshorizont vorgestellt wird.	❑	❑	❑	❑	❑	❑
…bei der Unterrichtsvorbereitung Schwierigkeiten vorhersehen, die bei der Aufgabenbearbeitung auftreten könnten, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.	❑	❑	❑	❑	❑	❑
…meine Schülerinnen und Schülern bei der Bearbeitung von Physikaufgaben unterstützen, auch ohne den Lösungsweg einfach vorzugeben.	❑	❑	❑	❑	❑	❑
…in meinem Physikunterricht die Auswertung einer Aufgabe strukturiert gestalten, auch wenn die Lösungswege der Schülerinnen und Schüler vielfältiger sind, als ich es erwartet hatte.	❑	❑	❑	❑	❑	❑
…bei der Unterrichtsvorbereitung für Physikaufgaben Lernhilfen erstellen, die den Schülerinnen und Schülern ein selbstständiges Weiterarbeiten ermöglichen, auch wenn es sich um komplexe Aufgaben handelt.	❑	❑	❑	❑	❑	❑
…eine Prüfungsaufgabe so verändern, dass unterschiedliche Versionen mit verschiedenen Schwierigkeitsstufen entstehen, auch wenn ich unter Zeitdruck bin.	❑	❑	❑	❑	❑	❑
…erkennen, weshalb meine Schülerinnen und Schüler Schwierigkeiten bei der Bearbeitung einer Physikaufgabe haben, auch wenn sie ihr Problem noch nicht selbst in Worte fassen können.	❑	❑	❑	❑	❑	❑
…bei der Planung einer längeren Übungsphase Aufgaben so zusammenstellen, dass die unterschiedlichen Fähigkeiten meiner Lerngruppe berücksichtigt werden, auch wenn die Fähigkeiten stark variieren.	❑	❑	❑	❑	❑	❑
…Physikaufgaben im Unterricht so besprechen, dass die Schülerinnen und Schüler auch in der Auswertungsphase noch etwas dazulernen, ohne sich zu langweilen.	❑	❑	❑	❑	❑	❑
…meine Schülerinnen und Schüler bei der Bearbeitung einer Physikaufgabe auch dann sinnvoll unterstützen, wenn sie Lösungswege einschlagen, die ich nicht vorhergesehen habe.	❑	❑	❑	❑	❑	❑
…bei Bedarf eine offene Physikaufgabe mit mehreren Lösungswegen für meinen Physikunterricht entwickeln, auch wenn ich unter Zeitdruck bin.	❑	❑	❑	❑	❑	❑
…bei der Planung meines Physikunterrichts beurteilen, ob eine komplexe Aufgabe zu meinen Lernzielen passt, auch wenn ich die Aufgabe noch nicht im Physikunterricht eingesetzt habe.	❑	❑	❑	❑	❑	❑
…in meiner Unterrichtsplanung eine Aufgabe entwickeln, mit der sich Schülerinnen und Schüler einen physikalischen Inhalt selbstständig erschließen können, auch wenn das Thema für die Lernenden schwierig ist.	❑	❑	❑	❑	❑	❑
…meinen Schülerinnen und Schülern einsichtig machen, weshalb die Bearbeitung einer Aufgabe für ihren eigenen Lernprozess sinnvoll ist, auch wenn ein Anwendungsbezug nicht direkt ersichtlich ist.	❑	❑	❑	❑	❑	❑
…mir in unvorhergesehenen Situationen spontan eine herausfordernde Aufgabe ausdenken, ohne meine Schülerinnen und Schüler zu überfordern.	❑	❑	❑	❑	❑	❑
…Prüfungsaufgaben entwickeln, sodass Fähigkeiten der Schülerinnen und Schüler im Kompetenzbereich Kommunikation erfasst werden, auch wenn es keine passenden Materialien gibt.	❑	❑	❑	❑	❑	❑

Ich kann …	stimmt nicht 1					stimmt genau 6
…ein Demonstrationsexperiment für meine Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine komplexe Versuchsanordnung handelt.	❑	❑	❑	❑	❑	❑
…bei meiner Unterrichtsplanung ein Experiment gegebenenfalls so variieren, dass ich es in einer Übungsphase einsetzen kann, auch wenn ich es bisher nur als Einstiegsexperiment genutzt habe.	❑	❑	❑	❑	❑	❑
…beim Experimentieren spontan mit den Schülerinnen und Schülern über das Wechselspiel von Theorie und Experiment reflektieren, auch wenn ich den Anlass nicht vorgesehen hatte.	❑	❑	❑	❑	❑	❑
…zu dem Kontext einer Unterrichtsreihe ein Experiment entwickeln, auch wenn es zu diesem Kontext keine fertigen Experimentieranweisungen gibt.	❑	❑	❑	❑	❑	❑
…bei der Unterrichtsplanung didaktisch begründet entscheiden, ob ein Experiment mit Hilfe von schultypischen Experimentiersets oder mit Alltagsgegenständen durchgeführt werden soll, auch wenn ich die Lerngruppe noch nicht lange kenne.	❑	❑	❑	❑	❑	❑
…physikalische Experimente an interessante Impulse meiner Schülerinnen und Schülern anpassen, auch wenn ich das vorher nicht geplant hatte.	❑	❑	❑	❑	❑	❑
…in meiner Unterrichtsplanung zu den Lernzielen passende Experimente aufbauen, auch wenn die Physiksammlung schlecht ausgestattet ist.	❑	❑	❑	❑	❑	❑
…unerwartete Messwerte aus einem Demonstrationsexperiment spontan als Lernanlass für meine Schülerinnen und Schüler nutzen, auch ohne „unpassende“ Werte einfach zu übergehen.	❑	❑	❑	❑	❑	❑
…für ein physikalisches Experiment begründet entscheiden, ob es didaktisch sinnvoller ist, es als Demonstrations- oder Schülerexperiment einzuplanen, auch wenn ich das Experiment noch nicht eingesetzt habe.	❑	❑	❑	❑	❑	❑
…spontan ein passendes Experiment einsetzen, um auf weiterführende physikalische Fragestellungen meiner Schülerinnen und Schüler zu reagieren, auch wenn ich das nicht vorgesehen hatte.	❑	❑	❑	❑	❑	❑
…mehrere Experimente so zusammenstellen, dass bei der Auswertung unterschiedliche Möglichkeiten des Umgangs mit Messdaten deutlich werden, auch wenn ich diesbezüglich keine Unterrichtsvorschläge kenne.	❑	❑	❑	❑	❑	❑
…auf unvorhergesehene Verständnisschwierigkeiten meiner Schülerinnen und Schüler beim Experimentieren so reagieren, dass sie selbstständig weiterarbeiten können, auch ohne einfach einen Lösungsweg vorzugeben.	❑	❑	❑	❑	❑	❑
…Schülerexperimente so zusammenstellen, dass die praktischen Fähigkeiten meiner Schülerinnen und Schüler auf verschiedenen Niveaus gefördert werden, auch wenn ich bei der Planung unter Zeitdruck stehe.	❑	❑	❑	❑	❑	❑
…in meiner Unterrichtsvorbereitung ein Experiment planen, welches meine Schülerinnen und Schüler dazu anregt, physikalische Fragestellungen zu entwickeln, auch wenn ich dieses Experiment neu entwickeln muss.	❑	❑	❑	❑	❑	❑
…ein Experiment, das im Physikunterricht nicht auf Anhieb funktioniert, zum Laufen bringen, auch wenn ich unter Zeitdruck stehe.	❑	❑	❑	❑	❑	❑
…meine Schülerinnen und Schüler bei der Planung ihres experimentellen Vorgehens unterstützen, auch wenn sie im Physikunterricht Ihren eigenen Fragestellungen nachgehen.	❑	❑	❑	❑	❑	❑
…ein Experiment so inszenieren, dass meine Schülerinnen und Schüler motiviert sind, eigene physikalische Fragestellungen zu entwickeln, auch wenn es sich um eine unbeliebte Randstunde handelt.	❑	❑	❑	❑	❑	❑
…ein Experiment planen, das meine Schülerinnen und Schüler begeistert, auch wenn sie sich sonst wenig für Physik interessieren.	❑	❑	❑	❑	❑	❑
…verschiedene Varianten eines Experimentes planen, mit denen sich Physikunterricht leistungsdifferenziert gestalten lässt, auch wenn es dazu keine Unterrichtsvorschläge gibt.	❑	❑	❑	❑	❑	❑

Ich kann …	stimmt nicht 1					stimmt genau 6
…in meiner Unterrichtsplanung ein physikalisches Thema in sinnvolle Lernschritte zerlegen, auch wenn dieses Thema bisher noch nicht didaktisch aufbereitet wurde.	❑	❑	❑	❑	❑	❑
…bei der Planung meines Physikunterrichts die Kernideen eines physikalischen Themas identifizieren, auch wenn es sich nicht um ein klassisches Thema der Schulphysik handelt.	❑	❑	❑	❑	❑	❑
…eine Physikstunde so durchführen, dass die Schülerinnen und Schüler die physikalischen Kernideen erkennen, auch wenn ich von meiner ursprünglichen Stundenplanung stark abweichen muss.	❑	❑	❑	❑	❑	❑
…eine Unterrichtssequenz so aufbereiten, dass neben dem physikalischen Fachwissen ein weiterer Kompetenzbereich (Kommunikation, Erkenntnisgewinnung, Bewertung) besonders gefördert wird, auch wenn ich diese Sequenz das erste Mal unterrichte.	❑	❑	❑	❑	❑	❑
…eine Sequenz für den Physikunterricht so planen, dass ein Alltagsphänomen den weiteren Unterrichtsverlauf strukturiert, auch wenn dadurch die gängige Fachsystematik nicht eingehalten wird.	❑	❑	❑	❑	❑	❑
…einen Alltagsbezug, den Schülerinnen und Schüler während des Physikunterrichts aufwerfen, spontan für den Lernprozess nutzbar machen, ohne meine Unterrichtsziele aus den Augen zu verlieren.	❑	❑	❑	❑	❑	❑
…bei der Planung meines Physikunterrichts die inhaltlichen Problemstellen der Schulbuchdarstellung eines physikalischen Themas erkennen, auch wenn ich nach diesem Vorschlag noch nicht unterrichtet habe.	❑	❑	❑	❑	❑	❑
…in einer Physikstunde die Basiskonzepte (System, Energie, Wechselwirkung und Materie) nutzen, mit meiner Lerngruppe Querverbindungen zwischen unterschiedlichen physikalischen Gebieten wie E-Lehre und Mechanik zu erarbeiten, auch wenn sich der Anlass überraschend ergibt.	❑	❑	❑	❑	❑	❑
…im Physikunterricht eine physikalische Erklärung, die die Schülerinnen und Schüler nicht verstehen, weiter vereinfachen, ohne das die Erklärung physikalisch falsch wird.	❑	❑	❑	❑	❑	❑
…eine gut strukturierte Unterrichtsstunde planen, auch wenn sie sich auf mehrere Teilgebiete der Physik (z.B. Energie in der Mechanik, Thermodynamik, …) bezieht.	❑	❑	❑	❑	❑	❑
…eine Physikstunde in einer sinnvollen inhaltlichen Reihenfolge durchführen, auch wenn ich mein Vorgehen spontan an Fragen der Schülerinnen und Schüler anpasse.	❑	❑	❑	❑	❑	❑
…ein physikalisches Thema beim Planen einer Unterrichtseinheit so vereinfachen, dass meine Schülerinnen und Schüler es verstehen können, auch wenn es sich dabei um Inhalte der modernen Physik handelt.	❑	❑	❑	❑	❑	❑
…in einer Physikstunde mit meinen Schülerinnen und Schülern die Verbindung zu vorangegangenen Themen erarbeiten, auch wenn sich die Gelegenheit dazu unerwartet ergibt.	❑	❑	❑	❑	❑	❑
…die im Schulbuch vorgeschlagene inhaltliche Strukturierung eines physikalischen Themas für meine Lerngruppe abwandeln, auch wenn ich wenig Planungszeit habe.	❑	❑	❑	❑	❑	❑
…eine Unterrichtssequenz planen, die auf ein bestimmtes Basiskonzept (System, Energie, Wechselwirkung und Materie) so fokussiert, dass dieses in späteren Lernprozessen sinnvoll erweitert werden kann, auch wenn der Rahmenlehrplan dazu keine Anregungen bietet.	❑	❑	❑	❑	❑	❑

Bitte geben Sie an, wie sehr sie den folgenden Aussagen zustimmen.

Aussagen	sehr gut	gut	mittel	schlecht	sehr schlecht
Ich verstehe den Stoff der Physikvorlesungen …	❑	❑	❑	❑	❑
Ich behalte den Stoff der Physikvorlesungen …	❑	❑	❑	❑	❑
Meine Leistungen in Physik sind nach meiner Einschätzung …	❑	❑	❑	❑	❑
Ich beteilige mich in den Physikvorlesungen und Übungen …	❑	❑	❑	❑	❑
Ich glaube, dass mich meine Kommilitonen für … halten.	❑	❑	❑	❑	❑
Ich glaube, dass meine Dozenten meine Leistungen in Physik als … einschätzen.	❑	❑	❑	❑	❑
Ich erwarte, dass in Zukunft meine Leistungen in Physik … sein werden.	❑	❑	❑	❑	❑

Aussagen	trifft nicht zu	trifft kaum zu	trifft eher zu	trifft genau zu
Wenn sich Widerstände auftun, finde ich Mittel und Wege, mich durchzusetzen.	❑	❑	❑	❑
Die Lösung schwieriger Probleme gelingt mir immer, wenn ich mich darum bemühe.	❑	❑	❑	❑
Es bereitet mir keine Schwierigkeiten, meine Absichten und Ziele zu verwirklichen.	❑	❑	❑	❑
In unerwarteten Situationen weiß ich immer, wie ich mich verhalten soll.	❑	❑	❑	❑
Auch bei überraschenden Ereignissen glaube ich, dass ich gut mit ihnen zurechtkommen kann.	❑	❑	❑	❑
Schwierigkeiten sehe ich gelassen entgegen, weil ich meinen Fähigkeiten immer vertrauen kann.	❑	❑	❑	❑
Was auch immer passiert, ich werde schon klarkommen.	❑	❑	❑	❑
Für jedes Problem kann ich eine Lösung finden.	❑	❑	❑	❑
Wenn eine neue Sache auf mich zukommt, weiß ich, wie ich damit umgehen kann.	❑	❑	❑	❑
Wenn ein Problem auftaucht, kann ich es aus eigener Kraft meistern.	❑	❑	❑	❑

Ich…	sehr unzu-treffend	eher unzu-treffend	weder noch	eher zu-treffend	sehr zu-treffend
...bin eher zurückhaltend, reserviert.	❑	❑	❑	❑	❑
...neige dazu, andere zu kritisieren.	❑	❑	❑	❑	❑
...erledige Aufgaben gründlich.	❑	❑	❑	❑	❑
...werde leicht deprimiert, niedergeschlagen.	❑	❑	❑	❑	❑
...bin vielseitig interessiert.	❑	❑	❑	❑	❑
...bin begeisterungsfähig und kann andere leicht mitreißen.	❑	❑	❑	❑	❑
...schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.	❑	❑	❑	❑	❑
...bin bequem, neige zur Faulheit.	❑	❑	❑	❑	❑
...bin entspannt, lasse mich durch Stress nicht aus der Ruhe bringen.	❑	❑	❑	❑	❑
...bin tiefsinnig, denke gerne über Sachen nach.	❑	❑	❑	❑	❑
...bin eher der „stille Typ“, wortkarg.	❑	❑	❑	❑	❑
...kann mich kalt und distanziert verhalten.	❑	❑	❑	❑	❑
...bin tüchtig und arbeite flott.	❑	❑	❑	❑	❑
...mache mir viele Sorgen.	❑	❑	❑	❑	❑
...habe eine aktive Vorstellungskraft, bin phantasievoll.	❑	❑	❑	❑	❑
...gehe aus mir heraus, bin gesellig.	❑	❑	❑	❑	❑
...kann mich schroff und abweisend anderen gegenüber verhalten	❑	❑	❑	❑	❑
...mache Pläne und führe sie auch durch.	❑	❑	❑	❑	❑
...werde leicht nervös und unsicher.	❑	❑	❑	❑	❑
...schätze künstlerische und ästhetische Eindrücke.	❑	❑	❑	❑	❑
...habe nur wenig künstlerisches Interesse.	❑	❑	❑	❑	❑

Aussagen	sehr falsch - - -	- -	-	+	+ +	sehr richtig + + +
Es hängt hauptsächlich von mir ab, ob sich andere Menschen nach meinen Wünschen richten oder nicht.	❑	❑	❑	❑	❑	❑
Ob ich einen Unfall habe oder nicht, hängt alleine von mir und meinem Verhalten ab.	❑	❑	❑	❑	❑	❑
Wenn ich Pläne schmiede, bin ich mir ganz sicher, dass das Geplante auch Wirklichkeit wird.	❑	❑	❑	❑	❑	❑
Ich kann mich am besten selbst durch mein Verhalten vor Krankheiten schützen.	❑	❑	❑	❑	❑	❑
Ich kann sehr viel von dem, was in meinem Leben passiert, selbst bestimmen.	❑	❑	❑	❑	❑	❑
Gewöhnlich kann ich meine Interessen selbst vertreten und erreiche dabei das, was ich will.	❑	❑	❑	❑	❑	❑
Wenn ich bekomme, was ich will, so ist das immer eine Folge meiner Anstrengung und meines persönlichen Einsatzes.	❑	❑	❑	❑	❑	❑
Mein Lebenslauf und mein Alltag werden alleine durch mein Verhalten und meine Wünsche bestimmt.	❑	❑	❑	❑	❑	❑

Herzlichen Dank für Ihre Teilnahme an der Befragung!

S: SV-A-Ex-El

M. Anreizsysteme und Aufrufe

M.1. Flyer

Abbildung M.1.: Flyer.

M.2. Foreneinträge

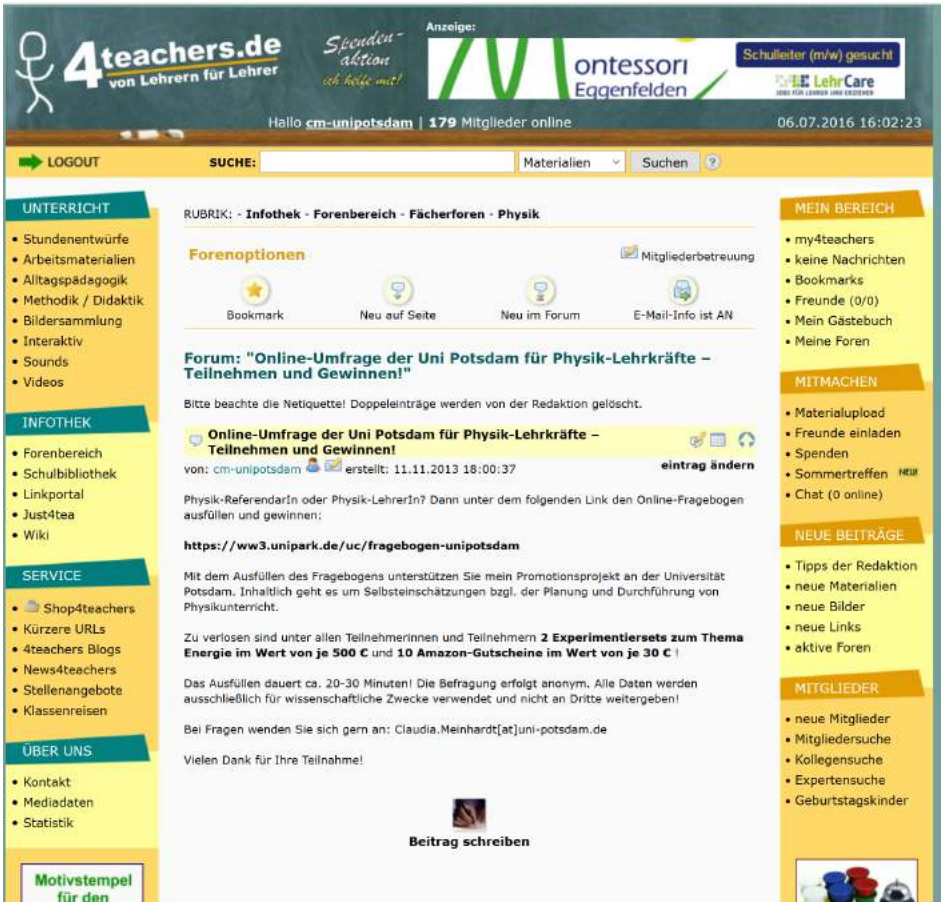


Abbildung M.2.: Foreneintrag auf www.4teachers.de.

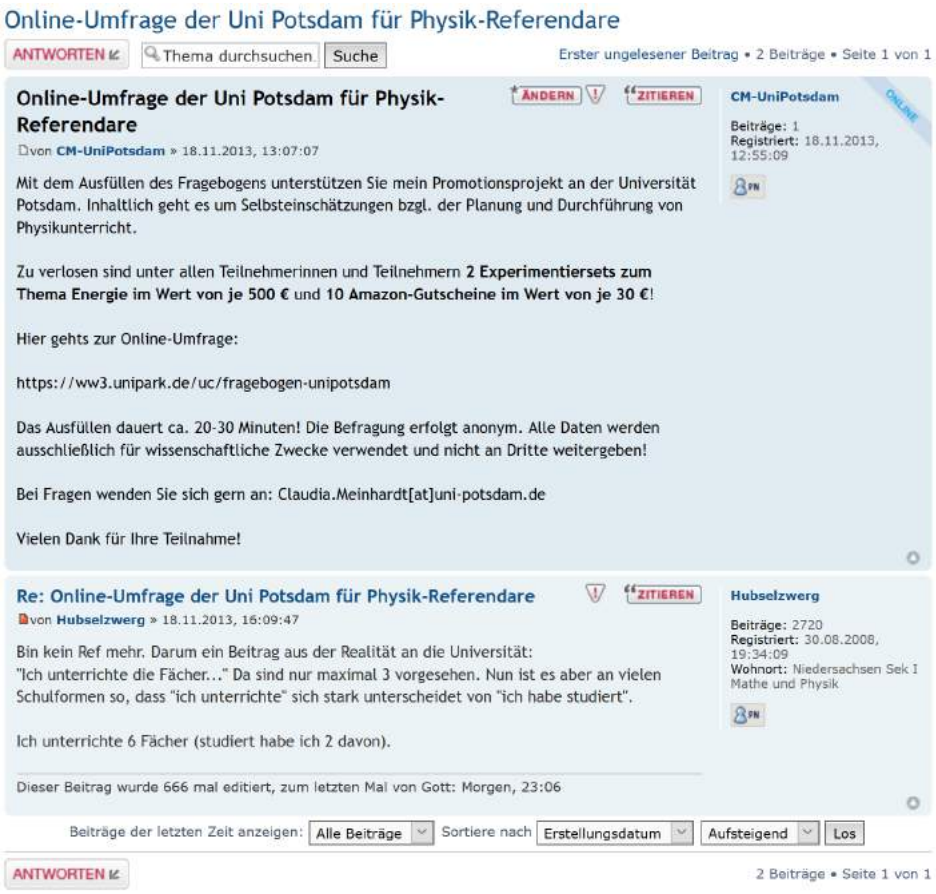


Abbildung M.3.: Foreneintrag auf www.referendar.de.

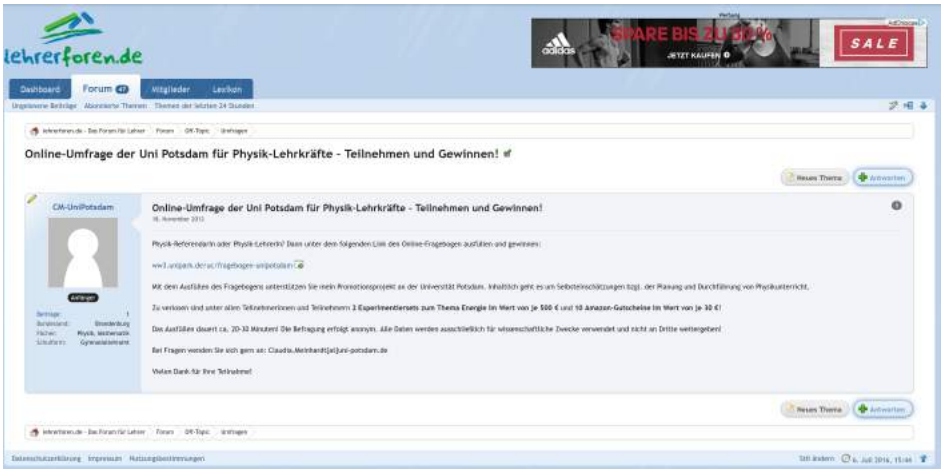


Abbildung M.4.: Foreneintrag auf www.lehrerforen.de.

N. Dokumentation Datenselektion

N.1. Dokumentation für die Stichprobe der Lehrkräfte

Aus inhaltlichen Gründen wurden 14 Fälle (vgl. Tabelle N.1), aus formalen Gründen sieben Fälle ausgeschlossen, wobei in fünf Fällen einige Skalen nicht beantwortet wurden und in zwei Fällen ein starker Verdacht des Musterkreuzens besteht (vgl. Tabelle N.2). Insgesamt werden folglich 21 Fälle ausgeschlossen.

Tabelle N.1.: Dokumentation ausgeschlossener Fälle (Lehrkräfte, inhaltlich).

Fall	**Begründung**
46	UFächer* ≠ Physik; SFächer** ≠ Physik
90	UFächer ≠ Physik; SFächer ≠ Physik; Grundschullehramt; Studium Primarstufe
91	UFächer ≠ Physik; Förderschule; Studium Sonderschule
94	UFächer ≠ Physik; Quereinstieg; Berufsschule
315	UFächer = Physik; SFächer ≠ Physik
319	UFächer = Physik; SFächer ≠ Physik; Berufsschullehramt
320	UFächer = Physik; SFächer ≠ Physik
326	UFächer ≠ Physik; SFächer ≠ Physik
376	UFächer ≠ Physik; SFächer ≠ Physik
396	UFächer = Physik; SFächer ≠ Physik
421	UFächer ≠ Physik; Quereinstieg
443	UFächer ≠ Physik; SFächer ≠ Physik
464	UFächer ≠ Physik; SFächer ≠ Physik
465	Waldorfschule, Quereinstieg

* UFächer: Unterrichtsfächer ** SFächer: Studienfächer

Tabelle N.2.: Dokumentation ausgeschlossener Fälle (Lehrkräfte, formal).

Fall	**#fW***	**nicht ausgefüllte Skalen**	**t ≈ 10 min**
37	91	SWE-SV, SWE-A, SWE-EX, A-SWE, BKI-K, K-I	✓
184	116	komplett	✓
328	15	SWE-EL	-1**
401	54	SWE-EL, A-SWE, BFI-K, K-I	-1
451	89	SWE-A, SWE-EX, SWE-EL, A-SWE, BFI-K, K-I	✓
Fall		**identisch ausgefüllte Skalen**	**t** [*sec*]
245		SWE-SV, SWE-A, SWE-EX, SWE-EL [6] (exp5=1); L-SWE [1]	1518
267		SWE-SV, SWE-A, SWE-EX, SWE-EL [6] (exd5, exp6=5)	-1

* Anzahl fehlender Werte ** Zeit durch System nicht ermittelt

N.2. Dokumentation für die Stichprobe der Referendare

Es mussten sechs Fälle aufgrund inhaltlicher Gründe von weiterführenden Analysen ausgeschlossen werden (vgl. Tabelle N.3) und aus formalen Gründen fünf Fälle, wobei in vier Fällen zu viele fehlende Werte und in einem Fall der Verdacht auf Musterkreuze ausschlaggebend waren (vgl. Tabelle N.4). Insgesamt werden folglich elf Fälle entfernt.

Tabelle N.3.: Dokumentation ausgeschlossener Fälle (Referendare, inhaltlich).

Fall	**o/p***	**Begründung**
1010	p	Grundschule, Quereinsteiger
1123	p	zu wenige demografische Angaben (Studiengang fehlt, etc.)
1166	p	Grundschule, UFach1 = Physik, HH (in HH wird kein Physik im Primarbereich)
1172	p	Studiengang: Haupt-, Real- & Grundschullehramt mit Schwerpunkt Primarstufe
1178	p	Studiengang: Haupt-, Real- & Grundschullehramt mit Schwerpunkt Primarstufe
1180	p	Studiengang: Haupt-, Real- & Grundschullehramt mit Schwerpunkt Primarstufe

* o: Online-Fragebogen, p: Papierfragebogen

Tabelle N.4.: Dokumentation ausgeschlossener Fälle (Referendare, formal).

Fall	**o/p***	**#fW****	**nicht ausgefüllte Skalen**	**t ≈ 10 min**
31	o	15	SWE-EL	-1***
306	o	48	SWE-SV, SWE-A, SWE-EL	✓
1077	p	29	A-SWE, SWE-EX	–
1168	p	38	SWE-SV, A-SWE, K-I	–
Fall	**o/p**		**identisch ausgefüllte Skalen**	**t** [*sec*]
295	o		SWE-SV, SWE-A, SWE-EX, SWE-EL [6] (exp3=5, exd4/6=5)	1317

* o: Online-Fragebogen, p: Papierfragebogen ** Anzahl fehlender Werte
*** Zeit durch System nicht ermittelt

N.3. Dokumentation für die Stichprobe der Studierenden

Aus inhaltlichen Gründen müssen 23 Fälle von weiteren Analysen ausgeschlossen werden (vgl. Tabelle N.6). Weitere zehn Fälle werden aufgrund zu vieler fehlender Werte und sechs Fälle aufgrund des Verdachts des Musterkreuzens aus dem Datensatz entfernt (vgl. Tabelle N.5). Insgesamt werden folglich 39 Fälle ausgeschlossen.

Tabelle N.5.: Dokumentation ausgeschlossener Fälle (formal, Studierende).

Fall	**#fW***	**nicht ausgefüllte Skalen**
60	17	2x SWE-SV, 9x SWE-A, 4x SWE-EX, 2x SWE-EL
101	80	SWE-SV, SWE-EL, A-SWE, BFI-K, K-I
169	32	SWE-EL , A-SWE
182	30	SWE-SV, SWE-EL
319	64	SWE-EX, PS, A-SWE, BFI-K, K-I
402	48	SWE-A, SWE-EX, K-I
433	22	7x SWE-EX, SWE-EL
519	79	SWE-SV, SWE-EX, SWE-EL , BFI-K
520	61	SWE-SV, SWE-A, SWE-EX, 5x A-SWE
521	26	7x SWE-SV, SWE-A
		identisch ausgefüllte Skalen/Muster
96	1	Wellenkreuzung z. B. SWE-EX
209	–	L-SWE[4], SWE-SV[6] (svp5=1), SWE-A[6], SWE-EX[6], SWE-EL[6], A-SWE[4]
233	–	L-SWE[3], SWE-SV[5], SWE-A[5], SWE-EX[5], A-SWE[4]
378	–	SWE-SV[5], SWE-A [5]
438	–	SWE-SV, SWE-A, SWE-EX [3 & 4 im Wechsel], PS[3]
477	–	SWE-SV[6], SWE-A [6], SWE-EX[6]

* Anzahl fehlender Werte

Tabelle N.6.: Dokumentation ausgeschlossener Fälle (Studierende, inhaltlich).

Fall	o/p*	Begründung
36	o	keine demografischen Angaben
71	p	Studiengang unbekannt
75	p	Lehramt an Sonderschulen
82	p	Lehramt an Sonderschulen
84	p	Lehramt an Berufsschulen
87	p	Lehramt an Sonderschulen
135	p	Lehramt an Berufsschulen
199	p	Studienort in Österreich
200	p	Studienort in Österreich
201	p	Studienort in Österreich
202	p	Studienort in Österreich
203	p	Studienort in Österreich
204	p	Studienort in Österreich
246	p	Lehramt an Berufsschulen
247	p	Lehramt an Berufsschulen
248	p	Lehramt an Berufsschulen
252	p	Lehramt an Primarschulen, Niedersachsen
340	p	Studiengang unbekannt
425	p	Studiengang unbekannt
451	p	Lehramt an Berufsschulen
470	p	Lehramt an Berufsschulen
478	p	Promotionsstudium, PH, Baden-Württemberg
511	p	Lehramt an Primarschulen, Bayern

* o: Online-Fragebogen, p: Papierfragebogen

O. Messinvarianz – zusätzliche Skalen

Die Messinvarianz für Subgruppen wird nur für die Skalen A-SWE und L-SWE geprüft, da nur für diese Skalen inhaltlich relevante Varianzanalysen durchgeführt werden.

O.1. Mehrgruppenvergleich: Personengruppe

Für die revidierte Skala der allgemeinen Selbstwirksamkeitserwartung (A-SWE ohne asw4) nach Schwarzer und Jerusalem (1999) kann partielle skalare Messinvarianz konstatiert werden (vgl. nachfolgende Tabelle O.1).

Tabelle O.1.: Mehrgruppenvergleich – Personengruppe (A-SWE).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
konfigural	1	179.81	81	.00	.945	.063*	.043					
metrisch	2	210.47	99	.00	.938	.060*	.109	1 vs. 2	30.74	18	.03	.007
metrisch[1]	3	202.09	97	.00	.942	.059*	.089	2 vs. 3	7.80	2	.02	.004
skalar	4	337.11	113	.00	.875	.080	.104	3 vs. 4	152.16	16	.00	.067
skalar[2]	5	301.63	111	.00	.894	.074	.098	4 vs. 5	47.79	2	.00	.019
skalar[3]	6	262.68	109	.00	.915	.067	.091	5 vs. 6	50.04	2	.00	.021
skalar[4]	7	244.97	107	.00	.923	.065	.093	6 vs. 7	18.87	2	.00	.008
skalar[5]	8	221.22	105	.00	.935	.060*	.090	7 vs. 8	27.09	2	.00	.012
skalar[6]	9	213.08	103	.00	.939	.059*	.091	8 vs. 9	8.90	2	.00	.004
								9 vs. 1	32.87	22	.07	.006

[1] Faktorladung asw10 frei geschätzt [2] Faktorladung asw10 und Intercept asw8 frei geschätzt
[3] Faktorladung asw10 und Intercept asw8 und asw7 frei geschätzt
[4] Faktorladung asw10 und Intercept asw8, asw7 und asw1 frei geschätzt
[5] Faktorladung asw10 und Intercept asw8, asw7, asw1 und asw2 frei geschätzt
[6] Faktorladung asw10 und Intercept asw8, asw7, asw1, asw2 und asw3 frei geschätzt

Für die revidierte Skala der Lehrer-Selbstwirksamkeitserwartung (L-SWE ohne lsw5 und lsw9) nach Schmitz und Schwarzer (2000) kann ebenfalls partielle skalare Messinvarianz konstatiert werden (vgl. nachfolgende Tabelle O.2).

Tabelle O.2.: Mehrgruppenvergleich – Personengruppe (L-SWE).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
konfigural	1	93.08	60	.00	.942	.042*	.040					
metrisch	2	110.43	76	.01	.939	.038*	.061	1 vs. 2	17.50	16	.35	.003
skalar	3	270.96	90	.00	.680	.080	.082	2 vs. 3	171.74	14	.00	.259
skalar[1]	4	228.45	88	.00	.752	.072	.072	3 vs. 4	47.67	2	.00	.072
skalar[2]	5	184.43	86	.00	.826	.061*	.071	4 vs. 5	29.17	2	.00	.074
skalar[3]	6	157.50	84	.00	.870	.053*	.067	5 vs. 6	24.44	2	.00	.044
skalar[4]	7	136.01	82	.00	.904	.046*	.069	6 vs. 7	18.87	2	.00	.034
skalar[5]	8	121.25	80	.00	.927	.041*	.063	7 vs. 8	15.27	2	.00	.023
								8 vs. 1	28.20	20	.11	.015

[1] Intercept lsw1 frei geschätzt [2] Intercept lsw1 und lsw6 frei geschätzt
[3] Intercept lsw1, lsw6 und lsw7 frei geschätzt
[4] Intercept lsw1, lsw6, lsw7 und lsw2 frei geschätzt
[5] Intercept lsw1, lsw6, lsw7, lsw2 und lsw8 frei geschätzt

Für die revidierte Skala der Kontrollüberzeugung-Internalität (K-I ohne ki2, ki4, ki8) nach Krampen (1991) kann skalare Messinvarianz konstatiert werden (vgl. nachfolgende Tabelle O.1).

Tabelle O.3.: Mehrgruppenvergleich – Personengruppe (K-I).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
konfigural	1	13.12	15	.00	1.000	.000*	.021					
metrisch	2	25.66	25	.00	.998	.009*	.070	1 vs. 2	12.73	10	.24	.002
skalar	3	34.54	33	.00	.996	.012*	.078	2 vs. 3	9.06	8	.34	.002
								3 vs. 1	21.86	18	.24	.004

Für die revidierten Skalen des Big-Five-Inventory (Kurzversion, BFI-K) nach Rammstedt und John (2005) kann jeweils (partielle) skalare Messinvarianz konstatiert werden (vgl. nachfolgende Tabellen O.4, O.5, O.6, O.7). Für die Verträglichkeitsskala konnte die Messinvarianz nicht überprüft werden, da kein Modellfit möglich war.

Tabelle O.4.: Mehrgruppenvergleich – Personengruppe (BFI-K(E)).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	|ΔCFI|
konfigural	1	22.66	6	.00	.979	.095	.030					
metrisch	2	38.33	14	.00	.970	.075*	.084	1 vs. 2	13.79	8	.09	.009
skalar	3	76.84	20	.00	.929	.096	.088	2 vs. 3	40.12	6	.00	.041
skalar[1]	4	52.30	18	.00	.957	.087	.092	3 vs. 4	25.98	2	.00	.028
skalar[2]	5	38.86	16	.00	.971	.068*	.083	4 vs. 5	13.97	2	.00	.014
								5 vs. 1	13.78	10	.18	.008

[1] Intercept bfi16 frei geschätzt [2] Intercept bfi18 und bfi1 frei geschätzt

Tabelle O.5.: Mehrgruppenvergleich – Personengruppe (BFI-K(G)).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	|ΔCFI|
konfigural	1	11.84	6	.07	.991	.000*	.019					
metrisch	2	16.90	14	.26	.996	.000*	.081	1 vs. 2	5.48	8	.07	.005
skalar	3	54.34	20	.00	.948	.074	.009	2 vs. 3	43.07	6	.00	.048
skalar[1]	4	21.67	18	.25	.994	.026*	.085	3 vs. 4	39.51	2	.00	.043
								4 vs. 1	9.90	10	.63	.003

[1] Intercept bfi8 frei geschätzt

Tabelle O.6.: Mehrgruppenvergleich – Personengruppe (BFI-K(N)).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	|ΔCFI|
konfigural	1	3.15	6	.79	1.010	.000*	.010					
metrisch	2	11.22	14	.67	1.010	.000*	.043	1 vs. 2	8.46	8	.39	.000
skalar	3	54.56	20	.00	.950	.075	.063	2 vs. 3	44.27	6	.00	.055
skalar[1]	4	27.93	18	.06	.980	.042*	.052	3 vs. 4	26.21	2	.00	.039
skalar[2]	5	13.65	16	.62	1.000	.000*	.046	4 vs. 5	14.55	2	.00	.016
								5 vs. 1	10.95	10	.36	.000

[1] Intercept bfi19 frei geschätzt [2] Intercept bfi19 und bfi9 frei geschätzt

Tabelle O.7.: Mehrgruppenvergleich – Personengruppe (BFI-K(O)).

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	|ΔCFI|
nkonfigural	1	2.67	6	.85	1.000	.000*	.012					
metrisch	2	17.90	14	.21	.983	.030*	.095	1 vs. 2	16.22	12	.18	.017
skalar	3	35.50	20	.02	.931	.050*	.110	2 vs. 3	18.65	6	.01	.052
skalar[1]	4	24.57	18	.14	.971	.034*	.101	3 vs. 4	11.88	2	.00	.040
								4 vs. 1	22.95	16	.12	.029

[1] Intercept bfi10 frei geschätzt

O.2. Mehrgruppenvergleich: Geschlecht

Tabelle O.8.: Mehrgruppenvergleich – Geschlecht (L-SWE)

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
Studierende: männlich (N=335), weiblich (N=188)												
konfigural	1	53.53	40	.07	.954	.036*	.040					
metrisch	2	61.62	48	.09	.954	.033*	.060	1 vs. 2	8.12	8	.42	.000
skalar	3	87.65	55	.00	.890	.048*	.062	2 vs. 3	26.40	7	.00	.064
skalar[1]	4	73.50	54	.04	.934	.037*	.058	3 vs. 4	15.10	1	.00	.044
								4 vs. 1	19.97	14	.13	.002
Referendare: männlich (N=167), weiblich (N=71)												
konfigural	1	48.67	40	.16	.937	.043*	.059					
metrisch	2	54.54	48	.24	.952	.034*	.088	1 vs. 2	6.26	8	.62	.015
skalar	3	62.22	55	.23	.947	.033*	.093	2 vs. 3	7.66	7	.36	.015
								3 vs. 1	13.74	15	.55	.010
Lehrkräfte: männlich (N=104), weiblich (N=63)												
konfigural	1	59.36	40	.02	.891	.076*	.068					
metrisch	2	65.01	48	.05	.904	.065*	.097	1 vs. 2	6.01	8	.44	.013
skalar	3	77.93	55	.02	.871	.071*	.100	2 vs. 3	12.97	7	.07	.033
								3 vs. 1	18.66	15	.23	.020

[1] Intercept von lsw7 frei geschätzt

Tabelle O.9.: Mehrgruppenvergleich – Geschlecht (A-SWE)

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
Studierende: männlich (N=335), weiblich (N=187)												
konfigural	1	135.81	54	.00	.929	.076	.048					
metrisch	2	147.02	63	.00	.927	.071	.079	1 vs. 2	10.77	9	.29	.002
skalar	3	195.61	71	.00	.892	.082	.085	2 vs. 3	52.61	8	.00	.355
skalar[1]	4	157.93	70	.00	.924	.069	.082	3 vs. 4	50.96	1	.00	.032
								4 vs. 1	20.99	16	.18	.005
Referendare: männlich (N=167), weiblich (N=70)												
konfigural	1	71.75	54	.05	.956	.053*	.056					
metrisch	2	75.82	63	.13	.968	.041*	.090	1 vs. 2	3.86	9	.92	.012
skalar	3	90.56	71	.06	.951	.048*	.097	2 vs. 3	15.05	8	.06	.017
								3 vs. 1	18.67	17	.35	.005
Lehrkräfte: männlich (N=104), weiblich (N=63)												
konfigural	1	80.62	54	.01	.918	.077*	.062					
metrisch	2	89.02	63	.02	.920	.070*	.137	1 vs. 2	8.80	9	.45	.002
skalar	3	95.67	71	.03	.924	.065*	.135	2 vs. 3	5.92	8	.66	.004
								3 vs. 1	14.97	17	.60	.006

[1] Intercept von asw6 frei geschätzt

O.3. Mehrgruppenvergleich: Studiengang und Schulform

Tabelle O.10.: Mehrgruppenvergleich – Studiengang und Schulform (L-SWE)

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
Studierende: Lehramt Gymnasium (N=385), Lehramt Sekundarstufe 1 (N=140)												
konfigural	1	63.41	40	.01	.918	.047*	.042					
metrisch	2	72.30	48	.01	.915	.044*	.058	1 vs. 2	9.07	8	.34	.003
skalar	3	81.13	55	.01	.909	.043*	.060	2 vs. 3	8.83	7	.27	.006
								3 vs. 1	17.88	15	.27	.009
Referendare: Lehramt Gymnasium (N=128), Quereinstieg (N=70)												
Modell konvergiert nicht												
Lehrpersonen: Gymnasium (N=100), Sekundarschule (N=68)												
konfigural	1	51.27	40	.11	.933	.058*	.065					
metrisch	2	58.40	48	.14	.938	.051*	.109	1 vs. 2	7.30	8	.51	.005
skalar	3	64.30	55	.18	.945	.045*	.112	2 vs. 3	5.83	7	.56	.007
								3 vs. 1	13.16	15	.59	.012

Tabelle O.11.: Mehrgruppenvergleich – Studiengang und Schulform (A-SWE)

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
Studierende: Lehramt Gymnasium (N=384), Lehramt Sekundarstufe 1 (N=140)												
konfigural	1	134.09	54	.00	.930	.075	.046					
metrisch	2	138.62	63	.00	.934	.068	.053	1 vs. 2	3.23	9	.96	.004
skalar	3	151.59	71	.00	.930	.066	.053	2 vs. 3	12.54	8	.13	.004
								3 vs. 1	15.60	17	.55	.000
Referendare: Lehramt Gymnasium (N=128), Quereinstieg (N=70)												
konfigural	1	71.60	54	.05	.943	.057*	.063					
metrisch	2	82.28	63	.05	.938	.056*	.151	1 vs. 2	10.67	9	.30	.005
skalar	3	92.48	71	.04	.931	.055*	.151	2 vs. 3	10.15	8	.25	.007
								3 vs. 1	20.76	17	.24	.012
Lehrpersonen: Gymnasium (N=100), Sekundarschule (N=68)												
konfigural	1	70.01	54	.07	.950	.060*	.059					
metrisch	2	77.20	63	.11	.956	.052*	.115	1 vs. 2	7.88	9	.55	.006
skalar	3	83.83	71	.14	.960	.046*	.117	2 vs. 3	6.36	8	.61	.004
								3 vs. 1	14.34	17	.64	.010

O.4. Mehrgruppenvergleich: Praxiserfahrung

Tabelle O.12.: Mehrgruppenvergleich – Praxiserfahrung (L-SWE)

Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
Studierende: keine institutionelle Praxiserfahrung (N=120), institutionelle Praxiserfahrung (N=385)												
konfigural	1	55.54	40	.05	.941	.039*	.042					
metrisch	2	65.16	48	.05	.935	.038*	.054	1 vs. 2	9.60	8	.30	.006
skalar	3	74.43	55	.04	.927	.037*	.059	2 vs. 3	9.26	7	.23	.008
								3 vs. 1	18.86	15	.22	.014

Tabelle O.13.: Mehrgruppenvergleich – Praxiserfahrung (A-SWE)

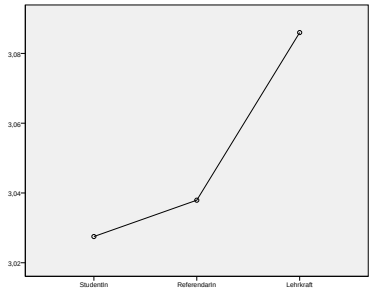
Modell	Nr.	χ^2	df	p	CFI	RMSEA	SRMR	Modell-vergleich	ΔSB-χ^2	Δdf	p	$\lvert\Delta$CFI$\rvert$
Studierende: keine institutionelle Praxiserfahrung (N=120), institutionelle Praxiserfahrung (N=385)												
konfigural	1	127.57	54	.00	.930	.074	.046					
metrisch	2	136.95	63	.00	.930	.068	.074	1 vs. 2	9.98	9	.35	.000
skalar	3	147.47	71	.00	.928	.065*	.082	2 vs. 3	9.64	8	.29	.002
								3 vs. 1	19.55	17	.30	.002

P. Mittelwertvergleiche

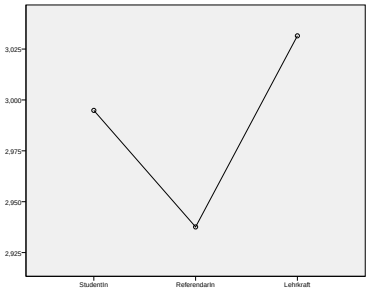
P.1. Zusätzliche Skalen

Tabelle P.1.: Mittelwerte nach Personengruppe: allgemeine Selbstwirksamkeitserwartungen (A-SWE), Lehrer-Selbstwirksamkeitserwartungen (L-SWE), Kontrollüberzeugungen/Internalität (K-I).

A-SWE



L-SWE



K-I

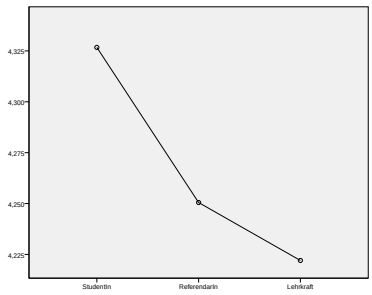
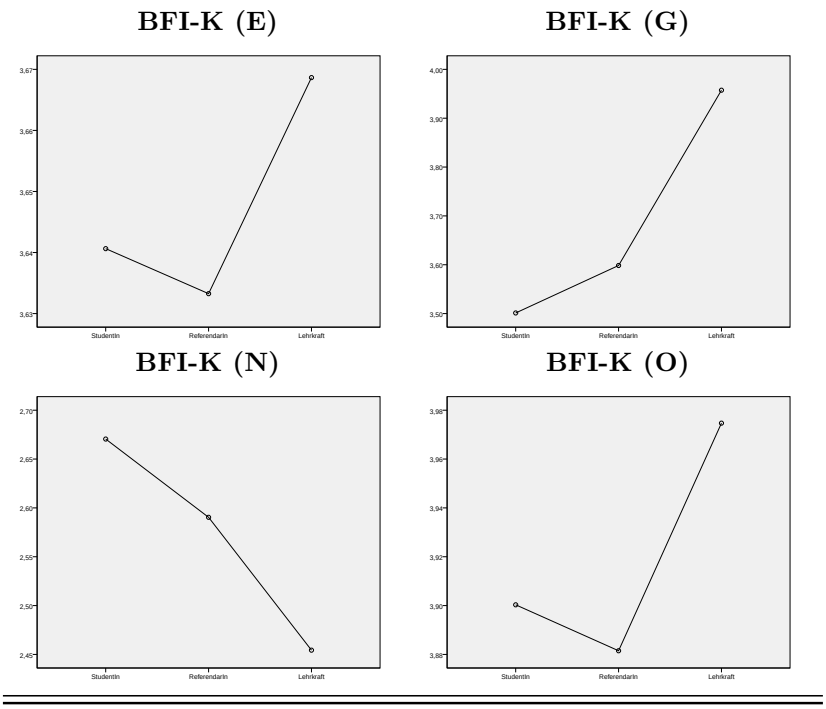


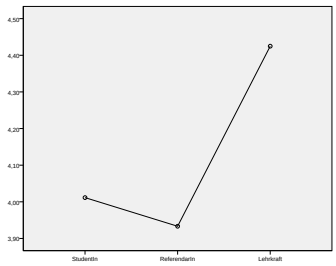
Tabelle P.2.: Mittelwerte nach Personengruppe: Persönlichkeitsvariablen Extraversion (E, Gewissenhaftigkeit (G), Neurotizismus (N), Offenheit (O)).



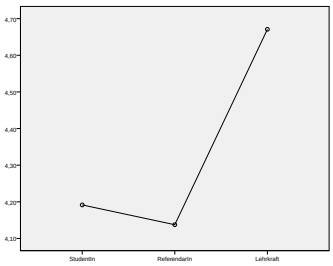
P.2. Neukonstruierte Skalen

Tabelle P.3.: Mittelwerte nach Personengruppe.

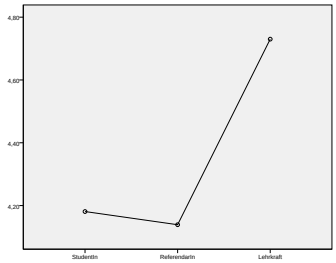
SWE-EX-P



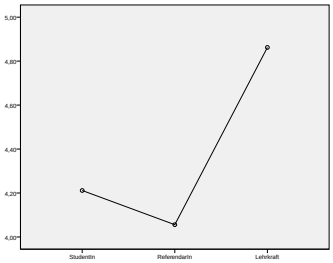
SWE-EX-D



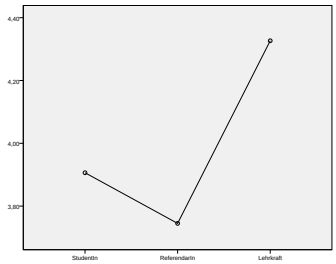
SWE-EL-P



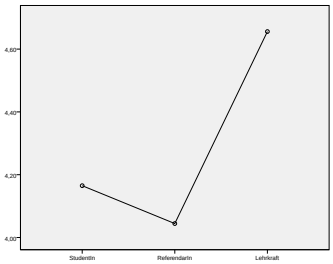
SWE-EL-D



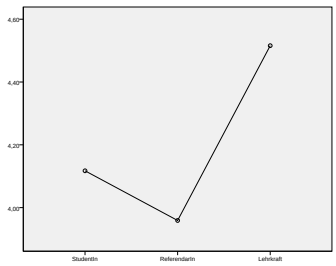
SWE-A-P



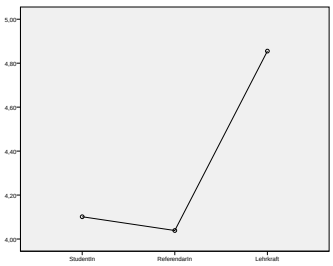
SWE-A-D



SWE-SV-P



SWE-SV-D



Literatur

AERA, APA, NCME. (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Ajzen, I. (2002). Perceived Behavioral Control, Self-Efficacy, Locus of Control, and the Theory of Planned Behavior. *Journal of Applied Social Psychology*, *32*(4), 665–683.

Allinder, R. M. (1994). The Relationship Between Efficacy and the Instructional Practices of Special Education Teachers and Consultants. *Teacher Education and Special Education*, *17*(2), 86–95.

Allinder, R. M. (1995). An Examination of the Relationship between Teacher Efficacy and Curriculum-Based Measurement and Student Achievement. *Remedial and Special Education*, *16*(4), 247–254.

Anderson, R. N., Greene, M. L. & Loewen, P. S. (1988). Relationships among Teachers' and Students' Thinking Skills, Sense of Efficacy, and Student Achievement. *Alberta Journal of Educational Research*, *34*(2), 148–165.

Arbuckle, J. L. (2010). Amos, Version 19.0. Chicago: IBM SPSS.

Arens, A. K., Trautwein, U. & Hasselhorn, M. (2011). Erfassung des Selbstkonzepts im mittleren Kindesalter: Validierung einer deutschen Version des SDQ I. *Zeitschrift fur Padagogische Psychologie*, *25*(2), 131–144.

Armor, D., Conry-Oseguera, P., Cox, M., King, N., McDonnell, L., Pascal, A., ... Zellmann, G. (1976). Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools. Santa Monica, CA: RAND.

Asendorpf, J. (2016). Persönlichkeit. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie*. https://portal.hogrefe.com/dorsch/persoenlichkeit/, abgerufen am 30.01.2016.

Asendorpf, J. B. & van Aken, M. A. G. (1993). Deutsche Versionen der Selbstkonzeptskalen von Harter. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, *25*(1), 64–86. Verfügbar unter https://www.psychologie.hu-berlin.de/de/prof/per/downloads/harterskalen.html, abgerufen am 30.07.2015.

Ashton, P. T., Webb, R. B. & Doda, N. (1983). *A Study of Teachers' Sense of Efficacy. Final Report, Executive Summary*. University of Florida. Gainsville.

Ashton, P., Buhr, D. & Crocker, L. (1984). Teachers' Sense of Efficacy: A Self- or Norm-Referenced Construct? *Florida Journal of Educational Research*, *26*(1), 29–41.

Ashton, P. & Webb, R. B. (1986). *Making a Difference: Teacher's Sense of Efficacy and Student Achievement*. New York: Longman.

Avanzi, L., Miglioretti, M., Velasco, V., Balducci, C., Vecchio, L., Fraccaroli, F. & Skaalvik, E. M. (2013). Cross-Validation of the Norwegian Teacher's Self-Efficacy Scale (NTSES). *Teaching and Teacher Education*, *31*, 69–78.

Avery, L. M. & Meyer, D. Z. (2012). Teaching Science as Science Is Practiced: Opportunities and Limits for Enhancing Preservice Elementary Teachers' Self-Efficacy for Science and Science Teaching. *School Science and Mathematics*, *112*(7), 395–409.

Baer, M., Dörr, G., Faefel, U., Kocher, M., Küster, O., Larcher, S., ... Wyss, C. (2007). Werden angehende Lehrpersonen durch das Studium kompetenter? *Unterrichtswissenschaft*, *35*(1), 15–47.

Bandura, A. (1977a). Self-Efficacy: Toward a Unifying Theory of Behavioral Change. *Psychological Review*, *84*(2), 191–215.

Bandura, A. (1977b). *Social Learning Theory*. Englewood Cliffs, N. J.: Prentice-Hall.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, N. J.: Prentice-Hall.

Bandura, A. (1989). Human Agency in Social Cognitive Theory. *The American Psychologist*, *44*(9), 1175–84.

Bandura, A. (1990). Multidimensional Scales of Perceived Academic Efficacy. Stanford University, Stanford, CA.

Bandura, A. (1997). *Self-Efficacy. The Exercise of Control*. New York: W. H. Freeman und Company.

Bandura, A. (2006). Guide for Constructing Self-Efficacy Scales. In F. Pajares & T. Urdan (Hrsg.), *Self-Efficacy Beliefs of Adolescents* (S. 307–337). Greenwich, CT: Information Age.

Bauer, J., Drechsel, B., Retelsdorf, J., Sporer, T., Rösler, L., Prenzel, M. & Möller, J. (2010). Panel zum Lehramtsstudium–PaLea: Entwicklungsverläufezukünftiger Lehrkräfte im Kontext der Reform der Lehrerbildung. *Beiträge zur Hochschulforschung*, *32*(2), 34–55.

Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U., ... Tsai, Y.-M. (2008). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV): Dokumentation der Erhebungsinstrumente*. Max-Planck-Institut für Bildungsforschung. Berlin.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520.

Baumert, J. & Kunter, M. (2011). Das Kompetenzmodell von COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle kompetenz von lehrkräften. ergebnisse des forschungsprogramms coactiv*. Münster [u.a.]: Waxmann.

Bencze, L. & Upton, L. (2006). Being Your Own Role Model for Improving Self-efficacy: An Elementary Teacher Self-Actualizes through Drama-Based Science Teaching. *Canadian Journal of Science, Mathematics and Technology Education*, *6*(3), 207–226.

Bergius, R. (2016). Selbstwertgefühl. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie*. https://portal.hogrefe.com/dorsch/selbstwertgefuehl/, abgerufen am 01.02.2016.

Berman, P., McLaughlin, M. W., Bass, G., Pauly, E. & Zellman, G. (1977). Federal Programs Supporting Educational Change, Vol VII: Factors Affecting Implementation and Continuation. Santa Monica, CA: RAND.

Betoret, F. D. (2006). Stressors, Self-Efficacy, Coping Resources, and Burnout among Secondary School Teachers in Spain. *Educational Psychology*, *26*(4), 519–539.

Betoret, F. D. (2009). Self-Efficacy, School Resources, Job Stressors and Burnout among Spanish Primary and Secondary School Teachers: a Structural Equation Approach. *Educational Psychology*, *29*(1), 45–68.

Bleicher, R. E. (2004). Revisiting the STEBI-B: Measuring Self-Efficacy in Preservice Elementary Teachers. *School Science and Mathematics*, *104*(8), 383–391.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2010a). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster [u.a.]: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2010b). *TEDS-M 2008: Professionelle Kompetenz und Lerngelegenheiten angehender Primarstufenlehrkräfte im internationalen Vergleich*. Münster [u.a.]: Waxmann.

Blömeke, S., Felbrich, A. & Müller, C. (2008). Theoretischer Rahmen und Untersuchungsdesign. In S. Blömeke, G. Kaiser & R. Lehrmann (Hrsg.), *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare* (S. 15–48). Münster [u.a.]: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare*. Münster [u.a.]: Waxmann.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. New York & London: Taylor & Francis.

Bong, M. (1998). Tests of the Internal/External Frames of Reference Model with Subject-Specific Academic Self-Efficacy and Frame-Specific Academic Self-Concepts. *Journal of Educational Psychology*, *90*(1), 102–110.

Bong, M. (2002). Predictive Utility of Subject-, Task-, and Problem-Specific Self-Efficacy Judgments for Immediate and Delayed Academic Performances. *Journal of Experimental Education*, *70*(2), 133–162.

Bong, M., Cho, C., Ahn, H. S. & Kim, H. J. (2012). Comparison of Self-Beliefs for Predicting Student Motivation and Achievement. *The Journal of Educational Research*, *105*(5), 336–352.

Bong, M. & Clark, R. E. (1999). Comparison Between Self-Concept and Self-Efficacy in Academic Motivation Research. *Educational Psychologist*, *34*(3), 139–153.

Bong, M. & Skaalvik, E. M. (2003). Academic Self-Concept and Self-Efficacy: How Different Are They Really? *Educational Psychology Review*, *15*(1), 1–40.

Boone, W. J., Staver, J. R. & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht, Heidelberg, New York, London: Springer Netherlands.

Boone, W. J., Townsend, J. S. & Staver, J. (2011). Using Rasch Theory to Guide the Practice of Survey Development and Ssurvey Data Analysis in Science Education and to Inform Science Reform Efforts: An Exemplar Utilizing STEBI Self-Efficacy Data. *Science Education*, *95*(2), 258–280.

Borowski, A., Fischer, H. E., Olszewski, J., Reinhold, P. & Riese, J. (2010). Ein Vergleich von Tests zum fachdidaktischen Wissen von Physiklehrkräften. In D. Höttecke (Hrsg.), *Entwicklung naturwissenschaftlichen Denkens zwischen Phänomen und Systematik. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Dresden 2009* (S. 377–379). Münster: LIT-Verlag.

Borowski, A., Neuhaus, B. J., Tepner, O., Wirth, J., Fischer, H. E., Leutner, D., ... Sumfleth, E. (2010). Professionswissen von Lehrkräften in den Naturwissenschaften (ProwiN). *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 341–350.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071.

Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler* (6. Auflage). Heidelberg: Springer Verlag.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. Heidelberg: Springer Medizin Verlag.

Bouffard-Bouchard, T., Parent, S. & Larivée, S. (1991). Influence of Self-Efficacy on Self-Regulation and Performance among Junior and Senior High-School Age Students. *International Journal of Behavioral Development*, *14*(2), 153–164.

Brand, B. R. & Wilkins, J. L. M. (2007). Using Self-Efficacy as a Construct for Evaluating Science and Mathematics Methods Courses. *Journal of Science Teacher Education*, *18*(2), 297–317.

Brígido, M., Borrachero, A. B., Bermejo, M. L. & Mellado, V. (2013). Prospective Primary Teachers' Self-Efficacy and Emotions in Science Teaching. *European Journal of Teacher Education*, *36*(2), 200–217.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (Kap. 4, S. 177–212). Göttingen [u.a.]: Hogrefe.

Brouwers, A. & Tomic, W. (2003). A Test of the Factorial Validity of the Teacher Efficacy Scale. *Research in Education*, *69*, 67–79.

Brovelli, D. (2014). Integrierte naturwissenschaftliche Lehrerbildung – Entwicklung professioneller Kompetenz bei Lehramtsstudierenden. *Zeitschrift für Didaktik der Naturwissenschaften*, 1–12.

Brovelli, D., Bölsterli, K., Rehm, M. & Wilhelm, M. (2013). Erfassen professioneller Kompetenzen für den naturwissenschaftlichen Unterricht: Ein Vignettentest mit authentisch komplexen Unterrichtssituationen und offenem Antwortformat. *Unterrichtswissenschaft*, *41*(1), 306–329.

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York & London: The Guilford Press.

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A. & Martin, R. (2010). The Structure of Academic Self-Concepts Revisited: The Nested Marsh/Shavelson Model. *Journal of Educational Psychology*, *102*(4), 964–981.

Brunner, M., Keller, U., Hornung, C., Reichert, M. & Martin, R. (2009). The Cross-Cultural Generalizability of a New Structural Model of Academic Self-Concepts. *Learning and Individual Differences*, *19*(4), 387–403.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3., aktualisierte und erweiterte Auflage). München: Pearson Studium.

Buss, R. R. (2010). Efficacy for Teaching Elementary Science and Mathematics Compared to Other Content. *School Science and Mathematics*, *110*(6), 290–297.

Byrne, B. M. (2012). *Structural Equation Modeling With Mplus. Basic Concepts, Applications, and Programming*. New York & London: Routledge.

Cakiroglu, J., Capa-Aydin, Y. & Woolfolk Hoy, A. (2012). Science Teaching Self Efficacy Beliefs. In B. J. Fraser, K. G. Tobin & C. J. McRobbie (Hrsg.), *Second international handbook of science education* (S. 449–462). Dordrecht, Heidelberg, London, New York: Springer.

Cappell, J. & von Aufschnaiter, C. (2011). Diagnosekompetenz angehender Physiklehrkräfte. In D. Höttecker (Hrsg.), *Naturwissenschaftliche Bildung als Beitrag zur Gestaltung partizipativer Demokratie, Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Potsdam 2010* (S. 78–80). Berlin: LIT Verlag.

Caprara, G. V., Barbaranelli, C., Steca, P. & Malone, P. S. (2006). Teachers' Self-Efficacy Beliefs as Determinants of Job Satisfaction and Students' Academic Achievement: A Study at the School Level. *Journal of School Psychology*, *44*(6), 473–490.

Carrier, S. J. (2009). The Effect of Outdoor Science Lessons with Elementary School Students on Preservice Teachers' Self-Efficacy. *Journal of Elementary Science Education*, *21*(2), 35–48.

Cartwright, T. J. & Atwood, J. (2014). Elementary Pre-Service Teachers' Response-Shift Bias: Self-Efficacy and Attitudes Toward Science. *International Journal of Science Education*, *36*(14), 2421–2437.

Cauet, E., Borowski, A. & Fischer, H. E. (2015). Testen wir relevantes Wissen? In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 136–138). Kiel: IPN.

Cauet, E., Liepertz, S., Borowski, A. & Fischer, H. E. (2015). Does it Matter What We Measure? Domain-specific Professional Knowledge of Physics Teachers. *Revue suisse des sciences de l'éducation*, *37*(3), 462–479.

Cauet, E., Liepertz, S., Kirschner, S., Borowski, A. & Fischer, H. E. (2014). Professionswissen von Physiklehrkräften und Schülerleistung. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013* (S. 141–143). Kie: IPN.

Chacón, C. T. (2005). Teachers' Perceived Efficacy Among English as a Foreign Language Teachers in Middle Schools in Venezuela. *Teaching and Teacher Education*, *21*(3), 257–272.

Chan, D. W. (2008). General, Collective, and Domain-Specific Teacher Self-Efficacy among Chinese Prospective and In-Service Teachers in Hong Kong. *Teaching and Teacher Education*, *24*(4), 1057–1069.

Chan, W.-Y., Lau, S., Nie, Y., Lim, S. & Hogan, D. (2008). Organizational and Personal Predictors of Teacher Commitment: The Mediating Role of Teacher Efficacy and Identification with School. *American Educational Research Journal*, *45*(3), 597–630.

Cheung, H. Y. (2008). Teacher Efficacy: A Comparative Study of Hong Kong and Shanghai Primary In-Service Teachers. *Australian Educational Researcher*, *35*(1), 103–123.

Cho, Y. & Shim, S. S. (2013). Predicting Teachers' Achievement Goals for Teaching: The Role of Perceived School Goal Structure and Teachers' Sense of Efficacy. *Teaching and Teacher Education*, *32*, 12–21.

Choi, N. (2005). Self-Efficacy and Self-Concept as Predictors of College Students' Academic Performance. *Psychology in the Schools*, *42*(2), 197–205.

Chong, W. H., Klassen, R. M., Huan, V. S., Wong, I. & Kates, A. D. (2010). The Relationships Among School Types, Teacher Efficacy Beliefs, and Academic Climate: Perspective from Asian Middle Schools. *The Journal of Educational Research*, *103*(3).

Chong, W. H. & Kong, C. A. (2012). Teacher Collaborative Learning and Teacher Self-Efficacy: The Case of Lesson Study. *The Journal of Experimental Education*, *80*(3), 263–283.

Christ, O. & Schlüter, E. (2012). *Strukturgleichungsmodelle mit Mplus. Eine praktische Einführung*. München: Oldenbourg Verlag.

Coladarci, T. & Fink, D. R. (1995). Correlations Among Measures of Teacher Efficacy: Are they Measuring the Same Thing? *Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco*.

Cone, N. (2009a). Community-Based Service-Learning as a Source of Personal Self-Efficacy: Preparing Preservice Elementary Teachers to Teach Science for Diversity. *School Science and Mathematics*, *109*(1), 20–30.

Cone, N. (2009b). Preservice Elementary Teachers' Self-Efficacy Beliefs About Equitable Science Teaching: Does Service Learning Make a Difference? *Journal of Elementary Science Education*, *21*(2), 25–34.

Corkett, J., Hatt, B. & Benevides, T. (2011). Student and Teacher Self-Efficacy and the Connection to Reading and Writing. *Canadian Journal of Education*, *34*(1), 65–98.

Crocker, L. (1997). Editorial: The Great Validity Debate. *Educational Measurement*, *16*(2), 4.

Cronbach, L. J. (1988). Five Perspectives on Validity Argument. In H. Wainer & H. I. Braun (Hrsg.), *Test Validity* (S. 3–17). Hillsday, New Jersey: Lawrence Erlbaum Associates.

Cronbach, L. J. & Meehl, P. E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin*, *52*(4), 281–302.

Deemer, S. A. & Minke, K. M. (1999). An Investigation of the Factor Structure of the Teacher Efficacy Scale. *Journal of Educational Research*, *93*(1), 3–10.

Dellinger, A. B., Bobbett, J. J., Olivier, D. F. & Ellett, C. D. (2008). Measuring teachers' self-efficacy beliefs: Development and use of the TEBS-Self. *Teaching and Teacher Education*, *24*, 751–766.

Dembo, M. H. & Gibson, S. (1985). Teachers' Sense of Efficacy: An Important Factor in School Improvement. *The Elementary School Journal*, *86*(2), 173.

Denzine, G. M., Cooney, J. B. & McKenzie, R. (2005). Confirmatory Factor Analysis of the Teacher Efficacy Scale for Prospective Teachers. *The British Journal of Educational Psychology*, *75*(4).

Deusinger, I. M. (1986). *Frankfurter Selbstkonzeptskalen (FSKN)*. Göttingen [u.a.]: Hogrefe.

Devos, C., Dupriez, V. & Paquay, L. (2012). Does the Social Working Environment Predict Beginning Teachers' Self-Efficacy and Feelings of Depression? *Teaching and Teacher Education*, *28*(2), 206–217.

Dicke, T., Parker, P. D., Holzberger, D., Kunina-Habenicht, O., Kunter, M. & Leutner, D. (2015). Beginning Teachers' Efficacy and Emotional Exhaustion: Latent Changes, Reciprocity, and the Influence of Professional Knowledge. *Contemporary Educational Psychology*, *41*, 62–72.

Dicke, T., Parker, P. D., Marsh, H. W., Kunter, M., Schmeck, A. & Leutner, D. (2014). Self-Efficacy in Classroom Management, Classroom Disturbances, and Emotional Exhaustion: A Moderated Mediation Analysis of Teacher Candidates. *Journal of Educational Psychology*, *106*(2), 569–583.

Dickhäuser, O., Schöne, C., Spinath, B. & Stiensmeier-Pelster, J. (2002). Die Skalen zum akademischen Selbstkonzept. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *23*(4), 393–405.

Djigić, G., Stojiljković, S. & Dosković, M. (2014). Basic Personality Dimensions and Teachers' Self-efficacy. *Procedia - Social and Behavioral Sciences*, *112*, 593–602.

DPG. (2014). *Zur fachlichen und fachdidaktischen Ausbildung für das Lehramt Physik. Eine Studie der Deutschen Physikalischen Gesellschaft e. V.* Deutsche Physikalische Gesellschaft.

Draude, M. & Wodzinski, R. (2015). Diagnosekompetenz von Physiklehrkräften beim Schülerexperimentieren. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 307–309). Kiel: IPN.

Duffin, L. C., French, B. F. & Patrick, H. (2012). The Teachers' Sense of Efficacy Scale: Confirming the Factor Structure with Beginning Pre-Service Teachers. *Teaching and Teacher Education*, *28*(6), 827–834.

Eggert, D., Reichenbach, C. & Bode, S. (2003). *Das Selbstkonzept Inventar (SKI) für Kinder im Vorschul- und Grundschulalter*. Dortmund: Verlag Modernes Lernen, Borgmann.

Emmer, E. T. & Hickman, J. (1991). Teacher Efficacy in Classroom Management and Discipline. *Educational and Psychological Measurement*, *51*(3), 755–765.

Enochs, L. G. & Riggs, I. M. (1990). Further Development of an Elementary Science Teaching Efficacy Belief Instrument: A Preservice Elementary Scale. *School Science and Mathematics*, *90*(8), 694–706.

Enochs, L. G., Smith, P. L. & Huinker, D. (2000). Establishing Factorial Validity of the Mathematics Teaching Efficacy Beliefs Instrument. *School Science and Mathematics*, *100*(4), 194–202.

Eren, A. (2009). Examining the Teacher Efficacy and Achievement Goals as Predictors of Turkish Student Teachers' Conceptions about Teaching and Learning. *Australian Journal of Teacher Education*, *34*(1), 69–87.

Evers, W., Tomic, W. & Brouwers, A. (2005). Does Equity Sensitivity Moderate the Relationship Between Self-Efficacy Beliefs and Teacher Burnout? *Representative Research in Social Psychology*, *28*, 35–46.

Ferla, J., Valcke, M. & Cai, Y. (2009). Academic Self-Efficacy and Academic Self-Concept: Reconsidering Structural Relationships. *Learning and Individual Differences*, *19*(4), 499–505.

Ferla, J., Valcke, M. & Schuyten, G. (2010). Judgments of Self-Perceived Academic Competence and Their Differential Impact on Students' Achievement Motivation, Learning Approach, and Academic Performance. *European Journal of Psychology of Education*, *25*(4), 519–536.

Fischer, H. E. & Borowski, A. (2010). Professionswissen und Fortbildung von Physiklehrern. In E. Kircher, R. Girwidz & P. Häußler (Hrsg.), *Physikdidaktik. Theorie und Praxis* (2. Auflage, S. 689–707). Berlin: Springer.

Fischer, H. E., Labudde, P., Neumann, K. & Viiri, J. (Hrsg.). (2014). *Quality of Instruction in Physics: Comparing Finland, Germany and Switzerland*. Münster [u.a.]: Waxmann.

Fives, H. & Buehl, M. M. (2010). Examining the Factor Structure of the Teachers' Sense of Efficacy Scale. *The Journal of Experimental Education*, *78*(1), 118–134.

Fives, H., Hamman, D. & Olivarez, A. (2007). Does Burnout Begin with Student-Teaching? Analyzing Efficacy, Burnout, and Support During the Student-Teaching Semester. *Teaching and Teacher Education*, *23*(6), 916–934.

Flores, I. M. (2015). Developing Preservice Teachers' Self-Efficacy Through field-based Science Teaching Practice with Elementary Students. *Research in Higher Education Journal*, *27*, 1–19.

Frey, A. (2006). Methoden und Instrumente zur Diagnose beruflicher Kompetenzen von Lehrkräften. Eine erste Standortbestimmung zu bereits publizierten Instrumenten. *Zeitschrift für Pädagogik, 51. Beiheft*, 30–46.

Frey, A. & Jung, C. (2011). Kompetenzmodelle und Standards in Lehrerbildung und Lehrerberuf. In T. Ewald, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 540–572). Münster [u.a.]: Waxmann.

Friedman, I. A. & Kass, E. (2002). Teacher Self-Efficacy: A Classroom-Organization Conceptualization. *Teaching and Teacher Education*, *18*(6), 675–686.

Gebauer, M. M. (2013). *Determinanten der Selbstwirksamkeitsüberzeugung von Lehrenden. Schulischer Berufsalltag an Gymnasien und Hauptschulen*. Wiesbaden: Springer VS.

Geiser, C. (2011). *Datenanalyse mit Mplus. Eine anwendungsorientierte Einführung*. Wiesbaden: VS Verlag für Sozialwissenschaften.

Gibson, S. & Dembo, M. H. (1984). Teacher Efficacy: A Construct Validation. *Journal of Educational Psychology*, *76*(4), 569–582.

Goddard, R. D., Hoy, W. K. & Hoy, A. W. (2004). Collective Efficacy Beliefs: Theoretical Developments, Empirical Evidence, and Future Directions. *Educational Researcher*, *33*(3), 3–13.

Göhring, A. & Schödl, A. (2015). Fachspezifische Lehrerkompetenzen (FALKO) – Teilprojekt Physik. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 364–366). Kiel: IPN.

Gramzow, Y. (2015). *Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion.* Berlin: Logos Verlag.

Gramzow, Y., Riese, J. & Reinhold, P. (2013). Modellierung fachdidaktischen Wissens angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 7–30.

Greenwald, A. G. & Farnham, S. D. (2000). Using the Implicit Association Test to Measure. *Jounal of Personality and Social Psychology*, *79*(6), 1022–1038.

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. (1998). Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Gromadecki-Thiele, U. & Priemer. (2015). Argumentationsmuster von Lehramtskandidaten in Physik. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 256–258). Kiel: IPN.

Gröschner, A. & Schmitt, C. (2010). Wirkt, was wir bewegen? – Ansätze zur Untersuchung der Qualität universitärer Praxisphasen im Kontext der Reform der Lehrerbildung. *Erziehungswissenschaft*, *21*(40), 89–97.

Gunning, A. M. & Moore Mensah, F. (2011). Preservice Elementary Teachers' Development of Self-Efficacy and Confidence to Teach Science: A Case Study. *Journal of Science Teacher Education*, *22*(2), 171–185.

Guo, Y., Dynia, J. M., Pelatti, C. Y. & Justice, L. M. (2014). Self-Efficacy of Early Childhood Special Education Teachers: Links to Classroom Quality and Children's Learning for Children with Language Impairment. *Teaching and Teacher Education*, *39*, 12–21.

Guskey, T. R. (1984). The Influence of Change in Instructional Effectiveness Upon the Affective Characteristics of Teachers. *American Educational Research Journal*, *21*(2), 245–259.

Guskey, T. R. (1981). Measurement of the Responsibility Teachers Assume For Academic Successes and Failures in the Classroom. *Journal of Teacher Education*, *32*(3), 44–51.

Guskey, T. R. (1988). Teacher Efficacy, Self-Concept, and Attitudes Toward the Implementation of Instructional Innovation. *Teaching and Teacher Education*, *4*(1), 63–69.

Guskey, T. R. (1998). Teacher Efficacy Measurement and Change. In *Paper presented at the annual meeting of the american educational research association (san diego, ca, april 13-17, 1998).*

Guskey, T. R. & Passaro, P. D. (1994). Teacher Efficacy: A Study of Construct Dimensions. *American Educational Research Journal*, *31*(3), 627–643.

Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development*, *53*(1), 87–97.

Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–171). Berlin, Heidelberg: Springer.

Hascher, T. (2006). Veränderungen im Praktikum - Veränderungen durch das Praktikum. Eine empirische Untersuchung zur Wirkung von schulpraktischen Studien in der Lehrerbildung. *Zeitschrift für Pädagogik, 51. Beiheft*, *52*, 130–148.

Head, C. F. (2012). Comparative Analyses of Mathematics Teachers' Efficacy Using Factor Analysis and the Rasch Model. *Dissertations, Theses and Capstone Projects*, *Paper 526*.

Hecht, P. (2013). Selbstwirksamkeitsüberzeugungen im Berufseinstieg von Lehrpersonen. *Unterrichtswissenschaft*, *41*(2), 108–124.

Helsper, W. (2007). Eine Antwort auf Jürgen Baumerts und Mareike Kunters Kritik am strukturtheoretischen Professionsansatz. *Zeitschrift für Erziehungswissenschaft*, *10*(4), 567–579.

Heneman, H. G., Kimball, S. & Milanowski, A. (2006). *The Teacher Sense of Efficacy Scale: Validation Evidence and Behavioral Prediction (WCER Working Paper No. 2006-7)*. Wisconsin Center for Education Research University, University of Wisconsin–Madison. verfügbar unter: http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED496313$\delimiter"026E30F$nhttp://www.eric.ed.gov/PDFS/ED496313.pdf, abgerufen am 26.02.2016. Madison.

Henson, R. K. (2001). Teacher Self-Efficacy: Substantive Implications and Measurement Dilemmas. In *Invited Keynote Address Given at the Annual Meeting of the Educational Research Exchange, January 26, 2001, Texas A&M University, College Station, Texas*.

Henson, R. K. (2002). From Adolescent Angst to Adulthood: Substantive Implications and Measurement Dilemmas in the Development of Teacher Efficacy Research. *Educational Psychologist*, *37*(3), 137–150.

Ho, I. T. & Hau, K.-T. (2004). Australian and Chinese Teacher Efficacy: Similarities and Differences in Personal Instruction, Discipline, Guidance Efficacy and Beliefs in External Determinants. *Teaching and Teacher Education*, *20*(3), 313–323.

Hoffmann, L., Häussler, P. & Lehrke, M. (1998). *Die IPN-Interessenstudie*. Kiel: IPN.

Holzberger, D., Philipp, A. & Kunter, M. (2014 April). Predicting Teachers' Instructional Behaviors: The Interplay between Self-Efficacy and Intrinsic Needs. *Contemporary Educational Psychology*, *39*(2), 100–111.

Hopf, M., Schecker, H. & Wiesner, H. (Hrsg.). (2011). *Physikdidaktik kompakt*. Hallbergmoos: Aulis Verlag.

Hoy, W. K. & Woolfolk, A. E. (1993). Teachers' Sense of Efficacy and the Organizational Health of Schools. *The Elementary School Journal*, *93*(4), 355–372.

Hu, L. & Bentler, P. M. (1999). Cutoff Criteria For Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

Huang, C. (2012). Discriminant and Incremental Validity of Self-Concept and Academic Self-Efficacy: A Meta-Analysis. *Educational Psychology*, *32*(6), 777–805.

IBM Corp. Released. (2010). IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp.

IBM Corp. Released. (2013). IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Jamil, F. M., Downer, J. T. & Pianta, R. C. (2012). Association of Pre-Service Teachers' Performance, Personality, and Beliefs with Teacher Self-Efficacy at Program Completion. *Teacher Educatio*, *39*(4), 119–138.

Jansen, M., Scherer, R. & Schroeders, U. (2015). Students ' Self-Concept and Self-Efficacy in the Sciences: Differential Relations to Antecedents and Educational Outcomes. *Contemporary Educational Psychology*, *41*, 13–24.

Jonkisz, E., Moosbrugger, H. & Brandt, H. (2012). Planung und Entwicklung von Tests und Fragebögen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 27–74). Berlin, Heidelberg: Springer.

Justice, L. M., Mashburn, A. J., Hamre, B. K. & Pianta, R. C. (2008). Quality of Language and Literacy Instruction in Preschool Classrooms Serving At-Risk Pupils. *Early Childhood Research Quarterly*, *23*(1), 51–68.

Kaiser, F. G. (2016). Einstellung; Struktur-, Prozess- und Verhaltensaspekte. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie*. https://portal.hogrefe.com/dorsch/einstellung-struktur-prozess-und-verhaltensaspekte/, abgerufen am 31.01.2016.

Kane, M. T. (1992). An Argument-Based Approach to Validity. *Quantitative Methods in Psychology*, *112*(3), 527–535.

Kane, M. T. (2001). Current Concerns in Validity Theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Karimi, M. N. (2011). The Effects of Professional Development Initiatives on EFL Teachers' Degree of Self Efficacy. *Australian Journal of Teacher Education*, *36*(6), 50–62.

Kastens, C. (2009). *Handlungsmotivation und Unterrichtsqualität – Erfassung affektiv-motivationaler Prädiktoren des Unterrichtshandelns von Lehrkräften* (Diss., Christian-Albrechts-Universität).

Kauper, T., Retelsdorf, J., Bauer, J., Rösler, L., Möller, J., Prenzel, M. & Drechsel, B. (2012). PaLea – Panel zum Lehramtsstudium. Skalendokumentation und Häufigkeitsauszählungen des BMBF-Projektes, 1. Welle; Herbst 2009. Kiel: IPN - Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik.

Kazempour, M. (2013). The Interrelationship of Science Experiences, Beliefs, Attitudes, and Self-Efficacy: A Case Study of a Pre-Service Teacher with Positive Science Attitude and High Science Teaching Self-Efficacy. *European Journal of Science and Mathematics Education*, *1*(3), 106–124.

Kelava, A. & Moosbrugger, H. (2012). Deskriptivstatistische Evaluation von Items (Itemanalyse) und Testwertverteilungen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Bd. 3, S. 75–102). Berlin, Heidelberg: Springer.

Kelava, A. & Schermelleh-Engel, K. (2012). Latent-State-Trait-Theorie (LST-Theorie). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 363–381). Berlin, Heidelberg: Springer.

Keller-Schneider, M. (2008). *Evaluation der Berufseinführung des Kantons St. Gallen, 2006-2008, Schlussbericht*. Pädagogische Hochschule. Zürich.

Kircher, E., Girwidz, R. & Häußler, P. (Hrsg.). (2010). *Physikdidaktik. Theorie und Praxis* (2. Auflage). Berlin: Springer.

Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Berlin: Logos Verlag.

Klassen, R. M., Bong, M., Usher, E. L., Chong, W. H., Huan, V. S., Wong, I. Y. F. & Georgiou, T. (2009 Januar). Exploring the Validity of a Teachers' Self-Efficacy Scale in Five Countries. *Contemporary Educational Psychology*, *34*(1), 67–76.

Klassen, R. M. & Chiu, M. M. (2010). Effects on Teachers' Self-Efficacy and Job Satisfaction: Teacher Gender, Years of Experience, and Job Stress. *Journal of Educational Psychology*, *102*(3), 741–756.

Klassen, R. M. & Tze, V. M. C. (2014). Teachers' Self-Efficacy, Personality, and Teaching Effectiveness: A Meta-Analysis. *Educational Research Review*, *12*, 59–76.

Klassen, R. M., Tze, V. M. C., Betts, S. M. & Gordon, K. A. (2011). Teacher Efficacy Research 1998–2009: Signs of Progress or Unfulfilled Promise? *Educational Psychology Review*, *23*(1), 21–43.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., ... Vollmer, H. J. (2007). Zur Entwicklung nationaler Bildungsstandards. Eine Expertise. Bonn, Berlin: Bundesministerium für Bildung und Forschung.

Klieme, E. & Hartig, J. (2007). Kompetenzdiagnostik in den Sozialwissenschaften und im erziehungswissenschaftlichen Diskurs. *Zeitschrift für Erziehungswissenschaft*, *10*(Sonderheft 8), 11–29.

Klös, H., Riese, J. & Reinhold, P. (2015). Entwicklung professioneller Handlungskompetenz im Praxissemester Physik. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 483–485). Kiel: IPN.

KMK. (2004). *Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

KMK. (2010). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung. Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 16.09.2010*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.

Knoblauch, D. & Chase, M. a. (2015). Rural, Suburban, and Urban Schools: The Impact of School Setting on the Efficacy Beliefs and Attributions of Student Teachers. *Teaching and Teacher Education*, *45*, 104–114.

Knoblauch, D. & Woolfolk Hoy, A. (2008). „Maybe I Can Teach Those Kids." The Influence of Contextual Factors on Student Teachers' Efficacy Beliefs. *Teaching and Teacher Education*, *24*(1), 166–179.

Kocher, M. (2014). *Selbstwirksamkeit und Unterrichtsqualität. Unterricht und Persönlichkeitsaspekte von Lehrpersonen im Berufsübergang*. Münster, New York: Waxmann.

König, J. & Seifert, A. (Hrsg.). (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen: Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung*. Münster [u.a.]: Waxmann.

Kopp, B. (2009). Inklusive Überzeugung und Selbstwirksamkeit im Umgang mit Heterogenität – Wie denken Studierende des Lehramts für Grundschulen? *Empirische Sonderpädagogik*, *1*(1), 5–25.

Korneck, F., Kohlenberger, M., Oettinghaus, L., Kunter, M. & Lamprecht, J. (2013). Lehrerüberzeugungen und Unterrichtshandeln im Fach Physik. *PhyDid B - Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung*, 1–9.

Korneck, F., Kunter, M., Oettinghaus, L., Lamprecht, J. & Sach, M. (2014). Analyse von Unterrichtshandeln in komplexitätsreduzierten Sequenzen. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013* (S. 138–140). Kiel: IPN.

Korneck, F. & Lamprecht, J. (2010). Quer- und SeiteneinsteigerInnen in den Lehrberuf – eine Analyse exemplarish für das Fach Physik. *journal für lehrerinnen- und lehrerbildung*, *3*, 8–21.

Korneck, F., Lamprecht, J., Wodzinski, R. & Schecker, H. (2010). *Quereinsteiger in das Lehramt Physik. Lage und Perspektiven der Physiklehrerausbildung in Deutschland*. Bad Honnef: DPG.

Krampen, G. (1991). *Fragebogen zu Kompetenz- und Kontrollüberzeugungen (FKK)*. Göttingen, Toronto, Zürich: Hogrefe.

Krauss, S. (2011). Das Experten-Paradigma in der Forschung zum Lehrerberuf. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 171–191). Münster [u.a.]: Waxmann.

Krauss, S., Neubrand, M., Blum, W., Baumert, J., Brunner, M., Kunter, M. & Jordan, A. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematikdidaktik*, *29*(3/4), 223–258.

Krofta, H. & Nordmeier, V. (2014). Bewirken Praxisseminare im Lehr-Lern-Labor Änderungen der Lehrerselbstwirksamkeitserwartung bei Studierenden? *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung*, 1–12.

Kröger, J., Neumann, K. & Petersen, S. (2013). Messung professioneller Kompetenz im Fach Physik. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen.*

Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012 (S. 533–535). Kiel: IPN.

Kröger, J., Neumann, K. & Petersen, S. (2015). Struktur und Entwicklung des Professionswissens angehender Physiklehrkräfte. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 106–108). Kiel: IPN.

Krüger, M., Korneck, F., Oettinghaus, L. & Kunter, M. (2015). Bewertung von Unterrichtsqualität. Schulart- und lehrkraftspezifische Unterschiede. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Jahrestagung in Bremen 2014* (S. 127–129). Kiel: IPN.

Kulgemeyer, C., Borowski, A., Fischer, H., Gramzow, Y., Reinhold, P., Riese, J., … Walzer, M. (2012). ProfiLe-P – Professionswissen in der Lehramtsausbildung Physik - Vorstellung eines Forschungsverbundes -. In V. Nordmeier & H. Grötzebauch (Hrsg.), *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung* (S. 1–6). Mainz.

Kulgemeyer, C. & Tomczyszyn, E. (2015). Physik erklären – Messung der Erklärensfähigkeit angehender Physiklehrkräfte in einer simulierten Unterrichtssituation. *Zeitschrift für Didaktik der Naturwissenschaften*.

Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S. & Neubrand, M. (Hrsg.). (2011). *Professionelle Kompetenz von Lehrkräften: Ergebnisse des Forschungsprogramms COACTIV*. Münster [u.a.]: Waxmann.

Kunter, M. & Klusmann, U. (2010). Kompetenzmessung bei Lehrkräften - Methodische Herausforderungen. *Unterrichtswissenschaft*, *38*(1), 68–86.

Kushner, S. N. (1993). Teacher Efficacy and Preservice Teachers: A Construct Validation. In *Paper presented at the annual meeting of the eastern educational research association (clearwater beach, fl, february 17-22, 1993)*.

Lamote, C. & Engels, N. (2010). The Development of Student Teachers' Professional Identity. *European Journal of Teacher Education*, *33*(1), 3–18.

Lange, K., Kleickmann, T., Tröbst, S. & Möller, K. (2012). Fachdidaktisches Wissen von Lehrkräften und multiple Ziele im naturwissenschaftlichen Sachunterricht. *Zeitschrift fur Erziehungswissenschaft*, *15*(1), 55–75.

Lau, I. C., Yeung, A. S., Jin, P. & Low, R. (1999). Toward a Hierarchical, Multidimensional English Self-Concept. *Journal of Educational Psychology*, *91*(4), 747–755.

Lee, J. (2009). Universals and Specifics of Math Self-Concept, Math Self-Efficacy, and Math Anxiety across 41 PISA 2003 Participating Countries. *Learning and Individual Differences*, *19*(3), 355–365.

Lee, M. H. & Tsai, C. C. (2010). Exploring Teachers' Perceived Self Efficacy and Technological Pedagogical Content Knowledge with Respect to Educational Use of the World wide Web. *Instructional Science*, *38*(1), 1–21.

Lee, V. E., Dedrick, R. F. & Smith, J. B. (1991). The Effect of the Social Organization of Schools on Teachers' Efficacy and Satisfaction. *Sociology of Education*, *64*(3), 190–208.

Lee, Y. & Lee, J. (2014). Enhancing Pre-Service Teachers' Self-Efficacy Beliefs for Technology Integration Through Lesson Planning Practice. *Computers & Education*, *73*, 121–128.

Lefcourt, H. M. (Hrsg.). (1981). *Research with the Locus of Control Construct: Vol. 1. Assessment Methods*. San Diego, C. A.: Academic Press.

Lefcourt, H. M. (1982). *Locus of Control: Current Trends in Theory and Research*. Hillsday, N. J.: Lawrence Erlbaum Associates.

Levenson, H. (1973). Perceived parental antecedents of internal, powerful others, and chance locus of control orientations. *Developmental Psychology*, *9*, 260–265.

Li, H. (2003). The Resolution of Some Paradoxes Related to Reliability and Validity. *Journal of Educational and Behavioral Statistics*, *28*(2), 89–95.

Lin, H.-L. & Gorrell, J. (1998). Pre-Service Teachers' Efficacy Beliefs in Taiwan. *Journal of Research and Development in Education*, *32*(1), 17–25.

Linacre, J. M. (2014). *A User's Guide to Winsteps/Ministeps. Rasch-Model Computer Programs. Program Manual 3.81.0*.

Linn, R. L. (1997). Evaluating the Validity of Assessments: The Consequences of Use. *Educational Measurement: Issues and Practice*, *16*(2), 14–16.

Linninger, C., Kunina-Habenicht, O., Emmenlauer, S., Dicke, T., Schulze-Stocker, F., Leutner, D., ... Kunter, M. (2015). Assessing Teachers' Educational Knowledge: Construct Specification and Validation Using Mixed Methods. *Zeitschrift fur Entwicklungspsychologie und Padagogische Psychologie*, *47*(2), 72–83.

Lipowsky, F. (2010). Lernen im Beruf – Empirische Befunde zur Wirksamkeit von Lehrerfortbildung. In F. H. Müller, A. Eichenberger, M. Lüders & J. Mayr (Hrsg.), *Lehrerinnen und Lehrer lernen – Konzepte und Befunde zur Lehrerfortbildung*. (S. 51–70). Münster [u.a.]: Waxmann.

Lissitz, R. W. & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher*, *36*(8), 437–448.

Liu, C. J., Jack, B. M. & Chiu, H. L. (2007). Taiwan Elementary Teachers' Views of Science teaching Self-Efficacy and Outcome Expectations. *International Journal of Science and Mathematics Education*, *6*(1), 19–35.

Lumpe, A., Czerniak, C., Haney, J. & Beltyukova, S. (2012). Beliefs about Teaching Science: The Relationship Between Elementary Teachers' Participation in Professional Development and Student Achievement. *International Journal of Science Education*, *34*(2), 153–166.

Maag Merki, K. & Werner, S. (2011). Erfassung und Bewertung professioneller Kompetenz von Lehrpersonen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 573–591). Münster [u.a.]: Waxmann.

Maddux, J. E. (2002). Self-Efficacy: The Power of Believing You Can. In C. R. Snyder & S. J. Lopez (Hrsg.), *Handbook of positive psychology* (S. 277–287). Oxford [u.a.]: Oxford University Press.

Markus, H. & Kunda, Z. (1986). Stability and Malleability of the Self-Concept. *Journal of Personality and Social Psychology*, *51*(4), 858–866.

Marsh, H. W. (1986). Verbal and Math Self-Concepts: An Internal/External Frame of Reference Model. *American Educational Research Journal*, *23*(1), 129–149.

Marsh, H. W. (1990a). A Multidimensional, Hierarchical Model of Self-Concept: Theoretical and Empirical Justification. *Educational Psychology Review*, *2*(2), 77–172.

Marsh, H. W. (1990b). Causal Ordering of Academic Self-Concept and Academic Achievement: A Multiwave, Longitudinal Panel Analysis. *Journal of Educational Psychology*, *82*(4), 646–656.

Marsh, H. W. (1990c). The Structure of Academic Self-Concept: The Marsh/Shavelson Model. *Journal of Educational Psychology*, *82*(4), 623–636.

Marsh, H. W., Byrne, B. M. & Shavelson, R. J. (1988). A Multifaceted Academic Self-Concept: Its Hierarchical Structure and Its Relation to Academic Achievement. *Journal of Educational Psychology*, *80*(3), 366–380.

Marsh, H. W., Dowson, M., Pietsch, J. & Walker, R. (2004). Why Multicollinearity Matters: A Reexamination of Relations Between Self-Efficacy, Self-Concept, and Achievement. *Journal of Educational Psychology*, *96*(3), 518–522.

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G. & Heubeck, B. G. (2005). A Short Version of the Self Description Questionnaire II: Operationalizing Criteria for Short-Form Evaluation With New Applications of Confirmatory Factor Analyses. *Psychological Assessment*, *17*(1), 81–102.

Marsh, H. W. & O'Neill, R. (1984). Self Description Questionnaire III: The Construct Validity of Multidimensional Self-Concept Ratings by Late Adolescents. *Journal of Educational Measurement*, *21*(2), 153–174.

Marsh, H. W., Walker, R. & Debus, R. (1991). Subject-Specific Components of Academic Self-Concept and Self-Efficacy. *Contemporary Educational Psychology*, *16*(4), 331–345.

Mayr, J. (2011). Der Persönlichkeitsansatz in der Lehrerforschung. Konzepte, Befunde, Folgerungen. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 125–148). Münster [u.a.]: Waxmann.

McKinnon, M., Moussa-Inaty, J. & Barza, L. (2014). Science Teaching Self-Efficacy of Culturally Foreign Teachers: A Baseline Study in Abu Dhabi. *International Journal of Educational Research*, *66*, 78–89.

Mehrens, W. A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice*, *16*(2), 16–18.

Meinhardt, C. (2011). *Pilotierung eines Instruments zur domänenspezifischen Selbstwirksamkeitserwartung von Physik-Lehramtsstudierenden, unveröffentlichte Masterarbeit*, Universität Potsdam.

Meinhardt, C., Krey, O. & Rabe, T. (2016a). Measurement Invariance as a Validity Argument of a Physics Teaching Efficacy Questionnaire. In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto & K. Hahl (Hrsg.), *Electronic Proceedings of the ESERA 2015 Conference. Science Education Research: Engaging Learners for a Sustainable Future, Part 13 (co-eds. M. Evagorou & M. Michelini)* (S. 2135–2140). Helsinki, Finland: University of Helsinki.

Meinhardt, C., Rabe, T. & Krey, O. (2014a). Pre-service teachers' self-efficacy beliefs for teaching physics. In C. P. Constantinou, N. Papadouris & A. Hadjigeorgiou (Hrsg.), *E-Book Proceedings of the ESERA 2013 Conference: Science Education Research For Evidence-based Teaching and Coherence in Learning. Part 13 (co-eds. L. Avraamidou*

& M. Michelini) (S. 244–250). Nicosia, Cyprus: European Science Education Research Association.

Meinhardt, C., Rabe, T. & Krey, O. (2014b). Qualitative Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen (SWE) in physikdidaktischen Handlungsfeldern. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in München 2013* (S. 558–560). Kiel: IPN.

Meinhardt, C., Rabe, T. & Krey, O. (2015). Quantitative Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern – Erste Ergebnisse. In S. Bernholt (Hrsg.), *Heterogenität und Diversität - Vielfalt der Voraussetzungen im naturwissenschaftlichen Unterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Bremen 2014* (S. 283–285). Kiel: IPN.

Meinhardt, C., Rabe, T. & Krey, O. (2016b). Selbstwirksamkeitserwartungen (angehender) Physiklehrkräfte: Ausgewählte Ergebnisse einer Validierungsstudie. In C. Maurer (Hrsg.), *Authentizität und Lernen - das Fach in der Fachdidaktik. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Berlin 2015* (S. 337–339). Universität Regensburg.

Meinhardt, C., Rabe, T. & Krey, O. (2016c). *Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern. Skalendokumentation.* Didaktik der Physik, Martin-Luther-Universität Halle-Wittenberg. Verfügbar unter: http://www.pedocs.de/frontdoor.php?source_opus=11818, abgerufen am 26.02.2016.

Meristo, M. & Eisenschmidt, E. (2014). Novice Teachers' Perceptions of School Climate and Self-Efficacy. *International Journal of Educational Research*, *67*, 1–10.

Messick, S. (1995). Validity of Psychological Assessment. *American Psychologist*, *50*(9), 741–749.

Midgley, C., Feldlaufer, H. & Eccles, J. S. (1989). Change in Teacher Efficacy and Student Self- and Task-Related Beliefs in Mathematics During the Transition to Junior High School. *Journal of Educational Psychology*, *81*(2), 247–258.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., … Urdan, T. (2000). *Manual for the Patterns of Adaptive Learning Scales.* University of Michigan.

Mikelskis, H. F. (Hrsg.). (2006). *Physikdidaktik. Praxishandbuch für die Sekundarstufe I und II.* Berlin: Cornelsen Scriptor.

Mislevy, R. J. (2004). Can There Be Reliability without „Reliability?" *Journal of Educational and Behavioral Statistics*, *29*(2), 241–244.

Moè, A., Pazzaglia, F. & Ronconi, L. (2010). When Being Able is not Enough. The Combined Value of Positive Affect and Self-Efficacy for Job Satisfaction in Teaching. *Teaching and Teacher Education*, *26*(5), 1145–1153.

Mohamadi, F. S. & Asadzadeh, H. (2012). Testing the Mediating Role of Teachers' Self-Efficacy Beliefs in the Relationship Between Sources of Efficacy Information and Students Achievement. *Asia Pacific Education Review*, *13*(3), 427–433.

Möller, J. & Köller, O. (2004). Die Genese akademischer Selbstkonzepte: *Psychologische Rundschau*, *55*(1), 19–27.

Möller, J., Pohlmann, B., Köller, O. & Marsh, H. W. (2009). A Meta-Analytic Path Analysis of the Internal/External Frame of Reference Model of Academic Achievement and Academic Self-Concept. *Review of Educational Research*, *79*(3), 1129–1167.

Möller, J. & Trautwein, U. (2015). Selbstkonzept. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (2., vollst. überarb. Auflage, Kap. 8, S. 329–342). Berlin, Heidelberg: Springer.

Mone, M. A., Baker, D. D. & Jeffries, F. (1995). Predictive Validity and Time Dependency of Self-Efficacy, Self-Esteem, Personal Goals, and Academic Performance. *Educational and Psychological Measurement*, *55*(5), 716–727.

Moosbrugger, H. (2012a). Item-Response-Theory (IRT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und fragebogenkonstruktion* (S. 227–274). Berlin Heidelberg: Springer.

Moosbrugger, H. (2012b). Klassische Testtheorie (KTT). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 103–117). Berlin, Heidelberg: Springer.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–25). Berlin, Heidelberg: Springer.

Moseley, C., Bilica, K., Wandless, A. & Gdovin, R. (2014). Exploring the Relationship Between Teaching Efficacy and Cultural Efficacy of Novice Science Teachers in High-Needs Schools. *School Science and Mathematics*, *114*(7), 315–325.

Moss, P. A. (1994). Can There Be Validity Without Reliability? *Educational Researcher*, *23*(2), 5–12.

Moss, P. A. (1995). Themes and Variations in Validity Theory. *Educational Measurement: Issues and Practice*, *14*(2), 5–13.

Moss, P. A. (2004). The Meaning and Consequences of „Reliability“. *Journal of Educational and Behavioral Statistics*, *29*(2), 245–249.

Moulding, L. R., Stewart, P. W. & Dunmeyer, M. L. (2014). Pre-Service Teachers' Sense of Efficacy: Relationship to Academic Ability, Student Teaching Placement Characteristics, and Mentor Support. *Teaching and Teacher Education*, *41*, 60–66.

Muijs, D. & Reynolds, D. (2002). Teachers' Beliefs and Behaviors: What Really Matters? *Journal of Classroom Interaction*, *37*(2), 3–15.

Mulholland, J., Dorman, J. P. & Odgers, B. M. (2004). Assessment of Science Teacher Efficacy of Preservice Teachers in an Australian University. *Journal of Science Teacher Education*, *15*(4), 313–331.

Mummendey, H. D. & Grau, I. (2014). *Die Fragebogen-Methode: Grundlagen und Anwendung in Persönlichkeits-, Einstellungs- und Selbstkonzeptforschung*. Göttingen [u.a.]: Hogrefe.

Muthén, L. K. & Muthén, B. O. (2012). *Mplus. User's Guide*. Los Angeles, CA: Muthén & Muthén.

Newmann, F. M., Rutter, R. A. & Smith, M. S. (1989). Organizational Factors that Affect School Sense of Efficacy, Community, and Expectations. *Sociology of Education*, *62*(4), 221–238.

Niebert, K. & Gropengießer, H. (2014). Leitfadengestützte Interviews. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen forschung* (S. 121–132). Berlin, Heidelberg: Springer Spektrum.

OECD. (2005). *PISA 2003. Technical Report.* Organisation for Economic Co-Operation und Development. Paris.

Oettinghaus, L. (2015). *Lehrerüberzeugungen und physikbezogenes Professionswissen. Vergleich von Absolventinnen und Absolventen verschiedener Ausbildungswege im Physikreferendariat* (Diss., Johann Wolfgang Goethe Universität Frankfurt am Main).

Oettinghaus, L., Lamprecht, J. & Korneck, F. (2014). Analyse der professionellen Kompetenz von Referendaren. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013* (S. 135–137). Kiel: IPN.

Ohle, A., Fischer, H. E. & Kauertz, A. (2011). Der Einfluss des physikalischen Fachwissens von Primarstufenlehrkräften auf Unterrichtsgestaltung und Schülerleistung. *Zeitschrift für Didaktik der Naturwissenschaften*, *17*, 357–389.

Olszewski, J. (2010). *The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes.* Berlin: Logos Verlag.

O'Neill, S. & Stephenson, J. (2012). Exploring Australian Pre-Service Teachers Sense of Efficacy, Its Sources, and Some Possible Influences. *Teaching and Teacher Education*, *28*(4), 535–545.

Oser, F. (1997). Standards in der Lehrerbildung. Teil 1: Berufliche Kompetenzen, die hohen Qualitätsmerkmalen entsprechen. *Beiträge zur Lehrerbildung*, *15*(1), 26–37.

Oser, F., Curcio, G.-P. & Düggeli, A. (2007). Kompetenzmessung in der Lehrerbildung als Notwendigkeit - Fragen und Zugänge. *Beiträge zur Lehrerbildung*, *25*(1), 14–26.

Oser, F. & Oelkers, J. (2001). *Die Wirksamkeit der Lehrerbildungssysteme: Von der Allrounderbildung zur Ausbildung professioneller Standards.* Chur: Rüegger.

Pajares, F. (1996). Self-Efficacy Beliefs in Academic Settings. *Review of Educational Research*, *66*(4), 543–578.

Pajares, F. & Miller, M. D. (1994). Role of Self-Efficacy and Self-Concept Beliefs in Mathematical Problem Solving: A Path Analysis. *Journal of Educational Psychology*, *86*(2), 193–203.

Pajares, F., Miller, M. D. & Johnson, M. J. (1999). Gender Differences in Writing Self-Beliefs of Elementary School Students. *Journal of Educational Psychology*, *91*(1), 50–61.

Palmer, D. (2011 Juli). Sources of Efficacy Information in an Inservice Program for Elementary Teachers. *Science Education*, *95*(4), 577–600.

Parker, P. D., Marsh, H. W., Ciarrochi, J., Marshall, S. & Abduljabbar, A. S. (2014). Juxtaposing Math Self-Efficacy and Self-Concept as Predictors of Long-Term Achievement Outcomes. *Educational Psychology*, *34*(1), 29–48.

Pelham, B. W. & Swann, W. B. (1989). From Self-Conceptions to Self-Worth: On the Sources and Structure of Global Self-Esteem. *Journal of Personality and Social Psychology*, *57*(4), 672–680.

Pendergast, D., Garvis, S. & Keogh, J. (2011). Pre-Service Student-Teacher Self-efficacy Beliefs: An Insight Into the Making of Teachers. *Australian Journal of Teacher Education*, *36*(12), 45–58.

Pietsch, J., Walker, R. & Chapman, E. (2003). The Relationship Among Self-Concept, Self-Efficacy, and Performance in Mathematics During Secondary School. *Journal of Educational Psychology*, *95*(3), 589–603.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *The Journal of Applied Psychology*, *88*(5), 879–903.

Pohlmann, B. & Möller, J. (2010). Fragebogen zur Erfassung der Motivation für die Wahl des Lehramtsstudiums (FEMOLA). *Zeitschrift für Pädagogische Psychologie*, *24*(1), 73–84.

Popham, J. W. (1997). Consequential Validity: Right Concern – Wrong Concept. *Educational Measurement: Issues and Practice*, *16*(2), 9–13.

Pötschke, M. (2010). Datengewinnung und Datenaufbereitung. In C. Wolf & H. Best (Hrsg.), *Handbuch der sozialwissenschaftlichen datenanalyse* (S. 41–64). Wiesbaden: VS Verlag für Sozialwissenschaften.

Poulou, M. (2007). Personal Teaching Efficacy and Its Sources: Student Teachers' Perceptions. *Educational Psychology*, *27*(2), 191–218.

Powell-Moman, A. D. & Brown-Schild, V. B. (2011). The Influence of a Two-Year Professional Development Institute on Teacher Self-Efficacy and Use of Inquiry-Based Instruction. *Science Educator*, *20*(2), 47–53.

Pruski, L. A., Blanco, S. L., Riggs, R. A., Grimes, K. K., Fordtran, C. W., Barbola, G. M., ... Lichtenstein, M. J. (2013). Construct Validation of the Self-Efficacy Teaching and Knowledge Instrument for Science Teachers-Revised (SETAKIST-R): Lessons Learned. *Journal of Science Teacher Education*, *24*(7), 1133–1156.

Puchner, L. D. & Taylor, A. R. (2006). Lesson Study, Collaboration and Teacher Efficacy: Stories From Two School-Based Math Lesson Study Groups. *Teaching and Teacher Education*, *22*(7), 922–934.

Rabe, T., Krey, O. & Meinhardt, C. (2013a). Physikdidaktische Selbstwirksamkeitserwartungen zukünftiger Physiklehrkräfte I. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012* (S. 635–637). Münster: LIT.

Rabe, T., Krey, O. & Meinhardt, C. (2013b). Physikdidaktische Selbstwirksamkeitserwartungen (zukünftiger) Physiklehrkräfte II – eine Projektskizze. In S. Bernholt (Hrsg.), *Inquiry-based Learning - Forschendes Lernen. Gesellschaft für Didaktik der Chemie und Physik. Jahrestagung in Hannover 2012* (S. 638–640). Münster: LIT.

Rabe, T., Meinhardt, C. & Krey, O. (2012). Entwicklung eines Instruments zur Erhebung von Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 293–315.

Rammstedt, B. & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklungund Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. *Diagnostica*, *51*(4), 195–206.

Rasch, B., Friese, M., Hofmann, W. & Naumann, E. (2006). *Quantitative Methoden. Einführung in die Statistik* (2. Auflage). Heidelberg: Springer Medizin Verlag.

Rath, V. & Reinhold, P. (2014). Diagnosekompetenz von Physiklehramtsstudierenden. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013* (S. 441–443). Kiel: IPN.

Raudenbush, S. W., Rowan, B. & Cheong, Y. F. (1992). Contextual Effects on the Self-Perceived Efficacy of High School Teachers. *Sociology of Education*, *65*(2), 150–167.

Reckase, M. D. (1998). Consequential Validity from the Test Developer's Perspective. *Educational Measurement: Issues and Practice*, *17*(2), 13–16.

Reilly, E., Dhingra, K. & Boduszek, D. (2014). Teachers' Self-Efficacy Beliefs, Self-Esteem, and Job Stress as Determinants of Job Satisfaction. *International Journal of Educational Management*, *28*(4), 365–378.

Retelsdorf, J., Bauer, J., Gebauer, S. K., Kauper, T. & Möller, J. (2014). Erfassung berufsbezogener Selbstkonzepte von angehenden Lehrkräften (ERBSE-L). *Diagnostica*, *60*(2), 98–110.

Rich, Y., Lev, S. & Fischer, S. (1996). Extending the Concept and Assessment of Teacher Efficacy. *Educational and Psychological Measurement*, *56*(6), 1015–1025.

Richter, D., Kunter, M., Lüdtke, O., Klusmann, U. & Baumert, J. (2011). Soziale Unterstützung beim Berufseinstieg ins Lehramt. *Zeitschrift für Erziehungswissenschaft*, *14*(1), 35–59.

Riese, J. (2009). *Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften*. Berlin: Logos Verlag.

Riggs, I. M. & Enochs, L. G. (1990). Toward the Development of an Elementary Teacher's Science Teaching Efficacy Belief Instrument. *Science Education*, *74*(6), 625–637.

Rimm-Kaufman, S. E. & Sawyer, B. E. (2004). Primary-Grade Teachers' Self-Efficacy Beliefs, Attitudes toward Teaching, and Discipline and Teaching Practice Priorities in Relation to the "Responsive ClassroomÄpproach. *The Elementary School Journal*, *104*(4), 321–341.

Ritter, J. M., Boone, W. J. & Rubba, P. A. (2001). Development of an Instrument to Assess Prospective Elementary Teacher Self-Efficacy Beliefs about Equitable Science Teaching an Learning (SEBEST). *Journal of Science Teacher Education*, *12*(3), 175–198.

Roberts, J. K. & Henson, R. K. (2000). Self-Efficacy Teaching and Knowledge Instrument for Science Teachers (SETAKIST): A Proposal for a New Efficacy Instrument. In *Paper presented at the annual meeting of the mid-south educational research association (bowling green, ky, november 17-19, 2000)*.

Roberts, J. K. & Henson, R. K. (2001). A Confirmatory Factor Analysis of a New Measure of Teacher Efficacy: Ohio State Teacher Efficacy Scale. In *Paper Presented at the Annual Meeting of the American Educational Research Association (Seattle, WA, April 10-14, 2001)*.

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, *9*, 222–245.

Rose, J. S. & Medway, F. J. (1981). Measurement of Teachers' Beliefs in Their Control over Students Outcome. *Journal of Educational Research*, *74*(3), 185–189.

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton, N J: Princeton University Press.

Rosenberg, M., Schooler, C., Schoenbach, C. & Rosenberg, F. (1995). Global Self-Esteem and Specific Self-Esteem: Different Concepts, Different Outcomes. *American Sociological Review*, *60*(1), 141.

Ross, J. A. (1992). Teacher Efficacy and the Effects of Coaching on Student Achievement. *Canadian Journal of Education*, *17*(1), 51–65.

Ross, J. A. (1998). The Antecedents and Consequences of Teacher Efficacy. In J. Brophy (Hrsg.), *Advances in research on teaching* (S. 49–73). Greenwich, CT: JAI Press.

Rost, D. H. & Lamsfuss, S. (1992). Entwicklung und Erprobung einer ökonomischen Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten (SKSLF). *Zeitschrift fur Pädagogische Psychologie*, (6), 239–250.

Rost, D. H. & Sparfeldt, J. R. (2002). Facetten des schulischen Selbstkonzepts: Ein Verfahren zur Messung des differentiellen Selbstkonzepts schulischer Leistungen und Fähigkeiten (DISK-Gitter). *Diagnostica*, *48*(3), 130–140.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Verlag Hans Huber.

Rost, J. (2013). Reliabilitäts-Validitätsdilemma. In M. A. Wirtz (Hrsg.), *Dorsch Lexikon der Psychologie* (S. 1413). Verlag Hans Huber.

Rots, I., Aelterman, A., Vlerick, P. & Vermeulen, K. (2007). Teacher Education, Graduates' Teaching Commitment and Entrance into the Teaching Profession. *Teaching and Teacher Education*, *23*(5), 543–556.

Rotter, J. B. (1966). Generalized Expectancies for Internal versus External Control of Reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28.

Ruys, I., Van Keer, H. & Aelterman, A. A. (2011). Student Teachers' Skills in the Implementation of Collaborative Learning: A Multilevel Approach. *Teaching and Teacher Education*, *27*(7), 1090–1100.

Ryan, A. M., Kuusinen, C. M. & Bedoya-Skoog, A. (2015). Managing Peer Relations: A Dimension of Teacher Self-Efficacy that Varies Between Elementary and Middle School Teachers and is Associated with Observed Classroom Quality. *Contemporary Educational Psychology*, *41*, 147–156.

Ryan, R. M. & Deci, E. L. (2000). Self-Determinaton Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist*, *55*(1), 68–78.

Sandmann, A. (2014). Lautes Denken – die Analyse von Denk-, Lern- und Problemlöseprozessen. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen forschung* (S. 179–188). Berlin, Heidelberg: Springer Spektrum.

Scherer, R. (2013). Further Evidence on the Structural Relationship Between Academic Self-Concept and Self-Efficacy: On the Effects of Domain Specificity. *Learning and Individual Differences*, *28*, 9–19.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119–141). Berlin, Heidelberg: Springer.

Schmalt, H.-D. (2016). Effektanzmotivation. In M. A. Wirtz (Hrsg.), *Dorsch – lexikon der psychologie*. https://portal.hogrefe.com/dorsch/effektanzmotivation/, abgerufen am 04.02.2016.

Schmiemann, P. & Lücken, M. (2014). Validität – Misst mein Test, was er soll? In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 107–118). Berlin, Heidelberg: Springer Spektrum.

Schmitz, G. S. (2001). Kann Selbstwirksamkeitserwartung Lehrer vor Burnout schützen? Eine Längsschnittstudie in zehn Bundesländern. *Psychologie in Erziehung und Unterricht*, *48*, 49–67.

Schmitz, G. S. & Schwarzer, R. (2000). Selbstwirksamkeitserwartung von Lehrern: Längsschnittbefunde mit einem neuen Instrument. *Zeitschrift für Pädagogische Psychologie*, *14*(1), 12–25.

Schulte, K. (2008). *Selbstwirksamkeitserwartungen in der Lehrerbildung – Zur Struktur und dem Zusammenhang von Lehrer- Selbstwirksamkeitserwartungen, Pädagogischem Professionswissen und Persönlichkeitseigenschaften bei Lehramtsstudierenden und Lehrkräften* (Dissertation, Georg-August-Universität Göttingen, Göttingen).

Schulte, K., Bögeholz, S. & Watermann, R. (2008). Selbstwirksamkeitserwartungen und padagogisches Professionswissen im Verlauf des Lehramtsstudiums. *Zeitschrift fur Erziehungswissenschaft*, *11*(2), 268–287.

Schunk, D. H. (1991). Self-Efficacy and Academic Motivation. *Educational Psychologist*, *26*(3&4), 207–231.

Schwanzer, A. D., Trautwein, U., Lüdtke, O. & Sydow, H. (2005). Entwicklung eines Instruments zur Erfassung des Selbstkonzepts junger Erwachsener. *Diagnostica*, *51*(4), 183–194.

Schwarzer, R. & Hallum, S. (2008). Perceived Teacher Self-Efficacy as a Predictor of Job Stress and Burnout: Mediation Analyses. *Applied Psychology: An International Review*, *57*, 152–171.

Schwarzer, R. & Jerusalem, M. (1999). Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Berlin.

Schwarzer, R. & Jerusalem, M. (2002). Das Konzept der Selbstwirksamkeit. *Zeitschrift für Pädagogik, Beiheft 44*, 28–53.

Schwarzer, R. & Schmitz, G. S. (1999). Kollektive Selbstwirksamkeitserwartung von Lehrern: Eine Längsschnittstudie in zehn Bundesländern. *Zeitschrift für Sozialpsychologie*, *30*(4), 262–274.

Schwarzer, R., Schmitz, G. S. & Daytner, G. T. (1999). Teacher Self-Efficacy. verfügbar unter: http://userpage.fu-berlin.de/~health/teacher_se.htm, abgerufen am 27.02.2016. Abteilung für Gesundheitspsychologie, Freie Universität Berlin.

Senemoglu, N., Demirel, M., Yagci, E. & Üstündag, T. (2009). Elementary School Teachers' Self-Efficacy Beliefs: A Turkish Case. *Humanity & Social Sciences Journal*, *4*(2), 164–171.

Settlage, J., Southerland, S. A., Smith, L. K. & Ceglie, R. (2009). Constructing a Doubt-Free Teaching Self: Self-efficacy, Teacher Identity, and Science Instruction within Diverse Settings. *Journal of Research in Science Teaching*, *46*(1), 102–125.

Sharma, U., Loreman, T. & Forlin, C. (2012). Measuring Teacher Efficacy to Implement Inclusive Practices. *Journal of Research in Special Educational Needs*, *12*(1), 12–21.

Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Self-Concept: Validation of Construct Interpretations. *Review of Educational Research*, *46*(3), 407–441.

Shepard, L. A. (1997). The Centrality of Test Use and Consequences for Test Validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24.

Sherer, M., Maddux, J. E., Mercandante, B., Prentice-Dunn, S., Jacobs, B. & Rogers, R. W. (1982). The Self-Efficacy Scale: Construction and Validation. *Psychological Reports*, *51*, 663–671.

Shulman, L. S. (1986). Those Who Understand: Knowledge Growth in Teaching. *Educational Researcher*, *15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and Teaching: Foundations of the New Reform. *Harvard Educational Review*, *57*(1), 1–21.

Sia, A. P. (1992). Preservice Elementary Teachers' Perceived Efficacy in Teaching Environmental Education: A Preliminary Study. In *Paper presented at the annual meeting oft he eco-ed north american association for environmental education (toronto, ontario, canada, october 20, 1992)*.

Siwatu, K. O. (2007). Preservice Teachers' Culturally Responsive Teaching Self-Efficacy and Outcome Expectancy Beliefs. *Teaching and Teacher Education*, *23*(7), 1086–1101.

Six, B. (2016). Einstellung. In M. A. Wirtz (Hrsg.), *Dorsch – Lexikon der Psychologie*. https://portal.hogrefe.com/dorsch/einstellung/, abgerufen am 31.01.2016.

Skaalvik, E. M. & Skaalvik, S. (2007). Dimensions of Teacher Self-Efficacy and Relations With Strain Factors, Perceived Collective Teacher Efficacy, and Teacher Burnout. *Journal of Educational Psychology*, *99*(3), 611–625.

Skaalvik, E. M. & Skaalvik, S. (2010). Teacher Self-Efficacy and Teacher Burnout: A Study of Relations. *Teaching and Teacher Education*, *26*(4), 1059–1069.

Skinner, E. A. (1996). A Guide to Constructs of Control. *Journal of Personality and Social Psychology*, *71*(3), 549–570.

Skinner, E. A., Chapman, M. & Baltes, P. B. (1988). Control, Means-Ends, and Agency Beliefs: A New Conceptualization and Its Measurement During Childhood. *Journal of Personality and Social Psychology*, *54*(1), 117–133.

Smolleck, L. A. & Mongan, A. M. (2011). Changes in Preservice Teachers' Self-Efficacy: From Science Methods to Student Teaching. *Journal of Educational and Developmental Psychology*, *1*(1), 133–146.

Smolleck, L. A. & Yoder, E. P. (2008). Further Development and Validation of the Teaching Science as Inquiry (TSI) Instrument. *School Science and Mathematics*, *108*(7), 291–298.

Smolleck, L. D., Zembal-Saul, C. & Yoder, E. P. (2006). The Development and Validation of an Instrument to Measure Preservice Teachers' Self-Efficacy in Regard to the Teaching of Science as Inquiry. *Journal of Science Teacher Education*, *17*(2), 137–163.

Smylie, M. A. (1988). The Enhancement Function of Staff Development: Organizational and Psychological Antecedents to Individual Teacher Change. *American Educational Research Journal*, *25*(1), 1–30.

Somech, A. & Drach-Zahavy, A. (2000). Understanding Extra-Role Behavior in Schools: The Relationships Between Job Satisfaction, Sense of Efficacy, and Teachers' Extra-Role Behavior. *Teaching and Teacher Education*, *16*, 649–659.

Soodak, L. C. & Podell, D. M. (1993). Teacher Efficacy and Student Problem as Factors in Special Education Referral. *The Journal of Special Education*, *27*(1), 66–81.

Soodak, L. C. & Podell, D. M. (1996). Teacher Efficacy: Toward the Understanding of a Multi-Faceted Construct. *Teaching and Teacher Education*, *12*(4), 401–411.

Southerland, S. A., Sowell, S. & Enderle, P. (2011). Science Teachers' Pedagogical Discontentment: Its Sources and Potential for Change. *Journal of Science Teacher Education*, *22*(5), 437–457.

Stankov, L., Lee, J., Luo, W. & Hogan, D. J. (2012). Confidence: A Better Predictor of Academic Achievement than Self-Efficacy, Self-Concept and Anxiety? *Learning and Individual Differences*, *22*(6), 747–758.

Stender, A. (2014). *Unterrichtsplanung: Vom Wissen zum Handeln. Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung* (Diss., Christian-Albrechts-Universität zu Kiel, Kiel).

Stender, A., Brückmann, M. & Neumann, K. (2014). Der Einfluss der professionellen Kompetenz auf die Qualität der Skripte. In S. Bernholt (Hrsg.), *Naturwissenschaftliche Bildung zwischen Science- und Fachunterricht. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in München 2013* (S. 123–125). Kiel: IPN.

Stevens, T., Aguirre-Munoz, Z., Harris, G., Higgins, R. & Liu, X. (2013). Middle Level Mathematics Teachers' Self-Efficacy Growth through Professional Development: Differences Based on Mathematical Background. *Australian Journal of Teacher Education*, *38*(4), 144–164.

Stiller, J., Straube, P., Mathesius, S., Hartmann, S., Tiemann, R., Nordmeier, V., ... Upmeier zu Belzen, A. (2014). Fachmethodische Kompetenzen im Lehramtsstudium - das Projekt Ko-WADiS. In S. Bernholt (Hrsg.), *Naturwissenschaftliche bildung zwischen science- und fachunterricht. gesellschaft für didaktik der chemie und physik, jahrestagung in münchen 2013* (S. 156–158). Kiel: IPN.

Straube, P. & Nordmeier, V. (2012). Ko-WADiS: Wohin geht es ? In V. Nordmeier & H. Grötzebauch (Hrsg.), *Phydid b, didaktik der physik, beiträge zur dpg-frühjahrstagung* (S. 1–4). Mainz.

Swann, W. B., Chang-Schneider, C. & Larsen McClarty, K. (2007). Do people's Self-Views Matter? Self-Concept and Self-Esteem in Everyday Life. *The American Psychologist*, *62*(2), 84–94.

Terhart, E. (2002). Standards für die Lehrerbildung. Eine Expertise für die Kultusministerkonferenz.

Terhart, E. (2007a). Standards in der Lehrerbildung - eine Einführung. *Unterrichtswissenschaft*, *35*(1), 2–14.

Terhart, E. (2007b). Was wissen wir über gute Lehrer? Ergebnisse aus der empirischen Lehrerforschung. *Friedrich Jahresheft*, *25*, 20–24.

Terhart, E., Czerwenka, K., Ehrich, K., Jordan, F. & Schmidt, H. J. (1994). *Berufsbiographien von Lehrern und Lehrerinnen*. Frankfurt a. M. [u. a.]: Peter Lang.

Throndsen, I. & Turmo, A. (2013). Primary Mathematics Teachers' Goal Orientations and Student Achievement. *Instructional Science*, *41*(2), 307–322.

Tournaki, N. & Podell, D. M. (2005). The Impact of Student Characteristics and Teacher Efficacy on Teachers' Predictions of Student Success. *Teaching and Teacher Education*, *21*(3), 299–314.

Trentham, L., Silvern, S. & Brogdon, R. (1985). Teacher Efficacy and Teacher Competency Ratings. *Psychology in the Schools*, *22*(3), 343–352.

Tschannen-Moran, M. & Hoy, A. W. (2007). The Differential Antecedents of Self-Efficacy Beliefs of Novice and Experienced Teachers. *Teaching and Teacher Education*, *23*(6), 944–956.

Tschannen-Moran, M. & Woolfolk Hoy, A. (2001). Teacher Efficacy: Capturing an Elusive Construct. *Teaching and Teacher Education*, *17*(7), 783–805.

Tschannen-Moran, M., Woolfolk Hoy, A. & Hoy, W. K. (1998). Teacher Efficacy: Its Meaning and Measure. *Review of Educational Research*, *68*(2), 202–248.

Tsigilis, N., Koustelios, A. & Grammatikopoulos, V. (2010). Psychometric Properties of the Teachers' Sense of Efficacy Scale Within the Greek Educational Context. *Journal of Psychoeducational Assessment*, *28*(2), 153–162.

Tuchman, E. & Isaacs, J. (2011). The Influence of Formal and Informal Formative Pre-Service Experiences on Teacher Self-Efficacy. *Educational Psychology*, *31*(4), 413–433.

Valentine, J. C., DuBois, D. L. & Cooper, H. (2004). The Relation Between Self-Beliefs and Academic Achievement: A Meta-Analytic Review. *Educational Psychologist*, *39*(2), 111–133.

van Dinther, M., Dochy, F., Segers, M. & Braeken, J. (2013). The Construct Validity and Predictive Validity of a Self-Efficacy Measure for Student Teachers in Competence-Based Education. *Studies in Educational Evaluation*, *39*(3), 169–179.

Velthuis, C., Fisser, P. & Pieters, J. (2014). Teacher Training and Pre-service Primary Teachers' Self-Efficacy for Science Teaching. *Journal of Science Teacher Education*, *25*(4), 445–464.

Villa, A. & Calvete, E. (2001). Development of the Teacher Self-Concept Evaluation Scale and Its Relation to Burnout. *Studies in Educational Evaluation*, *27*, 239–255.

Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*. Berlin: Logos Verlag.

Vogelsang, C. & Reinhold, P. (2013). Zur Handlungsvalidität von Tests zum professionellen Wissen von Lehrkräften. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 103–128.

von Aufschnaiter, C. & Blömeke, S. (2010). Professionelle Kompetenz von (angehenden) Lehrkräften erfassen – Desiderata. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 361–368.

Wainer, H. & Braun, H. I. (Hrsg.). (1988). *Test Validity*. Hillsday, New Jersey: Lawrence Erlbaum Associates.

Walan, S. & Chang Rundgren, S.-N. (2014). Investigating Preschool and Primary School Teachers' Self-Efficacy and Needs in Teaching Science: A Pilot Study. *CEPS Journal 4*, *1*, 51–67.

Walker, J. & Slear, S. (2011). The Impact of Principal Leadership Behaviors on the Efficacy of New and Experienced Middle School Teachers. *NASSP Bulletin*, *95*(1), 46–64.

Weiber, R. & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung. Eine anwendungsorientierte Einführung in die Kausalanalyse mit Hilfe von AMOS, SmartPLS und SPSS*. Berlin, Heidelberg: Springer.

Weinert, F. E. (2014). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (3., aktualisierte Auflage, S. 17–31). Weinheim, Basel: Beltz.

White, R. W. (1959). Motivation Reconsidered: The Concept of Competence. *Psychological Review*, *66*(5), 297–333.

Whitley, J. (2010). Modelling the Influence of Teacher Characteristics on Student Achievement for Canadian Students With and Without Learning Disabilities. *International Journal of Special Education*, *25*(3), 88–97.

Woitkowski, D. (2015). *Fachliches Wissen Physik in der Hochschulausbildung. Konzeptualisierung, Messung, Niveaubildung*. Berlin: Logos Verlag.

Wolters, C. A. & Daugherty, S. G. (2007). Goal Structures and Teachers' Sense of Efficacy: Their Relation and Association to Teaching Experience and Academic Level. *Journal of Educational Psychology*, *99*(1), 181–193.

Woodcock, S. (2011). A Cross Sectional Study of Pre-service Teacher Efficacy Throughout the Training Years. *Australian Journal of Teacher Education*, *36*(10), 23–34.

Woods, A. M. & Rhoades, J. (2013). Teaching Efficacy Beliefs of National Board Certified Physical Educators. *Teachers and Teaching: Theory and Practice*, *19*(5), 507–526.

Woolfolk Hoy, A. & Davis, H. A. (2006). Teacher Self-Efficacy and Its Influence on the Achievement of Adolescents. In F. Pajares & T. Urdan (Hrsg.), *Self-efficacy beliefs of adolescents* (S. 117–137). Greenwich, CT: Information Age Publishing.

Woolfolk Hoy, A., Hoy, W. K. & Davis, H. A. (2009). Teachers' Self-Efficacy Beliefs. In K. R. Wentzel & A. Wigfield (Hrsg.), *Handbook of motivation at school* (S. 627–653). New York & London: Routledge.

Woolfolk Hoy, A. & Spero, R. B. (2005). Changes in Teacher Efficacy During the Early Years of Teaching: A Comparison of Four Measures. *Teaching and Teacher Education*, *21*(4), 343–356.

Woolfolk, A. E. & Hoy, W. K. (1990). Prospective Teachers' Sense of Efficacy and Beliefs about Control. *Journal of Educational Psychology*, *82*(1), 81–91.

Woolfolk, A. E., Rosoff, B. & Hoy, W. K. (1990). Teachers' Sense of Efficacy and Their Beliefs about Managing Students. *Teaching and Teacher Education*, *6*(2), 137–148.

Wright, B. D. & Linacre, J. M. (1994). Reasonable Mean-Square Fit Values. *Rasch Measurement Transactions*, *8*(3), 370.

Yeung, A. S., Craven, R. G. & Kaur, G. (2014). Teachers' Self-Concept and Valuing of Learning: Relations with Teaching Approaches and Beliefs about Students. *Asia-Pacific Journal of Teacher Education*, *42*(3), 305–320.

Ziegler, M. & Bühner, M. (2009). Modeling Socially Desirable Responding and Its Effects. *Educational and Psychological Measurement*, *69*(4), 548–565.

Zimmerman, B. J. (2000). Self-Efficacy: An Essential Motive to Learn. *Contemporary Educational Psychology*, *25*, 82–91.

Zlatkin-Troitschanskaia, O. & Seidel, J. (2011). Kompetenz und ihre Erfassung – das neue „Theorie-Empirie-Problem" der empirischen Bildungsforschung? In O. Zlatkin-Troitschanskaia (Hrsg.), *Stationen Empirischer Bildungsforschung* (S. 218–233). Wiesbaden: Springer.

Skalendokumentation

Die Skalendokumentation wurde bereits online veröffentlicht (vgl. Meinhardt u. a., 2016c) und ist unter folgendem Link abrufbar:

http://www.pedocs.de/frontdoor.php?source_opus=11818

Bisher erschienene Bände der Reihe *„Studien zum Physik- und Chemielernen“*

ISSN 1614-8967 (vormals *Studien zum Physiklernen* ISSN 1435-5280)

1 Helmut Fischler, Jochen Peuckert (Hrsg.): Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie
ISBN 978-3-89722-256-4 40.50 EUR

2 Anja Schoster: Bedeutungsentwicklungsprozesse beim Lösen algorithmischer Physikaufgaben. *Eine Fallstudie zu Lernprozessen von Schülern im Physiknachhilfeunterricht während der Bearbeitung algorithmischer Physikaufgaben*
ISBN 978-3-89722-045-4 40.50 EUR

3 Claudia von Aufschnaiter: Bedeutungsentwicklungen, Interaktionen und situatives Erleben beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-89722-143-7 40.50 EUR

4 Susanne Haeberlen: Lernprozesse im Unterricht mit Wasserstromkreisen. *Eine Fallstudie in der Sekundarstufe I*
ISBN 978-3-89722-172-7 40.50 EUR

5 Kerstin Haller: Über den Zusammenhang von Handlungen und Zielen. *Eine empirische Untersuchung zu Lernprozessen im physikalischen Praktikum*
ISBN 978-3-89722-242-7 40.50 EUR

6 Michaela Horstendahl: Motivationale Orientierungen im Physikunterricht
ISBN 978-3-89722-227-4 50.00 EUR

7 Stefan Deylitz: Lernergebnisse in der Quanten-Atomphysik. *Evaluation des Bremer Unterrichtskonzepts*
ISBN 978-3-89722-291-5 40.50 EUR

8 Lorenz Hucke: Handlungsregulation und Wissenserwerb in traditionellen und computergestützten Experimenten des physikalischen Praktikums
ISBN 978-3-89722-316-5 50.00 EUR

9 Heike Theyßen: Ein Physikpraktikum für Studierende der Medizin. *Darstellung der Entwicklung und Evaluation eines adressatenspezifischen Praktikums nach dem Modell der Didaktischen Rekonstruktion*
ISBN 978-3-89722-334-9 40.50 EUR

10 Annette Schick: Der Einfluß von Interesse und anderen selbstbezogenen Kognitionen auf Handlungen im Physikunterricht. *Fallstudien zu Interessenhandlungen im Physikunterricht*
ISBN 978-3-89722-380-6 40.50 EUR

11 Roland Berger: Moderne bildgebende Verfahren der medizinischen Diagnostik. *Ein Weg zu interessanterem Physikunterricht*
ISBN 978-3-89722-445-2 40.50 EUR

12 Johannes Werner: Vom Licht zum Atom. *Ein Unterrichtskonzept zur Quantenphysik unter Nutzung des Zeigermodells*
ISBN 978-3-89722-471-1 40.50 EUR

13 Florian Sander: Verbindung von Theorie und Experiment im physikalischen Praktikum. *Eine empirische Untersuchung zum handlungsbezogenen Vorverständnis und dem Einsatz grafikorientierter Modellbildung im Praktikum*
ISBN 978-3-89722-482-7 40.50 EUR

14 Jörn Gerdes: Der Begriff der physikalischen Kompetenz. *Zur Validierung eines Konstruktes*
ISBN 978-3-89722-510-7 40.50 EUR

15 Malte Meyer-Arndt: Interaktionen im Physikpraktikum zwischen Studierenden und Betreuern. *Feldstudie zu Bedeutungsentwicklungsprozessen im physikalischen Praktikum*
ISBN 978-3-89722-541-1 40.50 EUR

16 Dietmar Höttecke: Die Natur der Naturwissenschaften historisch verstehen. *Fachdidaktische und wissenschaftshistorische Untersuchungen*
ISBN 978-3-89722-607-4 40.50 EUR

17 Gil Gabriel Mavanga: Entwicklung und Evaluation eines experimentell- und phänomenorientierten Optikcurriculums. *Untersuchung zu Schülervorstellungen in der Sekundarstufe I in Mosambik und Deutschland*
ISBN 978-3-89722-721-7 40.50 EUR

18 Meike Ute Zastrow: Interaktive Experimentieranleitungen. *Entwicklung und Evaluation eines Konzeptes zur Vorbereitung auf das Experimentieren mit Messgeräten im Physikalischen Praktikum*
ISBN 978-3-89722-802-3 40.50 EUR

19 Gunnar Friege: Wissen und Problemlösen. *Eine empirische Untersuchung des wissenszentrierten Problemlösens im Gebiet der Elektrizitätslehre auf der Grundlage des Experten-Novizen-Vergleichs*
ISBN 978-3-89722-809-2 40.50 EUR

20 Erich Starauschek: Physikunterricht nach dem Karlsruher Physikkurs. *Ergebnisse einer Evaluationsstudie*
ISBN 978-3-89722-823-8 40.50 EUR

21 Roland Paatz: Charakteristika analogiebasierten Denkens. *Vergleich von Lernprozessen in Basis- und Zielbereich*
ISBN 978-3-89722-944-0 40.50 EUR

22 Silke Mikelskis-Seifert: Die Entwicklung von Metakonzepten zur Teilchenvorstellung bei Schülern. *Untersuchung eines Unterrichts über Modelle mithilfe eines Systems multipler Repräsentationsebenen*
ISBN 978-3-8325-0013-9 40.50 EUR

23 Brunhild Landwehr: Distanzen von Lehrkräften und Studierenden des Sachunterrichts zur Physik. *Eine qualitativ-empirische Studie zu den Ursachen*
ISBN 978-3-8325-0044-3 40.50 EUR

24 Lydia Murmann: Physiklernen zu Licht, Schatten und Sehen. *Eine phänomenografische Untersuchung in der Primarstufe*
ISBN 978-3-8325-0060-3 40.50 EUR

25 Thorsten Bell: Strukturprinzipien der Selbstregulation. *Komplexe Systeme, Elementarisierungen und Lernprozessstudien für den Unterricht der Sekundarstufe II*
ISBN 978-3-8325-0134-1 40.50 EUR

26 Rainer Müller: Quantenphysik in der Schule
ISBN 978-3-8325-0186-0 40.50 EUR

27 Jutta Roth: Bedeutungsentwicklungsprozesse von Physikerinnen und Physikern in den Dimensionen Komplexität, Zeit und Inhalt
ISBN 978-3-8325-0183-9 40.50 EUR

28 Andreas Saniter: Spezifika der Verhaltensmuster fortgeschrittener Studierender der Physik
ISBN 978-3-8325-0292-8 40.50 EUR

29 Thomas Weber: Kumulatives Lernen im Physikunterricht. *Eine vergleichende Untersuchung in Unterrichtsgängen zur geometrischen Optik*
ISBN 978-3-8325-0316-1 40.50 EUR

30 Markus Rehm: Über die Chancen und Grenzen moralischer Erziehung im naturwissenschaftlichen Unterricht
ISBN 978-3-8325-0368-0 40.50 EUR

31 Marion Budde: Lernwirkungen in der Quanten-Atom-Physik. *Fallstudien über Resonanzen zwischen Lernangeboten und SchülerInnen-Vorstellungen*
ISBN 978-3-8325-0483-0 40.50 EUR

32 Thomas Reyer: Oberflächenmerkmale und Tiefenstrukturen im Unterricht. *Exemplarische Analysen im Physikunterricht der gymnasialen Sekundarstufe*
ISBN 978-3-8325-0488-5 40.50 EUR

33 Christoph Thomas Müller: Subjektive Theorien und handlungsleitende Kognitionen von Lehrern als Determinanten schulischer Lehr-Lern-Prozesse im Physikunterricht
ISBN 978-3-8325-0543-1 40.50 EUR

34 Gabriela Jonas-Ahrend: Physiklehrervorstellungen zum Experiment im Physikunterricht
ISBN 978-3-8325-0576-9 40.50 EUR

35 Dimitrios Stavrou: Das Zusammenspiel von Zufall und Gesetzmäßigkeiten in der nichtlinearen Dynamik. *Didaktische Analyse und Lernprozesse*
ISBN 978-3-8325-0609-4 40.50 EUR

36 Katrin Engeln: Schülerlabors: authentische, aktivierende Lernumgebungen als Möglichkeit, Interesse an Naturwissenschaften und Technik zu wecken
ISBN 978-3-8325-0689-6 40.50 EUR

37 Susann Hartmann: Erklärungsvielfalt
ISBN 978-3-8325-0730-5 40.50 EUR

38 Knut Neumann: Didaktische Rekonstruktion eines physikalischen Praktikums für Physiker
ISBN 978-3-8325-0762-6 40.50 EUR

39 Michael Späth: Kontextbedingungen für Physikunterricht an der Hauptschule. *Möglichkeiten und Ansatzpunkte für einen fachübergreifenden, handlungsorientierten und berufsorientierten Unterricht*
ISBN 978-3-8325-0827-2 40.50 EUR

40 Jörg Hirsch: Interesse, Handlungen und situatives Erleben von Schülerinnen und Schülern beim Bearbeiten physikalischer Aufgaben
ISBN 978-3-8325-0875-3 40.50 EUR

41 Monika Hüther: Evaluation einer hypermedialen Lernumgebung zum Thema Gasgesetze. *Eine Studie im Rahmen des Physikpraktikums für Studierende der Medizin*
ISBN 978-3-8325-0911-8 40.50 EUR

42 Maike Tesch: Das Experiment im Physikunterricht. *Didaktische Konzepte und Ergebnisse einer Videostudie*
ISBN 978-3-8325-0975-0 40.50 EUR

43 Nina Nicolai: Skriptgeleitete Eltern-Kind-Interaktion bei Chemiehausaufgaben. *Eine Evaluationsstudie im Themenbereich Säure-Base*
ISBN 978-3-8325-1013-8 40.50 EUR

44 Antje Leisner: Entwicklung von Modellkompetenz im Physikunterricht
ISBN 978-3-8325-1020-6 40.50 EUR

45 Stefan Rumann: Evaluation einer Interventionsstudie zur Säure-Base-Thematik
ISBN 978-3-8325-1027-5 40.50 EUR

46 Thomas Wilhelm: Konzeption und Evaluation eines Kinematik/Dynamik-Lehrgangs zur Veränderung von Schülervorstellungen mit Hilfe dynamisch ikonischer Repräsentationen und graphischer Modellbildung – mit CD-ROM
ISBN 978-3-8325-1046-6 45.50 EUR

47 Andrea Maier-Richter: Computerunterstütztes Lernen mit Lösungsbeispielen in der Chemie. *Eine Evaluationsstudie im Themenbereich Löslichkeit*
ISBN 978-3-8325-1046-6 40.50 EUR

48 Jochen Peuckert: Stabilität und Ausprägung kognitiver Strukturen zum Atombegriff
ISBN 978-3-8325-1104-3 40.50 EUR

49 Maik Walpuski: Optimierung von experimenteller Kleingruppenarbeit durch Strukturierungshilfen und Feedback
ISBN 978-3-8325-1184-5 40.50 EUR

50 Helmut Fischler, Christiane S. Reiners (Hrsg.): Die Teilchenstruktur der Materie im Physik- und Chemieunterricht
ISBN 978-3-8325-1225-5 34.90 EUR

51 Claudia Eysel: Interdisziplinäres Lehren und Lernen in der Lehrerbildung. *Eine empirische Studie zum Kompetenzerwerb in einer komplexen Lernumgebung*
ISBN 978-3-8325-1238-5 40.50 EUR

52 Johannes Günther: Lehrerfortbildung über die Natur der Naturwissenschaften. *Studien über das Wissenschaftsverständnis von Grundschullehrkräften*
ISBN 978-3-8325-1287-3 40.50 EUR

53 Christoph Neugebauer: Lernen mit Simulationen und der Einfluss auf das Problemlösen in der Physik
ISBN 978-3-8325-1300-9 40.50 EUR

54 Andreas Schnirch: Gendergerechte Interessen- und Motivationsförderung im Kontext naturwissenschaftlicher Grundbildung. *Konzeption, Entwicklung und Evaluation einer multimedial unterstützten Lernumgebung*
ISBN 978-3-8325-1334-4 40.50 EUR

55 Hilde Köster: Freies Explorieren und Experimentieren. *Eine Untersuchung zur selbstbestimmten Gewinnung von Erfahrungen mit physikalischen Phänomenen im Sachunterricht*
ISBN 978-3-8325-1348-1 40.50 EUR

56 Eva Heran-Dörr: Entwicklung und Evaluation einer Lehrerfortbildung zur Förderung der physikdidaktischen Kompetenz von Sachunterrichtslehrkräften
ISBN 978-3-8325-1377-1 40.50 EUR

57 Agnes Szabone Varnai: Unterstützung des Problemlösens in Physik durch den Einsatz von Simulationen und die Vorgabe eines strukturierten Kooperationsformats
ISBN 978-3-8325-1403-7 40.50 EUR

58 Johannes Rethfeld: Aufgabenbasierte Lernprozesse in selbstorganisationsoffenem Unterricht der Sekundarstufe I zum Themengebiet ELEKTROSTATIK. *Eine Feldstudie in vier 10. Klassen zu einer kartenbasierten Lernumgebung mit Aufgaben aus der Elektrostatik*
ISBN 978-3-8325-1416-7 40.50 EUR

59 Christian Henke: Experimentell-naturwissenschaftliche Arbeitsweisen in der Oberstufe. *Untersuchung am Beispiel des HIGHSEA-Projekts in Bremerhaven*
ISBN 978-3-8325-1515-7 40.50 EUR

60 Lutz Kasper: Diskursiv-narrative Elemente für den Physikunterricht. *Entwicklung und Evaluation einer multimedialen Lernumgebung zum Erdmagnetismus*
ISBN 978-3-8325-1537-9 40.50 EUR

61 Thorid Rabe: Textgestaltung und Aufforderung zu Selbsterklärungen beim Physiklernen mit Multimedia
ISBN 978-3-8325-1539-3 40.50 EUR

62 Ina Glemnitz: Vertikale Vernetzung im Chemieunterricht. *Ein Vergleich von traditionellem Unterricht mit Unterricht nach Chemie im Kontext*
ISBN 978-3-8325-1628-4 40.50 EUR

63 Erik Einhaus: Schülerkompetenzen im Bereich Wärmelehre. *Entwicklung eines Testinstruments zur Überprüfung und Weiterentwicklung eines normativen Modells fachbezogener Kompetenzen*
ISBN 978-3-8325-1630-7 40.50 EUR

64 Jasmin Neuroth: Concept Mapping als Lernstrategie. *Eine Interventionsstudie zum Chemielernen aus Texten*
ISBN 978-3-8325-1659-8 40.50 EUR

65 Hans Gerd Hegeler-Burkhart: Zur Kommunikation von Hauptschülerinnen und Hauptschülern in einem handlungsorientierten und fächerübergreifenden Unterricht mit physikalischen und technischen Inhalten
ISBN 978-3-8325-1667-3 40.50 EUR

66 Karsten Rincke: Sprachentwicklung und Fachlernen im Mechanikunterricht. *Sprache und Kommunikation bei der Einführung in den Kraftbegriff*
ISBN 978-3-8325-1699-4 40.50 EUR

67 Nina Strehle: Das Ion im Chemieunterricht. *Alternative Schülervorstellungen und curriculare Konsequenzen*
ISBN 978-3-8325-1710-6 40.50 EUR

68 Martin Hopf: Problemorientierte Schülerexperimente
ISBN 978-3-8325-1711-3 40.50 EUR

69 Anne Beerenwinkel: Fostering conceptual change in chemistry classes using expository texts
ISBN 978-3-8325-1721-2 40.50 EUR

70 Roland Berger: Das Gruppenpuzzle im Physikunterricht der Sekundarstufe II. *Eine empirische Untersuchung auf der Grundlage der Selbstbestimmungstheorie der Motivation*
ISBN 978-3-8325-1732-8 40.50 EUR

71 Giuseppe Colicchia: Physikunterricht im Kontext von Medizin und Biologie. *Entwicklung und Erprobung von Unterrichtseinheiten*
ISBN 978-3-8325-1746-5 40.50 EUR

72 Sandra Winheller: Geschlechtsspezifische Auswirkungen der Lehrer-Schüler-Interaktion im Chemieanfangsunterricht
ISBN 978-3-8325-1757-1 40.50 EUR

73 Isabel Wahser: Training von naturwissenschaftlichen Arbeitsweisen zur Unterstützung experimenteller Kleingruppenarbeit im Fach Chemie
ISBN 978-3-8325-1815-8 40.50 EUR

74 Claus Brell: Lernmedien und Lernerfolg - reale und virtuelle Materialien im Physikunterricht. *Empirische Untersuchungen in achten Klassen an Gymnasien (Laborstudie) zum Computereinsatz mit Simulation und IBE*
ISBN 978-3-8325-1829-5 40.50 EUR

75 Rainer Wackermann: Überprüfung der Wirksamkeit eines Basismodell-Trainings für Physiklehrer
ISBN 978-3-8325-1882-0 40.50 EUR

76 Oliver Tepner: Effektivität von Aufgaben im Chemieunterricht der Sekundarstufe I
ISBN 978-3-8325-1919-3 40.50 EUR

77 Claudia Geyer: Museums- und Science-Center-Besuche im naturwissenschaftlichen Unterricht aus einer motivationalen Perspektive. *Die Sicht von Lehrkräften und Schülerinnen und Schülern*
ISBN 978-3-8325-1922-3 40.50 EUR

78 Tobias Leonhard: Professionalisierung in der Lehrerbildung. *Eine explorative Studie zur Entwicklung professioneller Kompetenzen in der Lehrererstausbildung*
ISBN 978-3-8325-1924-7 40.50 EUR

79 Alexander Kauertz: Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben
ISBN 978-3-8325-1925-4 40.50 EUR

80 Regina Hübinger: Schüler auf Weltreise. *Entwicklung und Evaluation von Lehr-/ Lernmaterialien zur Förderung experimentell-naturwissenschaftlicher Kompetenzen für die Jahrgangsstufen 5 und 6*
ISBN 978-3-8325-1932-2 40.50 EUR

81 Christine Waltner: Physik lernen im Deutschen Museum
ISBN 978-3-8325-1933-9 40.50 EUR

82 Torsten Fischer: Handlungsmuster von Physiklehrkräften beim Einsatz neuer Medien. *Fallstudien zur Unterrichtspraxis*
ISBN 978-3-8325-1948-3 42.00 EUR

83 Corinna Kieren: Chemiehausaufgaben in der Sekundarstufe I des Gymnasiums. *Fragebogenerhebung zur gegenwärtigen Praxis und Entwicklung eines optimierten Hausaufgabendesigns im Themenbereich Säure-Base*
978-3-8325-1975-9 37.00 EUR

84 Marco Thiele: Modelle der Thermohalinen Zirkulation im Unterricht. *Eine empirische Studie zur Förderung des Modellverständnisses*
ISBN 978-3-8325-1982-7 40.50 EUR

85 Bernd Zinn: Physik lernen, um Physik zu lehren. *Eine Möglichkeit für interessanteren Physikunterricht*
ISBN 978-3-8325-1995-7 39.50 EUR

86 Esther Klaes: Außerschulische Lernorte im naturwissenschaftlichen Unterricht. *Die Perspektive der Lehrkraft*
ISBN 978-3-8325-2006-9 43.00 EUR

87 Marita Schmidt: Kompetenzmodellierung und -diagnostik im Themengebiet Energie der Sekundarstufe I. *Entwicklung und Erprobung eines Testinventars*
ISBN 978-3-8325-2024-3 37.00 EUR

88 Gudrun Franke-Braun: Aufgaben mit gestuften Lernhilfen. *Ein Aufgabenformat zur Förderung der sachbezogenen Kommunikation und Lernleistung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-2026-7 38.00 EUR

89 Silke Klos: Kompetenzförderung im naturwissenschaftlichen Anfangsunterricht. *Der Einfluss eines integrierten Unterrichtskonzepts*
ISBN 978-3-8325-2133-2 37.00 EUR

90 Ulrike Elisabeth Burkard: Quantenphysik in der Schule. *Bestandsaufnahme, Perspektiven und Weiterentwicklungsmöglichkeiten durch die Implementation eines Medienservers*
ISBN 978-3-8325-2215-5 43.00 EUR

91 Ulrike Gromadecki: Argumente in physikalischen Kontexten. *Welche Geltungsgründe halten Physikanfänger für überzeugend?*
ISBN 978-3-8325-2250-6 41.50 EUR

92 Jürgen Bruns: Auf dem Weg zur Förderung naturwissenschaftsspezifischer Vorstellungen von zukünftigen Chemie-Lehrenden
ISBN 978-3-8325-2257-5 43.50 EUR

93 Cornelius Marsch: Räumliche Atomvorstellung. *Entwicklung und Erprobung eines Unterrichtskonzeptes mit Hilfe des Computers*
ISBN 978-3-8325-2293-3 82.50 EUR

94 Maja Brückmann: Sachstrukturen im Physikunterricht. *Ergebnisse einer Videostudie*
ISBN 978-3-8325-2272-8 39.50 EUR

95 Sabine Fechner: Effects of Context-oriented Learning on Student Interest and Achievement in Chemistry Education
ISBN 978-3-8325-2343-5 36.50 EUR

96 Clemens Nagel: eLearning im Physikalischen Anfängerpraktikum
ISBN 978-3-8325-2355-8 39.50 EUR

97 Josef Riese: Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften
ISBN 978-3-8325-2376-3 39.00 EUR

98 Sascha Bernholt: Kompetenzmodellierung in der Chemie. *Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität*
ISBN 978-3-8325-2447-0 40.00 EUR

99 Holger Christoph Stawitz: Auswirkung unterschiedlicher Aufgabenprofile auf die Schülerleistung. *Vergleich von Naturwissenschafts- und Problemlöseaufgaben der PISA 2003-Studie*
ISBN 978-3-8325-2451-7 37.50 EUR

100 Hans Ernst Fischer, Elke Sumfleth (Hrsg.): nwu-essen – 10 Jahre Essener Forschung zum naturwissenschaftlichen Unterricht
ISBN 978-3-8325-3331-1 40.00 EUR

101 Hendrik Härtig: Sachstrukturen von Physikschulbüchern als Grundlage zur Bestimmung der Inhaltsvalidität eines Tests
ISBN 978-3-8325-2512-5 34.00 EUR

102 Thomas Grüß-Niehaus: Zum Verständnis des Löslichkeitskonzeptes im Chemieunterricht. *Der Effekt von Methoden progressiver und kollaborativer Reflexion*
ISBN 978-3-8325-2537-8 40.50 EUR

103 Patrick Bronner: Quantenoptische Experimente als Grundlage eines Curriculums zur Quantenphysik des Photons

ISBN 978-3-8325-2540-8 36.00 EUR

104 Adrian Voßkühler: Blickbewegungsmessung an Versuchsaufbauten. *Studien zur Wahrnehmung, Verarbeitung und Usability von physikbezogenen Experimenten am Bildschirm und in der Realität*
ISBN 978-3-8325-2548-4 47.50 EUR

105 Verena Tobias: Newton'sche Mechanik im Anfangsunterricht. *Die Wirksamkeit einer Einführung über die zweidimensionale Dynamik auf das Lehren und Lernen*
ISBN 978-3-8325-2558-3 54.00 EUR

106 Christian Rogge: Entwicklung physikalischer Konzepte in aufgabenbasierten Lernumgebungen
ISBN 978-3-8325-2574-3 45.00 EUR

107 Mathias Ropohl: Modellierung von Schülerkompetenzen im Basiskonzept Chemische Reaktion. *Entwicklung und Analyse von Testaufgaben*
ISBN 978-3-8325-2609-2 36.50 EUR

108 Christoph Kulgemeyer: Physikalische Kommunikationskompetenz. *Modellierung und Diagnostik*
ISBN 978-3-8325-2674-0 44.50 EUR

109 Jennifer Olszewski: The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes
ISBN 978-3-8325-2680-1 33.50 EUR

110 Annika Ohle: Primary School Teachers' Content Knowledge in Physics and its Impact on Teaching and Students' Achievement
ISBN 978-3-8325-2684-9 36.50 EUR

111 Susanne Mannel: Assessing scientific inquiry. *Development and evaluation of a test for the low-performing stage*
ISBN 978-3-8325-2761-7 40.00 EUR

112 Michael Plomer: Physik physiologisch passend praktiziert. *Eine Studie zur Lernwirksamkeit von traditionellen und adressatenspezifischen Physikpraktika für die Physiologie*
ISBN 978-3-8325-2804-1 34.50 EUR

113 Alexandra Schulz: Experimentierspezifische Qualitätsmerkmale im Chemieunterricht. *Eine Videostudie*
ISBN 978-3-8325-2817-1 40.00 EUR

114 Franz Boczianowski: Eine empirische Untersuchung zu Vektoren im Physikunterricht der Mittelstufe
ISBN 978-3-8325-2843-0 39.50 EUR

115 Maria Ploog: Internetbasiertes Lernen durch Textproduktion im Fach Physik
ISBN 978-3-8325-2853-9 39.50 EUR

116 Anja Dhein: Lernen in Explorier- und Experimentiersituationen. *Eine explorative Studie zu Bedeutungsentwicklungsprozessen bei Kindern im Alter zwischen 4 und 6 Jahren*
ISBN 978-3-8325-2859-1 45.50 EUR

117 Irene Neumann: Beyond Physics Content Knowledge. *Modeling Competence Regarding Nature of Scientific Inquiry and Nature of Scientific Knowledge*
ISBN 978-3-8325-2880-5 37.00 EUR

118 Markus Emden: Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens. *Eine vergleichende Studie zu Diagnoseinstrumenten zu Beginn der Sekundarstufe I*
ISBN 978-3-8325-2867-6 38.00 EUR

119 Birgit Hofmann: Analyse von Blickbewegungen von Schülern beim Lesen von physikbezogenen Texten mit Bildern. *Eye Tracking als Methodenwerkzeug in der physikdidaktischen Forschung*
ISBN 978-3-8325-2925-3 59.00 EUR

120 Rebecca Knobloch: Analyse der fachinhaltlichen Qualität von Schüleräußerungen und deren Einfluss auf den Lernerfolg. *Eine Videostudie zu kooperativer Kleingruppenarbeit*
ISBN 978-3-8325-3006-8 36.50 EUR

121 Julia Hostenbach: Entwicklung und Prüfung eines Modells zur Beschreibung der Bewertungskompetenz im Chemieunterricht
ISBN 978-3-8325-3013-6 38.00 EUR

122 Anna Windt: Naturwissenschaftliches Experimentieren im Elementarbereich. *Evaluation verschiedener Lernsituationen*
ISBN 978-3-8325-3020-4 43.50 EUR

123 Eva Kölbach: Kontexteinflüsse beim Lernen mit Lösungsbeispielen
ISBN 978-3-8325-3025-9 38.50 EUR

124 Anna Lau: Passung und vertikale Vernetzung im Chemie- und Physikunterricht
ISBN 978-3-8325-3021-1 36.00 EUR

125 Jan Lamprecht: Ausbildungswege und Komponenten professioneller Handlungskompetenz. *Vergleich von Quereinsteigern mit Lehramtsabsolventen für Gymnasien im Fach Physik*
ISBN 978-3-8325-3035-8 38.50 EUR

126 Ulrike Böhm: Förderung von Verstehensprozessen unter Einsatz von Modellen
ISBN 978-3-8325-3042-6 41.00 EUR

127 Sabrina Dollny: Entwicklung und Evaluation eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Chemielehrkräften
ISBN 978-3-8325-3046-4 37.00 EUR

128 Monika Zimmermann: Naturwissenschaftliche Bildung im Kindergarten. *Eine integrative Längsschnittstudie zur Kompetenzentwicklung von Erzieherinnen*
ISBN 978-3-8325-3053-2 54.00 EUR

129 Ulf Saballus: Über das Schlussfolgern von Schülerinnen und Schülern zu öffentlichen Kontroversen mit naturwissenschaftlichem Hintergrund. *Eine Fallstudie*
ISBN 978-3-8325-3086-0 39.50 EUR

130 Olaf Krey: Zur Rolle der Mathematik in der Physik. *Wissenschaftstheoretische Aspekte und Vorstellungen Physiklernender*
ISBN 978-3-8325-3101-0 46.00 EUR

131 Angelika Wolf: Zusammenhänge zwischen der Eigenständigkeit im Physikunterricht, der Motivation, den Grundbedürfnissen und dem Lernerfolg von Schülern
ISBN 978-3-8325-3161-4 45.00 EUR

132 Johannes Börlin: Das Experiment als Lerngelegenheit. *Vom interkulturellen Vergleich des Physikunterrichts zu Merkmalen seiner Qualität*
ISBN 978-3-8325-3170-6 45.00 EUR

133 Olaf Uhden: Mathematisches Denken im Physikunterricht. *Theorieentwicklung und Problemanalyse*
ISBN 978-3-8325-3170-6 45.00 EUR

134 Christoph Gut: Modellierung und Messung experimenteller Kompetenz. *Analyse eines large-scale Experimentiertests*
ISBN 978-3-8325-3213-0 40.00 EUR

135 Antonio Rueda: Lernen mit ExploMultimedial in kolumbianischen Schulen. *Analyse von kurzzeitigen Lernprozessen und der Motivation beim länderübergreifenden Einsatz einer deutschen computergestützten multimedialen Lernumgebung für den naturwissenschaftlichen Unterricht*
ISBN 978-3-8325-3218-5 45.50 EUR

136 Krisztina Berger: Bilder, Animationen und Notizen. *Empirische Untersuchung zur Wirkung einfacher visueller Repräsentationen und Notizen auf den Wissenserwerb in der Optik*
ISBN 978-3-8325-3238-3 41.50 EUR

137 Antony Crossley: Untersuchung des Einflusses unterschiedlicher physikalischer Konzepte auf den Wissenserwerb in der Thermodynamik der Sekundarstufe I
ISBN 978-3-8325-3275-8 40.00 EUR

138 Tobias Viering: Entwicklung physikalischer Kompetenz in der Sekundarstufe I. *Validierung eines Kompetenzentwicklungsmodells für das Energiekonzept im Bereich Fachwissen*
ISBN 978-3-8325-3277-2 37.00 EUR

139 Nico Schreiber: Diagnostik experimenteller Kompetenz. *Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*
ISBN 978-3-8325-3284-0 39.00 EUR

140 Sarah Hundertmark: Einblicke in kollaborative Lernprozesse. *Eine Fallstudie zur reflektierenden Zusammenarbeit unterstützt durch die Methoden Concept Mapping und Lernbegleitbogen*
ISBN 978-3-8325-3251-2 43.00 EUR

141 Ronny Scherer: Analyse der Struktur, Messinvarianz und Ausprägung komplexer Problemlösekompetenz im Fach Chemie. *Eine Querschnittstudie in der Sekundarstufe I und am Übergang zur Sekundarstufe II*
ISBN 978-3-8325-3312-0 43.00 EUR

142 Patricia Heitmann: Bewertungskompetenz im Rahmen naturwissenschaftlicher Problemlöseprozesse. *Modellierung und Diagnose der Kompetenzen Bewertung und analytisches Problemlösen für das Fach Chemie*
ISBN 978-3-8325-3314-4 37.00 EUR

143 Jan Fleischhauer: Wissenschaftliches Argumentieren und Entwicklung von Konzepten beim Lernen von Physik
ISBN 978-3-8325-3325-0 35.00 EUR

144 Nermin Özcan: Zum Einfluss der Fachsprache auf die Leistung im Fach Chemie. *Eine Förderstudie zur Fachsprache im Chemieunterricht*
ISBN 978-3-8325-3328-1 36.50 EUR

145 Helena van Vorst: Kontextmerkmale und ihr Einfluss auf das Schülerinteresse im Fach Chemie
ISBN 978-3-8325-3321-2 38.50 EUR

146 Janine Cappell: Fachspezifische Diagnosekompetenz angehender Physiklehrkräfte in der ersten Ausbildungsphase
ISBN 978-3-8325-3356-4 38.50 EUR

147 Susanne Bley: Förderung von Transferprozessen im Chemieunterricht
ISBN 978-3-8325-3407-3 40.50 EUR

148 Cathrin Blaes: Die übungsgestützte Lehrerpräsentation im Chemieunterricht der Sekundarstufe I. *Evaluation der Effektivität*
ISBN 978-3-8325-3409-7 43.50 EUR

149 Julia Suckut: Die Wirksamkeit von piko-OWL als Lehrerfortbildung. Eine Evaluation zum Projekt *Physik im Kontext* in Fallstudien
ISBN 978-3-8325-3440-0 45.00 EUR

150 Alexandra Dorschu: Die Wirkung von Kontexten in Physikkompetenztestaufgaben
ISBN 978-3-8325-3446-2 37.00 EUR

151 Jochen Scheid: Multiple Repräsentationen, Verständnis physikalischer Experimente und kognitive Aktivierung: *Ein Beitrag zur Entwicklung der Aufgabenkultur*
ISBN 978-3-8325-3449-3 49.00 EUR

152 Tim Plasa: Die Wahrnehmung von Schülerlaboren und Schülerforschungszentren
ISBN 978-3-8325-3483-7 35.50 EUR

153 Felix Schoppmeier: Physikkompetenz in der gymnasialen Oberstufe.*Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*
ISBN 978-3-8325-3502-5 36.00 EUR

154 Katharina Groß: Experimente alternativ dokumentieren. *Eine qualitative Studie zur Förderung der Diagnose- und Differenzierungskompetenz in der Chemielehrerbildung*
ISBN 978-3-8325-3508-7 43.50 EUR

155 Barbara Hank: Konzeptwandelprozesse im Anfangsunterricht Chemie. *Eine quasiexperimentelle Längsschnittstudie*
ISBN 978-3-8325-3519-3 38.50 EUR

156 Katja Freyer: Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie
ISBN 978-3-8325-3544-5 38.00 EUR

157 Alexander Rachel: Auswirkungen instruktionaler Hilfen bei der Einführung des (Ferro-)Magnetismus. *Eine Vergleichsstudie in der Primar- und Sekundarstufe*
ISBN 978-3-8325-3548-3 43.50 EUR

158 Sebastian Ritter: Einfluss des Lerninhalts Nanogrößeneffekte auf Teilchen- und Teilchenmodellvorstellungen von Schülerinnen und Schülern
ISBN 978-3-8325-3558-2 36.00 EUR

159 Andrea Harbach: Problemorientierung und Vernetzung in kontextbasierten Lernaufgaben
ISBN 978-3-8325-3564-3 39.00 EUR

160 David Obst: Interaktive Tafeln im Physikunterricht. *Entwicklung und Evaluation einer Lehrerfortbildung*
ISBN 978-3-8325-3582-7 40.50 EUR

161 Sophie Kirschner: Modellierung und Analyse des Professionswissens von Physiklehrkräften
ISBN 978-3-8325-3601-5 35.00 EUR

162 Katja Stief: Selbstregulationsprozesse und Hausaufgabenmotivation im Chemieunterricht
ISBN 978-3-8325-3631-2 34.00 EUR

163 Nicola Meschede: Professionelle Wahrnehmung der inhaltlichen Strukturierung im naturwissenschaftlichen Grundschulunterricht. *Theoretische Beschreibung und empirische Erfassung*
ISBN 978-3-8325-3668-8 37.00 EUR

164 Johannes Maximilian Barth: Experimentieren im Physikunterricht der gymnasialen Oberstufe. *Eine Rekonstruktion übergeordneter Einbettungsstrategien*
ISBN 978-3-8325-3681-7 39.00 EUR

165 Sandra Lein: Das Betriebspraktikum in der Lehrerbildung. *Eine Untersuchung zur Förderung der Wissenschafts- und Technikbildung im allgemeinbildenden Unterricht*
ISBN 978-3-8325-3698-5 40.00 EUR

166 Veranika Maiseyenka: Modellbasiertes Experimentieren im Unterricht. *Praxistauglichkeit und Lernwirkungen*
ISBN 978-3-8325-3708-1 38.00 EUR

167 Christoph Stolzenberger: Der Einfluss der didaktischen Lernumgebung auf das Erreichen geforderter Bildungsziele am Beispiel der W- und P-Seminare im Fach Physik
ISBN 978-3-8325-3708-1 38.00 EUR

168 Pia Altenburger: Mehrebenenregressionsanalysen zum Physiklernen im Sachunterricht der Primarstufe. *Ergebnisse einer Evaluationsstudie.*
ISBN 978-3-8325-3717-3 37.50 EUR

169 Nora Ferber: Entwicklung und Validierung eines Testinstruments zur Erfassung von Kompetenzentwicklung im Fach Chemie in der Sekundarstufe I
ISBN 978-3-8325-3727-2 39.50 EUR

170 Anita Stender: Unterrichtsplanung: Vom Wissen zum Handeln.
Theoretische Entwicklung und empirische Überprüfung des Transformationsmodells der Unterrichtsplanung
ISBN 978-3-8325-3750-0 41.50 EUR

171 Jenna Koenen: Entwicklung und Evaluation von experimentunterstützten Lösungsbeispielen zur Förderung naturwissenschaftlich-experimenteller Arbeitsweisen
ISBN 978-3-8325-3785-2 43.00 EUR

172 Teresa Henning: Empirische Untersuchung kontextorientierter Lernumgebungen in der Hochschuldidaktik. *Entwicklung und Evaluation kontextorientierter Aufgaben in der Studieneingangsphase für Fach- und Nebenfachstudierende der Physik*
ISBN 978-3-8325-3801-9 43.00 EUR

173 Alexander Pusch: Fachspezifische Instrumente zur Diagnose und individuellen Förderung von Lehramtsstudierenden der Physik
ISBN 978-3-8325-3829-3 38.00 EUR

174 Christoph Vogelsang: Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. *Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*
ISBN 978-3-8325-3846-0 50.50 EUR

175 Ingo Brebeck: Selbstreguliertes Lernen in der Studieneingangsphase im Fach Chemie
ISBN 978-3-8325-3859-0 37.00 EUR

176 Axel Eghtessad: Merkmale und Strukturen von Professionalisierungsprozessen in der ersten und zweiten Phase der Chemielehrerbildung. *Eine empirisch-qualitative Studie mit niedersächsischen Fachleiter_innen der Sekundarstufenlehrämter*
ISBN 978-3-8325-3861-3 45.00 EUR

177 Andreas Nehring: Wissenschaftliche Denk- und Arbeitsweisen im Fach Chemie. Eine kompetenzorientierte Modell- und Testentwicklung für den Bereich der Erkenntnisgewinnung
ISBN 978-3-8325-3872-9 39.50 EUR

178 Maike Schmidt: Professionswissen von Sachunterrichtslehrkräften. Zusammenhangsanalyse zur Wirkung von Ausbildungshintergrund und Unterrichtserfahrung auf das fachspezifische Professionswissen im Unterrichtsinhalt „Verbrennung“
ISBN 978-3-8325-3907-8 38.50 EUR

179 Jan Winkelmann: Auswirkungen auf den Fachwissenszuwachs und auf affektive Schülermerkmale durch Schüler- und Demonstrationsexperimente im Physikunterricht
ISBN 978-3-8325-3915-3 41.00 EUR

180 Iwen Kobow: Entwicklung und Validierung eines Testinstrumentes zur Erfassung der Kommunikationskompetenz im Fach Chemie
ISBN 978-3-8325-3927-6 34.50 EUR

181 Yvonne Gramzow: Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion
ISBN 978-3-8325-3931-3 42.50 EUR

182 Evelin Schröter: Entwicklung der Kompetenzerwartung durch Lösen physikalischer Aufgaben einer multimedialen Lernumgebung
ISBN 978-3-8325-3975-7 54.50 EUR

183 Inga Kallweit: Effektivität des Einsatzes von Selbsteinschätzungsbögen im Chemieunterricht der Sekundarstufe I. *Individuelle Förderung durch selbstreguliertes Lernen*
ISBN 978-3-8325-3965-8 44.00 EUR

184 Andrea Schumacher: Paving the way towards authentic chemistry teaching. *A contribution to teachers' professional development*
ISBN 978-3-8325-3976-4 48.50 EUR

185 David Woitkowski: Fachliches Wissen Physik in der Hochschulausbildung. *Konzeptualisierung, Messung, Niveaubildung*
ISBN 978-3-8325-3988-7 53.00 EUR

186 Marianne Korner: Cross-Age Peer Tutoring in Physik. *Evaluation einer Unterrichtsmethode*
ISBN 978-3-8325-3979-5 38.50 EUR

187 Simone Nakoinz: Untersuchung zur Verknüpfung submikroskopischer und makroskopischer Konzepte im Fach Chemie
ISBN 978-3-8325-4057-9 38.50 EUR

188 Sandra Anus: Evaluation individueller Förderung im Chemieunterricht.*Adaptivität von Lerninhalten an das Vorwissen von Lernenden am Beispiel des Basiskonzeptes Chemische Reaktion*
ISBN 978-3-8325-4059-3 43.50 EUR

189 Thomas Roßbegalle: Fachdidaktische Entwicklungsforschung zum besseren Verständnis atmosphärischer Phänomene. *Treibhauseffekt, saurer Regen und stratosphärischer Ozonabbau als Kontexte zur Vermittlung von Basiskonzepten der Chemie*
ISBN 978-3-8325-4059-3 45.50 EUR

190 Kathrin Steckenmesser-Sander: Gemeinsamkeiten und Unterschiede physikbezogener Handlungs-, Denk- und Lernprozesse von Mädchen und Jungen
ISBN 978-3-8325-4066-1 38.50 EUR

191 Cornelia Geller: Lernprozessorientierte Sequenzierung des Physikunterrichts im Zusammenhang mit Fachwissenserwerb. *Eine Videostudie in Finnland, Deutschland und der Schweiz*
ISBN 978-3-8325-4082-1 35.50 EUR

192 Jan Hofmann: Untersuchung des Kompetenzaufbaus von Physiklehrkräften während einer Fortbildungsmaßnahme
ISBN 978-3-8325-4104-0 38.50 EUR

193 Andreas Dickhäuser: Chemiespezifischer Humor. *Theoriebildung, Materialentwicklung, Evaluation*
ISBN 978-3-8325-4108-8 37.00 EUR

194 Stefan Korte: Die Grenzen der Naturwissenschaft als Thema des Physikunterrichts
ISBN 978-3-8325-4112-5 57.50 EUR

195 Carolin Hülsmann: Kurswahlmotive im Fach Chemie. Eine Studie zum Wahlverhalten und Erfolg von Schülerinnen und Schülern in der gymnasialen Oberstufe
ISBN 978-3-8325-4144-6 49.00 EUR

196 Caroline Körbs: Mindeststandards im Fach Chemie am Ende der Pflichtschulzeit
ISBN 978-3-8325-4148-4 34.00 EUR

197 Andreas Vorholzer: Wie lassen sich Kompetenzen des experimentellen Denkens und Arbeitens fördern? *Eine empirische Untersuchung der Wirkung eines expliziten und eines impliziten Instruktionsansatzes*
ISBN 978-3-8325-4194-1 37.50 EUR

198 Anna Katharina Schmitt: Entwicklung und Evaluation einer Chemielehrerfortbildung zum Kompetenzbereich Erkenntnisgewinnung
ISBN 978-3-8325-4228-3 39.50 EUR

199 Christian Maurer: Strukturierung von Lehr-Lern-Sequenzen
ISBN 978-3-8325-4247-4 36.50 EUR

200 Helmut Fischler, Elke Sumfleth (Hrsg.): Professionelle Kompetenz von Lehrkräften der Chemie und Physik
ISBN 978-3-8325-4523-9 34.00 EUR

201 Simon Zander: Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen
ISBN 978-3-8325-4248-1 35.00 EUR

202 Kerstin Arndt: Experimentierkompetenz erfassen.
Analyse von Prozessen und Mustern am Beispiel von Lehramtsstudierenden der Chemie
ISBN 978-3-8325-4266-5 45.00 EUR

203 Christian Lang: Kompetenzorientierung im Rahmen experimentalchemischer Praktika
ISBN 978-3-8325-4268-9 42.50 EUR

204 Eva Cauet: Testen wir relevantes Wissen? *Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten*
ISBN 978-3-8325-4276-4 39.50 EUR

205 Patrick Löffler: Modellanwendung in Problemlöseaufgaben. *Wie wirkt Kontext?*
ISBN 978-3-8325-4303-7 35.00 EUR

206 Carina Gehlen: Kompetenzstruktur naturwissenschaftlicher Erkenntnisgewinnung im Fach Chemie
ISBN 978-3-8325-4318-1 43.00 EUR

207 Lars Oettinghaus: Lehrerüberzeugungen und physikbezogenes Professionswissen. *Vergleich von Absolventinnen und Absolventen verschiedener Ausbildungswege im Physikreferendariat*
ISBN 978-3-8325-4319-8 38.50 EUR

208 Jennifer Petersen: Zum Einfluss des Merkmals Humor auf die Gesundheitsförderung im Chemieunterricht der Sekundarstufe I.
Eine Interventionsstudie zum Thema Sonnenschutz
ISBN 978-3-8325-4348-8 40.00 EUR

209 Philipp Straube: Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik
ISBN 978-3-8325-4351-8 35.50 EUR

210 Martin Dickmann: Messung von Experimentierfähigkeiten. *Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*
ISBN 978-3-8325-4356-3 41.00 EUR

211 Markus Bohlmann: Science Education. Empirie, Kulturen und Mechanismen der Didaktik der Naturwissenschaften
ISBN 978-3-8325-4377-8 44.00 EUR

212 Martin Draude: Die Kompetenz von Physiklehrkräften, Schwierigkeiten von Schülerinnen und Schülern beim eigenständigen Experimentieren zu diagnostizieren
ISBN 978-3-8325-4382-2 37.50 EUR

213 Henning Rode: Prototypen evidenzbasierten Physikunterrichts. *Zwei empirische Studien zum Einsatz von Feedback und Blackboxes in der Sekundarstufe*
ISBN 978-3-8325-4389-1 42.00 EUR

214 Jan-Henrik Kechel: Schülerschwierigkeiten beim eigenständigen Experimentieren. *Eine qualitative Studie am Beispiel einer Experimentieraufgabe zum Hooke'schen Gesetz*
ISBN 978-3-8325-4392-1 55.00 EUR

215 Katharina Fricke: Classroom Management and its Impact on Lesson Outcomes in Physics. *A multi-perspective comparison of teaching practices in primary and secondary schools*
ISBN 978-3-8325-4394-5 40.00 EUR

216 Hannes Sander: Orientierungen von Jugendlichen beim Urteilen und Entscheiden in Kontexten nachhaltiger Entwicklung. *Eine rekonstruktive Perspektive auf Bewertungskompetenz in der Didaktik der Naturwissenschaft*
ISBN 978-3-8325-4434-8 46.00 EUR

217 Inka Haak: Maßnahmen zur Unterstützung kognitiver und metakognitiver Prozesse in der Studieneingangsphase. *Eine Design-Based-Research-Studie zum universitären Lernzentrum Physiktreff*
ISBN 978-3-8325-4437-9 46.50 EUR

218 Martina Brandenburger: Was beeinflusst den Erfolg beim Problemlösen in der Physik? *Eine Untersuchung mit Studierenden*
ISBN 978-3-8325-4409-6 42.50 EUR

219 Corinna Helms: Entwicklung und Evaluation eines Trainings zur Verbesserung der Erklärqualität von Schülerinnen und Schülern im Gruppenpuzzle
ISBN 978-3-8325-4454-6 42.50 EUR

220 Viktoria Rath: Diagnostische Kompetenz von angehenden Physiklehrkräften. *Modellierung, Testinstrumentenentwicklung und Erhebung der Performanz bei der Diagnose von Schülervorstellungen in der Mechanik*
ISBN 978-3-8325-4456-0 42.50 EUR

221 Janne Krüger: Schülerperspektiven auf die zeitliche Entwicklung der Naturwissenschaften
ISBN 978-3-8325-4457-7 45.50 EUR

222 Stefan Mutke: Das Professionswissen von Chemiereferendarinnen und -referendaren in Nordrhein-Westfalen. *Eine Längsschnittstudie*
ISBN 978-3-8325-4458-4 37.50 EUR

223 Sebastian Habig: Systematisch variierte Kontextaufgaben und ihr Einfluss auf kognitive und affektive Schülerfaktoren
ISBN 978-3-8325-4467-6 40.50 EUR

224 Sven Liepertz: Zusammenhang zwischen dem Professionswissen von Physiklehrkräften, dem sachstrukturellen Angebot des Unterrichts und der Schülerleistung
ISBN 978-3-8325-4480-5 34.00 EUR

225 Elina Platova: Optimierung eines Laborpraktikums durch kognitive Aktivierung
ISBN 978-3-8325-4481-2 39.00 EUR

226 Tim Reschke: Lesegeschichten im Chemieunterricht der Sekundarstufe I zur Unterstützung von situationalem Interesse und Lernerfolg
ISBN 978-3-8325-4487-4 41.00 EUR

227 Lena Mareike Walper: Entwicklung der physikbezogenen Interessen und selbstbezogenen Kognitionen von Schülerinnen und Schülern in der Übergangsphase von der Primar- in die Sekundarstufe. *Eine Längsschnittanalyse vom vierten bis zum siebten Schuljahr*
ISBN 978-3-8325-4495-9 43.00 EUR

228 Stefan Anthofer: Förderung des fachspezifischen Professionswissens von Chemielehramtsstudierenden
ISBN 978-3-8325-4498-0 39.50 EUR

229 Marcel Bullinger: Handlungsorientiertes Physiklernen mit instruierten Selbsterklärungen in der Primarstufe. *Eine experimentelle Laborstudie*
ISBN 978-3-8325-4504-8 44.00 EUR

230 Thomas Amenda: Bedeutung fachlicher Elementarisierungen für das Verständnis der Kinematik
ISBN 978-3-8325-4531-4 43.50 EUR

231 Sabrina Milke: Beeinflusst *Priming* das Physiklernen?
Eine empirische Studie zum Dritten Newtonschen Axiom
ISBN 978-3-8325-4549-4 42.00 EUR

232 Corinna Erfmann: Ein anschaulicher Weg zum Verständnis der elektromagnetischen Induktion. *Evaluation eines Unterrichtsvorschlags und Validierung eines Leistungsdiagnoseinstruments*
ISBN 978-3-8325-4550-5 49.50 EUR

233 Hanne Rautenstrauch: Erhebung des (Fach-)Sprachstandes bei Lehramtsstudierenden im Kontext des Faches Chemie
ISBN 978-3-8325-4556-7 40.50 EUR

234 Tobias Klug: Wirkung kontextorientierter physikalischer Praktikumsversuche auf Lernprozesse von Studierenden der Medizin
ISBN 978-3-8325-4558-1 37.00 EUR

235 Mareike Bohrmann: Zur Förderung des Verständnisses der Variablenkontrolle im naturwissenschaftlichen Sachunterricht
ISBN 978-3-8325-4559-8 52.00 EUR

236 Anja Schödl: FALKO-Physik – Fachspezifische Lehrerkompetenzen im Fach Physik. *Entwicklung und Validierung eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Physiklehrkräften*
ISBN 978-3-8325-4553-6 40.50 EUR

237 Hilda Scheuermann: Entwicklung und Evaluation von Unterstützungsmaßnahmen zur Förderung der Variablenkontrollstrategie beim Planen von Experimenten
ISBN 978-3-8325-4568-0 39.00 EUR

238 Christian G. Strippel: Naturwissenschaftliche Erkenntnisgewinnung an chemischen Inhalten vermitteln. *Konzeption und empirische Untersuchung einer Ausstellung mit Experimentierstation*
ISBN 978-3-8325-4577-2 41.50 EUR

239 Sarah Rau: Durchführung von Sachunterricht im Vorbereitungsdienst. *Eine längsschnittliche, videobasierte Unterrichtsanalyse*
ISBN 978-3-8325-4579-6 46.00 EUR

240 Thomas Plotz: Lernprozesse zu nicht-sichtbarer Strahlung. *Empirische Untersuchungen in der Sekundarstufe 2*
ISBN 978-3-8325-4624-3 39.50 EUR

241 Wolfgang Aschauer: Elektrische und magnetische Felder. *Eine empirische Studie zu Lernprozessen in der Sekundarstufe II*
ISBN 978-3-8325-4625-0 50.00 EUR

242 Anna Donhauser: Didaktisch rekonstruierte Materialwissenschaft. *Aufbau und Konzeption eines Schülerlabors für den Exzellenzcluster Engineering of Advanced Materials*
ISBN 978-3-8325-4636-6 39.00 EUR

243 Katrin Schüßler: Lernen mit Lösungsbeispielen im Chemieunterricht. *Einflüsse auf Lernerfolg, kognitive Belastung und Motivation*
ISBN 978-3-8325-4640-3 42.50 EUR

244 Timo Fleischer: Untersuchung der chemischen Fachsprache unter besonderer Berücksichtigung chemischer Repräsentationen
ISBN 978-3-8325-4642-7 46.50 EUR

245 Rosina Steininger: Concept Cartoons als Stimuli für Kleingruppendiskussionen im Chemieunterricht. *Beschreibung und Analyse einer komplexen Lerngelegenheit*
ISBN 978-3-8325-4647-2 39.00 EUR

246 Daniel Rehfeldt: Erfassung der Lehrqualität naturwissenschaftlicher Experimentalpraktika
ISBN 978-3-8325-4590-1 40.00 EUR

247 Sandra Puddu: Implementing Inquiry-based Learning in a Diverse Classroom: Investigating Strategies of Scaffolding and Students' Views of Scientific Inquiry
ISBN 978-3-8325-4591-8 35.50 EUR

248 Markus Bliersbach: Kreativität in der Chemie. *Erhebung und Förderung der Vorstellungen von Chemielehramtsstudierenden*
ISBN 978-3-8325-4593-2 44.00 EUR

249 Lennart Kimpel: Aufgaben in der Allgemeinen Chemie. *Zum Zusammenspiel von chemischem Verständnis und Rechenfähigkeit*
ISBN 978-3-8325-4618-2 36.00 EUR

250 Louise Bindel: Effects of integrated learning: explicating a mathematical concept in inquiry-based science camps
ISBN 978-3-8325-4655-7 37.50 EUR

251 Michael Wenzel: Computereinsatz in Schule und Schülerlabor. *Einstellung von Physiklehrkräften zu Neuen Medien*
ISBN 978-3-8325-4659-5 38.50 EUR

252 Laura Muth: Einfluss der Auswertephase von Experimenten im Physikunterricht. *Ergebnisse einer Interventionsstudie zum Zuwachs von Fachwissen und experimenteller Kompetenz von Schülerinnen und Schülern*
ISBN 978-3-8325-4675-5 36.50 EUR

253 Annika Fricke: Interaktive Skripte im Physikalischen Praktikum. *Entwicklung und Evaluation von Hypermedien für die Nebenfachausbildung*
ISBN 978-3-8325-4676-2 41.00 EUR

254 Julia Haase: Selbstbestimmtes Lernen im naturwissenschaftlichen Sachunterricht. *Eine empirische Interventionsstudie mit Fokus auf Feedback und Kompetenzerleben*
ISBN 978-3-8325-4685-4 38.50 EUR

255 Antje J. Heine: Was ist Theoretische Physik? *Eine wissenschaftstheoretische Betrachtung und Rekonstruktion von Vorstellungen von Studierenden und Dozenten über das Wesen der Theoretischen Physik*
ISBN 978-3-8325-4691-5 46.50 EUR

256 Claudia Meinhardt: Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern
ISBN 978-3-8325-4712-7 47.00 EUR

257 Ann-Kathrin Schlüter: Professionalisierung angehender Chemielehrkräfte für einen Gemeinsamen Unterricht
ISBN 978-3-8325-4713-4 53.50 EUR

258 Stefan Richtberg: Elektronenbahnen in Feldern. Konzeption und Evaluation einer webbasierten Lernumgebung
ISBN 978-3-8325-4723-3 49.00 EUR

259 Jan-Philipp Burde: Konzeption und Evaluation eines Unterrichtskonzepts zu einfachen Stromkreisen auf Basis des Elektronengasmodells
ISBN 978-3-8325-4726-4 59.50 EUR

Alle erschienenen Bücher können unter der angegebenen ISBN direkt online (http://www.logos-verlag.de) oder per Fax (030 - 42 85 10 92) beim Logos Verlag Berlin bestellt werden.

Studien zum Physik- und Chemielernen

Herausgegeben von Hans Niedderer, Helmut Fischler und Elke Sumfleth

Die Reihe umfasst inzwischen eine große Zahl von wissenschaftlichen Arbeiten aus vielen Arbeitsgruppen der Physik- und Chemiedidaktik und zeichnet damit ein gültiges Bild der empirischen physik- und chemiedidaktischen Forschung in Deutschland.

Die Herausgeber laden daher Interessenten zu neuen Beiträgen ein und bitten sie, sich im Bedarfsfall an den Logos-Verlag oder an ein Mitglied des Herausgeberteams zu wenden.

Kontaktadressen:

Prof. Dr. Hans Niedderer
Institut für Didaktik der Naturwissenschaften,
Abt. Physikdidaktik, FB Physik/Elektrotechnik,
Universität Bremen,
Postfach 33 04 40, 28334 Bremen
Tel. 0421-218 2484/4695, e-mail:
niedderer@physik.uni-bremen.de

Prof. Dr. Helmut Fischler
Didaktik der Physik, FB Physik, Freie Universität Berlin,
Arnimallee 14, 14195 Berlin
Tel. 030-838 56712/55966, e-mail:
fischler@physik.fu-berlin.de

Prof. Dr. Elke Sumfleth
Didaktik der Chemie,
Fachbereich Chemie,
Universität Duisburg-Essen,
Schützenbahn 70, 45127 Essen
Tel. 0201-183 3757/3761, e-mail:
elke.sumfleth@uni-essen.de

Die vorliegende Dissertation beschreibt den Entwicklungs- und Validierungsprozess eines neuen Testinstruments zur Erfassung von Lehrer-Selbstwirksamkeitserwartungen in vier physikdidaktischen Handlungsfeldern (Umgang mit Aufgaben, Experimentieren, Elementarisieren, Umgang mit Schülervorstellungen), jeweils bezüglich der Dimensionen Planung und Durchführung von Physikunterricht.

Die Instrumentenentwicklung erfolgt in hohem Maße theoriebasiert, indem unter anderem auf der Grundlage der sozialkognitiven Theorie Banduras Regeln für die Itemkonstruktion abgeleitet werden, die außerdem eine kritische Analyse existierender Instrumente erlauben. Der mehrschrittige Validierungsprozess ist argumentbasiert angelegt.

Im Rahmen einer Interviewstudie und einer Expertenbefragung werden vor allem inhaltliche Aspekte, wie z.B. die definitionskonforme Formulierung der Items, die Abdeckung des Handlungsfeldes oder die wahrgenommene Authentizität der Items, untersucht. Im Zuge der auf eine quantitative Auswertung angelegten Hauptstudie ($N \approx 1000$) werden vor allem psychometrische Aspekte, wie die Eindimensionalität der Skalen oder die Passung des Antwortformates, in den Blick genommen. Zusammenfassend kann ein übergeordnetes Validitätsargument in Bezug auf die intendierte Testwertinterpretation formuliert werden. Weiterer Forschungsbedarf deutet sich unter anderem bezüglich der Unterscheidung der zwei genannten Dimensionen an.

Logos Verlag Berlin

ISBN 978-3-8325-4712-7