pedocs
Open Access Erziehungswissenschaften

DIPF

Cauet, Eva

# Testen wir relevantes Wissen? Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten

*Berlin : Logos Verlag 2016, XIX, 275 S. - (Studien zum Physik- und Chemielernen; 204) - (Dissertation, Universität Duisburg-Essen, 2016)*

Quellenangabe/ Reference:
Cauet, Eva: Testen wir relevantes Wissen? Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten. Berlin : Logos Verlag 2016, XIX, 275 S. - (Studien zum Physik- und Chemielernen; 204) - (Dissertation, Universität Duisburg-Essen, 2016) - URN: urn:nbn:de:0111-pedocs-333624 - DOI: 10.25656/01:33362

*https://nbn-resolving.org/urn:nbn:de:0111-pedocs-333624*
*https://doi.org/10.25656/01:33362*

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/4.0/deed.de - Sie dürfen das Werk bzw. den Inhalt vervielfältigen, verbreiten und öffentlich zugänglich machen sowie Abwandlungen und Bearbeitungen des Werkes bzw. Inhaltes anfertigen, solange Sie den Namen des Autors/Rechteinhabers in der von ihm festgelegten Weise nennen und das Werk bzw. den Inhalt nicht für kommerzielle Zwecke verwenden.
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/4.0/deed.en - You may copy, distribute and render this document accessible, make adaptations of this work or its contents accessible to 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, provided that the work or its contents are not used for commercial purposes.
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.]

204

Eva Cauet

# Testen wir relevantes Wissen?

## Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten

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

**Dissertation**
**zur Erlangung des Doktorgrades der**
**Naturphilosophie**
**(Dr. phil. nat.)**

# Testen wir relevantes Wissen?

**– Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und *gutem* und *erfolgreichem* Unterrichten –**

vorgelegt von
Eva Cauet
geboren am 11.10.1985
in Dortmund

Datum der Disputation:
5. Februar 2016

Lehrstuhl für Didaktik der Physik
Fakultät für Physik
Universität Duisburg-Essen

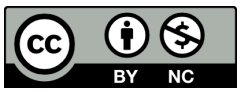
Erstgutachter: Prof. Dr. Hans E. Fischer
Zweitgutachter: Prof. Dr. Andreas Borowski

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 2016
Alle Rechte vorbehalten.

ISBN 978-3-8325-4276-4




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

## Zusammenfassung

Das Professionswissen von Lehrkräften wird seit Langem als wichtige Voraussetzung für gutes und erfolgreiches Unterrichten diskutiert. Empirisch abgesichert ist diese Annahme allerdings nicht. Schon auf theoretischer Ebene besteht weder Konsens über die Modellierung des Professionswissens noch herrscht Einigkeit bezüglich der Grundannahme über die Handlungsrelevanz explizierbaren Wissens. Im Zuge der Formulierung von Standards für die Lehrerbildung wurde die Entwicklung standardisierter Testinstrumente zur Erfassung des Fachwissens, fachdidaktischen Wissens und pädagogischen Wissens angehender oder ausgebildeter Lehrkräfte vorangetrieben. Derartige Testinstrumente werden meist über Expertenbefragungen, Abgleich mit Fachcurricula, den Vergleich bekannter Gruppen mit zu erwartenden Fähigkeitsunterschieden oder durch Zusammenhangsanalysen zwischen den Dimensionen des Professionswissens validiert. Genutzt werden sie oftmals auch um Aussagen über die Güte der Lehrerausbildung zu treffen – die Validität dieser Aussagen ist allerdings fraglich, sofern nicht gezeigt wird, dass das erhobene Wissen tatsächlich relevant für gutes und erfolgreiches Unterrichten ist.

Ziel der vorliegenden Arbeit ist die Überprüfung der prädiktiven Validität der im Rahmen des Projektes „Professionswissen in den Naturwissenschaften“ (ProwiN) entwickelten schriftlichen Testinstrumente zur Erfassung des Fachwissens, fachdidaktischen und pädagogischen Wissens von *Physik*lehrkräften in Bezug auf gutes und erfolgreiches Unterrichten. In einer quasiexperimentellen Feldstudie wurden Test-, Fragebogen- und Videodaten von 23 Gymnasiallehrkräften und ihren Klassen der Jahrgangsstufe 8/9 erhoben. Das Professionswissen der Lehrkräfte wurde in Bezug zu ihrem Unterrichtserfolg gesetzt, der über den Schülerfachwissenserwerb im Rahmen einer mehrmonatigen Unterrichtseinheit zur Mechanik sowie über das situationale Interesse der Lernenden in zwei, innerhalb dieser Einheit videographierten, Unterrichtsstunden modelliert wurde. Mehrebenenanalysen zeigten lediglich einen Zusammenhang zwischen dem pädagogischen Wissen der Lehrkräfte und dem Fachwissenserwerb der Lernenden. Als Maß für Unterrichtsqualität wurde die kognitiv aktivierende Gestaltung des videographierten Unterrichts beurteilt. Im Rahmen der Mehrebenenanalysen konnten Zusammenhänge zwischen der kognitiven Aktivierung und den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit gezeigt werden. Korrelationsanalysen zeigten signifikante Zusammenhänge zwischen dem Fachwissen sowie dem pädagogischen Wissen der Lehrkräfte und der kognitiv aktivierenden Gestaltung ihres Unterrichts. Das fachdidaktische Wissen der Lehrkräfte korrelierte nicht signifikant mit kognitiver Aktivierung. Bei der Interpretation dieser Ergebnisse müssen sowohl designbedingte als auch messtheoretische Einschränkungen sowie die geringe Stichprobengröße berücksichtigt werden – eindeutige Aussagen über die Relevanz des mit den ProwiN-Testinstrumenten gemessenen Wissens können auf Basis der Ergebnisse nicht getroffen werden. Die vorliegende Arbeit zeigt, wie wichtig – aber auch wie problematisch – die Untersuchung der Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg ist.

## Summary

Teachers' professional knowledge has long been discussed as an important precondition for good and successful teaching. However, this assumption has not been empirically verified yet. Even from a theoretical point of view, there is neither a consensus on how to model professional knowledge nor do researchers agree upon the question if a relationship between explicable knowledge and acting exists at all. In the course of formulating standards for teacher education, standardised test instruments—for measuring teachers' content knowledge, pedagogical content knowledge and pedagogical knowledge—have been developed. Those test instruments usually are validated by using expert ratings, by aligning content with subject-specific curricula, by verifying expected differences between different groups, or by analysing the correlations between dimensions of professional knowledge. Statements based on data gathered with such test instruments often include statements on the quality of teacher education; however, without proving that these tests measure knowledge which matters for good and successful teaching, the validity of those statements has to be questioned.

This thesis aims to examine the predictive validity of a written test on content knowledge, pedagogical content knowledge and pedagogical knowledge for *physics* teachers—developed in the scope of the project Professional Knowledge in Science (ProwiN)—regarding good and successful teaching. In a quasi-experimental field study, the test results, questionnaire responses and video data of 23 teachers teaching physics in grades eight and nine at grammar schools (Gymnasium) and their 13 to 15 years old students were gathered. In order to analyse the relationship between teachers' professional knowledge and teaching success, teachers' knowledge was related to their students' content knowledge gained within a several-month course on mechanics and to their students' situational interest in two videotaped lessons within this course. Multi-level analyses showed significant relationships only between teachers' pedagogical knowledge and students' content knowledge gains. As a measure of the quality of instruction, teacher actions supporting students' cognitive activation in the two videotaped lessons were rated. Multi-level analyses showed significant relationships between cognitively activating lesson designs and students' content knowledge gains. Teachers' content knowledge and pedagogical knowledge—but not teachers' pedagogical content knowledge—correlated significantly with the measure of cognitive activation. When interpreting these results, limitations due to study design, measurement problems and small sample size have to be taken into account, for example. Unambiguous statements on the relevance of the knowledge measured with the ProwiN test instruments cannot be made. This thesis shows how important—yet how problematic—the investigation of relationships between professional knowledge, quality of instruction and teaching success is.

## Danksagung

Vor fünf Jahren hat mich der Zufall aus der Fachphysik in die Fachdidaktik getragen (Danke an Willi Roer, den Architekten dieses Zufalls). Das „Professionswissen“ des Professionswissensforschers musste ich mir erst aneignen – Fachwissen alleine reichte bei weitem nicht aus. Doch Dank der fachdidaktischen und pädagogischen Expertise zahlreichen lieber Kollegen, die mich auf meinem Weg als Doktorandin begleitet haben, und Dank der Infrastruktur, die mir die nwu Essen in dieser Zeit geboten hat, fühlte ich mich schon bald mit den fachdidaktischen Fragestellungen ebenso vertraut wie zuvor mit den physikalischen.

Meinem Doktorvater Hans Fischer möchte ich dafür danken, dass er mich stets unterstützt, gefördert und gewähren lassen hat, immer hinter mir stand und es zudem geduldig ertrug, wenn ich auch noch das letzte Ergebnis meiner Arbeit hinterfragte.

Andreas Borowski danke ich dafür, dass er immer ein offenes Ohr für mich hatte, mich oftmals zurück auf die Spur brachte und immer um mein Wohlergehen bemüht war – und mir darüber hinaus mit Rat und Tat zur Seite stand.

Für spannende und leidenschaftliche Diskussionen möchte ich mich bei Sophie Kirschner, Cornelia Geller und Katharina Fricke bedanken – ihr habt dafür gesorgt, dass die Kaffeepausen nie langweilig wurden! Mein besonderer Dank gilt auch meinem Doktorzwilling Sven Liepertz sowie meiner ProwiN-Mitstreiterin Linda Lenske: Geteiltes Leid, ist halbes Leid! Eine tolle Zeit – sei es bei Tagungen oder Workshops, im Inland oder im Ausland – habe ich auch mit Silke Schiffhauer, Nora Stanke, Jenna Koenen, Luisa Friedrich, David Buschhüter, Florian Gigl und vielen anderen verbringen dürfen.

Ein ganz großer Dank geht zudem an meine Hilfskräfte Sarah van Vörden, Judith Janes, Evelin Mross, Julia Alwin, Daniel Wieltsch, Roman Lettmann, Ben Kisudi, Florian Gigl und Jenny Siegmund sowie an die tollen Lehrerinnen und Lehrer, die ihren Klassenzimmertüren für uns geöffnet haben: Ohne Euch und Sie alle wäre diese Studie nicht möglich gewesen.

Auch für die Unterstützung meines Homeoffice-Teams aus dem Café Asemann möchte ich mich bedanken: Es schreibt sich doch gleich viel leichter, wenn man so gut umsorgt wird! Vor allem Christian und Christof haben Sonnenschein in den Schreibprozess gebracht.

Danken möchten ich auch meinen Mädels, die mich immer herrlich abgelenkt haben und einfach die Besten sind!

Ohne meine Familie, wäre diese Arbeit jedoch nie zu einem Ende gekommen. Ihr habt mich immer wieder aufgefangen, bestärkt und beschützt! DANKE! Besonders meiner Mutter danke ich, die mir unermüdlich zur Seite gestanden hat und auch noch das zwanzigste Mal Korrektur gelesen hat.

Zum Schluss danke ich Christophe, der jeden Höhen- und Sturzflug dieser Arbeit mitgeflogen ist: Du bist mein bester Freund, bester Ratgeber, und der tollste Ehemann, den ich mir wünschen kann!

# Inhaltsverzeichnis


**1. Einleitung** **1**
1.1. Struktur der Arbeit . . . . . 2

**2. Professionswissen als Konstrukt in der Unterrichtsforschung** **5**
2.1. Von der Lehrerpersönlichkeit über Prozess-Produkt Modelle zu den Lehrerkognitionen . . . . . 5
2.2. Professionswissen als Bestandteil professioneller Handlungskompetenz 9
2.3. Ein Konstrukt – viele Modelle: Modellierung von Professionswissen 12
2.3.1. Fachwissen - CK . . . . . 15
2.3.2. Fachdidaktisches Wissen - PCK . . . . . 16
2.3.3. Pädagogisches Wissen - PK . . . . . 17

**3. Professionswissen als Voraussetzung für erfolgreiches und gutes Unterrichten** **21**
3.1. Hängen Wissen und Handeln zusammen? Eine kontroverse Diskussion . . . . . 22
3.2. Kriterien erfolgreichen Unterrichts . . . . . 24
3.3. Unterrichtsqualität . . . . . 25
3.3.1. Klassenführung . . . . . 27
3.3.2. Konstruktive Unterstützung . . . . . 27
3.3.3. Kognitive Aktivierung . . . . . 28
3.3.3.1. Merkmale eines kognitiv aktivierenden Unterrichts 29
3.3.3.2. Zusammenhang von kognitiv aktivierendem Unterricht und Zielkriterien von Unterricht . . . . . 31

**4. Herausforderungen in der empirischen Professionswissensforschung** **33**
4.1. Erfassung von Professionswissen . . . . . 33
4.2. Validität in der Professionswissensforschung . . . . . 36
4.3. Empirische Studien zur prädiktiven Validität von Professionswissenstests . . . . . 41

**5. Ableitung des eigenen Forschungsansatzes** **53**
5.1. Das „ProwiN“-Projekt . . . . . 54
5.1.1. Professionswissen in „ProwiN“ . . . . . 55
5.1.1.1. Fachwissen . . . . . 56
5.1.1.2. Fachdidaktisches Wissen . . . . . 57
5.1.1.3. Pädagogisches Wissen . . . . . 58
5.1.2. Validierung der „ProwiN“-Testinstrumente . . . . . 58
5.1.3. Ziele der ProwiN-Videostudie . . . . . 60

5.2. Auswahl der Kriterien für erfolgreiches Unterrichten: Fachwissenserwerb und situationales Interesse . . . . . 63
5.3. Auswahl eines Merkmals guten Unterrichts: Kognitive Aktivierung 65
5.3.1. Kognitive Aktivierung und Fachwissen der Lernenden . . . 67
5.3.2. Kognitive Aktivierung und situationales Interesse . . . . . 68
5.3.3. CK und kognitive Aktivierung . . . . . 69
5.3.4. PCK und kognitive Aktivierung . . . . . 71
5.3.5. PK und kognitive Aktivierung . . . . . 73
5.4. Einordnung der vorliegenden Studie in das ProwiN-Projekt . . . . 74

**6. Forschungsfragen und Hypothesen 75**
6.1. Forschungsfrage 1: Professionswissen und Unterrichtserfolg . . . . 76
6.2. Forschungsfrage 2: Professionswissen und Unterrichtsqualität . . . 77

**7. Methoden und Anlage der Studie 81**
7.1. Untersuchungsdesign . . . . . 81
7.2. Durchführung der Studie . . . . . 82
7.2.1. Auswahl der Jahrgangsstufe . . . . . 83
7.2.2. Teilnehmerakquise und Teilnahmeanreize . . . . . 84
7.2.3. Ablauf der Erhebungen . . . . . 85
7.2.3.1. Prä-Erhebung . . . . . 85
7.2.3.2. Post-Erhebung . . . . . 86
7.2.3.3. Video-Erhebung . . . . . 87
7.2.3.4. Zeitraum zwischen den Erhebungen . . . . . 88
7.2.4. Maßnahmen zur Sicherung der Datenqualität . . . . . 88
7.3. Stichprobe . . . . . 89
7.4. Statistische Methoden . . . . . 90
7.4.1. Allgemeine Hinweise zur Datenanalyse . . . . . 90
7.4.2. Die Rasch-Analyse . . . . . 93
7.4.3. Reliabilitätsberechnungen . . . . . 96
7.4.4. Beurteilung von Interrater-Übereinstimmungen . . . . . 98
7.4.5. Mehrebenenanalysen . . . . . 99
7.4.6. Messfehlerbereinigte Korrelationen . . . . . 101
7.5. Beschreibung der schriftlichen Erhebungsinstrumente . . . . . 102
7.5.1. Tests zur Messung des fachspezifischen Professionswissens 102
7.5.1.1. PCK-Test . . . . . 103
7.5.1.2. CK-Test . . . . . 103
7.5.1.3. Technische Details zur Auswertung . . . . . 104
7.5.1.4. Unterschiede zum Testinstrument aus ProwiN I . 105
7.5.1.5. Objektivität . . . . . 106
7.5.1.6. Reliabilität . . . . . 108
7.5.1.7. Validität . . . . . 109
7.5.2. Test zur Messung des pädagogischen Wissens . . . . . 112
7.5.2.1. Beschränkung der Auswertung auf den Test zum deklarativen Wissen . . . . . 112
7.5.2.2. PK-Test . . . . . 113

7.5.2.3. Technische Details zur Auswertung . . . . . 114
7.5.2.4. Objektivität . . . . . 115
7.5.2.5. Reliabilität . . . . . 115
7.5.2.6. Validität . . . . . 115
7.5.3. Schülerfachwissenstest . . . . . 116
7.5.3.1. Entwicklung und Pilotierung . . . . . 116
7.5.3.2. Technische Details zur Auswertung . . . . . 120
7.5.3.3. Objektivität . . . . . 122
7.5.3.4. Reliabilität . . . . . 122
7.5.3.5. Validität . . . . . 123
7.5.4. Fragebogen zum situationalen Interesse am Unterricht . . 128
7.5.4.1. Technische Details zur Auswertung . . . . . 129
7.5.4.2. Objektivität, Reliabilität, Validität . . . . . 129
7.5.5. Erhebung der Kontrollvariablen . . . . . 130
7.5.5.1. Kognitive Fähigkeiten der Lernenden . . . . . 132
7.5.5.2. Zuhause gesprochene Sprache der Lernenden . . 133
7.5.5.3. Unterrichtszeit . . . . . 133
7.5.5.4. Repräsentativität des videographierten Unterrichts 134
7.6. Beschreibung des videobasierten Ratinginstruments . . . . . 134
7.6.1. Rating zur kognitiven Aktivierung im Unterricht . . . . . 135
7.6.2. Unterschiede zum Paderborner Ratinginstrument . . . . . 135
7.6.3. Beschreibung des Ratertrainings . . . . . 137
7.6.4. Beschreibung des Ratingverfahrens . . . . . 138
7.6.5. Technische Details zur Auswertung . . . . . 139
7.6.6. Objektivität . . . . . 142
7.6.7. Reliabilität . . . . . 145
7.6.8. Validität . . . . . 146

**8. Ergebnisse** **159**
8.1. Deskriptive Ergebnisse . . . . . 159
8.1.1. Beschreibung der Lehrerstichprobe . . . . . 159
8.1.1.1. Demographischer Hintergrund und Lehrerfahrung 159
8.1.1.2. Professionswissen . . . . . 160
8.1.2. Beschreibung des Unterrichts . . . . . 162
8.1.2.1. Unterrichtszeit in der Unterrichtseinheit Mechanik 162
8.1.2.2. Kognitive Aktivierung im Unterricht . . . . . 163
8.1.3. Beschreibung der Schülerstichprobe . . . . . 163
8.1.3.1. Demographischer Hintergrund . . . . . 163
8.1.3.2. Fachwissensleistungen und kognitive Fähigkeiten 165
8.1.3.3. Situationales Interesse . . . . . 165
8.2. Fachwissenszuwachs der Lernenden . . . . . 168
8.3. Mehrebenenanalysen . . . . . 170
8.3.1. Prädiktoren für die Fachwissensleistungen im Posttest . . 171
8.3.1.1. Kontrollvariablenmodell (KV-Modell) . . . . . 171
8.3.1.2. Professionswissensmodelle (Modelle 1a-c) . . . . 173

8.3.1.3. Modelle zur kognitiven Aktivierung (Modelle $2.1a_{1M/2M/1M\&2M}$) . . . . . 174
8.3.2. Prädiktoren für das situationale Interesse der Lernenden . 177
8.3.2.1. Professionswissensmodelle (Modelle 1d-f) . . . . 177
8.3.2.2. Modelle zur kognitiven Aktivierung (Modelle $2.1b_{1M/2M}$) . . . . . 180
8.4. Professionswissen und kognitiv aktivierend gestalteter Unterricht . 181

**9. Diskussion und Ausblick 187**
9.1. Kurzzusammenfassung der Ergebnisse . . . . . 188
9.2. Voraussetzungen für eine valide Interpretation der Ergebnisse . . 189
9.2.1. Diskussion der internen Validität der Untersuchung . . . . 190
9.2.1.1. Diskussion der Messfehler . . . . . 192
9.2.2. Diskussion der externen Validität der Untersuchung . . . . 193
9.2.2.1. Fehler 1. Art . . . . . 194
9.2.2.2. Fehler 2. Art . . . . . 197
9.2.3. Diskussion der Bedeutsamkeit der Varianz im Unterrichtserfolg und in der Unterrichtsqualität . . . . . 199
9.2.3.1. Schülerfachwissen . . . . . 199
9.2.3.2. Situationales Interesse der Lernenden . . . . . 202
9.2.3.3. Kognitiv aktivierende Unterrichtsgestaltung . . . 204
9.3. Diskussion der zentralen Ergebnisse . . . . . 206
9.3.1. Fachwissen der Lehrkräfte . . . . . 206
9.3.2. Fachdidaktisches Wissen der Lehrkräfte . . . . . 208
9.3.3. Pädagogisches Wissen der Lehrkräfte . . . . . 210
9.4. Fazit und Ausblick . . . . . 211
9.4.1. Empfehlungen für künftige Untersuchungen . . . . . 213

**Appendizes**

**A. Manuale und Testhefte 217**

**B. Ergänzende Tabellen und Abbildungen 241**

**Literatur 255**

# Abbildungsverzeichnis


2.1. Angebots-Nutzungs-Modell . . . . . 8
2.2. Modell professioneller Handlungskompetenz . . . . . 10

5.1. ProwiN-Modell für das Professionswissen von Physiklehrkräften . 55

6.1. Forschungsfrage 1: Professionswissen und Unterrichtserfolg . . . . 76
6.2. Forschungsfrage 2: Professionswissen und Unterrichtsqualität . . . 78

7.1. Untersuchungsdesign . . . . . 83
7.2. Behandlung des Themas Kraft in verschiedenen Jahrgangsstufen . 84
7.3. Beispielhafte Anordnung der Videokameras im Klassenraum . . . 87
7.4. Beispielaufgabe PCK . . . . . 103
7.5. Beispielaufgabe PK . . . . . 113
7.6. Scatterplots für den Zusammenhang zwischen Klassenführung und kognitiver Aktivierung . . . . . 154

8.1. Fachwissenszuwächse zwischen Prä- und Post-Test . . . . . 169
8.2. Scatterplots für den Zusammenhang zwischen den Professionswissensdimensionen und kognitiver Aktivierung . . . . . 183

# Tabellenverzeichnis


2.1. Übersicht über die in Operationalisierungen von PCK einbezogenen Facetten . . . . . 18

5.1. Übersicht über die Ergebnisse aus ProwiN I zur Validierung der ProwiN-Professionswissenstests . . . . . 61

7.1. Beschreibung der Schülerstichprobe . . . . . 89
7.2. Kriterien zur Prüfung der Modellpassung im Rasch-Modell . . . . 96
7.3. Übersicht über die in ProwiN I und ProwiN II zur Berechnung der Lehrerfähigkeiten im fachspezifischen Professionswissen hinzugezogenen Aufgaben . . . . . 107
7.4. Reliabilität des CK- und PCK-Tests . . . . . 109
7.5. Korrelationen zwischen den Dimensionen des Professionswissens in der ersten und zweiten Projektphase . . . . . 111
7.6. Verteilung der Schülerfachwissenstestsaufgaben auf Testhefte und Unterthemen der Mechanik . . . . . 117
7.7. Beschreibung der Pilotierungsstichprobe für den Schülerfachwissenstest . . . . . 118
7.8. Reliabilität des Schülerfachwissenstest . . . . . 122
7.9. Von den Lehrkräften behandelte Unterthemen der Mechanik . . . 125
7.10. Korrelationen zwischen den Prä- und Post-Testwerten und den kognitiven Fähigkeiten der Lernenden . . . . . 126
7.11. Korrelationen zwischen den Prä- und Post-Testwerten und den Schulnoten der Lernenden . . . . . 127
7.12. Korrelationen zwischen den Maßen für das situationale Interesse der Lernenden in der 1. und 2. Unterrichtsstunde . . . . . 131
7.13. Reliabilität des Kognitive Fähigkeitentests . . . . . 133
7.14. Subskalen und Handlungsindikatoren zur Beurteilung der kognitiven Aktivierung . . . . . 136
7.15. Rekodierte Missings in den Handlungsindikatoren des Ratings zur kognitiven Aktivierung und maximale Fehler auf die Qualitätsmaße 141
7.16. Auffällige Handlungsindikatoren im Rating zur kognitiv aktivierenden Gestaltung der 1. und 2. Unterrichtsstunden . . . . . 142
7.17. Interrater-Übereinstimmung für die Subskalenmittelwerte und den Gesamtskalenmittelwert zur kognitiven Aktivierung für die 1. und 2. Unterrichtsstunde . . . . . 144
7.18. Reliabilität des Ratings zur kognitiven Aktivierung . . . . . 145
7.19. Korrelationen zwischen den Qualitätsmaßen zur kognitiv aktivierenden Gestaltung der 1. und 2. Unterrichtsstunde . . . . . 151

7.20. Korrelationen zwischen den Subskalenmittelwerten und den Subskalengesamteindrücken in der 1. und 2. Unterrichtsstunde . . . . . 152
7.21. Korrelationen zwischen kognitiver Aktivierung und Klassenführung bzw. Vernetztheit der Sachstruktur im Unterricht . . . . . 156

8.1. Vergleich ProwiN I/ProwiN II: Demographischer Hintergrund und Lehrerfahrung der Lehrkräfte . . . . . 160
8.2. Vergleich ProwiN I/ProwiN II: Professionswissen der Lehrkräfte . 161
8.3. Unterschiede zwischen den CK-, PCK- und PK-Testwerten der ProwiN I- und ProwiN II-Lehrkräften . . . . . 162
8.4. Korrelationen zwischen den Dimensionen des Professionswissens in der Stichprobe der ProwiN II-Lehrkräfte . . . . . 162
8.5. Deskriptive Statistik für die Qualitätsmaße zur kognitiven Aktivierung 164
8.6. Deskriptive Statistik für die Fachwissensleistungen und die kognitiven Fähigkeiten der Lernenden . . . . . 166
8.7. Deskriptive Statistik für das situationale Interesse der Lernenden am Unterricht . . . . . 167
8.8. Ergebnisse der Mehrebenenregressionen auf die Post-Testwerte der Lernenden im Fachwissen . . . . . 176
8.9. Residualvarianzen in den Nullmodellen für das situationale Interesse der Lernenden . . . . . 178
8.10. Ergebnisse der Mehrebenenregressionen auf das situationale Interesse der Lernenden . . . . . 179
8.11. Korrelationen zwischen dem Professionswissen der Lehrkräfte und kognitiver Aktivierung . . . . . 185

# Abkürzungsverzeichnis

## Terminologie

| | |
|---|---|
| 1M/2M | 1./2. Unterrichtsstunde Mechanik |
| 1M&2M | 1. und 2. Unterrichtsstunde Mechanik; Index für über beide Unterrichtsstunden gemittelte Maße |
| CK | Fachwissen (Content Knowledge) |
| DIF | Differential Item Functioning |
| FAM | Fragebogen zur aktuellen Motivation |
| F | Forschungsfrage |
| GL | Gymnasiales Lehramt |
| GS | Gesamtskala |
| GyGe | Gymnasien und Gesamtschulen |
| Gym | Gymnasium |
| HR | Haupt- und Realschulen |
| HS | Hauptschule |
| H | Hypothese |
| ICC | Intraklassenkorrelation |
| ID | Identifikationsnummer |
| Ind. | (Handlungs-)Indikatoren |
| IRT | Item Response Theory |
| JS | Jahrgangsstufe |
| KA | Kognitive Aktivierung |
| KF | Klassenführung |
| KFT | Kognitive Fähigkeitentest |

| | |
|---|---|
| KOSM | Wissen über Schülerfehlvorstellungen (Knowledge of Students Misconceptions) |
| KV | Kontrollvariablen |
| LiV | Lehrkräfte im Vorbereitungsdienst |
| LK | Lehrkräfte |
| LZW | Leistungszuwachs |
| MLR | Maximum-Likelihood mit robusten Standardfehlern |
| n.b. | nicht beurteilbar |
| NRW | Nordrhein-Westfalen |
| n.s | nicht signifikant |
| n.u. | nicht untersucht |
| NW | Naturwissenschaften |
| OBAS | Ordnung zur berufsbegleitenden Ausbildung von Seiteneinsteigerinnen und Seiteneinsteigern und der Staatsprüfung |
| PCK | Fachdidaktisches Wissen (Pedagogical Content Knowledge) |
| PCK-CxK | Contextual Knowledge |
| Ph | Physik |
| PK | Pädagogisches Wissen (Pedagogical Knowledge) |
| $PK_D$ | Deklaratives pädagogisches Wissen |
| $PK_{KP}$ | Konditional-prozedurales pädagogisches Wissen |
| Sit. Interesse | Situationales Interesse |
| SS | Subskala |
| St. | (Unterrichts-)Stunde |
| StdY/StdYX | Index für halb-/vollstandardisierte Steigungskoeffizienten in den Mehrebenenmodellen |
| SuS | Schülerinnen und Schüler |
| TH | Testheft |
| UML | Unbedingte Maximum-Likelihood |
| V | Vernetzung der Sachstruktur |

## Organisationen

| | |
|---|---|
| AERA | American Educational Research Association |
| APA | American Psychological Association |
| BMBF | Bundesministerium für Bildung und Forschung |
| KMK | Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland |
| MSW | Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen |
| NCME | National Council on Measurement in Education |
| OECD | Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (Organization for Economic Cooperation and Development) |

## Forschungsprogramme und -projekte

| | |
|---|---|
| COACTIV | Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz |
| COACTIV-R | Ergänzungsstudie COACTIV-Referendariat |
| KiL | Messung professioneller Kompetenzen in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen |
| MT21 | Mathematics Teaching in the 21st Century |
| PISA | Programme for International Student Assessment |
| PLUS | Professionswissen von Lehrkräften, naturwissenschaftlicher Unterricht und Zielerreichung im Übergang von der Primar- zur Sekundarstufe |
| Profile-P | Professionswissen in der Lehramtsausbildung Physik |
| ProPäda | Entwicklung von Professionalität des pädagogischen Personals in Bildungseinrichtungen |
| ProwiN | Professionswissen in den Naturwissenschaften |
| Pythagoras | Unterrichtsqualität und mathematisches Verständnis in verschiedenen Unterrichtskulturen |
| QuiP | Quality of Instruction in Physics |

| | |
|---|---|
| SII | Study of Instructional Improvement |
| TEDS-M | The Teacher Education Study in Mathematics |
| TIMSS | Third International Mathematics and Science Study |

## Symbole

| | |
|---|---|
| $\alpha_{\mathrm{C}}$ | Reliabilitätskoeffizient Cronbachs Alpha |
| $\beta^{\mathrm{StdY/StdYX}}$ | Halb-/Vollstandardisierter Steigungskoeffizient für einen Prädiktor auf Schülerebene in einer Mehrebenenregression |
| $\gamma^{\mathrm{StdYX}}$ | Vollstandardisierter Steigungskoeffizient für einen Prädiktor auf Klassenebene in einer Mehrebenenregression |
| $d$ | Effektstärke Cohens $d$ |
| $df$ | Anzahl der Freiheitsgerade (Degrees of Freedom) |
| *DIF.Contrast* | Statistik des Differential Item Functioning: Differenz der in zwei unterschiedlichen Personengruppen bestimmten Aufgabenschwierigkeiten |
| $ICC_{\text{2-fakt.,unjust}}$ | Unjustierte Intraklassenkorrelation als Übereinstimmungsmaß für die Skalenwerte zufällig gezogener Rater im zweifaktoriellen Modell |
| $ICC_{\text{1-fakt.,unjust}}$ | Unjustierte Intraklassenkorrelation im einfaktoriellen Modell als Maß für den Anteil der zwischen den Klassen liegende Varianz von Testwerten an der Gesamtvarianz der Testwerte |
| $\mathrm{KI}_{95\,\%}$ | 95 %-Konfidenzinterval |
| $M$ | Mittelwert |
| *Max* | Maximum |
| *Min* | Minimum |
| *MnSq* | Mean-Square-Statistik im Rasch-Modell |
| $N$ | Stichprobengröße oder Anzahl |
| $p$ | Signifikanzwert bei zweiseitiger Testung |
| $p_{\text{1-seitig}}$ | Signifikanzwert bei einseitiger Testung |
| $R^2$ | Determinationskoeffizient, Maß für die aufgeklärte Varianz |
| $r$ | Korrelation |

| | |
|---|---|
| $r_{\mathrm{MW}}$ | Effektstärke für den Mann-Whitney-U-Tests |
| $r_{\mathrm{Pearson}}$ | Pearson-Korrelationskoeffizient |
| $r_{\mathrm{Spearman}}$ | Spearman-Rangkorrelationskoeffizient |
| $r_{\mathrm{W}}$ | Effektstärke für den Wilkoxon-Vorzeichen-Rang-Test |
| $\sigma$ | Fehler auf einen Wert |
| $SD$ | Standardabweichung |
| $\tau_{\mathrm{Kendall}}$ | Rangkorrelationskoeffizient nach Kendall |
| $t$ | Statistik des t-Tests |
| $T$ | Statistik des Wilkoxon-Vorzeichen-Rang-Tests |
| $U$ | Statistik des Mann-Whitney-U-Tests |
| $\Phi$ | Korrelationsmaß zwischen zwei dichotomen Merkmalen |
| $W$ | Statistik des Shapiro-Wilk-Tests |
| $z$ | z-standardisierter Wert einer Statistik oder Maßzahl |
| $Z$ | z-standardisierte Differenz zwischen zwei Korrelationen |
| ♀ | Weiblich |
| ♂ | Männlich |

# 1. Einleitung

Das Interesse am Professionswissen von Lehrkräften erwächst aus der Grundannahme über die Relevanz von Professionswissen für gutes und erfolgreiches Unterrichten, die auch der universitären Lehrerausbildung zugrunde liegt. Fachwissen, fachdidaktisches Wissen und pädagogisches Wissen werden als Teil der professionellen Handlungskompetenz von Lehrkräften angesehen (Baumert & Kunter, 2011, S. 32). Was genau Lehrkräfte wissen müssen, um erfolgreich unterrichten zu können und inwieweit das im Rahmen der Lehrerausbildung vermittelte Wissen als handlungsleitend für die Unterrichtspraxis angenommen werden kann, ist allerdings weitestgehend ungeklärt. Schon auf theoretischer Ebene herrscht in dieser Frage Uneinigkeit, was sich in der Heterogenität der Modellierungen des Professionswissens von Lehrkräften und der Operationalisierungen der Professionswissensdimensionen widerspiegelt (vergl. z. B. Baumert & Kunter, 2006, S. 481; Kirschner, 2013, S. 8). Darüber hinaus ist auch der grundsätzliche Zusammenhang zwischen explizierbarem Wissen und Handeln Gegenstand kontroverser theoretischer Diskussionen (Kolbe, 2004). Ein zentrales Forschungsdesiderat der Professionswissensforschung ist daher die Untersuchung der Zusammenhänge zwischen Professionswissen und gutem und erfolgreichem Unterrichten (Abell, 2007, S. 1134; Abell, 2008, S. 1412; Borowski et al., 2010, S. 344; Fischler, 2008, S. 46; Gess-Newsome, 2013, S. 259).

In Deutschland wächst die Anzahl an Instrumenten, die das Professionswissen von Lehrkräften, Referendaren oder Lehramtsstudierenden schriftlich erfassen sollen (z. B. Blömeke et al., 2010; Brovelli, Bölsterli, Rehm & Wilhelm, 2013; Kirschner, 2013; Krauss, Neubrand et al., 2008; Kröger, Neumann & Petersen, 2015; Riese, 2009; Riese et al., 2015; Schmelzing, 2010). Die Entwicklung derartiger Testinstrumente ist nicht zuletzt auch mit dem Ziel verbunden, die Wirksamkeit der Lehrerausbildung überprüfen zu können. In der Regel basieren diese Testinstrumente auf in der Forschungsgemeinschaft breit akzeptierten – aber dennoch normativ gesetzten – Wissensfacetten. Ob die Testinstrumente das Professionswissen von (angehenden) Lehrkräften valide erfassen, wird meist über Expertenbefragungen, Abgleich mit Fachcurricula, Analysen der Zusammenhänge zwischen den Professionswissensdimensionen oder den Vergleich bekannter Gruppen mit zu erwartenden Fähigkeitsunterschieden überprüft. Valide Aussagen, ob die im Rahmen der Ausbildung gelehrten Inhalte auch wirklich gelernt werden, können auf Basis solcher Testinstrumente getroffen werden. Ziel der Lehrerausbildung ist allerdings die Ausbildung guter und erfolgreich unterrichtender Lehrkräfte. Solange nicht gezeigt wird, dass die Grundannahme über die Handlungsrelevanz von Professionswissen gerechtfertigt ist und das mit solch einem Instrument erfasste Wissen mit Unterrichtsqualität oder Unterrichtserfolg in Zusammenhang steht, kann das ge-

messene Wissen allerdings nicht als Handlungsressource für gutes und erfolgreiches Unterrichten angenommen werden.

Die Frage, die in diesem Kontext gestellt werden sollte, lautet daher: Testen wir eigentlich relevantes Wissen? Hinter dieser Frage verbirgt sich zum einen die Frage nach der Validität der Modellierung von Professionswissen als Voraussetzung für erfolgreiches und gutes Unterrichten und zum anderen die Frage nach der Validität der Testinstrumente in diesem Zusammenhang.

Die hier vorgestellte Studie wurde im Rahmen der zweiten Projektphase des vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmenprogramm „Entwicklung von Professionalität des pädagogischen Personals in Bildungseinrichtungen“ (ProPäda) geförderten fächerübergreifenden Projekts „Professionswissen in den Naturwissenschaften“ (ProwiN) im Fach Physik durchgeführt (Borowski et al., 2010). In der ersten Projektphase wurden Testinstrumente zur Erfassung des Fachwissens, fachdidaktischen Wissens und pädagogischen Wissens von Lehrkräften der Naturwissenschaften in den Fächern Physik, Chemie und Biologie entwickelt und – zunächst ohne Bezug zu Unterrichtsqualität oder Unterrichtserfolg – auf die zuvor beschriebene Weise validiert. In der zweiten Projektphase wurden diese Instrumente im Rahmen von Unterrichtsanalysen eingesetzt.

Ziel der vorliegenden Arbeit ist es, die oben gestellte Frage für die ProwiN-Testinstrumente zur Erfassung des Professionswissens von *Physik*lehrkräften durch die Überprüfung der prädiktiven Validität dieser Testinstrumente in Bezug auf gutes und erfolgreiches Unterrichten zu beantworten. Hierfür werden Zusammenhänge zwischen dem mit den ProwiN-Testinstrumenten erfassten Fachwissen, dem fachdidaktischen Wissen und dem pädagogischen Wissen von Physiklehrkräften, der kognitiv aktivierenden Gestaltung ihres Unterrichts (als Merkmal *guten* Unterrichts) und dem Fachwissenserwerb sowie dem situationalen Interesse ihrer Schülerinnen und Schüler (als Kriterien *erfolgreichen* Unterrichts) untersucht.

## 1.1. Struktur der Arbeit

In den Kapiteln 2 bis 6 werden der theoretische Hintergrund der Arbeit aufgearbeitet, der aktuelle Forschungsstand beschrieben und die Forschungsfragen und Hypothesen abgeleitet.

Kapitel 2 geht auf das Professionswissen von Lehrkräften als Konstrukt in der Unterrichtsforschung ein. Ziel dieses Kapitels ist es, herauszuarbeiten, welche Entwicklungen in der Unterrichtsforschung zu der Annahme führten, dass das Professionswissen von Lehrkräften eine wichtige Voraussetzung für gutes und erfolgreiches Unterrichten darstellt und damit einen zentralen Bestandteil der professionellen Handlungskompetenz von Lehrkräften bildet. Darüber hinaus soll aufgezeigt werden, wie wenig Konsens darüber herrscht, wie Professionswissen zu modellieren ist und welches Wissen aus theoretischer Sicht als notwendig für erfolgreiches und gutes Unterrichten angenommen werden kann.

Da zudem auch keine Einigkeit darüber herrscht, ob grundsätzlich ein Zusammenhang zwischen explizierbarem Wissen und Handeln besteht, werden in Kapitel 3 zunächst die unterschiedlichen Positionen hierzu vorgestellt. Hiermit soll deutlich

gemacht werden, dass die Annahme über die Bedeutung des Professionswissens für qualitätsvolles Unterrichten der Überprüfung bedarf. Die Überprüfbarkeit setzt allerdings eine Definition erfolgreichen und guten Unterrichtens voraus. Daher werden in diesem Kapitel Zielkriterien für Unterrichtserfolg formuliert und Merkmale der Unterrichtsqualität vorgestellt. Insbesondere wird auf die kognitiv aktivierende Gestaltung des Unterrichts eingegangen.

In Kapitel 4 werden die Herausforderungen thematisiert, mit denen sich die Professionswissensforschung bei der Erfassung von Professionswissen auseinandersetzen muss. Insbesondere wird die Problematik diskutiert, dass ohne Überprüfung der prädiktiven Validität von Testinstrumenten zur Erfassung des Professionswissens in Bezug auf Unterrichtsqualität nicht davon ausgegangen werden kann, dass Wissen erfasst wird, das als handlungsrelevant für erfolgreiches und gutes Unterrichten angesehen werden kann. Über die Vorstellung der heterogenen Ergebnisse der wenigen empirischen Studien, die Zusammenhänge zwischen dem Professionswissen von Lehrkräften und gutem und erfolgreichem Unterricht untersuchen, wird deutlich gemacht, dass insbesondere für den Physikunterricht noch nicht hinreichend geklärt ist, welches Wissen als unterrichtsrelevant angenommen werden kann.

In Kapitel 5 wird der Forschungsansatz abgeleitet, den die vorliegende Arbeit zur Untersuchung der Zusammenhänge zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten wählt. Zunächst wird das ProwiN-Projekt näher beschrieben und anschließend die Auswahl des Fachwissenserwerbs und des situationalen Interesses der Lernenden als Kriterien erfolgreichen Unterrichts sowie die Auswahl der kognitiven Aktivierung als Merkmal guten Unterrichts ausführlich begründet. Den Abschluss der theoretischen Aufarbeitung bildet Kapitel 6, in dem die Forschungsfragen und Hypothesen formuliert werden.

Der empirische Teil der vorliegenden Arbeit beginnt mit Kapitel 7 zu Methoden und Anlage der Studie. In diesem Kapitel werden zunächst das Design, die Durchführung und die Stichprobe der vorliegenden Studie beschrieben. Anschließend werden die in der vorliegenden Arbeit angewendeten statistischen Methoden erläutert und die Testinstrumente zur Erfassung des Professionswissens, des Schülerfachwissens, des situationalen Interesses der Lernenden und der Kontrollvariablen sowie das Videoinstrument zur Erfassung der kognitiv aktivierenden Gestaltung des Unterrichts vorgestellt und deren Güte ausführlich diskutiert.

Die Ergebnisse der vorliegenden Studie werden in Kapitel 8 dargestellt. Kapitel 9 bildet den Abschluss der Arbeit. In diesem Kapitel werden die Ergebnisse zusammengefasst und diskutiert. Es wird ein besonderes Augenmerk auf die Voraussetzungen für eine valide Interpretation der Ergebnisse gelegt und der Beitrag der vorliegenden Arbeit für den wissenschaftlichen Diskurs diskutiert.

# 2. Professionswissen als Konstrukt in der Unterrichtsforschung

Das Professionswissen von Lehrkräften wird als wichtige Voraussetzung für erfolgreiches Unterrichten angesehen (vergl. z. B. Abell, 2007; Fischer, Borowski & Tepner, 2012; Peterson, Carpenter & Fennema, 1989; Shulman, 1987). Zunächst handelt es sich hierbei allerdings um eine Annahme – die Relevanz des Professionswissens für gutes und erfolgreiches Unterrichten ist nicht hinreichend empirisch abgesichert. Um nachvollziehen zu können, warum dennoch weitestgehend Einigkeit über die Gültigkeit dieser Annahme herrscht, wird in diesem Kapitel zunächst beschrieben, wie es dazu kam, dass das Professionswissen von Lehrkräften in den Fokus der empirischen Bildungsforschung rückte und inzwischen als fester Bestandteil der professionellen Handlungskompetenz von Lehrkräften gilt. Anschließend soll geklärt werden, was sich hinter dem Konstrukt „Professionswissen“ verbirgt. Hier soll insbesondere die Heterogenität in der Modellierung des Professionswissens von Lehrkräften deutlich gemacht werden – so besteht zwar Konsens darüber, dass das Professionswissen von Lehrkräften kein eindimensionales Konstrukt darstellt, allerdings herrscht weder Einigkeit über die Anzahl der Dimensionen des Professionswissens noch darüber, wie diese zu operationalisieren sind. Letzteres ist für die vorliegende Arbeit insofern von Relevanz, als dass selbst theoretisch nicht geklärt ist, welches Wissen als notwendig für erfolgreiches und gutes Unterrichten angenommen werden kann.

## 2.1. Von der Lehrerpersönlichkeit über Prozess-Produkt Modelle zu den Lehrerkognitionen

Historisch gesehen, steht dem heutigen Begriff der Lehrer*aus*bildung, der impliziert, dass die für erfolgreiches Unterrichten erforderlichen Kenntnisse, Fähigkeiten und Fertigkeiten erlernbar sind, der Begriff der Lehrerbildung entgegen, der eher auf Persönlichkeitsmerkmale von Lehrkräften und deren Weiterentwicklung fokussiert (Blömeke, 2009). Diese Sichtweise auf die Lehrkraft spiegelt sich im *Persönlichkeitsparadigma* der Unterrichtsforschung wider. In den 1950er und 1960er Jahren konzentrierte sich die empirische Unterrichtsforschung auf die Untersuchung der Zusammenhänge zwischen allgemeinen Persönlichkeitsmerkmalen von Lehrkräften und Schülervariablen, wie beispielsweise Lernerfolg (vergl. z. B. Getzels & Jackson, 1963). Das Problem vieler Forschungsarbeiten in diesem Bereich waren allerdings entweder triviale oder inkonsistente Ergebnisse zum Einfluss der Per-

sönlichkeitsmerkmale auf Schülervariablen (Bromme, 1997; Bromme & Rheinberg, 2006; Helmke, 2009). Auch kann die Persönlichkeitsforschung lediglich Aufschluss darüber geben, welche Eigenschaften angehende Lehrkräfte zum erfolgreichen Unterrichten benötigen; sie eröffnet damit aber keine Perspektiven für die Ausbildung von erfolgreich Unterrichtenden.

Im Zuge des *Prozess-Produkt-Paradigmas* richtete die Unterrichtsforschung daher den Blick auf den Unterricht und das Lehrerhandeln. Es wurde untersucht, welche Verhaltensweisen von Lehrkräften, unabhängig vom Unterrichtsfach, einen direkten Einfluss auf Schülervariablen hatten. Die Forschungsarbeiten in diesem Bereich lieferten viele wichtige Erkenntnisse (vergl. z. B. Brophy & Good, 1986; Rosenshine, 1983), indem sie lernwirksame Unterrichtsmerkmale identifizierten und damit ein Fundament für das heutige Verständnis erfolgreichen Unterrichtens legten (vergl. z. B. Fischer, Labudde, Neumann & Viiri, 2014b; Helmke, 2009; Klieme & Rakoczy, 2008). Nach Bromme und Rheinberg (2006, S. 301-302) lassen sich die wichtigsten empirischen Ergebnisse für erfolgreiche Lehrerverhaltensweisen (bezogen auf die Lernleistung als Erfolgskriterium) wie folgt zusammenfassen: Erfolgreiche Lehrkräfte verfügen über ein reichhaltiges, flexibel einsetzbares Repertoire an Methoden, aktivieren die Lernenden und geben ihnen kontinuierlich die Möglichkeit Erfolgserfahrungen zu sammeln; sie nutzen die Unterrichtszeit effektiv aus, stimmen Tempo und Abfolge der Beschäftigung mit dem Unterrichtsgegenstand auf die Lernenden ab, teilen den Lernenden bei Gruppenarbeiten angemessen Aufgaben zu und strukturieren den Gruppenarbeitsprozess; sie äußern sich klar und konsistent und machen ihre jeweiligen Ziele sowie die Struktur des Unterrichts transparent, erkennen mögliche Störungen und beugen ihnen vor, gestalten fließende Übergänge zwischen Unterrichtsthemen oder -methoden und übermitteln den Lernenden glaubhaft eine optimistische Haltung in Bezug auf deren Lernfähigkeiten.

Das Prozess-Produkt-Paradigma ließ allerdings den Lerner als aktiven Akteur im Unterrichtsgeschehen außer Acht. Darüber hinaus vernachlässigte es die durch den Fachinhalt bedingten Unterschiede in der Wirksamkeit von Lehrerverhaltensweisen und dass „[die] Wirkung einzelner Lehrerverhaltensweisen in erheblichem Maße von der Abstimmung des Lehrerverhaltens auf die konkrete Situation (Unterschiede zwischen den Schülern, Unterrichtsinhalt, didaktische Intention, verfügbare Medienarrangements) abhängt“ (Bromme, 2008, S. 160). Damit werden zentrale Herausforderungen des Lehrerberufs im Prozess-Produkt-Paradigma nicht beschrieben. Außerdem werden kognitive Strukturen und Prozesse nicht berücksichtigt, die die Adaptivität und Flexibilität des Handelns erst ermöglichen (Bromme, 2008, S. 161). Bromme und Rheinberg (2006, S. 302-303) nennen als Beispiel für die Notwendigkeit den Fachinhalt und das professionelle Wissen der Lehrkräfte in Betrachtungen zur Unterrichtsqualität einzubeziehen, die Auswahl angemessener Aufgaben für unterschiedliche Schülergruppen. Ohne fachliches und fachdidaktischen Wissen kann nicht konkretisiert werden, was als „angemessen“ gelten kann. Des Weiteren wurde in Prozess-Produkt-Modellen nicht berücksichtigt, dass das Handeln der Lehrkraft zwar die sichtbaren Verhaltensweisen der Lernenden im Unterricht direkt beeinflussen kann, der Einfluss auf Lernleistung und Verstehen aber lediglich indirekt erfolgt (Bromme, 1997, S. 186).

Als Synthese aus dem Persönlichkeitsparadigma und dem Prozess-Produkt-Paradigma entwickelte sich daher das *Expertenparadigma*, in dem Lehrkräfte als „kompetente Fachleute für die Kunst des Unterrichtens“ mit ihrem Wissen und Handeln wieder mehr im Fokus standen (Bromme, 1997, S. 186). Analog zum Persönlichkeitsparadigma wird nach erfolgreichen Lehrkräften gesucht, die aber nicht mehr über Persönlichkeitseigenschaften, sondern vielmehr über ihr professionelles Wissen und Können und ihre subjektiven und intuitiven Theorien zum Lehren und Lernen charakterisiert werden (Helmke, 2009, S. 49). Über die vergleichende Analyse des Unterrichtens und des Wissens von erfolgreichen und weniger erfolgreichen Lehrpersonen oder erfahrenen und unerfahrenen Lehrkräften, können so Erkenntnisse über Lehrerwissen, -einstellungen, -wahrnehmungen und -handlungen gewonnen werden, die für die Bewältigung beruflicher Anforderungen notwendig sind (Bromme & Rheinberg, 2006, S. 304-307). Im Zuge der Expertenforschung erfolgte auch eine stärkere Berücksichtigung fachspezifischer Unterschiede.

Ausgehend von dem bereits 1963 eingeführten Carroll-Modell des schulischen Lernens (vergl. Carroll, 1989), das erstmals eine Unterscheidung zwischen Lehrangebot und Nutzung der Lerngelegenheiten durch die Lernenden vornahm, wurden Prozess-Produkt-Modelle um Variablen zur Beschreibung individueller Lernvoraussetzungen auf Schülerseite (vergl. z. B. Slavin, 1994) und um Persönlichkeitseigenschaften auf Lehrerseite ergänzt (vergl. Wiley & Harnischfeger, 1974). Das auf theoretischen Überlegungen von Fend (1980) und einem Modell von Helmke und Weinert (1997) aufbauende *Angebots-Nutzungsmodell der Wirkungsweise von Unterricht* von Helmke (2009, S. 73) modelliert sowohl den Einfluss der Lernenden mit ihren individuellen Eingangsvoraussetzungen auf das Unterrichtsgeschehen als auch den Einfluss der Lehrkraft, die den Erkenntnissen der Expertenforschung folgend, beispielsweise durch ihr Professionswissen oder ihre pädagogischen Orientierungen charakterisiert wird (vergl. Abbildung 2.1 auf der nächsten Seite). Das Modell basiert auf der Vorstellung von Lerngelegenheiten als Möglichkeitsraum und berücksichtigt damit die Erfolgsunsicherheit des Lehrerhandelns (Baumert & Kunter, 2006, S. 476-477). Die Lehrkraft kann lediglich ein Lehrangebot zur Verfügung stellen. Die Nutzung des Angebots und damit verbundene Lernerfolge können aber nur durch die Schülerinnen und Schüler selbst realisiert werden. Wichtig hierfür ist, dass die Lehrkraft die Teilnahmemotivation der Lernenden am Unterricht sichert, indem sie z. B. das situationale Interesse und die Aufmerksamkeit der Schülerinnen und Schülern weckt (Baumert & Kunter, 2006, S. 476).

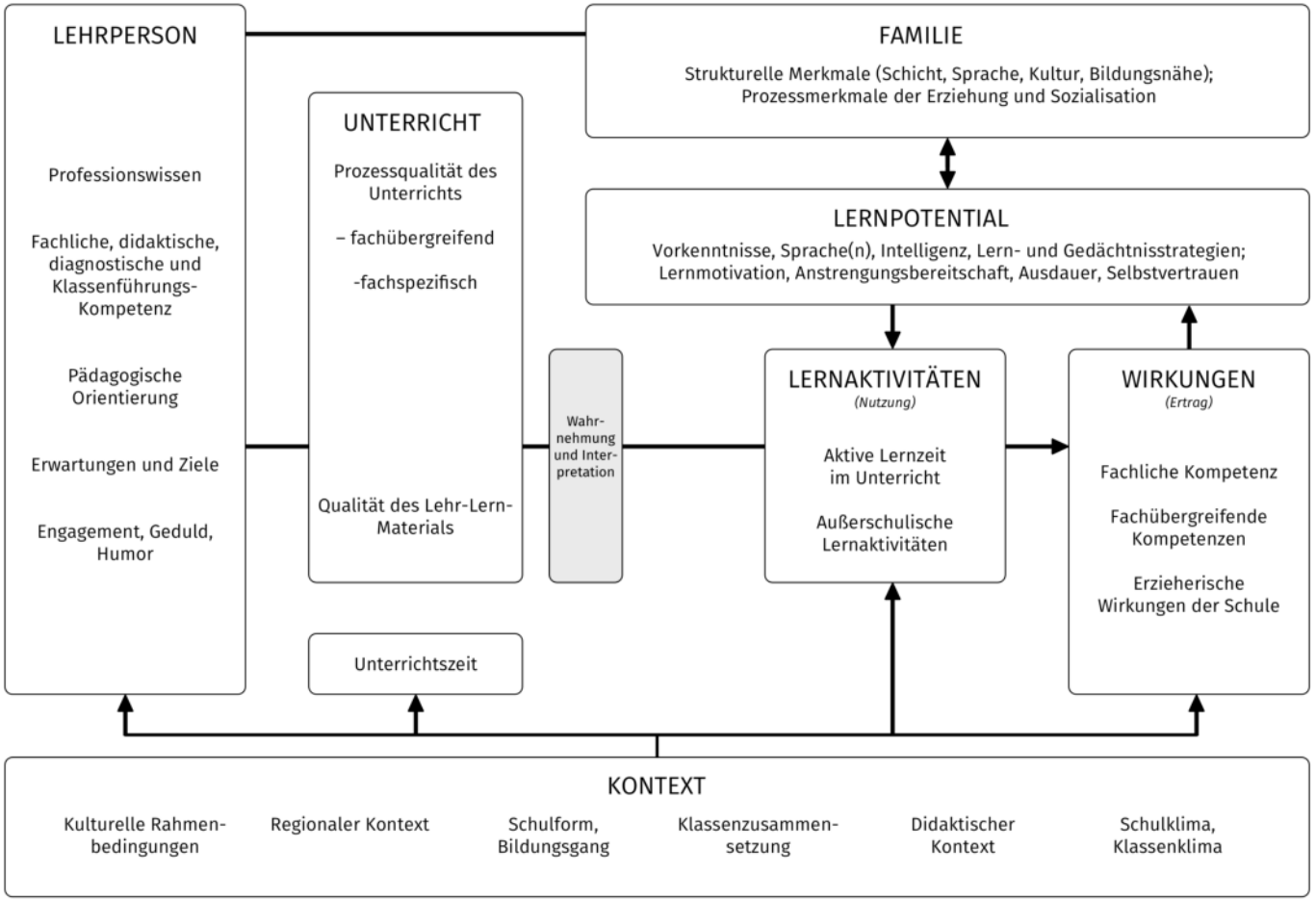



Abbildung 2.1.
*Angebots-Nutzungsmodell der Wirkungsweise von Unterricht (Helmke, 2009, S. 73).*

## 2.2. Professionswissen als Bestandteil professioneller Handlungskompetenz

Nach Baumert und Kunter (2006, S. 477) vernachlässigt das Angebots-Nutzungsmodell der Unterrichtsforschung die doppelte Unsicherheit im Lehrerhandeln, die darin besteht, dass Lerngelegenheiten im Unterricht oftmals das Ergebnis sozialer Ko-Konstruktion und nur schwer planbar sind. So kann die Lehrkraft einerseits nicht sicherstellen, dass ihr Lehrangebot von den Schülerinnen und Schülern genutzt wird, andererseits ist sie schon beim Bereitstellen des Angebots darauf angewiesen, dass die Lernenden sich an bestimmte soziale Grundregeln halten. Aus dieser prinzipiellen Erfolgsunsicherheit des Lehrerhandelns und dem Umstand, dass Lehrerhandeln nicht standardisierbar, sondern situationsspezifisch ist, zogen Vertreter der Auffassung, dass Lehrerhandeln als quasi-therapeutische Tätigkeit zu betrachten ist, den Schluss auf ein Technologiedefizit des Lehrerberufs. Sie hielten das professionelle Handlungsrepertoire von Lehrkräften weder für beschreibbar, noch für erlernbar (vergl. z. B. Luhmann & Schorr, 1979; Oevermann, 1996, zur Diskussion dieser Standpunkte siehe Baumert & Kunter, 2006; Tenorth, 2006). Nach Luhmann und Schorr (1979, S. 353) kann man nicht wissen, „ob im Unterricht richtig oder falsch gehandelt wird“. Tenorth (2006) hingegen schreibt:

> Diese Technologie existiert, ich würde sie „paradoxe Technologie“ nennen, weil sie angesichts der Struktur von Unterricht und Lernen ganz besondere Probleme zu lösen hat: das Nicht-Planbare zu planen, einen festen Rahmen für offene Ereignisse zu geben, mit der Alltäglichkeit von Überraschungen zu rechnen und das [...] zur Routine werden zu lassen. (S. 587-588)

Dieser Sichtweise folgend ziehen Baumert und Kunter (2006) aus der doppelten Unsicherheit des Lehrerhandelns vielmehr Rückschlüsse auf die Struktur des professionellen Wissens von Lehrkräften, das zentraler Bestandteil ihres Modells professioneller Handlungskompetenz ist (vergl. Abbildung 2.2 auf der nächsten Seite).
„Will man wissen, warum Lehrkräfte auf eine bestimmte Weise handeln (und manchmal auch: warum sie wünschenswerte Handlungen unterlassen), so muss man sich genauer mit den kognitiven [...], motivationalen und emotionalen [...] Bedingungen des beruflichen Handelns befassen“, stellen Bromme und Rheinberg (2006, S. 307) fest. Mit der Modellierung der professionellen Handlungskompetenz wird der im Zuge der Expertenforschung von Bromme (1997) entwickelte, primär wissensbasierte Begriff der Lehrerexpertise daher um motivational-selbstregulative Merkmale ergänzt (Baumert & Kunter, 2011). Neben dem Professionswissen enthält das Modell zur professionellen Handlungskompetenz drei weitere Kompetenzfacetten:

- *Überzeugungen und Werthaltungen*, die sowohl langfristig als auch im unmittelbaren Unterrichtskontext handlungsteuernde Funktionen haben können (Brunner et al., 2006). Hierunter werden Wertbindungen und Berufsmoral, subjektive Überzeugungen über Wissen und Wissenserwerb bzw. über die Struktur, Verlässlichkeit, Genese, Validierung und Rechtfertigung von

Wissensbeständen (epistemologische Überzeugungen) sowie subjektive Theorien über das Lehren und Lernen und Zielvorstellungen für Curricula und Unterricht zusammengefasst.

- *Motivationale Orientierungen*, da erfolgreiches Unterrichten voraussetzt, dass die Lehrkraft motiviert ist ihr Wissen auch im Unterricht einzusetzen (Brunner et al., 2006). Die motivationalen Orientierungen werden zum einen durch die Kontrollüberzeugungen und die Selbstwirksamkeitserwartung der Lehrkräfte – nach Bandura (1997) definiert als die Überzeugung einer Person über Mittel und Fähigkeiten zur Erzielung gewünschter Effekte durch die eigenen Handlungen zu verfügen – und zum anderen durch den Lehrerenthusiasmus als Komponente der intrinsischen motivationalen Orientierung modelliert, der den Grad des positiven emotionalen Erlebens während der Ausübung der Lehrtätigkeit beschreibt.
- *Selbstregulative Fähigkeiten*, da der verantwortungsvolle Umgang mit den persönlichen Ressourcen nur durch das richtige Maß an Engagement und Distanzierungsfähigkeit realisiert werden kann (Brunner et al., 2006).

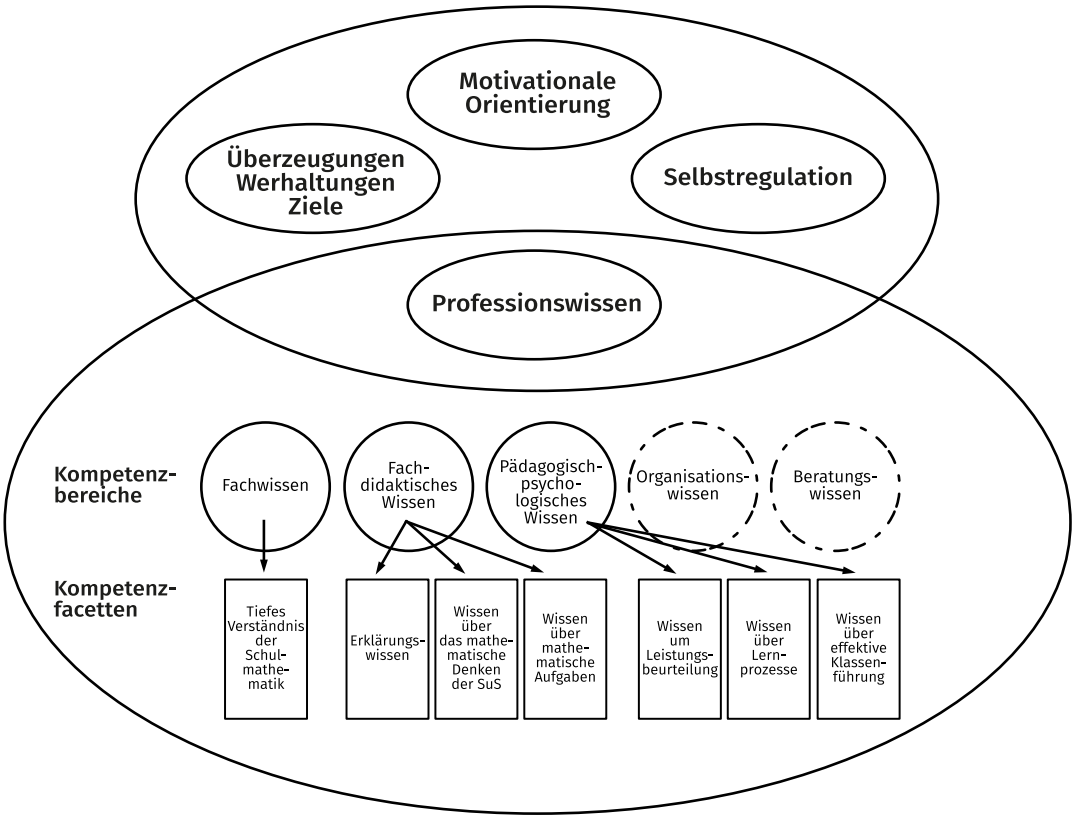



Abbildung 2.2.
*Modell professioneller Handlungskompetenz von Baumert und Kunter (2011, S. 32).*

Das Herzstück des Modells professioneller Handlungskompetenz und die aus Perspektive der Lehrerausbildung zugänglichste Kompetenzfacette stellt das professio-

nelle Wissen und Können der Lehrkräfte in Form von deklarativem, prozeduralem und strategischem Wissen dar (Baumert & Kunter, 2006, S. 481). Das Professionswissen von Lehrkräften wurde bereits in den 1960er Jahren als potenzieller Prädiktor für erfolgreiches Unterrichten erwähnt (Morris, 1989; Yamamoto, 1963) und wird spätestens seit den 1980er Jahren als Voraussetzung für erfolgreiches Unterrichten diskutiert (Abell, 2007; Fischer et al., 2012; Peterson et al., 1989; Shulman, 1987).

Maßgeblich angestoßen wurde diese Diskussion im Zuge der Reformen zur Professionalisierung des Lehrerberufs in den USA (vergl. z. B. Cascio, 1995; Olson, 1987; Shulman, 1987). Politische Entscheidungsträger forderten damals die Entwicklung von Standards für die Lehrerausbildung und die Beschreibung der für den Lehrberuf erforderlichen Wissensbasis auf Basis der empirischen Forschungsergebnisse aus der Prozess-Produkt-Forschung. Dabei wurde allerdings die für die Forschung unerlässliche Reduzierung der Komplexität realer Unterrichtssituationen vernachlässigt: Um eine Forschungsfrage beantworten zu können, muss ein Forscher seinen Blickwinkel auf einen präzise definierten Sachverhalt fokussieren, mit dem Preis der Komplexität von Unterrichtssituationen nicht immer gerecht zu werden (Shulman, 1986). So sind die im Zuge des Prozess-Produkt-Paradigmas identifizierten Lehrerverhaltensmerkmale für sich allein genommen nicht per se notwendig für erfolgreiches Unterrichten, sondern in gewissem Ausmaß gegenseitig kompensierbar (Reusser, 2009, S. 892). Außerdem wurde außer Acht gelassen, dass die Prozess-Produkt-Forschung, wie bereits erwähnt, lediglich allgemeine Verhaltensmerkmale identifizierte und keine Berücksichtigung fachspezifischer Unterschiede erfolgte. Die Beschreibung der für den Lehrerberuf erforderlichen Wissensbasis auf Basis dieser Forschungsergebnisse fokussierte daher auf allgemein-pädagogisches Wissen und ignorierte weitestgehend das Fachwissen und das fachspezifische pädagogische Wissen. Shulman (1987) warnte eindringlich vor diesem Vorgehen:

> Critical features of teaching, such as the subject matter being taught, the classoom context, the physical and psychological characteristics of the students, or the accomplishment of purposes not readily assessed on standardized tests, are typically ignored in the quest for general principles of effective teaching. When policymakers have sought „research-based“ definitions of good teaching to serve as the basis for teacher tests or systems of classroom observation, the list of teacher behaviors that had been identified as effective in the empirical research were translated into the desirable competencies for classroom teachers. They became items on tests or on classroom-observation scales. They were accorded legitimacy because they had been „confirmed by research.“ While the researchers understood the findings to be simplified and incomplete, the policy community accepted them as sufficient for the definitions of standards. [...] Thus, what may have been an acceptable strategy for research became an unacceptable policy for teacher evaluation. (S. 6-7)

Als Konsequenz daraus entwickelte Shulman einen der ersten Ansätze, das Professionswissen von Lehrkräften zu beschreiben.

## 2.3. Ein Konstrukt – viele Modelle: Modellierung von Professionswissen

Auch wenn weitestgehend Einigkeit über die Wichtigkeit des Professionswissens für erfolgreiches Unterrichten herrscht, ist diese weder empirisch abgesichert noch besteht Konsens darüber, wie das Professionswissen von Lehrkräften zu operationalisieren ist und welches Wissen nachweislich als unterrichtsrelevant erachtet werden kann (vergl. Abell, 2007; Baumert & Kunter, 2006; Kirschner, 2013).

Die ersten systematischen Versuche das Professionswissen von Lehrkräften zu operationalisieren, wurden von Shulman (1986) unternommen. Shulman (1987) beschrieb zunächst vier und später sieben Kategorien des Professionswissens:

- „content knowledge,
- general pedagogical knowledge, with special reference to those broad principles and strategies of classroom management and organization that appear to transcend subject matter;
- curriculum knowledge, with particular grasp of the materials and programs that serve as ‚tools of the trade' for teachers;
- pedagogical content knowledge, that special amalgam of content and pedagogy that is uniquely the province of teachers, their own special form of professional understanding;
- knowledge of learners and their characteristics;
- knowledge of educational context, ranging from the workings of the group or classroom, the governance and finance of school districts, to the character of communities and cultures; and
- knowledge of educational ends, purposes, and values, and their philosophical and historical grounds." (S.8)

Ausgehend von einer Analyse der beruflichen Anforderungen von Lehrkräften und in direkter Anlehnung an Shulman schlägt Bromme (1992 und 1997, S. 96-98 bzw. S. 196-198) eine Topologie des professionellen Lehrerwissens vor, die fünf Bereiche des Lehrerwissens beschreibt:

Das Fachwissen wird differenziert in *fachliches Wissen über die Fachdisziplin*, über das auch Fachwissenschaftler verfügen können und *schulfachliches* bzw. *curriculares Wissen*. Letzteres umfasst die Logik des Schulfaches, die nicht allein aus der Logik der wissenschaftlichen Fachdisziplin zu erklären ist, da die Lerninhalte des Schulfaches nicht nur Vereinfachungen fachwissenschaftlicher Zusammenhänge darstellen. In dieses Wissen können auch Zielvorstellungen über Schule einfließen.

Die *Philosophie des Schulfachs* beschreibt normativ geprägte Auffassungen über die Nützlichkeit des Fachinhalts und seine Beziehung zu anderen Bereichen „menschlichen Lebens und Wissens" und stellt eine bewertende Perspektive auf den Unterrichtsinhalt dar. Im Sinne des Modells professioneller Handlungskompetenz

von Baumert und Kunter (2006) würde man die Philosophie des Schulfaches eher den Überzeugungen und Werthaltungen zuordnen (vergl. auch Bromme, 1997, S. 198).

Das *pädagogische Wissen* beschreibt das fachunspezifische pädagogisch-psychologische Wissen von Lehrkräften, wie z. B. das Wissen über Lehr-Lern-Prozesse, Klassenführung oder den Umgang mit Disziplinproblemen.

In Anlehnung an Shulmans *pedagogical content knowledge*, beschreibt der Bereich des *fachspezifisch-pädagogischen Wissens* das „integrierte Wissen, in dem psychologisch-pädagogische Kenntnisse sowie eigene Erfahrungen des Lehrers auf den Fachinhalt bezogen werden“ und damit das Wissen darüber, wie ein Fachinhalt in spezifischen Unterrichtssituationen zu unterrichten ist (Bromme, 1992, S. 97). Bromme (1992, S. 102) grenzt das fachspezifische-pädagogische Wissen vom fachdidaktischen Wissen, wie es an deutschen Universitäten unterrichtet wird, ab. Im fachdidaktischen Wissen sieht er Hilfestellungen für die Integration von fachlichem und pädagogisch-psychologischem Wissen, die allerdings erst an die der Lehrkraft vorliegende Unterrichtssituation angepasst werden müssen. In der Regel werden die Begriffe aber synonym verwendet. Das professionelle Wissen einer Lehrkraft betrachtet Bromme (1992, S. 102) als „eine ganz besondere, von den Lehrern selbst entwikkelte [sic!] Mischung curricular-fachlichen und pädagogisch-psychologischen Wissens mit eigenen Erfahrungen über Unterrichtssituationen“.

Nahezu alle großen empirischen Forschungsarbeiten und Übersichtsartikel zum Professionswissen von Lehrkräften stellen als zentrale Wissensdimensionen des Professionswissens das *Fachwissen* (content knowledge: CK), das *fachdidaktische Wissen* (pedagogical content knowledge: PCK) und das *pädagogische Wissen* (pedagocical content knowledge: PK) dar und machen einzelne oder alle drei Wissensdimensionen zum Gegenstand ihrer Forschung (vergl. z. B. Baumert & Kunter, 2006; Blömeke, Kaiser & Lehmann, 2008; Borko & Putnam, 1996; Borowski et al., 2010; Fischer et al., 2012; Fischer, Labudde, Neumann & Viiri, 2014a; Hill, Rowan & Ball, 2005; Kröger et al., 2015; Kulgemeyer et al., 2012; Lange, 2010; Lipowsky, 2006; Ohle, 2010). In Deutschland besteht damit eine direkte Anknüpfung an die drei Säulen der universitäre Lehrerausbildung: Fachwissenschaft, Fachdidaktik und Pädagogik.

Im angloamerikanischen Sprachraum wird als vierte Dimension des Professionswissens häufig auch das *knowledge about context* oder *contextual knowledge* genannt, das nach einer Definition von Grossman (1990, S. 9) als Kombination aus den von Shulman (1987) beschriebenen Kategorien *knowledge of learners and their characteristics* und *knowledge of educational context* verstanden werden kann (vergl. z. B. Gess-Newsome, 1999; Grossman, 1990; Magnusson, Krajcik & Borko, 1999; Park & Oliver, 2008). In deutschen Forschungsarbeiten wird dieses auch als *Organisationswissen* bezeichnete Wissen zwar vereinzelt in die Modellierung des Professionswissens mit einbezogen, aber nur selten als zentraler Forschungsgegenstand betrachtet (vergl. Baumert & Kunter, 2006, S. 482). Darüber hinaus wird das *contextual knowledge* oder Aspekte davon teils explizit (Gess-Newsome, Carlson, Gardner & Taylor, 2010), teils implizit (Magnusson et al., 1999) in die Modellierung von PCK einbezogen (vergl. auch Park & Oliver, 2008). Auch das von Shulman (1987) beschriebene *curriculum knowledge* findet sich in einigen Arbeiten

als Facette von PCK wieder (Baumert & Kunter, 2006; Blömeke et al., 2008; Ergönenç, Neumann & Fischer, 2014; Grossman, 1990; Magnusson et al., 1999; Riese, 2009).

Wie heterogen die Modellierung des Professionswissens oder der Professionswissensdimensionen ist, sieht man auch daran, dass die Dimensionen CK und PK in einigen Modellen nicht nur neben PCK betrachtet werden, sondern als Teil von PCK. Gess-Newsome et al. (2010) modellieren Professionswissen mit den Wissensdimensionen *akademisches* CK, PK und PCK. PCK wird allerdings wiederum durch die drei Faktoren PCK-CK, PCK-PK und *contextual knowledge* (PCK-CxK) beschrieben. Empirisch konnten Gess-Newsome et al. (2010) allerdings nur die beiden Faktoren PCK-CK und PCK-PK nachweisen. PCK-CxK ließ sich nicht von PCK-PK trennen.

In der deutschsprachigen Forschung zum Professionswissen von Lehrkräften werden CK, PCK und PK meist als separate, für erfolgreiches Unterrichten relevante Wissensdimensionen modelliert. Diese hängen zwar zusammen, stellen aber dennoch unterschiedliche Wissensbereiche dar (vergl. z. B. Baumert & Kunter, 2006; Blömeke et al., 2008; Fischer et al., 2014b; Lange, 2010; Ohle, 2010; Riese, 2009; Schmelzing, 2010; Tepner et al., 2012). Ob die Dimensionen empirisch trennbar sind, hängt allerdings von den in den jeweiligen Untersuchungen vorgenommenen Operationalisierungen und in einigen Fällen von der Expertise der untersuchten Lehrkräfte ab. So konnten Krauss, Brunner et al. (2008, S. 724) das CK und PCK von Mathematiklehrkräften, die nicht am Gymnasium unterrichteten, getrennt erfassen. Dies galt allerdings nicht für das CK und PCK von Gymnasiallehrkräften, die über ein wesentlich höheres CK und PCK verfügten. Die Autoren erklärten die unterschiedliche Dimensionalität des fachspezifischen Professionswissens in den verschiedenen Gruppen von Mathematiklehrkräften mit Ergebnissen aus der Expertiseforschung, die darauf hinweisen, dass das Wissen von Experten gegenüber dem von Novizen vernetzter und besser integriert ist (vergl. z. B. Berliner, 2001). Arbeiten, in denen das Professionswissen von Physiklehrkräften untersucht wurde, konnten diese Ergebnisse allerdings nicht replizieren (Kirschner, 2013, S. 83-85; Riese, 2009, S. 151). Die hier gefundenen Zusammenhänge deuteten eher auf getrennte Wissensdimensionen hin. Die dreidimensionale Struktur des Professionswissens wurde von Kirschner (2013) explizit untersucht und empirisch bestätigt. Außerdem wird davon ausgegangen, dass CK eine notwendige, jedoch nicht hinreichende Bedingung für PCK darstellt (Krauss, Neubrand et al., 2008, S. 228; Riese, 2009, S. 180; Sadler et al., 2013, S. 1036).

In diesem Abschnitt wurde deutlich gemacht, wie unterschiedlich das Professionswissen von Lehrkräften modelliert wird. Wie bereits angedeutet wurde, bestehen auch deutliche Unterschiede in der Operationalisierung der einzelnen Wissensdimensionen. In den folgenden Abschnitten wird daher näher auf die Professionswissensdimensionen CK, PCK und PK eingegangen und es werden unterschiedliche Ansätze für deren Operationalisierung beschrieben.

### 2.3.1. Fachwissen - CK

Ein wichtiger Bestandteil der Wissensbasis von Lehrkräften ist ihr *Fachwissen*. Eine Lehrkraft, die Englisch unterrichtet, wird Schwierigkeiten haben Physik zu unterrichten, selbst wenn sie sich das Fachwissen, das den Lernenden in einer Schulstunde vermittelt werden soll, vorher angeeignet hat. Eine Lehrkraft muss über Fachwissen verfügen, das über das zu lehrende Wissen hinausgeht, um die Anforderungen des Lehrberufs zu bewältigen (Baumert & Kunter, 2006, S. 495). Sie muss wissen, wie Fachinhalte zusammenhängen, welchen Stellenwert bestimmte Konzepte innerhalb der Fachdisziplin einnehmen und unterschiedliche Zugangswege zu einem Fachinhalt kennen.

> [A teacher] must understand the structures of subject matter, the principles of conceptual organization, and the principles of inquiry that help answer two kinds of questions in each field: What are the important ideas and skills in this domain? and How are new ideas added and deficient ones dropped by those who produce knowledge in this area? That is, what are the rules and procedures of good scholarship or inquiry? (Shulman, 1987, S. 9)

Neben dem Wissen über die Fachinhalte benötigt eine Lehrkraft also Wissen über die Struktur des Faches. Hierbei wird nach Schwab (1964, zitiert nach Grossman, 1990, S. 6) Wissen über die *substantive* Struktur und die *syntaktische Struktur* des Faches unterschieden. Als substantive Strukturen werden die verschiedenen Arten und Weisen bezeichnet, auf die die Konzepte und leitenden Prinzipien des Faches organisiert werden können (Grossman, 1990, S. 6). Unterschiedliche Sichtweisen auf die konzeptuelle Organisation eines Faches führen zu unterschiedlichen Fragestellungen innerhalb der Fachdisziplin. Die syntaktische Struktur des Faches beschreibt hingegen die Regeln der Erkenntnisgewinnung innerhalb eines Faches und damit „[...] the set of rules for determining what is legitimate to say in a disciplinary domain and what ‚breaks' the rules." (Shulman, 1987, S. 9). Wissen über die syntaktische Struktur des Faches beinhaltet also das Wissen darüber, was in der Fachdisziplin als Evidenz oder Beweis anerkannt wird und wie neue Erkenntnisse in der Fachdisziplin gewonnen und evaluiert werden (vergl. Grossman, 1990, S. 6-7; Shulman, 1987, S. 9) und weist in seiner Definition damit eine Nähe zum epistemologischen Wissen auf (vergl. z. B. Phillips, 2003, S. 423-424). Das Wissen über die substantive und syntaktische Struktur des Faches versetzt Lehrkräfte in die Lage, ihren Schülerinnen und Schülern unterschiedliche Zugangswege zu den Fachinhalten zu ermöglichen, Verbindungen aufzuzeigen und ihnen zu erklären, warum bestimmte fachliche Positionen eingenommen wurden oder werden, warum diese als richtig oder falsch erachtet werden und warum es wichtig ist, die verschiedenen Positionen zu kennen (Shulman, 1987, S. 9). In der Topologie des Professionswissens von Bromme (1997, S. 196) lässt sich dieses Wissen am ehesten im Bereich des curricularen Wissens und zu einem gewissen Teil im Wissen über die Philosophie des Schulfachs verorten.

Auf welchem Niveau eine Lehrkraft über Fachwissen verfügen muss, um erfolgreich zu unterrichten, ist allerdings nicht geklärt (Baumert & Kunter, 2006).

Viele Arbeiten sehen neben dem zu lehrenden Schulwissen und dem universitären Fachwissen vor allem ein *vertieftes Schulwissen* als zentralen Wissensbereich des Fachwissens von Lehrkräften an (Baumert & Kunter, 2011; Blömeke et al., 2008; Kirschner, 2013; Riese, 2009).[1] Diese Annahme erscheint so naheliegend und plausibel, dass für dieses Wissen oftmals nur eine allgemeine, nicht sehr präzise Arbeitsdefinition existiert, zumal die Benennung zunächst selbsterklärend erscheint. Aktuelle Arbeiten unternehmen daher den Versuch diesen Wissensbereich weiter auszuschärfen (Gigl, Zander, Borowski & Fischer, 2015; Woitkowski, Riese & Reinhold, 2011). Beispielsweise operationalisieren Gigl et al. (2015) vertieftes Schulwissen über die folgenden fünf Aspekte:

- „Verschiedene Wege zur Lösung einer Aufgabe identifizieren und anwenden
- Lösung einer Aufgabe aus theoretischer Sicht planen
- Randbedingungen einer Schulaufgabe erkennen
- Aufgaben fachlich korrekt vereinfachen
- Zusammenhänge, Gemeinsamkeiten und Unterschiede physikalischer Phänomene erkennen“ (S. 112)

### 2.3.2. Fachdidaktisches Wissen - PCK

Die Dimension des „pedagogocal content knowledge“ wurde erstmals von Shulman (1987) eingeführt:

> [...] pedagogical content knowledge [...] identifies the distinctive bodies of knowledge for teaching. It represents the blending of content and pedagogy into an understanding of how particular topics, problems, or issues are organized, represented, and adapted to the diverse interests and abilities of learners, and presented for instruction. Pedagogical content knowledge is the category most likely to distinguish the understanding of the content specialist from that of the pedagogue. (S.8)

Was genau unter PCK zu verstehen ist, ist allerdings bis heute nicht einheitlich definiert. Das PCK von Lehrkräften wird je nach Modell durch bis zu acht Unterfacetten beschrieben. Tabelle 2.1 auf Seite 18 zeigt eine Übersicht der in die verschiedenen Modellierungen einbezogenen Facetten.
Wie bereits erwähnt, modellieren einige Autoren auch CK und PK als Facetten von PCK. Shulman (1986, S. 9-10) beschreibt PCK über die Facetten Wissen über Schülervorstellungen sowie Wissen über Instruktionsstrategien und Repräsentationen, die in nahezu alle Arbeiten in die Modellierung von PCK einbezogen werden. Beim Unterrichten eines Fachinhalts muss eine Lehrkraft also zum einen

[1] Nicht immer wird hierfür der Begriff „vertieftes Schulwissen“ verwendet. Baumert und Kunter (2011, S. 37) sprechen von „einem profunden mathematischem Verständnis des Hintergrunds des Schulstoffs“ und Blömeke et al. (2008, S. 107) von „Schulmathematik vom höheren Standpunkt“.

über Wissen darüber verfügen, welche vorunterrichtlichen Vorstellungen über ein Konzept die Lernenden in den Unterricht mitbringen könnten, welche Schwierigkeiten sich daraus für das Verständnis des Lerngegenstands ergeben und wie damit umgegangen werden kann. Zum anderen muss die Lehrkraft über ein breites Repertoire an Vermittlungsstrategien, Darstellungsformen, Beispielen und Analogien verfügen, um den Fachinhalt auf verständliche Weise zu unterrichten. Das Magnusson-Modell erweitert diese Definition um das Wissen über die Beurteilung von Scientific Literacy und Wissen über Fachcurricula (Magnusson et al., 1999). Außerdem bezieht das Modell die Orientierungen zum Unterrichten von Naturwissenschaften mit ein, die nach dem Modell für professionelle Handlungskompetenz von Baumert und Kunter (2006) allerdings in den Bereich der Überzeugungen und Werthaltungen fallen. Nach Ergebnissen einer qualitativen Studie von Park und Chen (2012, S. 937) zur Vernetzung der fünf PCK-Facetten des Magnusson-Modells stellen sich allerdings die zwei von Shulman ursprünglich eingeführten Facetten als zentral für die Struktur von PCK heraus.

Borowski, Olszewski und Fischer (2010, S. 262) fanden Unterschiede im fachdidaktischen Wissen von Physikreferendare und -referendarinnen bezüglich unterschiedlicher Inhaltsbereiche (Mechanik/Elektrizitätslehre) und Sadler et al. (2013, S. 1041) stellten wenig Transfer zwischen dem Wissen über Schülerfehlvorstellungen (als Aspekt von PCK) in verschiedenen Inhaltsbereichen fest. Nach dem aktuellen Forschungsstand wird PCK daher nicht nur als fachspezifisches, sondern vielmehr als themenspezifisches Wissen modelliert (vergl. z. B. Gess-Newsome, 2015; Rollnick & Mavhunga, 2014).

PCK wird in der Regel handlungsnah operationalisiert und umfasst daher zusätzlich zum fachdidaktischen universitären Wissen auch fachspezifisch-pädagogisches Wissen im Sinne von Bromme (1997).[2]

### 2.3.3. Pädagogisches Wissen - PK

Das pädagogische Wissen von Lehrkräften ist das fachunspezifische Wissen über das Lehrkräfte aller Fächer gleichermaßen verfügen können. Auch für das pädagogische Wissen gilt, das weder theoretisch noch empirisch genau geklärt ist, wie und aus welchen Facetten dieses Wissen aufgebaut ist (König & Blömeke, 2009, S. 501). Allerdings erscheint der Konsens über die zu PK gehörenden Wissensfacetten wesentlich ausgeprägter als in der Diskussion über PCK zu sein. Ganz allgemein wird unter dem pädagogischen Wissen das Wissen um Strategien und Mittel zur Erzeugung und Aufrechterhaltung lernförderlicher Bedingungen im Unterricht verstanden (Lenske, Thillmann, Wirth, Dicke & Leutner, 2015). Shulman (1987, S. 8) zählt zum pädagogischen Wissen in erster Linie das Wissen über Prinzipien der Klassenführung und Organisation.

Nach Helmke (2009, S. 174) stellt Klassenführung eine Basiskompetenz des Lehrberufs und damit eine „unabdingbare Voraussetzung für die Sicherung anspruchsvollen Unterrichts [dar], indem sie einen geordneten Rahmen für die eigentlichen

[2] Die Bezeichnung fachdidaktisches Wissen und *pedagocical content knowledge* werden in den meisten Arbeiten und auch in der vorliegenden Arbeit synonym verwendet (siehe kritisch hierzu Gramzow & Reinhold, 2013, S. 10-11).

Tabelle 2.1.
*Übersicht über die in Operationalisierungen von PCK einbezogenen Facetten (Übernommen, erweitert und angepasst aus Kirschner (2013, S. 32) und Kirschner, Borowski, Fischer, Gess-Newsome und von Aufschnaiter (in Druck))*

| | Wissen über ... | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Projekt/Autoren | Fachwissen | Pädagogik | Kontext | Schülerverständnis | Instruktionsstrategien und Repräsentationen | Curriculum | Unterrichtsziele | Leistungsbeurteilung |
| Shulman (1986) | - | - | - | + | + | - | - | |
| Tamir (1988) | - | - | | + | + | + | | + |
| Smith und Neale (1989) | - | | | + | + | | ? | |
| Grossman (1990) | - | | | + | + | + | + | |
| Geddis et al. (1993) | | | | + | + | + | | |
| Magnusson, Krajcik und Borko (1999) | | | | + | + | + | + | + |
| Park und Oliver (2008) | - | - | | + | + | + | + | + |
| Riese (2009) | - | - | | + | + | + | + | + |
| MT21 (Blömeke et al., 2008) | - | - | | + | + | + | | |
| TEDS-M (Döhrmann et al., 2010) | - | - | | + | + | + | | + |
| COACTIV (Baumert & Kunter, 2011) | - | - | - | + | + | | | - |
| PLUS[1](Lange, 2010) | - | - | | + | + | + | + | + |
| ProwiN (Tepner et al., 2012) | - | - | | + | + | | | |
| KiL (Kröger et al., 2013) | - | - | | + | + | + | | + |
| QuiP (Ergönenç et al., 2014) | - | - | | + | + | + | | |
| Marks (1990) | + | | | + | + | | | |
| Cochran et al. (1993) | + | + | + | + | | | | |
| Fernández-Balboa und Stiehl (1995) | + | | + | + | + | | + | |
| Hashweh (2005) | + | + | + | + | + | + | + | + |
| Rollnick et al. (2008) | + | + | + | + | + | + | + | + |
| Loughran et al. (2012) | + | + | + | + | + | + | + | + |
| ProfiLe-P[1](Gramzow et al., 2013) | - | - | + | + | + | + | + | + |
| Malcolm und Mavhunga (2015) | + | + | + | + | + | + | + | + |

*Legende:* + =Facette wird explizit PCK zugeordnet; - =Facette wird explizit nicht PCK zugeordnet

[1] Das in dieser Studie entwickelte Testinstrument zur Messung von PCK erfasst nur einen Teil der angegebenen Facetten.

Lehr- und Lernaktivitäten schafft und insbesondere die aktive Lernzeit steuert". Dies kann eine Lehrkraft zum Beispiel über die Einführung von Regeln und Ritualen, Maßnahmen zur Störungsprävention und einen angemessenen Umgang mit Disziplinproblemen realisieren (Evertson & Emmer, 1982; Helmke, 2009). Die präventive Steuerungsleistung der Lehrkraft wird dabei als wesentlicher erachtet als der reaktive Umgang mit Unterrichtsstörungen (Kunter & Voss, 2011, S. 88). In diesem Kontext steht auch das von Kounin (2006, S. 148) als wesentlicher Bestandteil von Klassenführung eingeführte Prinzip der *Allgegenwärtigkeit* und *Überlappung*. Eine Lehrkraft sollte einerseits über die Fähigkeit verfügen, den Lernenden zu zeigen, dass sie über ihr Verhalten informiert ist und andererseits dazu in der Lage sein, ihre Aufmerksamkeit simultan auftretenden Störungen gleichermaßen zuzuwenden.

Neben dem Wissen über Unterrichtsführung geben Baumert und Kunter (2006, S. 485) die folgenden Facetten des pädagogischen Wissens als weitestgehend konsensfähig an: Konzeptuelles bildungswissenschaftliches Grundlagenwissen, das auch Wissen über Entwicklungs-, Lern- und Motivationspsychologie umfasst, allgemeindidaktisches Konzeptions- und Planungswissen, Wissen über die Orchestrierung von Lerngelegenheiten und Wissen über fachübergreifende Prinzipien des Diagnostizierens, Prüfens und Bewertens. Die Autoren verweisen allerdings auf die unterschiedliche Entfernung der einzelnen Facetten zum Handeln im Unterricht und die damit einhergehenden zu erwartenden Unterschiede in deren Relevanz für erfolgreiches Unterrichten.

Das pädagogische Wissen, dem direkte Relevanz für erfolgreiches Unterrichten zugesprochen wird, operationalisieren Voss und Kunter (2011, S. 195) als „Wissen über Klassenprozesse", zu dem Wissen über effektive Klassenführung, Unterrichtsmethoden und deren zieladäquate Orchestrierung und Wissen über Leistungsbeurteilung zählen und als „Wissen über Schüler und Quellen für Heterogenität der Schülerschaft", zu dem Wissen über (individuelle) Lernprozesse, Unterschiede und Besonderheiten und sich daraus ergebene Anforderungen an die Unterrichtsgestaltung zählen. Um der Vielseitigkeit von Lehr-Lern-Situationen und den Voraussetzungen unterschiedlicher Lernender im Klassenkontext gerecht zu werden, sind Kenntnisse über eine Vielfalt an Unterrichtmethoden unabdingbar (Voss & Kunter, 2011, S. 197). Weitaus wichtiger als die reine Methodenkenntnis ist allerdings das Wissen darüber, wie Unterrichtsmethoden passend zum jeweiligen Unterrichtziel ausgewählt, umgesetzt und orchestriert werden können (Oser & Baeriswyl, 2001; Tepner et al., 2012; Voss & Kunter, 2011).

Ein weiterer zentraler Aspekt der Lehrertätigkeit ist die Leistungsbeurteilung. Wissen über das Potenzial verschiedener Formen der Leistungsbeurteilung (summativ, am Ende einer Unterrichtseinheit, oder formativ, innerhalb einer Unterrichtseinheit) verbunden mit Wissen darüber, wie diese verständlich und motivierend an die Lernenden rückgemeldet werden können (vergl. z. B. Narciss & Huth, 2004), bieten der Lehrkraft die Möglichkeit Informationen zum Leistungsstand und zum Verständnis der Lernenden zu generieren, den Unterricht an die Bedürfnisse der Lernenden anzupassen und die Schülerinnen und Schüler motivierend beim Lernen

zu unterstützen (vergl. Tepner et al., 2012, S. 197; Voss & Kunter, 2011, S. 11-12). Hierfür ist auch Wissen über die Psychologie des Lernens nötig.[3]

Nach dem Angebots-Nutzungsmodell von Helmke (2009, S. 73) sind für den Lernerfolg die individuellen kognitiven und motivationalen Voraussetzungen der Lernenden von großer Bedeutung. Daher ist auch das Wissen über individuelle Lernprozesse und Besonderheiten von Schülerinnen und Schülern (z. B. Schwächen und Stärken) und deren Berücksichtigung im Unterricht ein wichtiger Bestandteil des pädagogischen Wissens von Lehrkräften (Voss & Kunter, 2011). Die Facetten Klassenführung, Unterrichtsmethoden, Leistungsbeurteilung und individuelle Lernprozesse finden sich auch in anderen Operationalisierungen des pädagogischen Wissens wieder (vergl. z. B. König & Blömeke, 2009; Tepner et al., 2012).

Die Konzeptualisierung des pädagogischen Wissens ist eng mit Betrachtungen zur Unterrichtsqualität verbunden. Die Wahl der Wissensfacetten wird meist aus empirischen Ergebnissen der Unterrichtsqualitätsforschung abgeleitet (vergl. König & Blömeke, 2009, S. 503-507). Anders als im Fall der fachspezifischen Wissensdimensionen kann dabei auf die zahlreichen Ergebnisse aus der Prozess-Produkt-Forschung zurückgegriffen werden, die sich in der Regel mit fachunabhängigen Merkmalen von Unterricht beschäftigte (vergl. z. B. die Ausführungen von Shulman (1986) zu der als „Missing Paradigm“ bezeichneten fehlenden Berücksichtigung fachlicher Aspekte).

[3] Vor diesem Hintergrund wird das pädagogische Wissen von Lehrkräften in einigen Arbeiten als pädagogisch-psychologisches Wissen bezeichnet (Voss, Kunter & Baumert, 2011a). In der vorliegenden Arbeit werden die Begriffe allerdings synonym verwendet.

# 3. Professionswissen als Voraussetzung für erfolgreiches und gutes Unterrichten

Im Zuge des Paradigmenwechsels in der Unterrichtsqualitätsforschung rückte das Professionswissen von Lehrkräften zunehmend in den Fokus der fachdidaktischen und pädagogisch-psychologischen Bildungsforschung. Die intensive Beschäftigung mit dem Professionswissen von Lehrkräften ist nicht zuletzt auch mit der Hoffnung verbunden, dass Lehrkräfte im Rahmen von Aus- und Weiterbildungsangeboten handlungsrelevantes Wissen aufbauen können, das sie in die Lage versetzt, qualitätvoll zu unterrichten. Die Annahme über die Relevanz des Professionswissens für qualitätvolles Unterrichten, auf der auch die Modellierung des Professionswissens als Bestandteil der professionellen Handlungskompetenz von Lehrkräften beruht, impliziert die Annahme eines grundsätzlichen Zusammenhangs zwischen Wissen und Handeln. Schon aus theoretischer Perspektive herrscht in dieser Frage allerdings keineswegs Einigkeit. Vielmehr ist der Zusammenhang zwischen Wissen und Handeln selbst Gegenstand kontroverser Diskussionen (vergl. Kolbe, 2004). In Abschnitt 3.1 auf der nächsten Seite werden daher kurz die verschiedenen diesbezüglich eingenommenen Positionen vorgestellt.

Will man die Annahme, dass das Professionswissen von Lehrkräften eine wichtige Voraussetzung für qualitätvolles Unterrichten darstellt, nicht einfach hinnehmen, sondern auch überprüfen, stellt sich zunächst die Frage: Was heißt das eigentlich? Wann kann der Unterricht einer Lehrkraft als qualitätvoll bezeichnet werden? Fenstermacher und Richardson (2005) nähern sich dem Begriff des qualitätvollen Unterrichtens über die Unterscheidung zwischen *good teaching* und *succesful teaching*:

> By *good teaching* we mean that the content taught accords with disciplinary standards of adequacy and completeness, and that the methods employed are age-appropriate, morally defensible, and undertaken with the intention of enhancing the learner's competence with respect to the content studied [...]. By *successful teaching* we mean that the learner actually acquires, to some reasonable and acceptable level of proficiency, what the teacher is engaged in teaching. (S. 191)

Mit dieser Unterscheidung tragen die Autoren zum einen dem Umstand Rechnung, dass der Erfolg des Unterrichtens (im Sinne des Erreichens eines intendierten Bildungsziels) nicht allein von der Lehrkraft abhängt, sondern in nicht unerheblichem Maße auch von der Lernbereitschaft der Schülerinnen und Schüler, einem Lehren

und Lernen unterstützenden sozialen Umfeld sowie der Zeit und den Ressourcen, die zum Erreichen des Bildungsziels zur Verfügung stehen (Fenstermacher & Richardson, 2005, S. 190). Zum anderen weisen die Autoren darauf hin, dass erfolgreiches Unterrichten nicht zwingend mit gutem Unterricht einhergeht: Als extremes Beispiel sei ein Unterricht genannt, in dem die Lernenden durch Androhung von Gewalt und Züchtigung Wissen erwerben – dieser Unterricht wäre zwar „erfolgreich“, aber hier würde niemand von gutem Unterricht sprechen (Fenstermacher & Richardson, 2005, S. 189). Die Definition „guten“ Unterrichts ist allerdings stets mit normativen Annahmen darüber, was als gut erachtet werden kann, verbunden. Nach Fenstermacher und Richardson (2005, S. 192) ist Unterrichten erst dann *qualitätvoll*, wenn es sowohl gut, als auch erfolgreich ist.

Um die Begriffe *erfolgreich* und *gut* mit Inhalt zu füllen, wird in Abschnitt 3.2 auf Seite 24 zunächst auf die Definition von Unterrichtserfolg über das Erreichen von Zielkriterien eingegangen. Anschließend wird in Abschnitt 3.3 auf Seite 25 auf Merkmale guten Unterrichts und Unterrichtsqualität eingegangen – hier soll insbesondere das aus fachdidaktischer Perspektive interessante Merkmal der kognitiven Aktivierung näher vorgestellt werden, da dieses Merkmal im späteren Verlauf der vorliegenden Arbeit als besonders geeignet identifiziert wird, um Zusammenhänge zwischen dem fachspezifischen Professionswissen von Lehrkräften und Unterrichtsqualität zu untersuchen (vergl. Abschnitt 5.3 auf Seite 65 im Kapitel zur Ableitung des eigenen Forschungsansatzes).

## 3.1. Hängen Wissen und Handeln zusammen? Eine kontroverse Diskussion

> Wissen und Handeln stehen schon nach dem Alltagsverständnis in enger Verbindung und weisen doch ein kompliziertes Verhältnis zueinander auf. Wissensbestände entstehen in Entwicklung und in der individuellen Auseinandersetzung mit den Erfahrungen durch eigenes Handeln, gleichzeitig aber stellen sie immer mit auch die Basis dieses Handelns dar, (Kolbe, 2004, S. 206)

schreibt Kolbe (2004) und skizziert die sehr inkonsistente Forschungslage zum Zusammenhang zwischen Wissen und Handeln. So herrscht beispielsweise keine Einigkeit, ob Wissen als implizites Wissen oder als explizierbares, im Handlungsprozess transformiertes, wissenschaftliches Wissen beim Handeln wirksam wird. Während das Modell zur professionellen Handlungskompetenz von Baumert und Kunter (2006) Wissen als Bestandteil von Handlungskompetenz versteht und damit impliziert, dass erfolgreiches Handeln durch Anwendung von Wissen realisiert werden kann, gehen andere Forscher davon aus, dass Expertenkönnen sich nicht als Wissensanwendung rekonstruieren lässt (Neuweg, 2002, S. 22). Begründet wird diese Sichtweise zum Beispiel damit, dass Lehrkräfte in zeitkritischen Handlungssituationen nicht bewusst auf Wissen zurückgreifen, auf dessen Basis sie eine Handlungsentscheidung fällen, und auch bei einer anschließenden Reflexion ihrer Handlungen nicht benennen können auf welcher Wissensbasis sie gehandelt haben

– „so, wie wir beim Sprechen Regeln der Grammatik zu befolgen vermögen, ohne sie explizit zu kennen“ (Neuweg, 2002, S. 13). Im Handeln unter Zeitdruck, von dem viele Unterrichtssituationen geprägt sind, würde demnach lediglich auf ein Repertoire an Handlungsmustern zurückgegriffen, das in vergleichbaren Situationen erworben wurde (Fischler, 2008, S. 32). Neuweg (2002, S. 123) spricht in diesem Zusammenhang von einem intuitiv-improvisierten Handeln und Polanyi (1985, S. 14, zitiert nach Neuweg, 2002, S. 13) darüber, „dass wir mehr wissen, als wir zu sagen wissen“. Dieses „mehr wissen“ stellt das implizite Wissen dar, das den Handelnden nicht immer bewusst ist (vergl. z. B. Fischler, 2008, S. 32). Neuweg (2002) ergänzt:

> [Wir] sagen oft auch mehr, als wir wissen können [...]. [Das] nachträgliche Angeben von handlungssteuerndem Wissen [ist] immer die Bewältigung einer Rekonstruktionsaufgabe, der Versuch einer ex-post-Rationalisierung eines ursprünglich mehr oder weniger spontanen Verhaltens. Insofern ist recht fraglich, ob [...] tatsächlich Auskunft über die Wissensbasis unseres Handelns [gegeben wird]. (S. 14)

Eine radikale Auslegung dieser Sichtweise würde die theoretisch ausgerichtete erste Phase der universitären Lehrerausbildung infrage stellen. Hier schränkt Neuweg (2002, S. 22) aber ein, dass das implizite Wissen „in hohem Maße theorieimprägniert ist“, da beim Handeln zwar nicht „*an*“, aber dennoch „*mit*“ dem expliziten Wissen gedacht wird. Die Suche nach handlungsleitendem Wissen bezeichnet er als Kategorienfehler und schlägt anstatt der Frage nach dem unbewusst angewandten Wissen die Frage danach, wie gut „[explizites] Wissen Können simuliert“ oder „inwieweit sich dieses Wissen zur *Instruktion* des fraglichen Könnens eignet“ (S. 17), vor. Insgesamt vertritt Neuweg (2002, S. 11) allerdings die Auffassung, dass explizites Wissen weder immer notwendig, noch hinreichend für intelligentes Handeln ist. Eine ausführliche Darstellung der verschiedenen Positionen und der in der Diskussion um handlungsleitendes Wissen eingeführten Konstrukte zur Beschreibung von Lehrerkognitionen wie *implizites Wissen*, *subjektive Theorien* oder *Habitus* findet sich in Vogelsang (2014, S. 110-129).

Zusammenfassend lässt sich feststellen, dass nicht hinreichend geklärt ist, ob und wie das explizierbare Professionswissen von Lehrkräften beim Unterrichten wirksam wird. Hinzu kommt, dass Wissen zwar vorhanden sein kann, aber dennoch nicht angewendet wird. So kann es vorkommen, dass Anfänger zwar theoretisch *wissen*, wie sie im Unterricht handeln sollten, es aber dennoch nicht *können* (Bromme, 1992, S. 131). Dieses ungenutzte Wissen wird nach Mandl, Gruber und Renkl (1993, S. 64) auch als „träges Wissen“ bezeichnet. Vor dem Hintergrund der Diskussion um die „Kluft zwischen Wissen und Handeln“ (Gruber, Mandl & Renkl, 2000, S. 139) drängt sich die Frage auf, ob bei der Erfassung des professionellen Wissens von Lehrkräften Wissen erfasst wird, das als handlungsleitend und damit als Voraussetzung für erfolgreiches und gutes Unterrichten angenommen werden kann. Will man diese Annahme überprüfen, muss man zum einen Kriterien für erfolgreiches Unterrichten formulieren und zum anderen die Qualität des Unterrichts selbst betrachten, in dem die Lehrkraft als handelnde Person in Erscheinung tritt.

## 3.2. Kriterien erfolgreichen Unterrichts

Der Unterricht einer Lehrkraft kann als „erfolgreich“ bezeichnet werden, wenn die Ziele, die mit Unterricht verfolgt werden sollen, auch erreicht werden. Ein zentrales Ziel des Unterrichts ist neben allgemeinen, fächerübergreifenden Bildungszielen (wie z. B. Erziehung zu Freiheit, Demokratie und Toleranz, Verständlichmachen kultureller und religiöser Werte) ein Kompetenzaufbau aufseiten der Schülerinnen und Schülern (KMK, 2005a, S. 7). Nach Weinert (2001, S. 27-28) sind Kompetenzen definiert als „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“.
Auf Grundlage dieser Definition können zahlreiche Perspektiven für die Bewertung erfolgreichen Unterrichts eingenommen werden und sowohl kognitive, als auch motivationale und volitionale Zielkriterien für Unterricht formuliert werden. In Anlehnung an die Bildungsstandards der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (KMK) können als Kriterien für erfolgreiches Unterrichten von Physik in der Schule beispielsweise Schülerleistungen in den Kompetenzbereichen Fachwissen, Erkenntnisgewinnung, Kommunikation und Bewertung betrachtet werden (z. B. KMK, 2005b, S. 7).
Neben Leistungszuwächsen kann auch der Abbau von Leistungsunterschieden zwischen leistungsschwachen und leistungsstarken Schülerinnen und Schülern von Interesse sein (Helmke, 2009, S. 40).

Als weitere „[wichtige] Zielgrößen des Bildungssystems“ bezeichnen Jansen, Schroeders und Stanat (2013, S. 347) motivationale Aspekte schulischer Kompetenzen wie Selbstkonzept oder Interesse. Auch Weinert und Helmke (1996, S. 226) nennen neben leistungsbezogenen, motivationale und affektive Kriterien wie Lernfreude und Selbstkonzept der Lernenden als Zielkriterien für erfolgreiches Unterrichten. Für den Physikunterricht stellt mit Blick auf den Fachkräftemangel in naturwissenschaftlich-technischen Berufen insbesondere die Förderung des Fachinteresses der Lernenden ein wichtiges Ziel dar:

> Eine hohe Kompetenz in Naturwissenschaften und Mathematik, die in der Schule erworben wurde, erleichtert den Einstieg in ein MINT-Studium. Doch wer als Schüler eine hohe MINT-Kompetenz erreicht, muss sich später noch lange nicht für ein MINT-Studium entscheiden. Für eine solche Entscheidung ist insbesondere auch ein hohes Interesse an MINT-Fragestellungen vonnöten, (Hetze, 2011, S. 8)

stellt auch Hetze (2011) fest. In Bezug auf Interesse kann man zwischen *individuellem Interesse* („Wertschätzung eines spezifischen Gegenstands oder Themas“) und *situationalem Interesse* („durch äußere Umstände hervorgerufene Zustand des Interessiertseins“) unterscheiden (Schiefele, 2008, S. 46-47). Beim Fachinteresse handelt es sich um individuelles Interesse, das als relativ stabiles Personenmerkmal gilt (Schiefele, 2008, S. 46). Es wird allerdings angenommen, dass individuelles Interesse durch situationales Interesse beeinflusst werden kann (Krapp, 2002,

S. 406). Situationales Interesse begünstigt zudem die intrinsische Motivation der Lernenden (Schiefele, 2008, S. 46), die wiederum als „unerlässliche Voraussetzung des Wissenserwerbs“ gilt (Edelmann, 2003, S. 32). Schiefele und Schreyer (1994) geben einen guten Überblick über zahlreiche Studien, die einen positiven Einfluss der intrinsischen Motivation auf Lernerfolg belegen. Insbesondere vor dem Hintergrund der Erfolgsunsicherheit des Lehrerhandelns – die Lehrkraft kann lediglich ein Lehrangebot zur Verfügung stellen, die Nutzung des Angebots obliegt den Lernenden – kann daher auch das situationale Interesse der Lernenden als Zielkriterium für erfolgreiches Unterrichten betrachtet werden, da erfolgreiches Unterrichten die Teilnahmemotivation der Lernenden am Unterricht erfordert. Unterricht verfolgt niemals nur ein einzelnes Ziel. Daher sollte man sich bei der Identifizierung erfolgreichen Unterrichts nicht auf ein Zielkriterium beschränken, sondern stattdessen eine multikriteriale Perspektive einnehmen (vergl. z. B. Helmke, 2009, S. 84-85, Weinert & Helmke, 1996, S. 226).

Da es schwierig ist im Unterricht zu beobachten, ob Schülerinnen und Schüler die ihnen bereitgestellten Lerngelegenheiten auch nutzen, beziehen sich die beschriebenen Zielkriterien erfolgreichen Unterrichts lediglich auf die Ergebnisse von Unterricht, den „Output“. Das Professionswissen von Lehrkräften wirkt allerdings nicht zwingend direkt auf Schüleroutputvariablen. Nur weil eine Lehrkraft mehr weiß als andere Lehrkräfte, wissen die Schülerinnen und Schüler dieser Lehrkraft nicht automatisch mehr oder sind motivierter. Vielmehr liegt dem Wirkzusammenhang die Annahme zu Grunde, dass Lehrkräfte ihr Wissen dazu nutzen das Lehrangebot optimal zu gestalten und die Lernenden bei dessen Nutzung zu unterstützen. Dies spiegelt sich auch im Angebots-Nutzungsmodell von Helmke (2009, S. 73) (vergl. Abbildung 2.1 auf Seite 8) wieder. Im Hinblick auf die Untersuchung des Zusammenhangs zwischen Wissen und Handeln und vor dem Hintergrund der Erfolgsunsicherheit des Lehrerhandelns ist es daher von Interesse auch die Qualität des Unterrichts zu betrachten.

## 3.3. Unterrichtsqualität

Bei der Betrachtung von Merkmalen guten Unterrichts unterscheidet man zunächst zwischen *Oberflächenmerkmalen* (wie z. B. Sozial- und Inszenierungsformen oder Methoden und Gestaltungsformen von Unterricht) und *Tiefenstrukturmerkmalen* von Unterricht: „Bezieht sich die Oberflächenstruktur auf die variablen und daher austauschbaren Handlungs- und Formelemente des Unterrichts, so bezieht sich die Tiefenstruktur auf dessen invariante, psychologisch notwendige Basisprozesse und Elemente“ erklärt Reusser (2009, S. 888). Bedeutung für eine Kompetenzentwicklung aufseiten der Schülerinnen und Schülern wird eher den Tiefenstrukturmerkmalen von Unterricht zugesprochen (vergl. z. B. Neumann, Kauertz & Fischer, 2012, S. 256), da Oberflächenmerkmale[1] in der Regel keinen Einfluss auf Schülervariablen haben (vergl. z. B. Seidel & Prenzel, 2006, S. 238, Olszewski, 2010, S. 94, für einen Überblick vergl. Vogelsang, 2014, S. 208-211). So stellt auch Reusser (2009, S. 888) fest, dass es „[als] sicher [gelten] kann [...], dass es

[1] mit Ausnahme des Merkmals *Time on Task*

auf der Ebene spezifischer Unterrichtsmethoden keinen »Königsweg« des Lehrens gibt". Als Merkmale von Unterrichtsqualität werden daher Tiefenstrukturmerkmale von Unterricht betrachtet.

Als fächerübergreifende allgemeine Merkmale von gutem Unterricht nennt Helmke (2009, S. 168-169) *Klassenführung*, *Klarheit und Strukturiertheit*, *Konsolidierung und Sicherung*, *Aktivierung*, *Motivierung*, *lernförderliches Klima*, *Schülerorientierung*, *Kompetenzorientierung*, *Umgang mit Heterogenität* und *Angebotsvariation*. Da die Bedeutsamkeit der Merkmale vom jeweiligen Bildungsziel abhängig sein kann und die Merkmale zudem untereinander konkurrieren können, ist qualitätvoller Unterricht allerdings nicht automatisch mit maximalen Ausprägungen in allen Merkmalen gleichzusetzen. Zudem ist die empirische Wirksamkeit der einzelnen Merkmale unterschiedlich gut belegt (Helmke, 2009, S. 170).

In Anlehnung an Helmkes Angebots-Nutzungsmodell schlagen Fischer et al. (2014b) für die physikdidaktische Unterrichtforschung ein Unterrichtsqualitätsmodell vor, das die Tiefenstruktur von Unterricht über die Merkmale *Klassenführung*, *experimentelles Handeln*, *Sachstruktur*, *Motivierung*, *enthusiastisches Lehrerhandeln*, *Interaktion zwischen Lehrenden und Lernenden* und *kognitive Aktivierung* modelliert.

Im Rahmen der „Third International Mathematics and Science Study" (TIMSS) konnten Klieme, Schümer und Knoll (2001, S. 51) zeigen, dass die Qualität des Mathematikunterrichts über die drei empirisch aus Videodaten gewonnenen Faktoren *Klassenführung*, *Schülerorientierung* und *kognitive Aktivierung* beschrieben werden kann. Diese Faktoren werden als „Grunddimensionen der Unterrichtsqualität" bezeichnet (S. 51). Vergleicht man unterschiedliche Konzeptualisierungen der Merkmale fällt schnell auf, dass ein gewisser Überlapp zwischen den Konstrukten besteht und die Merkmale nicht zwingend als disjunkt angenommen werden können. Daher stellt sich die Frage, ob die Bezeichnung der Merkmale als „Dimensionen" von Unterrichtsqualität als angemessen erachtet werden kann. Da in der Literatur in Bezug auf diese drei Merkmale stets der Begriff „Dimension" verwendet wird, wird diese (wenn auch nicht ganz saubere) Bezeichnung auch hier gewählt. Auf die gleichen Dimensionen bezieht sich auch die deutsch-schweizerische Studie „Unterrichtsqualität und mathematisches Verständnis in verschiedenen Unterrichtskulturen" (auch bekannt als „Pythagoras"-Studie), wobei hier anstatt von *Schülerorientierung* von *unterstützendem Unterrichtsklima* gesprochen wird (Klieme, Lipowsky, Rakoczy & Ratzka, 2006, S. 131). Auch nach Voss, Kunter, Seiz, Hoehne und Baumert (2014, S. 186) lassen sich viele Aspekte „guten" Unterrichts den Dimensionen *Klassenführung*, *konstruktive Unterstützung* und *kognitive Aktivierung* zuordnen. Aus fachdidaktischer Sicht ist insbesondere das Merkmal *kognitive Aktivierung* interessant:

> While supportive climate and effective classroom management can be identified as more general qualities of the learning environment on the classroom or even the school level, or possibly as a general component of teacher competence, cognitive activation can only be judged with respect to the specific content that is being taught, the

> way it is implemented, and how the instructional process is related to students' prerequisites. (Klieme, Pauli & Reusser, 2009, S. 142)

Im Folgenden soll kurz auf die Konstrukte *Klassenführung* und *konstruktive Unterstützung* und ausführlicher auf das Konstrukt der *kognitiven Aktivierung* eingegangen werden.

### 3.3.1. Klassenführung

Das Konstrukt der Klassenführung bezieht sich darauf, wie Unterricht organisiert wird. Ganz allgemein versteht man unter Klassenführung eine proaktive Steuerungsleistung der Lehrkraft, die einen störungsarmen Unterricht mit möglichst wenigen Unterbrechungen ermöglichen soll und so die Bereitstellung zeitlicher Ressourcen für das Initiieren von Lernprozessen im Unterricht sicherstellt (Voss et al., 2014, S. 187).

Nach Kounin (2006, S. 10,148-149)[2] kann eine Lehrkraft eine gute Klassenführung realisieren, indem sie *Allgegenwärtigkeit* gegenüber den Lernenden demonstriert, simultan auf gleichzeitig auftretende Probleme reagiert (*Überlappung*), den Unterrichtsablauf und Übergänge reibungslos gestaltet und Sprunghaftigkeit und Inkonsistenz vermeidet (*Reibungslosigkeit und Schwung*), einen Gruppenfokus bewahrt und sich gleichzeitig Freiraum für eine systematische Berücksichtigung individueller Unterschiede schafft (*Gruppenmobilisierung und Rechenschaftsprinzip*) und außerdem durch *Abwechslung und Herausforderung* zur Mitarbeit im Unterricht motiviert und Überdruss aufseiten der Lernenden vermeidet. Letztere Kategorie wird allerdings in aktuellen Konzeptualisierungen von Klassenführung meist nicht berücksichtigt (vergl. z. B. Baumert & Kunter, 2006; Fricke, van Ackeren, Kauertz & Fischer, 2012; Klieme et al., 2001; Seidel & Shavelson, 2007; Voss et al., 2014) und kann zudem eher der Dimension der kognitiven Aktivierung zugeordnet werden.

Während der Umgang mit Disziplinproblemen ebenfalls als Merkmal der Klassenführung gilt, wird der Prävention von Störungen ein weitaus höherer Stellenwert zugeschrieben (Voss et al., 2014, S. 187). Um diesem Umstand Rechnung zu tragen, betrachten aktuelle Arbeiten auch die Klarheit von Regeln und Ritualen als Merkmal guter Klassenführung (Fricke, 2015, S. 19). Klieme et al. (2001, S. 53) können empirisch belegen, dass Klassenführung eine „notwendige, wenngleich nicht hinreichende Vorbedingung für die kognitive Aktivierung" darstellt.

### 3.3.2. Konstruktive Unterstützung

Mit dem Konstrukt der konstruktiven Unterstützung wird beschrieben, inwiefern die Lehrkraft die Lernenden im Unterricht in ihrem Streben nach Autonomie, Kompetenz und sozialer Eingebundenheit unterstützt und so die Entwicklung intrinsischer Motivation begünstigt (Klieme et al., 2006, S. 129). Theoretisch verankert ist dieses Konstrukt in der Selbstbestimmungstheorie von Deci und Ryan

[2] Hierbei handelt es sich um eine Neuauflage von Kounins 1970 erschienenem Standardwerk „Techniken der Klassenführung"

(1993). Ein konstruktiv-unterstützender, schülerorientierter Unterricht zeichnet sich durch eine wertschätzende Schüler-Lehrerbeziehung, Unterstützung bei persönlichen Problemen, Förderung sozialer Integration, konstruktiven Umgang mit Fehlern, adaptiven Umgang mit Verständnisproblemen, Geduld und positive Rückmeldungen aus (vergl. z. B. Klieme et al., 2006, S. 129-132; Kunter et al., 2006, S. 166-167; Voss et al., 2014, S. 187).

### 3.3.3. Kognitive Aktivierung

Der Begriff der kognitiven Aktivierung ist geprägt durch die Arbeiten der Arbeitsgruppe um Jürgen Baumert und taucht erstmals explizit im Rahmen einer Veröffentlichung zu TIMSS 1995 auf (Klieme et al., 2001, S. 50-51). Stellt man die Frage, *ob* Lernende im Unterricht kognitiv aktiviert sind, kann der Begriff der kognitive Aktivierung zur Beschreibung der Nutzung des Lehrangebots verwendet werden. Mit Bezug zur Qualität des Lehrangebotes, beschreibt der Begriff „kognitive Aktivierung“ hingegen, *inwieweit* versucht wird die Lernenden zu einer kognitiv aktiven Auseinandersetzung mit dem Lerngegenstand anzuregen (Kunter et al., 2006, S. 165).

Theoretisch basiert die Konzeptualisierung der kognitiven Aktivierung auf einem sozial-konstruktivistischen Lehr-Lernverständnis und den Grundsätzen verständnisvollen Lernens (Baumert & Köller, 2000; Hugener, Rakoczy, Pauli & Reusser, 2006; Pauli & Reusser, 2003). „Verständnisvolles Lernen ist ein aktiver individueller Konstruktionsprozess, in dem Wissensstrukturen verändert, erweitert, vernetzt, hierarchisch geordnet oder neu generiert werden. [...] Die soziale Rahmung von Lernprozessen ist demnach unter der Perspektive zu beurteilen, inwieweit sie diese mentale Aktivität stützt, fördert oder erschwert“, fassen Baumert und Köller (2000, S. 273-274) zusammen. Dabei ist die Unterscheidung zwischen Aktivität und *mentaler* Aktivität zentral. So weist auch Meyer (2004) darauf hin, dass

> [activity] may help promote meaningful learning, but instead of behavioral activity per se (e.g., hands-on activity, discussion, and free exploration), the kind of activity that really promotes meaningful learning is cognitive activity (e.g., selecting, organizing, and integrating knowledge). Instead of depending solely on learning by doing or learning by discussion, the most genuine approach to constructivist learning is learning by thinking. (S. 17)

In der Literatur findet sich keine klare Definition des Konstruktes der kognitiven Aktivierung. Vielmehr wird beschrieben, wie eine kognitiv aktivierende Unterrichtsgestaltung aussehen sollte. Bei der kognitiven Aktivierung handelt es sich um ein Merkmal der Tiefenstruktur von Unterricht – Merkmale einer kognitiv aktivierenden Unterrichtsgestaltung werden also daraus abgeleitet, inwiefern sie sich an den Lernprozessen der Lernenden orientieren. Die Lernprozesse selbst sind nicht direkt beobachtbar, es können aber Lehrerhandlungen beschrieben werden, die sogenannte „Gelegenheitsstrukturen“ für verschiedene Funktionen im Lernprozess der Schülerinnen und Schüler schaffen und damit bestimmte Lernaktivitäten wahrscheinlich machen (Hugener, 2008, S. 56; Seidel, 2003, S. 137). Die meisten Arbeiten greifen

hierfür auf Merkmale eines problemlösenden oder konstruktivistisch-orientierten Unterrichts zurück (vergl. z. B. Hugener, 2008; Klieme et al., 2001; Kunter, 2005; Rakoczy & Pauli, 2006; Widodo & Duit, 2004).

In der Tat ist es so, dass Merkmale eines solchen Unterrichts, wie beispielsweise der angemessene Umgang mit Schülervorstellungen und dem Vorwissen der Lernenden, nach und nach in immer mehr Arbeiten unter dem Label „kognitive Aktivierung“ diskutiert wurden (vergl. z. B. Hugener, 2008; Vogelsang, 2014). Der in der Literatur zu Schülervorstellungen oder zum konstruktivistisch-orientierten Unterricht bewandte Leser sei daher auf gewisse Überschneidungen des Konstruktes mit anderen Konstrukten der Unterrichtsforschung hingewiesen.

#### 3.3.3.1. Merkmale eines kognitiv aktivierenden Unterrichts

Zentral für eine kognitiv aktivierende Unterrichtgestaltung ist die Schaffung herausfordernder Lerngelegenheiten. Kunter (2005, S. 91) vermutet im Erleben von Herausforderungen im Unterricht sogar einen der wichtigsten Prozesse für die Unterstützung des Kompetenzerlebens von Schülerinnen und Schülern und eines bedeutungsvollen und selbstbestimmten Lernens. Herausfordernde Lerngelegenheiten können durch Fragestellungen realisiert werden, die die Lernenden zum Nachdenken anregen, ohne jedoch eine Überforderung darzustellen (Rakoczy & Pauli, 2006, S. 227). Als nicht kognitiv aktivierend werden hingegen ein enges, kleinschrittiges Frageverhalten der Lehrkraft und rezeptartige Anleitungen zur Bearbeitung von Aufgaben angesehen (Rakoczy & Pauli, 2006, S. 227). Die Lernenden werden hierbei weder zum Nachdenken angeregt, noch haben sie die Möglichkeit eigene Ideen und Vorstellungen in den Unterricht einzubringen.

Die intuitiven Vorstellungen physikalischer Konzepte, mit denen Lernende in den Unterricht kommen, stimmen oftmals nicht mit der wissenschaftlichen Sichtweise überein (Duit & Treagust, 2003, S. 671). Diese Vorstellungen beeinflussen allerdings, ebenso wie das Vorwissen der Lernenden, die Verarbeitung neuer Informationen – Lernen kann demnach als individueller Verstehensprozess aufgefasst werden, der nicht für alle Schülerinnen und Schüler gleich abläuft (Kunter, 2005, S. 31). Um die Konstruktion von Wissen zu unterstützen und Lernprozesse in Gang zu setzen, muss an die vorunterrichtlichen Vorstellungen und das Vorwissen der Lernenden angeknüpft werden (Kunter, 2005, S. 55).

Besonders herausfordernde Lerngelegenheiten und eine aktive Auseinandersetzung mit dem Unterrichtsgegenstand können durch die Erzeugung kognitiver Konflikte realisiert werden: durch das Provozieren von Situationen, in denen den Lernenden bewusst wird, dass ihre eigenen Interpretationen nicht ausreichen, um bestimmte Sachverhalte zu erklären (Kunter, 2005, S. 91; Rakoczy & Pauli, 2006, S. 227). Die Konfrontation der Lernenden mit den Grenzen ihrer eigenen Interpretationen kann auch indirekt durch ein genetisch-sokratisches Vorgehen der Lehrkraft im Unterricht erfolgen, indem die Lehrkraft die Lernenden auf ihren Vorstellungen aufbauend argumentieren und schlussfolgern lässt – auch wenn diese Vorstellungen falsch sind – und sie solange in die Irre laufen lässt, bis sie es selbst merken (vergl. Clausen, 2002, S. 114; Klieme et al., 2001, S. 51). Eine kognitiv

aktivierende Unterrichtsgestaltung ist also auch durch einen evolutionären oder revolutionären Umgang mit den Vorstellungen der Lernenden gekennzeichnet.

Durch das Anknüpfen an die Vorstellungen der Lernenden kann die Lehrkraft versuchen bestehende Konzepte im Sinne einer Konzepterweiterung oder im Sinne eines Konzeptwechsels zu verändern (Rakoczy & Pauli, 2006, S. 226). Hierzu müssen die Vorstellungen und das Vorwissen der Lernenden zunächst durch die Lehrkraft exploriert werden. Diese Exploration erfüllt zwei Funktionen: Zum einen kann die Lehrkraft Aufschluss darüber bekommen, wo sie die Lernenden abholen muss, zum anderen wird das Vorwissen der Lernenden aktiviert und ihnen selbst bewusst gemacht. Letzteres wird auch im Rahmen der Basismodelltheorie von Oser und Baeriswyl (2001) als notwendiger erster Schritt für einen an den Lernprozessen der Lernenden orientierten Konzeptaufbau betrachtet.

Die Aktivierung des Vorwissens kann zudem durch ein Bewusstmachen des Lernstatus im jeweiligen Thema unterstützt werden. Indem die Lehrkraft deutlich macht, worauf neue Lerninhalte aufbauen und worauf sie abzielen und Verbindungen zu früher Gelerntem und neu zu Lernenden aufzeigt, fördert sie die vertikale Vernetzung des Wissens und regt die Lernenden zur Integration des neu zu lernenden Wissens in ihr bestehendes Wissenssystem an (Rakoczy & Pauli, 2006, S. 224). Durch das Bewusstmachen des Lernstatus wird zudem ein zielgerichtetes Lernen der Schülerinnen und Schüler möglich.

Neben den Vorstellungen der Lernenden können auch ihre Denkweisen von den wissenschaftlichen Denkweisen abweichen. Dies kann beispielsweise dazu führen, dass sie Experimente oder Phänomene, anders als von der Lehrkraft intendiert, erklären. Um zu klären, worauf bestimmte Schülervorstellungen zurückzuführen sind oder um zu verstehen, warum die Lernenden bestimmte Erklärungen ablehnen, sollte die Lehrkraft versuchen die Denkweisen der Lernenden nachzuvollziehen (Widodo & Duit, 2004, S. 239). Auf dieser Basis können dann angemessene Aktivitäten oder besser an das Verständnis der Lernenden anknüpfende Repräsentationsformen für den weiteren Unterrichtsverlauf ausgewählt werden. Nur wenn die Lehrkraft mit den Denkweisen der Lernenden vertraut ist, kann sie neue Konzepte auf eine Art einführen, die nachvollziehbar für die Schülerinnen und Schüler ist (Rakoczy & Pauli, 2006, S. 226). Durch das Einfordern von Begründungen und Erklärungen werden gleichzeitig die durch einzelne Schülerinnen und Schüler vertretende Standpunkte für den Rest der Klasse transparent, was die soziale Ko-Konstruktion des Wissens begünstigt.

Grundsätzlich kann eine Lehrkraft die Bedingungen für das soziale Aushandeln von Bedeutungen schaffen, indem sie im Unterricht die Rolle eines Mediators einnimmt (Rakoczy & Pauli, 2006, S. 228). Dadurch, dass sie die Äußerungen der Lernenden moderiert, zueinander in Bezug setzt, den Lernenden Zeit gibt Ideen und Antworten zu äußern oder zu finden und sie bei deren Ausformulierung unterstützt, ohne direkte Bewertungen vorzunehmen, fördert sie den Austausch von Ideen und damit einen aktiven Diskurs der Lernenden im Unterricht (Klieme & Clausen, 1999, S. 6).

#### 3.3.3.2. Zusammenhang von kognitiv aktivierendem Unterricht und Zielkriterien von Unterricht

Nach der Definition von Unterrichtsqualität von Fenstermacher und Richardson (2005, S. 192) – Unterricht muss sowohl gut, als auch erfolgreich sein – kann kognitive Aktivierung nur dann als Merkmal der Unterrichtsqualität bezeichnet werden, wenn ein kognitiv aktivierend gestalteter Unterricht mit der erfolgreichen Erreichung von Zielkriterien des Unterrichts verbunden ist. Während zwar durchaus empirische Evidenz für einen Zusammenhang zwischen Merkmalen kognitiv aktivierenden Unterrichts und Zielkriterien erfolgreichen Unterrichts wie Schülerleistung oder Interesse existiert, können diese Zusammenhänge dennoch nicht als empirisch abgesichert gelten.

Baumert und Köller (2000) konnten auf Basis von Fragebogendaten aus TIMSS III 1995 Zusammenhänge zwischen der von den Lernenden wahrgenommenen „Verständnisorientierung durch kognitive Herausforderung“ und Schülerleistung im Mathematik- wie auch im Physikunterricht der Sekundarstufe II nachweisen. Unter Kontrolle des Kursniveaus (Leistungskurs vs. Grundkurs) zeigten sich allerdings keine Zusammenhänge mehr. Im Physikunterricht wurden die Leistungsunterschiede zwischen Grund- und Leistungskursen zwar über Unterrichtmerkmale mediiert (Baumert & Köller, 2000, S. 295), der Mediatoreffekt des Merkmals Verständnisorientierung, das der kognitiven Aktivierung zugeordnet werden kann, wurde allerdings nicht separat untersucht. Gruehn (2000, zitiert nach Hugener, 2008, S. 76) konnte Zusammenhänge zwischen dem Einsatz kognitiv anspruchsvoller Übungsaufgaben und einem genetisch-sokratischen Vorgehen der Lehrkraft (erhoben durch Schülerfragebögen) und Schülerleistung im Mathematik- und Physikunterricht nachweisen.

Videostudien, in denen Merkmale kognitiv aktivierenden Unterrichts durch externe Beobachter auf Basis hoch-inferenter Ratings beurteilt wurden, zeigen widersprüchliche Ergebnisse zum Zusammenhang zwischen kognitiver Aktivierung und Schülervariablen. Während Klieme et al. (2001) und Lipowsky et al. (2009) kleine Zusammenhänge zwischen kognitiv aktivierender Unterrichtsgestaltung und Schülerleistungszuwächsen in Mathematik nachweisen konnten, fanden Klieme und Clausen (1999), Kunter (2005) und Olszewski (2010) keine signifikanten Zusammenhänge. Klieme und Clausen (1999, S. 12) konnten allerdings Zusammenhänge zur Interessenentwicklung der Lernenden nachweisen. In der Studie „Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz“ (COACTIV) erwies sich das Potential zur kognitiven Aktivierung von im Unterricht eingesetzten Aufgaben als signifikanter Prädiktor für die Mathematikleistung der Lernenden, nicht aber für die Freude an Mathematik (Kunter & Voss, 2011, S. 104).

Eine detailliertere Übersicht über die zitierten Studien findet sich in Hugener (2008) und Vogelsang (2014). Zu beachten ist, dass die genannten Studien zum Teil recht unterschiedliche Operationalisierungen des Konstrukts der kognitiven Aktivierung vornehmen. Zudem werden sehr unterschiedlich große Stichprobengrößen untersucht und die Ergebnisse resultieren teils aus Korrelationsanalysen auf Klassenebene und teils aus Mehrebenenenanalysen. Darüber hinaus stehen

die gemessenen Schülerleistungen nicht immer in direktem Bezug zum beurteilten Unterricht (beispielsweise weil internationale Leistungstests aus TIMSS oder dem „Programme for International Student Assessment“ (PISA) eingesetzt wurden).

In diesem Kapitel wurde zunächst ein kurzer Einblick in die Diskussion um den Zusammenhang zwischen dem Wissen und Handeln von Lehrkräften gegeben, da es sich hierbei um eine Grundvoraussetzung für die Annahme handelt, dass das Professionswissen von Lehrkräften relevant für erfolgreiches und gutes Unterrichten ist. Um diese Annahme zu diskutieren, ist es notwendig, zunächst zu klären, was man unter erfolgreichem und gutem Unterrichten versteht. Es wurden daher mögliche Zielkriterien erfolgreichen Unterrichts sowie Merkmale guten Unterrichts vorgestellt. Viele dieser Merkmale lassen sich den drei Dimensionen *Klassenführung*, *konstruktive Unterstützung* und *kognitive Aktivierung* zuordnen. Die fachdidaktische Unterrichtsforschung beschäftigt sich insbesondere mit der kognitiven Aktivierung, da dieses Merkmal einen höheren Fachbezug aufweist – so ist beispielsweise die Schaffung herausfordernder Lerngelegenheiten durch eine Lehrkraft, die nur über pädagogisches, nicht aber über Fachwissen oder fachdidaktisches Wissen verfügt, nur schwer vorstellbar (ausführliche Überlegungen hierzu werden in Abschnitt 5.3 auf Seite 65 vorgenommen). Da die empirischen Ergebnisse zum Zusammenhang zwischen einer kognitiv aktivierenden Unterrichtsgestaltung und Schülerleistung oder Schülerinteresse nicht unabhängig von dem in den jeweiligen Studien gewähltem methodischen Vorgehen zu sein scheinen, kann allerdings nicht per se davon ausgegangen werden, dass man bei der Erhebung der kognitiven Aktivierung ein Merkmal der Unterrichtsqualität im Sinne von Fenstermacher und Richardson (2005, S. 192) erfasst.

# 4. Herausforderungen in der empirischen Professionswissensforschung

Die Forschung zum Professionswissen von Lehrkräften beschränkt sich nicht auf die theoretische Modellierung dieses Wissens und der Wissensdimensionen. Vielmehr gibt es zahlreiche Studien, die sich mit der Erfassung von Professionswissen beschäftigen (für einen Überblick siehe z. B. Abell, 2007). Am Anfang dieses Kapitel wird zunächst kurz begründet, warum in den letzten Jahren gerade die Entwicklung von schriftlichen Testinstrumenten zur Erfassung des Professionswissens vorangetrieben wurde. Bisher ist nicht hinreichend empirisch abgesichert, *ob* Professionswissen als Voraussetzung für gutes und erfolgreiches Unterrichten gelten kann und falls ja, für *welches* Wissen dies gilt – schließlich besteht keineswegs Konsens darüber, wie Professionswissen modelliert werden sollte. Daher wird auf die Problematik hingewiesen, die mit der Interpretation von Daten einhergeht, die auf Grundlage derartiger Professionswissenstests erhoben wurden. So werden diese oft genutzt, um Aussagen über die Güte der Lehrerausbildung zu treffen – die Validität dieser Aussagen ist allerdings fraglich, sofern nicht gezeigt wird, dass das erhobene Wissen auch wirklich relevant für den Lehrberuf ist.

In Abschnitt 4.2 auf Seite 36 wird diskutiert, was Validität im Kontext der Professionswissensforschung eigentlich bedeutet. Abschließend werden in Abschnitt 4.3 auf Seite 41 Studien aus der Mathematik und Physik vorgestellt, die Zusammenhänge zwischen dem Professionswissen von Lehrkräften und gutem und erfolgreichem Unterricht untersuchen und damit auch die prädiktive Validität ihrer Testinstrumente überprüfen. Die Kontrastierung der Studien in den zwei Fächern soll deutlich machen, dass insbesondere für den Physikunterricht noch nicht hinreichend geklärt ist, welches Wissen als unterrichtsrelevant angenommen werden kann.

Das Ziel dieses Kapitel ist es, ein Problembewusstsein dafür zu schaffen, dass nicht davon ausgegangen werden kann, dass mit Testinstrumenten zur Erfassung des Professionswissens Wissen erfasst wird, das prädiktiv für gutes und erfolgreiches Unterrichten ist.

## 4.1. Erfassung von Professionswissen

Das CK, PCK oder PK von Lehrkräften wurde anfangs lediglich indirekt über distale Indikatoren wie staatliche Zertifizierungen, Abschlüsse, Ausbildungsdauer oder die Anzahl besuchter Fachkurse gemessen (vergl. Abell, 2007, S. 1110; Baumert & Kunter, 2006, S. 485/490; Fischer et al., 2012, S. 10). Baumert und

Kunter (2006, S. 490) beklagen diesbezüglich, dass „[diese] Indikatoren [...] keine Auskunft über Inhalt, Struktur und Qualität des fachlichen Wissens [geben] und der Erklärungsabstand zu Unterrichtsprozessen sowie zum Lernfortschritt von Schülerinnen und Schülern [...] groß“ ist. Die direkte Erfassung des Professionswissens von Lehrkräften ist zwar wesentlich aufwendiger, stellt aber eine wichtige Aufgabe für die empirische Bildungsforschung dar.

Loughran, Mulhall und Berry (2004, S. 373) führen aus, warum insbesondere die Erfassung des fachdidaktischen Wissens von Lehrkräften keine einfache Aufgabe darstellt: Bei PCK handelt es sich, wie bei CK und PK auch, um ein internales Konstrukt, das nicht direkt im Unterricht sichtbar werden muss. Die Erfassung von PCK muss also über die direkte Befragung von Lehrkräften erfolgen. Während Lehrkräfte es meist gewohnt sind, ihr Fachwissen und pädagogisches Wissen explizit zu artikulieren, ist das für ihr PCK eher nicht der Fall. So fehlt den Lehrkräften zum Teil das entsprechende Vokabular um ihr implizites Wissen zu explizieren. Loughran et al. (2004, S. 373) versuchen das PCK von Naturwissenschaftslehrkräften über speziell für diesen Zweck entwickelte Interviewtechniken, sogenannte *Content Representations* (CoRes) und *Professional and Pedagogical experience Repertoires* (PaP-eRs), im Rahmen von Gruppen- und Einzelinterviews zu erfassen. Damit verfolgen die Autoren wie viele andere Forscher auch (vergl. z. B. De Jong & Van Driel, 2004; De Jong, Van Driel & Verloop, 2005; Drechsler & Van Driel, 2008; Grossman, 1990; Loughran et al., 2004; Park & Chen, 2012) einen qualitativen Ansatz um das PCK von Lehrkräften zu ergründen. „Both the CoRe and the PaPers are qualitative in nature thus are more suitable for capturing than measurement“ stellen (Rollnick & Mavhunga, 2014, S. 356) diesbezüglich fest, da die Erfassung von PCK über qualitative Ansätze sehr zeitintensiv ist.

Im Zuge der Formulierung von Standards für die Lehrerbildung in den USA und später auch in Deutschland (KMK, 2004, 2008) wurde, mit Blick auf sich daraus ergebene Möglichkeiten zur Evaluation der Lehrerausbildung, die Entwicklung standardisierter Testinstrumente zur Erfassung von Professionswissen im Rahmen von Large-Scale Assessments angestoßen. Vorreiter hierfür war die Studie „Mathematics Teaching in the 21st Century“ (MT21), in deren Rahmen erstmals schriftliche Testinstrumente zur standardisierten Erfassung des Fachwissens, des fachdidaktischen Wissens und des pädagogischen Wissens von angehenden Mathematiklehrkräften entwickelt wurden (Blömeke et al., 2010, S. 30). Genutzt wurden diese Instrumente, um Unterschiede im Umfang mathematischer, mathematikdidaktischer und pädagogischer Lerngelegenheiten in der Lehrerausbildung zu untersuchen und zu analysieren, wie diese mit dem CK, PCK und PK angehender Mathematiklehrkräfte zusammenhängen (Blömeke et al., 2008; Blömeke et al., 2010). Ähnliche Fragestellungen werden in der in 17 Ländern durchgeführten Studie „The Teacher Education Study in Mathematics“ (TEDS-M) untersucht (Blömeke, Kaiser & Lehmann, 2010; Tatto et al., 2008, 2012).

Auch in den Naturwissenschaften wurden inzwischen zahlreiche schriftliche Testinstrumente zur Erhebung des Professionswissens von (angehenden) Lehrkräften

entwickelt (z. B. Brovelli et al., 2013; Kröger et al., 2015; M. Ndlovu, 2014; Riese, 2009; Riese et al., 2015; Schmelzing, 2010).[1]

An dieser Stelle sei angemerkt, dass die Erfassung des Professionswissens sich in der Regel nicht auf die Erfassung dekontextualisierten Faktenwissens beschränkt – vielmehr wird versucht Kompetenzen im Sinne des auf kognitive Bereiche eingeschränkten Kompetenzbegriffs von Klieme und Leutner (2006, S. 879) zu erfassen, die Kompetenzen „als *kontextspezifische kognitive Leistungsdispositionen*, die sich funktional auf Situationen und Anforderungen in bestimmten *Domänen* beziehen", definieren. Hierfür werden sowohl Aufgaben im Multiple-Choice Format als auch offene Aufgaben oder sogenannte Unterrichtsvignetten genutzt, die authentische Unterrichtssituationen beschreiben, in denen spezifische Anforderungen bewältigt werden müssen.

Allen diesen Studien ist gemein, dass sie von der Annahme ausgehen, dass das Professionswissen von Lehrkräften Voraussetzung für erfolgreiches Unterrichten darstellt, ohne diese Annahme empirisch zu überprüfen. Solange kein Bezug zur Qualität des Lehrangebots oder zu Zielkriterien von Unterricht wie beispielsweise Leistungszuwachs hergestellt wird, kann allerdings nicht entschieden werden, ob das erhobene Wissen relevant für erfolgreiches und gutes Unterrichten ist. Gerade vor dem Hintergrund, dass keineswegs Einigkeit darüber besteht, wie die Dimensionen des Professionswissens modelliert werden und welche Wissensfacetten als relevant erachtet werden (vergl. Abschnitt 2.3 auf Seite 12), ist die Frage nach der Unterrichtsrelevanz von großer Bedeutung. Während die Facetten des pädagogischen Wissens zumindest deduktiv aus den umfangreichen Befunden aus der Prozess-Produkt-Forschung abgeleitet werden können, erfolgt die Auswahl der als relevant für erfolgreiches Unterrichten erachteten Facetten des fachdidaktischen Wissens eher normativ. Hill et al. (2005) stellen diesbezüglich fest:

> Despite conventional wisdom that [...] teachers' subject-matter knowledge influences student achievement, no large-scale studies have demonstrated this empirically [...]. Nor is the situation ameliorated by examining process-product research on teaching, in which both measurement of subject-specific teaching behaviors and direct measurement of teachers' subject-matter knowledge have been notably absent. (S. 372)

In den meisten der genannten Studien steht die Modellierung und die Beschreibung der Struktur oder der Entwicklung des Professionswissens im Vordergrund. Die Interpretation der Ergebnisse geht allerdings oftmals weit darüber hinaus. Beispielsweise weist Schmelzing (2010) in der Diskussion seiner Ergebnisse zwar durchaus darauf hin, „dass auf Basis der erbrachten Ergebnisse nicht geschlussfolgert werden kann, inwieweit die erfassten fachdidaktischen Kenntnisse und Fertigkeiten von tatsächlicher Relevanz für die Praxis des Biologieunterrichts sind" (S. 126). Dennoch sieht er den Ertrag seiner Arbeit als „wesentlichen Beitrag [...] zu einer empirisch

[1] Studien wie z. B. COACTIV, SII, PLUS und QuiP, die überprüfen, ob das mit derartigen Testinstrumenten gemessene Wissen relevant für erfolgreiches und gutes Unterrichten ist, werden in Abschnitt 4.3 auf Seite 41 beschrieben. Das ProwiN-Projekt, in das die vorliegende Arbeit eingebettet ist, wird in Kapitel 6 auf Seite 75 beschrieben.

fundierten Diskussion über mögliche Verbesserungen der Biologielehrerbildung“ (S.5). Auch Riese (2009) schränkt ein, dass „die getätigten Aussagen in Bezug auf die Relevanz der betrachteten Konstrukte für das Handeln der Lehrpersonen im Unterricht [...] nur als eingeschränkt fundiert angesehen werden [können]“ und dass „[eine] Ausweitung des Geltungsbereichs der Aussagen zur Relevanz einzelner Kompetenzbereiche im Hinblick auf die Lernentwicklung der Schüler [...] in keinster Weise getätigt werden [kann]“ (S. 183), erkennt aber „mögliche Ansatzpunkte für die Verbesserung der Lehrerbildung in Deutschland [...], womit das eigentliche Ziel, die Verbesserung des Unterrichts im Hinblick auf bessere Schülerleistungen, ein weiteres Stück näher rückt“ (S. 11).

Die Frage nach der Relevanz des Professionswissens für erfolgreiches und gutes Unterrichten ist auch eine Frage nach der Validität der eingesetzten Testinstrumente und insbesondere nach der Validität der Interpretation der mit diesen Testinstrumenten erhobenen Daten.

> If [professional knowledge] measures do not predict gains in student achievement, the information provided is of little use to test consumers. Showing that teachers improved on a multiple-choice assessment, for instance, is of little interest to policy-makers if the assessment is itself not predictive of student gains. (Hill, Ball, Blunk, Goffney & Rowan, 2007, S. 108)

## 4.2. Validität in der Professionswissensforschung

Was bedeutet Validität im Kontext der Professionswissensforschung und insbesondere in Bezug auf schriftliche Professionswissenstests?

Zunächst einmal bezeichnet Validität „den Umfang, in dem ein Test tatsächlich das Merkmal erfasst, das er erfassen soll“ (Hartig & Jude, 2007, S. 22). Nach dieser Definition stellt die Validität ein Qualitätsmerkmal und damit eine Eigenschaft des Testinstruments dar. In der Regel spielt Validität eine Rolle, wenn nicht direkt sichtbare, sogenannte *latente* Merkmale gemessen werden sollen. Diese Merkmale werden auch als theoretische Konstrukte bezeichnet (Schmiemann & Lücken, 2014, S. 108). In Anlehnung an die Empfehlungen der American Psychological Association (APA) von 1954 wird nach Cronbach und Meehl (1955) üblicherweise zwischen der *Konstruktvalidität*, der *Inhaltsvalidität* und der *Kriteriumsvalidität* unterschieden, auf die später noch ausführlicher eingegangen werden soll. Letztere umfasst die *Vorhersage-* und *Übereinstimmungsvalidität*. Aktuellere Validitätskonzepte betrachten Validität als Eigenschaft der Testwerte. Validität ist dann definiert als der „Grad[...], zu dem die empirischen Belege und theoretischen Sachverhalte die beabsichtigte Interpretation der Testwerte unterstützen“ (O. Wilhelm & Kunina, 2009, S. 318). Diese Sichtweise geht auf Messick (1987) zurück:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores. As such, validity is an inductive summary of both the existing evidence for and

> the potential consequences of test interpretation and use. Hence, what is to be validated is not the test as such, but the inferences derived from test scores – inferences about score meaning or interpretation and about the implications for action that the interpretation entails. (S.1)

Validität ist keine statische, sondern vielmehr eine dynamische Eigenschaft. Validierung ist daher ein kontinuierlicher und niemals abgeschlossener Prozess: Evidenz für Validität kann durch neue Forschungsergebnisse verstärkt, aber auch abgeschwächt werden und auch potenzielle Konsequenzen, die sich aus der Interpretation von Testdaten ergeben, ändern sich mit sich verändernden sozialen Rahmenbedingungen (Messick, 1987, S. 1). In Anlehnung an Messick verzichten die aktuellen gemeinsamen Standards der APA, der American Educational Research Association (AERA) und des National Council on Measurement in Education (NCME) auf die Unterscheidung verschiedener Validitätsarten und unterscheiden stattdessen zwischen Arten der Validierung (O. Wilhelm & Kunina, 2009, S. 318-319).

In der psychologischen Kompetenzdiagnostik und in der Didaktik der Naturwissenschaften wird dennoch gemeinhin von den drei Validitätsarten Konstruktvalidität, Inhaltsvalidität und Kriteriumsvalidität gesprochen (Hartig & Jude, 2007; Schmiemann & Lücken, 2014). Dabei handelt es sich allerdings in erster Linie um ein sprachliches Mittel, um Überlegungen zur Testvalidierung zu strukturieren, nicht aber um Ausdruck des Standpunktes, dass Validität ausschließlich eine Eigenschaft des Testinstruments ist (vergl. Schmiemann & Lücken, 2014).

- Die *Inhaltsvalidität* bezieht sich darauf, wie gut der Merkmalsbereich, der ein theoretisches Konstrukt definiert, durch die Testaufgaben repräsentiert wird (Hartig & Jude, 2007, S. 23). Hierfür ist eine klare Definition des zu messenden Konstruktes notwendig (Schmiemann & Lücken, 2014, S. 109). Zur Beurteilung der Inhaltsvalidität kann zum einen auf Grundlage subjektiver Einschätzungen (z.B. durch Expertenbefragungen) entschieden werden, ob die Testaufgaben die wichtigsten Aspekte des zu messenden Konstrukts erfassen (Bortz & Döring, 2006, S. 200). Zum anderen kann die Inhaltsvalidität durch eine modellbasierte Testentwicklung gewährleistet werden (vergl. Fischer, Boone & Neumann, 2014, S. 22). Verfahren wie die „Methode des lauten Denkens“ können zudem Aufschluss darüber geben, ob die Bearbeitung der Testaufgaben durch die Versuchspersonen unter Rückgriff auf das zu messende Konstrukt oder unter Rückgriff auf andere, nicht zum Konstrukt gehörende, Ressourcen erfolgt (Schmiemann & Lücken, 2014, S. 111).

- Die *Konstruktvalidität* bezieht sich auf die Einbettung des zu messenden Konstrukts in ein sogenanntes *nomologisches Netz* – ein konzeptionelles Rahmensystem aus theoretischen Überlegungen über die Struktur und den Aufbau des zu messenden Konstrukts und seine Definition und Abgrenzung zu anderen Konstrukten. Evidenz für Konstruktvalidität ergibt sich aus der empirischen Bestätigung theoretisch abgeleiteter Hypothesen über Zusammenhänge innerhalb des nomologischen Netzes, die gelten müssen, wenn die Testwerte wie beabsichtigt interpretierbar sein sollen (O. Wilhelm & Kunina, 2009, S. 318). Die Konstruktvalidität eines Testinstruments kann auch im

Rahmen einer konvergenten Validierung über den parallelen Einsatz eines bereits validierten Testinstruments zur Messung des gleichen Konstrukts und anschließende Korrelationsanalysen untersucht werden. Außerdem kann im Zuge einer diskriminanten Validierung die Abgrenzung zu anderen erhobenen Konstrukten Hinweise auf die Konstruktvalidität liefern (Bortz & Döring, 2006, S. 203). Darüber hinaus kann eine Validierung auch ohne Rückgriff auf externe Kriterien über die Überprüfung konstruktimmanenter Annahmen erfolgen (Schmiemann & Lücken, 2014, S. 116). Diese rein methodischen Vorgehensweisen zur Konstruktvalidierung werden allerdings auch kritisch diskutiert (Borsboom, Mellenbergh & van Heerden, 2004). Nicht bestätigte Hypothesen über Zusammenhänge des Konstruktes mit anderen Variablen können nicht eindeutig interpretiert werden. So können deratige Befunde entweder die Validität des zu untersuchenden Testinstrumentes in Frage stellen, oder aber die Validität des zugrunde liegenden theoretischen Modells (Schmiemann & Lücken, 2014, S. 117).

- Die *Kriteriumsvalidität* bezieht sich auf die praktische Bewährung des Testinstruments bzw. auf den korrelativen Zusammenhang der Testwerte mit einer oder mehreren, für das Testinstrument praktisch bedeutsamen Variablen (Schmiemann & Lücken, 2014, S. 108). In Bezug auf die *Übereinstimmungsvalidität* wird überprüft, inwieweit ein Test mit korrespondierenden manifesten (also direkt messbaren) Variablen korreliert. Hierzu zählt auch die „Technik der bekannten Gruppen“, mit der überprüft wird, ob ein Testinstrument zu erwartende Fähigkeitsunterschiede im Vergleich bestimmter Gruppen misst (Bortz & Döring, 2006, S. 201). Die *Vorhersagevalidität* oder auch *prädiktive Validität* bezieht sich darauf, ob ein Test in der Lage ist Verhalten, Erfolg oder Misserfolg außerhalb der Testsituation zu prognostizieren (Hartig & Jude, 2007, S. 23).[2]

Aus diesen Definitionen wird deutlich, dass eigentlich Validierungsarten definiert werden, auch wenn von Validitätsarten gesprochen wird.
Es soll nicht unerwähnt bleiben, dass genau dieser Umstand von einzelnen Autoren kritisiert wird. So plädieren Borsboom et al. (2004, S. 1068) für ein gänzlich anderes, konstruktbezogenes Validitätskonzept, das die Validität wieder als Eigenschaft eines Testinstruments definiert. Validität ist demnach gegeben, wenn das zu messende Konstrukt existiert und Variation in den Messwerten kausal (und nicht korrelativ) durch Variation im Konstrukt verursacht wird.

Zentral für die Beurteilung der Validität ist die Frage: „*Was* und *wozu* soll überhaupt gemessen werden?“ (Schmiemann & Lücken, 2014, S. 108). Welche Aspekte müssen also bei der Validierung von Testinstrumenten zur Erfassung von Professionswissen berücksichtigt werden?

**Was soll gemessen werden?** Schon diese, auf den ersten Blick einfache Frage ist im Kontext der Professionswissensforschung nicht leicht zu beantworten. Wie bereits in Abschnitt 2.3 auf Seite 12 dargestellt wurde, herrscht keine Einigkeit

[2] Kriteriums- und Konstruktvalidität überschneiden sich in dieser Definition zum Teil.

darüber, wie genau die drei Professionswissensdimensionen CK, PCK und PK zu operationalisieren sind. Unabhängig von den als wichtig erachteten Wissensfacetten, sollte allerdings Wissen erfasst werden, das spezifisch für Lehrkräfte ist. Evidenz hierfür kann durch den Vergleich der Testwerte von Lehrkräften und Fachkräften oder Lehramtsstudierenden und Studierenden anderer Fächer ohne pädagogisch-psychologischen Hintergrund gewonnen werden (Kirschner, Taylor, Rollnick, Borowski & Mavhunga, 2015, S. 236). Ebenso können Expertenbefragungen oder der Abgleich mit Fachcurricula und Standards für die Lehrerausbildung Hinweise darauf liefern, ob lehrerspezifisches Wissen erhoben wird.

Für die Dimensionen CK und PCK sollte zudem nachgewiesen werden, dass fachspezifisches Wissen erhoben wird – Lehrkräfte der Physik sollten also beispielsweise besser in Tests zur Erfassung des physikspezifischen Professionswissens abschneiden als Lehrkräfte anderer Fächer (Kirschner et al., 2015, S. 236). Das Gegenteil gilt für Tests zum pädagogischen Wissen, hier sollten Lehrkräfte verschiedener Fächer gleichermaßen gut abschneiden können (Lenske et al., 2015, S. 7). Als Hinweis auf die Erfassung unterrichtsrelevanten Wissens können außerdem bessere Testwerte von Versuchspersonen mit Unterrichtserfahrung gegenüber Versuchspersonen ohne Unterrichtserfahrung interpretiert werden (Kirschner et al., 2015, S. 236).

In Bezug auf die Konstruktvalidität kann zudem die Struktur des Professionswissens näher betrachtet werden. Aus theoretischer Sicht wären getrennte Wissensbereiche, aber dennoch Zusammenhänge zwischen PCK und CK - schließlich handelt es sich in beiden Fällen um fachspezifisches Wissen - und zwischen PCK und PK zu erwarten, da PCK fachspezifisch-pädagogisches Wissen umfasst (Kirschner, 2013, S. 81). Alle diese Aspekte liefern Hinweise für die valide Erfassung von Professionswissen und Antworten auf die Frage *was* gemessen wird.

In der TEDS-M Studie wird eine valide Erfassung des Professionswissens auf Grundlage von Untersuchungen zur curricularen Validität angenommen (Blömeke & König, 2010; Döhrmann, Kaiser & Blömeke, 2010). In den bereits erwähnten Studien von Riese (2009) und Schmelzing (2010) oder in der Arbeit von Kirschner (2013) wird auf Grundlage von Zusammenhangsanalysen zwischen den Professionswissensdimensionen, konvergenten oder diskriminanten Validierungen oder dem Vergleich bekannter Gruppen von einer validen Erfassung von Professionswissen ausgegangen.

***Wozu* soll gemessen werden?** Eine Evaluation der Lehrerausbildung, wie sie in TEDS-M oder MT21 erfolgt, hat das Ziel zu erheben, ob die im Rahmen der Ausbildung gelehrten Inhalte auch wirklich gelernt werden. Die curriculare Validität ist hier also zentral. Ziel der Lehrerausbildung ist allerdings die Ausbildung erfolgreicher Lehrkräfte. So schreiben Blömeke et al. (2010, S. 46) „Es ging in MT21 um die Erfassung jenes Wissens, das die erfolgreiche Bewältigung konkreter beruflicher Aufgaben erwarten lässt, und zwar fokussiert auf das Unterrichten und Diagnostizieren." Ob dieses Ziel erreicht wird, kann nur überprüft werden, wenn Testinstrumente eingesetzt werden, deren prädiktive Validität im Hinblick auf erfolgreiches Unterrichten gezeigt wurde.

Blömeke, Kaiser, Döhrmann und Lehmann (2010, S. 237) weisen auf Grundlage der schlechten Ergebnisse angehender Mathematiklehrkräfte mit Lehrbefähigung bis zur Klasse 10 im TEDS-M Fachwissenstest auf einen „dringende[n] Reformbedarf“ der Primar- und Sekundarstufen-I-Ausbildung hin. Es stellt sich die Frage, ob eine derartige Interpretation der Testwerte ohne Überprüfung der prädiktiven Validität als angemessen bzw. valide erachtet werden kann. Zwar führen Blömeke et al. (2010, S. 237) als Hinweise auf die prädiktive Validität des Fachwissenstests an, dass sich die in TEDS-M gefundenen Länderunterschiede im Fachwissen angehender Lehrkräfte zum Teil in den im Rahmen von TIMSS 2007 gefundenen Länderunterschieden im mathematischen Fachwissen von Schülerinnen und Schülern widerspiegeln. Hierbei handelt es sich allerdings um zwei völlig unabhängige Untersuchungen und die Autoren nennen selbst zahlreiche Einschränkungen für die Vergleichbarkeit der Ergebnisse der beiden Studien (wie z. B. „unterschiedliche Länderzusammensetzung, Erfassung unterschiedlicher Konstrukte mit unterschiedlichen Instrumenten, Einflüsse einer Vielzahl an Drittvariablen“, Blömeke et al., 2010, S. 237).

Die Arbeit von Schmelzing (2010, S. 35) zielt darauf ab ein valides Testinstrument für die Erfassung des fachdidaktischen Wissens von Biologielehrkräften zu entwickeln, um „die fachdidaktische Biologielehrerbildung zu evaluieren und verallgemeinerbare Einsichten zum fachdidaktischen Wissen von Biologielehrkräften zu gewinnen“ und „mögliche Optimierungen der Biologielehrerbildung durch eine empirische Datenbasis zu stützen“. Ein ähnliches Ziel verfolgt Riese (2009, S. 70) mit der Entwicklung eines Professionswissenstests für angehende Physiklehrkräfte „um Erkenntnisse zu Ausmaß und Entwicklung professioneller Handlungskompetenz und damit zur Wirksamkeit der Lehrerbildung zu gewinnen“. Auch hier kann ohne die Überprüfung der prädiktiven Validität der Testinstrumente nicht entschieden werden, ob diese Ziele erreicht werden.

Grundlegende Voraussetzung dafür, dass die prädiktive Validität von Professionswissenstests in Bezug auf gutes und erfolgreiches Unterrichten überhaupt nachgewiesen werden kann, ist der Zusammenhang zwischen dem Wissen und Handeln einer Lehrkraft. So modelliert Riese (2009, S. 26) in Anlehnung an Baumert und Kunter (2006) Professionswissen als Teil der professionellen Handlungskompetenz von (angehenden) Lehrkräften (vergl. Abschnitt 2.2 auf Seite 9). Wie bereits erwähnt, liegt diesem Modell die Annahme zugrunde, dass ein Zusammenhang zwischen Wissen und Handeln existiert. Umso wichtiger ist die Untersuchung der prädiktiven Validität von Professionswissenstests, da hiermit gleichzeitig auch die Validität des zugrunde liegenden Modells geprüft wird. Im Umkehrschluss können nicht gefundene Zusammenhänge zwischen dem Professionswissen von Lehrkräften und erfolgreichem Unterrichten allerdings nicht eindeutig interpretiert werden – sie können das Resultat einer nicht validen Erfassung des Professionswissens sein (z. B. weil die bei der Operationalisierung berücksichtigten Wissensfacetten nicht die angenommene Relevanz für erfolgreiches Unterrichten haben) oder daraus resultieren, dass kein Zusammenhang zwischen dem (explizierbaren) Wissen von Lehrkräften und ihrem Handeln existiert.

## 4.3. Empirische Studien zur prädiktiven Validität von Professionswissenstests

Die Grundannahme über die Relevanz des Professionswissens für gutes und erfolgreiches Unterrichten, die das CK, PCK und PK von Lehrkräften zu einem viel betrachteten Forschungsgegenstand macht, wurde bisher nur in wenigen Studien empirisch überprüft. Die Mehrzahl der Studien, die das Professionswissen von Lehrkräften quantitativ erheben, begnügt sich mit einer Validierung der eingesetzten Testinstrumente auf „herkömmliche“ Weise: über Expertenbefragungen, Abgleich mit Fachcurricula, den Vergleich bekannter Gruppen mit zu erwartenden Fähigkeitsunterschieden oder durch Zusammenhangsanalysen zwischen den Dimensionen des Professionswissens (vergl. z. B. Blömeke et al., 2010; Brovelli et al., 2013; Großschedl, Mahler, Kleickmann & Harms, 2014; Kirschner, 2013; Kröger, Neumann & Petersen, 2013; Kulgemeyer et al., 2012; Riese, 2009; Schmelzing, 2010). Die prädiktive Validität im Hinblick darauf, ob mit den Testinstrumenten Wissen erhoben wird, das mit gutem und erfolgreichem Unterrichten einhergeht und damit als unterrichtsrelevant angenommen werden kann, wird nicht untersucht.

Dies ist insofern problematisch, dass beispielsweise eine Evaluation der Lehrerausbildung auf Basis von Testinstrumenten zur Erfassung des Professionswissens, wie sie in MT21 und TEDS-M erfolgt, nur Aufschluss darüber geben kann, ob das Wissen, das in der universitären Lehrerausbildung gelehrt wird und in den Standards für die Lehrerbildung festgeschrieben ist, von Lehramtsstudierenden tatsächlich erworben wird. Es können jedoch keine Rückschlüsse darüber gezogen werden, ob über die Vermittlung dieses Wissens erfolgreiche Lehrkräfte ausgebildet werden können und das eigentliche Ziel der Lehrerausbildung erreicht wird. Auch Hill et al. (2005) stellen fest:

> Because teachers' knowledge has not been adequately measured, the existing educational production function research could be limited in terms of its conclusions, not only regarding the magnitude of the effect of teachers' knowledge on student learning but also regarding the kinds of teacher knowledge that matter most in producing student learning. (S. 372)

In den 1980er- und 1990er-Jahren wurde in ersten qualitativen Fallstudien der Zusammenhang zwischen Fachwissen und Unterrichten in Stichproben von bis zu sechs Lehrkräften beleuchtet (z. B. Carlsen, 1993; Gess-Newsome & Lederman, 1995; Hashweh, 1987; Sanders, Borko & Lockard, 1993). In den einzelnen Studien wurden jeweils verschiedene Verhaltensweisen und Unterrichtsmerkmale identifiziert, die mit höherem Fachwissen einhergingen: z. B. die Thematisierung komplexerer Fragestellungen im Unterricht, stärkerer Einbezug der Lernenden, bessere Diagnose von inadäquaten Schülervorstellungen sowie weniger rezeptartiges Vorgehen in Experimentiersituationen (für eine ausführlichere Darstellung vergl. Abell, 2007, S. 1117-1120). Diese Studien enthalten allerdings keine Aussage darüber, ob das beobachtete Verhalten lernförderlich ist.

Im Folgenden sollen einige Studien aus der Mathematik und aus der Physik vorgestellt werden, die Zusammenhänge zwischen dem mit schriftlichen Testinstru-

menten erhobenem Professionswissen von (angehenden) Lehrkräften, Merkmalen guten Unterrichts und Zielkriterien erfolgreichen Unterrichts untersucht haben. Die in diesen Studien eingesetzten Testinstrumente basieren auf unterschiedlichen Operationalisierungen des Fachwissen, fachdidaktischen Wissen und pädagogischen Wissen von Lehrkräften. Daher wird auch kurz auf die in den jeweiligen Studien vorgenommenen Operationalisierungen eingegangen.

## Studien in der Mathematik

**Carpenter, Fennema, Peterson und Carey (1988)**
In Bezug auf fachdidaktisches Wissen untersuchten Carpenter et al. (1988) eine Stichprobe von 40 Grundschulmathematiklehrkräften. 20 Lehrkräften wurde im Rahmen eines vierwöchigen Workshops forschungsbasiertes Wissen über das Lernen, die Entwicklung von Additions- und Subtraktionskonzepten bei Kindern und die von Kindern genutzten Lösungsstrategien beim Bearbeiten von Aufgaben vermittelt. Der Vergleich von Experimental- und Kontrollgruppe ergab Unterschiede im allgemeinen Wissen der Lehrkräfte über Aufgabeschwierigkeiten und Problemlöse- und Rechenstrategien von Lernenden, im Unterrichten (die Lehrkräfte der Experimentalgruppe fragten zum Beispiel öfter nach den Lernprozessen der Schülerinnen und Schüler, stellten öfter Problemlöseaufgaben und erlaubten den Lernenden verschiedene Lösungsstrategien beim Bearbeiten von Aufgaben anzuwenden) und in den Leistungsergebnissen der Schülerinnen und Schüler in Bezug auf komplexe Additions-und Subtraktionsaufgaben zugunsten der Experimentalgruppe (Carpenter, Fennema, Peterson, Chiang & Loef, 1989).

Sowohl in der Gesamtstichprobe als auch innerhalb der Experimentalgruppe ging die Fähigkeit der Lehrkräfte die Problemlöse- und Rechenfähigkeit ihrer Schülerinnen und Schüler einzuschätzen mit besseren Leistungsergebnissen auf Schülerseite einher. Diese korrelierte allerdings zur Verwunderung der Autoren nicht mit dem Wissen der Lehrkräfte über Aufgabeschwierigkeiten und Problemlöse- und Rechenstrategien von Lernenden (Carpenter et al., 1989; Peterson et al., 1989). Auch unterschieden sich Experimental- und Kontrollgruppe nicht bezüglich ihrer Fähigkeit die Problemlöse- und Rechenfähigkeit ihrer Schülerinnen und Schüler einzuschätzen (Carpenter et al., 1988).

Obwohl die Autoren ihre Ergebnisse als Hinweise auf einen Zusammenhang zwischen PCK und lernförderlichem Unterrichten interpretieren und ihre Arbeiten beispielsweise von Baumert und Kunter (2006, S. 493) auch in diesem Kontext zitiert werden, scheint das Wissen, das zu den Leistungsunterschieden zwischen den Experimental- und Kontrollklassen geführt haben könnte, nicht durch die Erhebungsinstrumente zur Messung des allgemeinen Wissen über Aufgabeschwierigkeiten und Problemlöse- und Rechenstrategien von Lernenden erfasst worden zu sein.

In der Mathematik existieren bisher zwei Large-Scale Studien, die den Zusammenhang zwischen dem fachspezifischen Professionswissen von Lehrkräften und Schülerleistungszuwachs untersucht haben und damit auch die prädiktive Validität

ihrer Messinstrumente überprüft haben.

**Study of Instructional Improvement (SII)**
Im Rahmen der SII-Studie konnten Hill et al. (2005, S. 396) in einer Stichprobe von über 300 Grundschulmathematiklehrkräften einen Zusammenhang zwischen CK und dem Leistungszuwachs ihrer Schülerinnen und Schüler über den Zeitraum eines Jahres nachweisen. Eine Standardabweichung im CK der Lehrkräfte führte zu einem Lernvorsprung von etwa zwei bis drei Wochen Unterricht aufseiten der Schülerinnen und Schüler. Der Effekt war damit in vergleichbarer Größenordnung wie der Einfluss des sozioökonomischen Index der Lernenden. Die Ergebnisse legten allerdings einen nicht-linearen Zusammenhang nahe. Nur die Lernleistungen der Klassen der $20-30\%$ der Lehrkräfte, die am schlechtesten im CK-Test abgeschnitten hatten, unterschieden sich signifikant von den restlichen Klassen. Ab einem gewissen Schwellenwert des Fachwissens, zeigte sich kein systematischer Zusammenhang zwischen CK und Lernleistung mehr. Hill et al. (2007) konnten außerdem im Rahmen einer qualitativen Studie zeigen, dass das CK von zehn Mathematiklehrkräften mit der mathematischen Qualität ihres Unterrichtens einherging.

Das Fachwissen der Lehrkräfte operationalisierten die Autoren als *common content knowledge* über das auch gute Schüler, Banker, Krankenschwestern oder Mathematiker verfügen, und *specialised content knowledge*, über das nur Lehrkräfte verfügen und das beim Unterrichten von Mathematik genutzt wird. Letzteres umfasst Wissen über Erklärungen, alternative Repräsentationen mathematischer Konzepte und Wissen über das Potenzial ungewöhnlicher Lösungsstrategien. So mussten die Lehrkräfte erklären, warum bestimmte Rechenoperationen funktionieren und für den Fall, dass Lernende eigene Strategien für die Lösung einer Aufgabe anwenden, bewerten, ob die angewendete Strategie verallgemeinerbar ist oder nur in bestimmten Fällen funktioniert (Hill & Ball, 2004; Hill, Schilling & Ball, 2004).

Ob die Autoren das *specialised content knowledge* als Teil von CK oder PCK interpretieren, wird zunächst nicht klar. In späteren Veröffentlichungen erfolgt allerdings eine klare Einordnung in den Bereich CK und eine Abgrenzung zu PCK (Hill et al., 2007; Hill, Ball & Schilling, 2008; Hill et al., 2008). Im deutschsprachigen Raum werden ihre Arbeiten allerdings auch in Bezug auf den Zusammenhang von fachdidaktischem Wissen und Schülerleistung zitiert (Baumert & Kunter, 2006, S. 494; Ergönenç et al., 2014, S. 145).

In einer Videostudie von Kersting, Givvin, Thompson, Santagata und Stigler (2012) wurden ein Teil der Items aus der SII-Studie in einer Stichprobe von 36 Mathematiklehrkräften, die in den Jahrgangsstufen $5-7$ unterrichteten, eingesetzt und Zusammenhänge zur Qualität des Unterrichts in den Dimensionen Konzeptentwicklung, angemessener Einsatz von Repräsentationen zur Erklärung von Algorithmen und Verknüpfung von mathematischen Konzepten und Inhalten, sowie zum Leistungszuwachs der Lernenden untersucht. Die Ergebnisse der SII-Studie konnten in dieser Stichprobe nicht repliziert werden – zwischen dem CK der Lehrkräfte, Unterrichtsqualität und Schülerleistungszuwachs existierten keine signifikanten Zusammenhänge. Geht man davon aus, dass die in dieser Studie untersuchten Lehrkräfte über ein höheres CK als die Grundschullehrkräfte in der

SII-Studie verfügen, könnte es sein, dass der CK-Test in dieser Stichprobe nicht ausreichend in dem für das erfolgreiche Unterrichten relevanten Wissensbereich differenzierte.

**Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz (COACTIV)**

In Deutschland gilt die COACTIV-Studie (Baumert et al., 2010) als Vorreiter in den Bemühungen den Zusammenhang zwischen dem Professionswissen von Mathematiklehrkräften, Aspekten der Unterrichtsqualität und Schülerleistungen zu beleuchten. Die COACTIV-Studie war in die nationale PISA 2003/2004 Erhebung (Prenzel et al., 2005) integriert. Im Zuge dieser Studie entwickelten Baumert und Kunter (2006) ihr Modell zur professionellen Handlungskompetenz von Lehrkräften.

In COACTIV wurde das fachspezifische Professionswissen von Mathematiklehrkräften mit separaten Testinstrumenten zur Erfassung von CK und PCK erhoben. Zudem wurde der Leistungszuwachs der Schülerinnen und Schüler zwischen dem Ende der Jahrgangsstufe 9 und 10 erfasst. Die mathematische Kompetenz der Lernenden am Ende der Jahrgangsstufe 9 wurde mit Aufgaben aus dem internationalen und nationalen PISA-Test erhoben. Am Ende der Jahrgangsstufe 10 wurden zusätzliche, an das Curriculum dieser Jahrgangsstufe angepasste, Aufgaben eingesetzt (vergl. Baumert & Kunter, 2011, S. 174; Carstensen, 2006, S. 313; Ehmke et al., 2006, S. 69; Löwen et al., 2011, S. 78). Der zwischen den Erhebungen stattfindende Unterricht wurde auf Basis von im Unterricht eingesetzten Aufgaben, Hausaufgaben und Klassenarbeiten rekonstruiert. Eine direkte Erfassung des Unterrichtsgeschehens über Videoaufnahmen erfolgte nicht. Als Merkmale der Unterrichtsqualität wurden kognitive Aktivierung (operationalisiert als die kognitive und curriculare Passung der Aufgaben zum Lehrplan der Jahrgangsstufe 10), die durch die Lernenden wahrgenommene individuelle Lernunterstützung und die durch Lernende und Lehrkräfte beurteilte Effektivität der Klassenführung betrachtet. Insbesondere kognitive Aktivierung und Klassenführung erwiesen sich als bedeutsame Prädiktoren für Schülerleistung. Der Effekt der individuellen Lernunterstützung war mit dem Effekt der Klassenführung konfundiert. Unter Kontrolle der Klassenführung war die individuelle Lernunterstützung kein signifikanter Prädiktor für Schülerleistung (Baumert et al., 2010, S. 162). Insgesamt wurde eine repräsentative Stichprobe von 181 Mathematiklehrkräften (mit 80 Gymnasialklassen und 114 Nicht-Gymnasialklassen[3]) untersucht.

Das Fachwissen der Lehrkräfte wurde in COACTIV in Abgrenzung vom mathematischem Alltagswissen, dem Schulwissen, über das durchschnittliche bis gute Schüler verfügen, und dem vom Curriculum der Schule losgelösten reinen universitären Wissen als „tieferes Verständnis der Fachinhalte des Curriculums der Sekundarstufe (z.B. auch ‚Elementarmathematik vom höheren Standpunkt aus', wie sie an der Universität gelehrt wird)" operationalisiert (Krauss, Neu-

[3]Einige Lehrkräfte nahmen mit mehreren Klassen an der COACTIV-Studie teil, daher addieren sich die Anzahl der Klassen nicht zu 181. In Baumert et al. (2010) finden sich keine Angaben darüber, wie viele der 181 Mathematiklehrkräfte am Gymnasium oder an anderen Schulformen unterrichteten.

brand et al., 2008, S. 237). PCK operationalisieren Krauss, Neubrand et al. (2008, S. 234-237) über Wissen über Erklären und Repräsentieren, Wissen über typische Schülerfehler- und -schwierigkeiten und Wissen über das Potenzial für multiple Lösungsansätze von Mathematikaufgaben. Damit überschneidet sich die Operationalisierung des fachdidaktischen Wissens in COACTIV zu einem gewissen Teil mit der Operationalisierung des *specialised content knowledge* in der SII-Studie.

Die Ergebnisse der COACTIV-Studie zeigten einen deutlichen Einfluss von CK und PCK auf die Schülerleistung, wobei letzterer wesentlich größer ausfällt. Zwei Standardabweichungen im PCK der Lehrkräfte führten zu Unterschieden in den Schülerleistungen, die in einer vergleichbaren Größenordnung lagen, wie der Lernzuwachs der Schülerinnen und Schüler über das gesamte Schuljahr. Dabei handelte es sich um einen linearen Zusammenhang, der vollständig über die kognitive Aktivierung und die individuelle Lernunterstützung mediiert wurde. Das CK der Lehrkräfte wirkte lediglich über die curriculare Passung der Aufgaben auf die Schülerleistung. Die Linearität des Zusammenhangs zwischen CK und Schülerleistung wurde nicht untersucht (Baumert et al., 2010, S. 165-166).

Im Rahmen der Studie COACTIV-Referendariat (COACTIV-R) konnte außerdem gezeigt werden, dass das pädagogisch-psychologische Wissen über das angehende Mathematiklehrkräfte im Referendariat verfügten, prädiktiv dafür ist, wie Merkmale der Qualität ihres Unterrichts zwei Jahre später durch ihre Schülerinnen und Schüler wahrgenommen wurden. Untersucht wurde eine Stichprobe von 181 Mathematikreferendaren und -referendarinnen. Während sich substantielle Zusammenhänge zur Klassenführung und zur konstruktiven Lernunterstützung zeigten, galt dies nicht für das Potenzial zur kognitiven Aktivierung (Voss et al., 2014). Unterschiede in der Schülerleistung wurden in COACTIV-R nicht betrachtet. Das PK der Lehrkräfte wurde in dieser Studie über die Dimensionen Wissen über Klassenführung, Unterrichtsmethoden, Leistungsbeurteilung und Schülerheterogenität operationalisiert und mit Hilfe eines Videovignettentests erhoben.

Zusammenfassend lässt sich feststellen, dass die Studien in der Mathematik, die eine unterrichtsnahe Operationalisierung des Fachwissens und fachdidaktischen Wissens vornehmen, einigermaßen konsistente Ergebnisse für den Zusammenhang dieses Wissens mit Merkmalen der Unterrichtsqualität und mit der Lernleistung von Schülerinnen und Schülern ergeben. Im Grundschulbereich scheint sich ein Mindestmaß an mathematischem Schulwissen, über das auch gute Schülerinnen und Schüler verfügen könnten, positiv auf die Leistungszuwächse der Lernenden auszuwirken. Dies gilt allerdings nur bis zu einem bestimmten Schwellenwert. In weiterführenden Schulen scheint ein vertieftes Verständnis der Fachinhalte mit Lernerfolgen auf Schülerseite zusammenzuhängen. Im Bezug auf das fachdidaktische Wissen von Mathematiklehrkräften scheint das Wissen über Erklärungen und alternative Repräsentationen, Wissen über Schülerfehler- und -schwierigkeiten sowie Wissen über das Potenzial ungewöhnlicher Lösungsstrategien und das Potenzial für multiple Lösungsansätze von Mathematikaufgaben relevant für erfolgreiches Unterrichten zu sein. Außerdem zeigen sich erste Hinweise für die Bedeutung des

pädagogischen Wissens von Lehrkräften für gutes Unterrichten.

Motiviert durch die vielversprechenden Ergebnisse der COACTIV-Studie wurden auch in der Physik einige Studien zur prädiktiven Validität von fachspezifischen Professionswissenstests durchgeführt. Hier ergibt sich aber ein wesentlich weniger eindeutiges Bild als in der Mathematik.

## Studien in der Physik

**Professionswissen von Lehrkräften, naturwissenschaftlicher Unterricht und Zielerreichung im Übergang von der Primar- zur Sekundarstufe (PLUS)**

In der Videostudie PLUS wurden Zusammenhänge zwischen dem fachspezifischen Professionswissen von 60 Sachunterrichtslehrkräften, Merkmalen der Unterrichtsqualität und Schülerleistungszuwächsen von Grundschülern über eine sechsstündige Unterrichtseinheit zum Thema „Aggregatzustände und ihre Übergänge am Beispiel Wasser“ untersucht (Kauertz & Kleickmann, 2009; Lange, Kleickmann, Tröbst & Möller, 2012).

Das Fachwissen wurde im PLUS-Projekt auf unterschiedlichen Komplexitätsniveaus modelliert und bezog sich auf Inhalte aus der Grundschule, der Sekundarstufe und der Universität. Ihre Operationalisierung des Fachwissens halten Ohle, Fischer und Kauertz (2011, S. 396) für vergleichbar mit der Operationalisierung des *specialised content knowledge* in der SII-Studie. Inwieweit ähnliches Wissen adressiert wurde, lässt sich allerdings auf Grundlage der angegebenen Informationen nicht beurteilen (dies gilt auch für die Ausführungen zur Operationalisierung in Ohle, 2010). Das fachdidaktische Wissen wurde im PLUS-Projekt als Wissen über Bedingungen des Lernens (Nennen von typischen Schülervorstellungen und -schwierigkeiten, Analyse von Schülerantworten in Bezug auf Inhalt oder Anschlussfähigkeit) und als Wissen über instruktionale Aktivitäten (Skizzierung und Beurteilung von Versuchen zur zielorientierten Unterstützung von Verständnisprozessen, Identifikation sinnvoller Sequenzierungen von Inhalten und Konzepten) operationalisiert (Lange et al., 2012, S. 61).

In einer Teilstichprobe von 58 Lehrkräften zeigte sich kein direkter Zusammenhang zwischen dem CK der Lehrkräfte und der Schülerleistung (Ohle, 2010, S. 105). Anzumerken ist allerdings, dass lediglich die Hälfte der Lehrkräfte einen naturwissenschaftlichen Schwerpunkt in ihrem Studium gewählt hatten und nur zwei der Lehrkräfte als Schwerpunktfach Physik studiert hatten (Ohle, 2010, S. 91). Die Varianz im Fachwissen der Lehrkräfte war daher eingeschränkt (Lange et al., 2015, S. 34-35). Für eine Teilstichprobe von 30 Lehrkräften (15 Lehrkräfte mit hohem CK und 15 Lehrkräfte mit niedrigem CK) wurden zusätzlich Zusammenhänge zu Merkmalen der Unterrichtsqualität untersucht. Es ergab sich kein Zusammenhang des Fachwissens der Lehrkräfte mit der inhaltlichen Sachstruktur des Unterrichts oder der Sequenzierung von Lernprozessen im Unterricht. Allerdings ergab sich für diese Stichprobe ein mittlerer Effekt des Fachwissens auf Schülerleistung, der durch die Sequenzierung der Lernprozesse und die Selbstwirksamkeitserwartung der Lehrkraft moderiert wurde. Ob überhaupt ein Effekt vorlag, hing also von

diesen zwei Variablen ab (Ohle et al., 2011, S. 382-383). Unklar ist, wie dieser Befund inhaltlich interpretiert werden kann.

In der Gesamtstichprobe klärte das PCK der Lehrkräfte einen Anteil von 13% der zwischen den Klassen liegenden Varianz in den Schülerleistungen auf, sofern auf Klassenebene die durch die Lernenden wahrgenommene Klassenführung, die tatsächliche Unterrichtszeit und die Lehrerfahrung der Lehrkräfte kontrolliert wurde. Dabei zeigten sich signifikante Zusammenhänge zum konzeptuellen Wissen der Lernenden, jedoch nicht zum begrifflichen Wissen. Ein direkter Effekt von PCK auf Schülerleistung, ohne Kontrolle der genannten Variablen auf Klassenebene, konnte nicht nachgewiesen werden (Lange, 2010, S. 168). Darüber hinaus konnten unter Kontrolle der gleichen Variablen signifikante (kleine) Effekte von PCK auf das situative Fachinteresse und das Kompetenzerleben der Lernenden nachgewiesen werden. Insgesamt fallen die Ergebnisse in der PLUS-Studie wesentlich weniger eindeutig aus als in den Studien in der Mathematik.

**Quality of Instruction in Physics (QuiP)**

Die QuiP-Videostudie untersuchte das fachspezifische Professionswissen von Physiklehrkräften, den Unterricht in der Jahrgangsstufe 9/10 und die Leistungs-, Interessen- und Selbstkonzeptentwicklung der Lernenden im Rahmen einer mehrmonatigen Unterrichtseinheit zur Elektrizitätslehre im Ländervergleich Finnland, Deutschland, Schweiz (Fischer et al., 2014a). Da leider keine Ergebnisse in Bezug auf die prädiktive Validität des CK-Tests veröffentlicht wurden, wird hier lediglich auf Befunde zum fachdidaktischen Wissen eingegangen.

Das PCK der Lehrkräfte wurde im QuiP-Projekt als Wissen über Schülerfehlvorstellungen (Prognostizieren von Schülerantworten und -fehlvorstellungen, Wissen über Conceptual Change), Wissen über das Curriculum (Zuordnung von Inhalten zu Inhaltsgebieten und Schulstufen) und Wissen über Schwierigkeiten (Benennung und Evaluation verschiedener Repräsentationen von Inhalten, Erkennen von inhaltsspezifischen Schwierigkeiten) operationalisiert (Ergönenç et al., 2014, S. 148).

PCK zeigte einen kleinen bis mittleren Effekt auf die kognitive Aktivierung (definiert über die Passung zwischen dem kognitiven Level von Lehrerfragen und Schülerantworten) im Unterricht, die ihrerseits allerdings keinen signifikanten Zusammenhang zu den Schülerleistungen zeigte. Außerdem zeigte sich ein mittlerer direkter Effekt von PCK auf die Schülerleistung (Ergönenç et al., 2014, S. 153-154). Diese Ergebnisse beziehen sich auf die Teilstichprobe von 33 deutschen und 20 schweizer Physiklehrkräften, die finnischen Lehrkräfte wurden nicht in die Analyse mit einbezogen.

Die Schülerinnen und Schüler der finnischen Lehrkräfte zeigten die größten Leistungszuwächsen und schnitten signifikant besser ab als die deutschen und schweizer Schülerinnen und Schüler (Spoden & Geller, 2014, S. 56). Auch die kognitive Aktivierung im Unterricht war im finnischen Unterricht am stärksten ausgeprägt. Im PCK schnitten die Lehrkräfte allerdings signifikant schlechter ab als die deutschen Lehrkräfte. Ergönenç et al. (2014, S. 153) zweifeln aufgrund des schlechten Abschneidens der finnischen Lehrkräfte die Validität des PCK-Tests für die finnische Stichprobe an. Zur Untermauerung dieser Interpretation führen

die Autoren an, dass 60% der PCK-Items ein *Differential Item Functioning* (DIF) zwischen der finnischen und der deutsch-schweizerischen Stichprobe zeigten, was als Hinweis auf die Erfassung unterschiedlicher Konstrukte interpretiert wird. Die Analyse wurde allerdings mit einer für den Kontext von DIF-Analysen in Raschmodellen eher kleinen Stichprobe von insgesamt 92 Lehrkräften (FI: 25, DE: 41, CH:26) durchgeführt und es wird nicht darüber berichtet, ob die Unterschiede zwischen den Stichproben signifikant werden (Ergönenç et al., 2014, vergl. auch Olszewski, 2010).

Eine andere mögliche Interpretation der Ergebnisse besteht darin, dass die finnischen Lehrkräfte über anderes Wissen verfügten, welches mit dem PCK-Test im QuiP-Projekt nicht erfasst wurde, das aber ausschlaggebend für erfolgreiches Unterrichten sein könnte. Dieser Interpretation liegt natürlich die Annahme zu Grunde, dass grundsätzlich ein Zusammenhang zwischen Wissen und erfolgreichem Unterrichten besteht.

**Vogelsang (2014)**

In der Studie von Vogelsang (2014) wurde die prädiktive Validität des Testinstruments von Riese (2009) im Hinblick auf die Qualität des durch die Probanden bereitgestellten Lehrangebots bezüglich der Merkmale Motivierung, kognitive Aktivierung, Strukturierung, Adaptivität, Klassenführung, Umgang mit Experimenten und lernprozessorientierte Sequenzierung untersucht. Hierfür wurde eine Videostudie mit 14 Lehramtsstudierenden und 8 Lehramtsanwärtern durchgeführt.

Riese (2009, S. 84) operationalisierte das pädagogische Wissen (in dieser Studie als erziehungswissenschaftliches Wissen bezeichnet) als Wissen in den Inhaltsbereichen Erziehung und Bildung, Unterricht sowie allgemeine Didaktik und Schulentwicklung und Gesellschaft. Das fachdidaktische Wissen wurde als deklaratives Wissen über (allgemeine) Aspekte physikalischer Lernprozesse sowie über den Einsatz von Experimenten und als prozedurales Wissen zur Gestaltung und Planung sowie zur Beurteilung, Analyse und Reflexion von Lernprozessen und zur adäquaten Reaktion in kritischen Unterrichtssituationen operationalisiert (Riese, 2009, S. 82-83). Mit dem Fachwissenstest wurden Schulwissen, vertieftes Schulwissen und universitäres Wissen erfasst (Riese, 2009, S. 77).

Das pädagogische Wissen der Lehrenden korrelierte positiv mit nahezu allen betrachteten Unterrichtsqualitätsmerkmalen (mit Ausnahme der kognitiven Aktivierung, der Adaptivität und der lernprozessorientierten Sequenzierung) in mittlerer Höhe ($\tau_{\text{Kendall}} = .36 - .49$). Während sich keinerlei Zusammenhang zum fachdidaktischen Wissen zeigte, korrelierte das Fachwissen der Lehrenden signifikant negativ mit der Motivierung ($\tau_{\text{Kendall}} = -.31$), Adaptivität ($\tau_{\text{Kendall}} = -.45$) und Klassenführung ($\tau_{\text{Kendall}} = -.33$) (Vogelsang, 2014, S. 487). Eine detailliertere Analyse zeigte, dass die beobachteten Korrelationen auf negative Zusammenhänge zum vertieften Schulwissen der Lehrenden zurückgingen – zum Schulwissen und universitärem Wissen zeigten sich keine signifikanten Korrelationen (Vogelsang, 2014, S. 489).

Problematisch an der Studie von Vogelsang ist die Heterogenität der untersuchten Stichprobe. Es wurden sowohl Lehramtsstudierende als auch Lehramtsanwärter untersucht, die Unterrichtserfahrung der Probanden variierte also in erheblichem

Maße. Auch innerhalb der Teilstichproben existierten große Unterschiede in der Unterrichtserfahrung – so gab es Studierende, die zum Zeitpunkt der Videoaufnahmen bereits 20 Stunden Physik unterrichtet hatten und Studierende, deren erste Unterrichtsstunde im Fach Physik aufgezeichnet wurde. Die Hälfte der Probanden unterrichtete an der Haupt- oder Realschule (HR), die andere Hälfte am Gymnasium oder der Gesamtschule (GyGe) (Vogelsang, 2014, S. xl im Anhang).

Eine Analyse der Zusammenhänge zwischen dem Professionswissen und den Unterrichtsqualitätsdimensionen in den Substichproben (Studierende/Lehramtsanwärter bzw. HR/GyGe) zeigte in der Tat erhebliche Unterschiede: Während das pädagogische Wissen, abhängig von der betrachteten Substichprobe, mit unterschiedlichen Qualitätsdimensionen korrelierte, die Korrelationen aber stets positiv waren, ergaben sich für das fachdidaktische Wissen mal positive, mal negative Korrelationen. Zum Fachwissen zeigten sich ebenfalls unterschiedliche, aber durchweg negative Korrelationen (Vogelsang, 2014, S. 487/489). Aufgrund der geringen Stichprobengrößen der Substichproben ist es allerdings fraglich, ob diese Korrelationen überhaupt interpretiert werden sollten.

Die positiven Zusammenhänge zum pädagogischen Wissen in der Gesamtstichprobe führt Vogelsang auf das hohe (und innerhalb dieser Substichprobe sehr homogene) Wissen der Lehramtsanwärter zurück, deren Lehrangebot erwartungsgemäß besser beurteilt wurde als das der Studierenden. Auch die negativen Zusammenhänge zwischen Fachwissen und den Unterrichtsqualitätsdimensionen resultieren nach Vogelsang aus den Unterschieden zwischen diesen beiden Subgruppen, da die Studierenden tendenziell besser im Fachwissen abschneiden als die Lehramtsanwärter (Vogelsang, 2014, S. 487-488).[4]

Die Ergebnisse zusammenfassend zu bewerten, erweist sich aufgrund der Heterogenität der Stichprobe und der fraglichen Aussagekraft der Analyse der Zusammenhänge in den Substichproben als schwierig. Ungünstig für die Interpretierbarkeit der Ergebnisse ist zudem, dass die Anzahl der für die Beurteilung des Lehrangebots hinzugezogenen Unterrichtsstunden zwischen den Lehrenden zwischen $1-3$ Unterrichtsstunden variierte, das Stundenthema nicht konstant gehalten wurde und der Unterricht in verschiedenen Jahrgangsstufen (6-11) aufgezeichnet wurde (Vogelsang, 2014, S. xl im Anhang). Es ist fraglich, ob die Beurteilung der Unterrichtsqualitätsmerkmale über diese unterschiedlichen Unterrichtssettings hinweg vergleichbar ist. Da keine Variablen auf Schülerseite (wie z. B. Leistung oder Motivation) untersucht wurden, kann zudem keine Aussage darüber getroffen werden, ob es sich bei den betrachteten Merkmalen guten Unterrichts um Maße für Unterrichtsqualität im Sinne von Fenstermacher und Richardson (2005, S. 192) handelt.

**Sadler et al. (2013)**

Die einzige Large-Scale Studie zur Untersuchung des Zusammenhangs zwischen dem

[4]Die Gruppenunterschiede im Fachwissen sind allerdings nicht signifikant. Die deskriptiven Ergebnisse bezüglich der Fachwissensniveaus zeigen, dass die Studierenden im vertieften Schulwissen und universitären Wissen besser und im Schulwissens schlechter abschneiden als die Lehramtsanwärter (Vogelsang, 2014, S. 469-470).

fachspezifischen Professionswissen von Lehrkräften und Schülerleistungszuwachs in der Physik wurde von Sadler et al. (2013) durchgeführt. In einer Stichprobe von 181 Naturwissenschaftslehrkräften und ihren Klassen (Jahrgangsstufe 7/8, American Middle School) wurden Zusammenhänge zwischen dem Fachwissen und fachdidaktischen Wissen der Lehrkräfte und den Leistungszuwächsen der Lernenden über den Zeitraum eines (bzw. eines halben) Schuljahres untersucht. Dabei wählten die Autoren einen gänzlich anderen Ansatz als die bisher vorgestellten Studien.

Das Fachwissen der Lehrkräfte und das Fachwissen ihrer Schülerinnen und Schüler wurde mit dem gleichen Testinstrument erhoben. In 20 Multiple-Choice-Aufgaben wurden chemisches und physikalisches Fachwissen und typische Fehlvorstellungen von Lernenden zu allen in den amerikanischen Bildungsstandards für die Naturwissenschaften für diese Schulstufe vorgesehenen Konzepten gemessen.[5] Zusätzlich zu der korrekten Antworten für eine Aufgabe sollten die Lehrkräfte außerdem die Antwortmöglichkeit ankreuzen, die ihrer Einschätzung nach die meisten ihrer Schülerinnen und Schüler ankreuzen würden. Auf Basis aller Schülerantworten wurden 12 Aufgaben identifiziert, die sogenannte starke Fehlvorstellungen abtesteten – die Mehrheit der Lernenden, die solch eine Aufgabe falsch beantworteten, kreuzte die gleiche falsche Antwortmöglichkeit an. Die Fähigkeit der Lehrkräfte diese falsche Antwortmöglichkeit zu identifizieren bezeichnen die Autoren als *Knowledge of Students Misconceptions* (KOSM), das sie als Teilaspekt von PCK ansehen.[6]

Die Zusammenhänge zwischen den Lehrerantworten und den Lernzuwächsen der Schülerinnen und Schüler in den entsprechenden Aufgaben wurden auf Aufgabenebene untersucht. Ob die Lernenden die Antwort auf eine Aufgabe ohne starke Fehlvorstellung zwischen Prä- und Post-Erhebung gelernt hatten, hing stark damit zusammen, ob ihre Lehrkraft die Aufgabe korrekt beantworten konnte. Für Aufgaben mit einer populären falschen Antwortmöglichkeit, spielte das CK der Lehrkräfte keine Rolle, wohl aber ihr KOSM: Lernende, deren Lehrkraft die unter den Antwortmöglichkeiten einer Aufgabe die häufigste Fehlvorstellung identifizieren konnten, zeigten wesentlich höhere Leistungszuwächse in der entsprechenden Aufgabe.

Sadler et al. (2013, S. 1041) heben hervor, dass die gefundenen Zusammenhänge erst im Rahmen von Analysen auf Aufgabenebene sichtbar wurden – zwischen den CK- und KOSM-Gesamtscores der Lehrkräfte und den Testergebnissen der Lernenden zeigten sich nur äußerst geringe Zusammenhänge (ob die Zusammenhänge signifikant wurden, wird nicht erläutert). Als Grund hierfür sehen die Autoren die Themenabhängigkeit des fachspezifischen Professionswissens an - so scheint nur wenig Transfer zwischen dem CK und KOSM der Lehrkräfte zu verschiedenen Konzepten stattzufinden.

Zusammenfassend lässt sich feststellen, dass die in der Physik durchgeführten Studien zur prädiktiven Validität von Professionswissenstests bisher inkonsistente und noch unvollständige Ergebnisse liefern. In der Grundschule scheint das Fach-

[5] Da 60% der Aufgaben physikalische Wissen testeten, wird diese Studie hier als Beispielstudie aus der Physik aufgeführt.

[6] In Sadler et al. (2013) finden sich keine Informationen darüber, wie viele der KOSM-Aufgaben *physikalisches* Wissen abtesten.

wissen von Lehrkräften auf Grundschul-, Sekundarstufen- und Universitätsniveau nur unter bestimmten Voraussetzungen positiv mit Schülerleistungszuwächsen zusammenzuhängen (Ohle et al., 2011). An weiterführenden Schulen zeigten sich negative Zusammenhänge zwischen Fachwissen und Merkmalen guten Unterrichts, wenn Wissen erfasst wurde, das über reines Schulwissen hinaus geht (Vogelsang, 2014). Diese Ergebnisse könnten konsistent mit Ergebnissen aus der Mathematik sein, wenn man davon ausgeht, dass zwar ein gewisses Mindestmaß an Fachwissen nötig ist, um erfolgreich zu unterrichten, dass das Fachwissen von Lehrkräften oberhalb eines bestimmten Schwellenwertes allerdings keinen positiven Einfluss auf Unterrichtserfolg oder Unterrichtsqualität mehr hat (vergl. hierzu auch Darling-Hammond, 2000, S. 3-4). Möglicherweise verfügten die in der PLUS-Studie untersuchten Lehrkräfte, mangels einer Ausbildung im Fach Physik, nicht über dieses Mindestmaß an Fachwissen oder aber der im PLUS-Projekt eingesetzte Fachwissenstest, in dem auch Wissen auf Sekundarstufen- und Universitätsniveau abgefragt wurde, differenzierte nicht ausreichend in dem für das erfolgreiche Unterrichten relevanten Wissensbereich. Letzteres könnte (in Bezug auf *gutes* Unterrichten) auch für den in der Studie von Vogelsang eingesetzten Fachwissenstest gelten. Betrachtet man den Einfluss von Fachwissen auf Schulniveau, wie er auch in COACTIV betrachtet wurde, konnten bedeutsame Zusammenhänge zum Schülerleistungszuwachs bisher nur auf Aufgabenebene nachgewiesen werden. Auf Testebene scheinen diese Zusammenhänge sehr gering zu sein (Sadler et al., 2013).

Bezüglich des pädagogischen Wissens von Physiklehrkräften deuten sich, wie auch in der Mathematik, Zusammenhänge zu Merkmalen guten Unterrichts an. Diese Zusammenhänge wurden allerdings erst in einer Studie untersucht (Vogelsang, 2014). Zusammenhänge zwischen pädagogischem Wissen und Zielkriterien von Unterricht wurden bisher nicht betrachtet.

Eine Zusammenfassung der Ergebnisse bezüglich PCK ist schwierig: Kleine Zusammenhänge zwischen PCK und erfolgreichem Unterrichten konnten in zwei Studien zwar nachgewiesen werden, allerdings nur unter Kontrolle zahlreicher anderer Variablen auf Klassenebene (PLUS) oder unter Ausschluss von Teilstichproben (QuiP). In der Studie von Vogelsang (2014) fanden sich gar keine Zusammenhänge. PCK wurde in allen Studien durch unterschiedliche Facetten operationalisiert, die aber zum Teil überlappen. Rückschlüsse auf die Relevanz bestimmter Facetten von PCK zu ziehen, ist auf dieser Grundlage nicht möglich. In allen vier genannten Studien beinhalten die eingesetzten Testinstrumente Aufgaben zur Abfrage von konkreten Schülerfehlvorstellungen. Dass dieses Wissen durchaus eine Rolle spielen kann, belegen bisher nur die Ergebnisse von Sadler et al. (2013) in einer US-amerikanischen Stichprobe: Lernende, die von einer Lehrkraft unterrichtet wurden, die in einer Aufgabe unter den Antwortmöglichkeiten die typische Schülerfehlvorstellung erkannte, kannten mit höherer Wahrscheinlichkeit am Ende einer Unterrichtseinheit die richtige Antwort auf diese Aufgabe als wenn die Lehrkraft die Fehlvorstellung nicht erkannt hatte. Dieses Wissen scheint allerdings sehr isoliert zu sein, so dass sich dieser Zusammenhang nur auf Aufgabenebene zeigt, nicht aber, wenn der Zusammenhang zwischen der Gesamtzahl erkannter Schülerfehlvorstellungen und dem Leistungszuwachs über alle Aufgaben betrachtet wird.

Die Ergebnisse aus den Studien in der Mathematik scheinen nicht einfach übertragbar auf den Physikunterricht zu sein. An dieser Stelle sei darauf hingewiesen, dass sich sowohl das in der SII-Studie beschriebene *specialised content knowledge* als auch das in COACTIV gemessene PCK primär auf Aufgaben bezieht. Aufgaben spielen im Mathematikunterricht allerdings eine wesentlich größere Rolle als im Physikunterricht. Darüber hinaus wird Mathematik in der Schule mit einem wesentlich höheren Stundenumfang unterrichtet als Physik – die Leistungsentwicklung der Lernenden könnte daher im Physikunterricht in geringerem Maße durch die Lehrkraft beeinflusst sein. Dazu kommt, das Physiklehrkräfte sich weitaus stärker mit einem geringen Fachinteresse der Lernenden auseinandersetzen müssen als dies bei Mathematiklehrkräften der Fall ist – damit ist es auch schwieriger die Lernenden zu einer aktiven Teilnahme am Unterricht zu motivieren.

Insgesamt kann festgehalten werden, dass bezogen auf den Physikunterricht, noch nicht hinreichend geklärt ist, welches Wissen als relevant für gutes und erfolgreiches Unterrichten angesehen werden kann und dass die Rahmenbedingungen des Physikunterrichts den Nachweis dieses Wissens erschweren könnten.

# 5. Ableitung des eigenen Forschungsansatzes

Das Bestreben der Professionswissensforschung liegt in der Identifikation von Wissen, auf dessen Grundlage sich Lehrkräfte zu erfolgreich Unterrichtenden entwickeln können. Bei der Suche danach muss sich die Forschung zum Professionswissen allerdings noch mit zahlreichen Herausforderungen auseinandersetzen. So existiert weder ausreichend empirische Evidenz für die Annahme, dass das Professionswissen von Lehrkräften eine wichtige Voraussetzung für qualitätvolles Unterrichten darstellt (Abschnitt 4.3 auf Seite 41), noch besteht Einigkeit darüber, wie das Professionswissen von Lehrkräften zu modellieren ist – sowohl die Anzahl an Dimensionen als auch die in den jeweiligen Dimensionen als relevant erachteten Facetten variieren zwischen verschiedenen Modellen (vergl. Abschnitt 2.3 auf Seite 12). Insbesondere für den Physikunterricht ist noch nicht hinreichend geklärt, welches Wissen als unterrichtsrelevant angenommen werden kann.[1] Darüber hinaus ist selbst der grundsätzliche Zusammenhang zwischen Wissen und Handeln von Lehrkräften Gegenstand von Diskussionen (vergl. Abschnitt 3.1 auf Seite 22). Ein zentrales Ziel der Professionswissensforschung muss daher die Untersuchung der Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg sein.

Um diese Zusammenhänge quantitativ zu untersuchen, werden valide und reliable Testinstrumente für die zeitökonomische Erfassung des Professionswissens von Lehrkräften benötigt. Bei der Entwicklung solcher Testinstrumente müssen sich Forschende allerdings wieder auf ein bestimmtes Modell, bestimmte Dimensionen des Professionswissens und bestimmte Facetten innerhalb dieser Dimensionen beziehen. Werden nun Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Lehrerfolg untersucht, diese aber nicht gefunden, lassen sich kaum Rückschlüsse auf die Ursache hierfür ziehen. Mögliche Gründe könnten sein, dass Professionswissen nicht die ihm zugeschriebene Rolle für gutes und erfolgreiches Unterrichten spielt, dass kein Zusammenhang zwischen dem in schriftlichen Testinstrumenten explizierbaren Wissen und dem Handeln von Lehrkräften im Unterricht besteht oder aber, dass die Testinstrumente nicht valide sind – sei es, weil in die Modellierung nicht die relevanten Wissensfacetten einbezogen wurden oder weil diese in den Testinstrumenten nicht adäquat umgesetzt wurden. Unabhängig von der tatsächlichen Ursache sollte ein solches Ergebnis allerdings ernst genommen werden: Bei der Interpretation von Daten, die mit solchen Testinstrumenten er-

[1] Ein ähnlicher Forschungsbedarf besteht in den naturwissenschaftlichen Fächern Chemie und Biologie. Auch hier existieren bisher wenige Studien zur Untersuchung des Zusammenhang zwischen Professionswissen, Unterrichtsqualität und Lehrerfolg. Ausnahmen stellen z. B. die Arbeiten von Gess-Newsome et al. (2010) oder Mahler, Großschedl und Harms (2015) aus der Biologie dar.

hoben wurden, sollte man Vorsicht walten lassen, da offensichtlich kein Wissen erhoben wird, das mit Unterrichtsqualität oder Lehrerfolg einhergeht.[2] Dies ist insbesondere vor dem Hintergrund wichtig, dass Instrumente zur Erfassung des Professionswissens oftmals mit dem Ziel entwickelt werden, die Wirksamkeit der Lehrerausbildung zu überprüfen. Lassen sich allerdings Zusammenhänge nachweisen, liefert dies zum einen empirische Evidenz für die prädiktive Validität der eingesetzten Testinstrumente und zum anderen für die Relevanz des Professionswissens für gutes und erfolgreiches Unterrichten – letzteres gilt natürlich nur unter der Voraussetzung, dass über weitere Verfahren der Validierung sicher gestellt wird, dass nicht ein anderes für Unterrichtsqualität und Lehrerfolg prädiktives Konstrukt erfasst wurde.

Auch wenn aus theoretischer Sicht noch viele offene Fragen bezüglich der Modellierung von Professionswissen existieren und man zudem aus den genannten Gründen Gefahr läuft, Ergebnisse zum Zusammenhang zwischen Professionswissen, Unterrichtsqualität und Lehrerfolg nicht eindeutig interpretieren zu können, ist die Untersuchung dieser Zusammenhänge von großem Interesse. Schließlich gibt es kaum eine andere Möglichkeit, um der Antwort auf die Frage näher zu kommen, welches Wissen relevant für gutes und erfolgreiches Unterrichten ist.

Die vorliegende Studie wurde im Rahmen des Projekts „Professionswissen in den Naturwissenschaften“ (ProwiN) durchgeführt. In der ersten Projektphase von ProwiN (ProwiN I) wurden schriftliche Testinstrumente zur Erfassung des Professionswissens von Physik-, Chemie- und Biologielehrkräften entwickelt und zunächst ohne Bezug zu Unterrichts- oder Schüleroutputvariablen validiert. Die vorliegende Arbeit ist Teil der zweiten Projektphase, und verfolgt das Ziel, die prädiktive Validität der Testinstrumente für die Erfassung des Professionswissens von *Physik*lehrkräften in Bezug auf gutes und erfolgreiches Unterrichten zu überprüfen.

Im Folgenden wird zunächst das ProwiN-Projekt vorgestellt. Um die Ausgangslage der vorliegenden Arbeit zu beschreiben, wird ausführlich auf das ProwiN-Professionswissensmodell und dessen Umsetzung in den Testinstrumenten eingegangen. Außerdem werden die Ergebnisse aus den in der ersten Projektphase durchgeführten Validierungsstudien zusammengefasst. Im Anschluss daran wird abgeleitet, welche Zielkriterien für erfolgreiches Unterrichten und welches Merkmal guten Unterrichts für die Untersuchung der prädiktiven Validität der ProwiN-Testinstrumente betrachtet werden. Abschließend erfolgt eine Einordnung der vorliegenden Studie in das ProwiN-Projekt.

## 5.1. Das „ProwiN“-Projekt

Das Projekt „Professionswissen in den Naturwissenschaften“ (Borowski et al., 2010) ist ein vom Bundesministerium für Bildung und Forschung (BMBF) im Rahmenprogramm „Entwicklung von Professionalität des pädagogischen Personals in Bildungseinrichtungen“ (ProPäda) gefördertes fächerübergreifendes Projekt. In

[2] Die zum Professionswissen in Beziehung gesetzten Merkmale der Unterrichtsqualität oder Zielkriterien von Unterricht müssen natürlich ihrerseits valide und reliabel erfasst werden.

der ersten Phase des Projekts wurden in den beteiligten Fachdidaktiken (Physik: Kirschner (2013), Chemie: Dollny (2011), Biologie: Jüttner (2013)) auf Grundlage eines gemeinsamen Professionswissensmodells Papier-und-Bleistift-Tests zur Erfassung des Fachwissens und fachdidaktischen Wissens von Physik-, Chemie- und Biologielehrkräften entwickelt und validiert. In der Lehr-Lernpsychologie wurde außerdem ein Papier-und-Bleistift-Test zur Erfassung des pädagogischen Wissens von Naturwissenschaftslehrkräften entwickelt und validiert (Lenske et al., 2015).

### 5.1.1. Professionswissen in „ProwiN“

Als Grundlage für die Testkonstruktion wurde im Rahmen der ersten Phase des ProwiN-Projekts ein Modell für die Erfassung des Professionswissens von Naturwissenschaftslehrkräften entwickelt (Tepner et al., 2012), das von Kirschner (2013) für das Professionswissen von Physiklehrkräften spezifiziert wurde (siehe Abbildung 5.1). Die folgenden drei Abschnitte zur Modellierung des Professionswissens im ProwiN-Projekt basieren auf der Beschreibung des fächerübergreifenden ProwiN-Modells von Tepner et al. (2012) und der Beschreibung des physikspezifischen ProwiN-Modells von Kirschner (2013). Der Ausdruck ProwiN-Modell wird im Folgenden stets für das in der Physik entwickelte Modell zur Aufgabenkonstruktion verwendet.

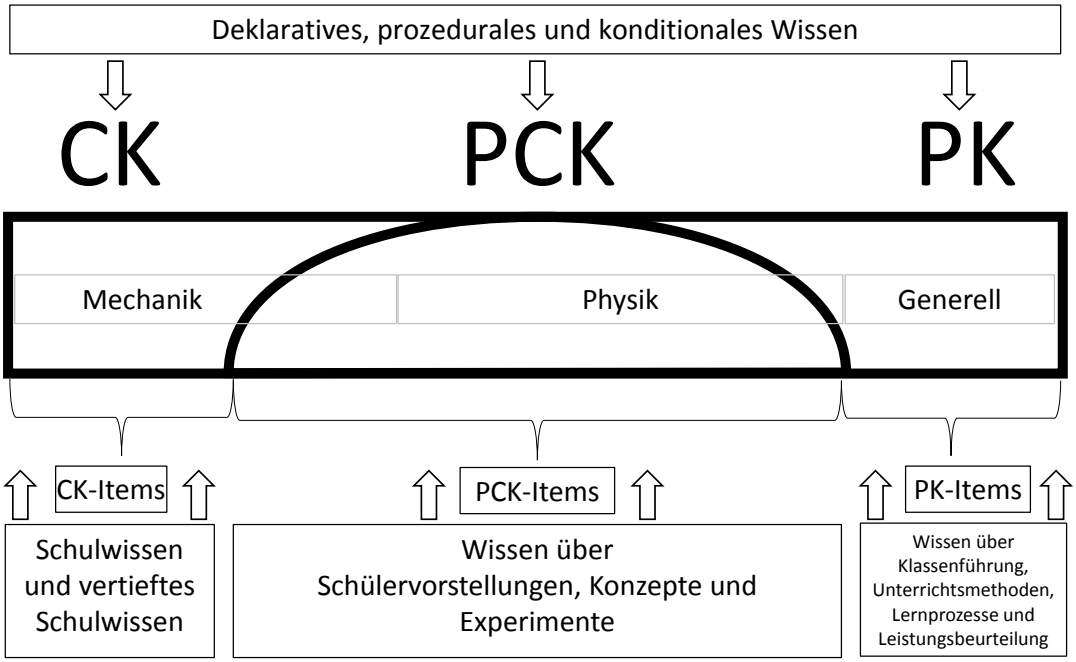



Abbildung 5.1.
*ProwiN-Modell für das Professionswissen von Physiklehrkräften (Kirschner, 2013, S. 36).*

Im ProwiN-Modell wird das Professionswissen von Lehrkräften durch die drei als besonders wichtig für erfolgreiches Unterrichten erachteten Dimensionen Fachwissen

(CK), fachdidaktisches Wissen (PCK) und pädagogisches Wissen (PK) modelliert. Die dem Modell zugrunde liegende Annahme, dass es sich dabei um disjunkte, aber zusammenhängende Dimensionen handelt, konnte von Kirschner (2013, S. 76) für die anhand des Modells entwickelten Testinstrumente empirisch bestätigt werden. In allen drei Dimensionen wird Wissen über Tatsachen und Inhalte (deklaratives Wissen, „knowing *that*“), Wissen über Handlungen (prozedurales Wissen, „knowing *how*“) und Wissen über Begründungen, Beurteilungen von Unterrichtssituationen und Bedingungen, unter denen eine Handlung als angemessen erachtet werden kann (konditionales Wissen, „knowing *when* and *why*“), erfasst (Kirschner, 2013, S. 26; Paris et al., 1983, S. 302-304; Tepner et al., 2012, S. 17).[3] Die Betrachtung des konditionalen Wissens berücksichtigt die Situationsabhängigkeit von unterrichtlichen Entscheidungsprozessen (Tepner et al., 2012, S. 17). Im Folgenden wird näher erläutert, wie die Dimensionen CK, PCK und PK in ProwiN operationalisiert und in Testaufgaben umgesetzt wurden.

#### 5.1.1.1. Fachwissen

Physikalisches Fachwissen ist nicht nur domänenspezifisch, sondern auch themenabhängig. Schließlich ist nicht zwingend davon auszugehen, dass Fachwissen in einem Inhaltsbereich mit Fachwissen in einem anderen Inhaltsbereich der Physik einhergeht (vergl. auch Sadler et al., 2013, S. 1041). In ProwiN wird daher primär das Fachwissen im Inhaltsbereich *Mechanik* erfasst, da diesem Thema sowohl in der Schule als auch im Fachstudium an den Universitäten eine hohe Relevanz zugesprochen wird (Kirschner, 2013, S. 29). Ergänzend beinhaltet der CK-Test eine Aufgabe aus dem ebenfalls im Physikunterricht an Schulen behandelten Inhaltsbereich der Elektrizitätslehre (Dieser Inhaltsbereich taucht im ProwiN-Modell auf Seite 55 nicht explizit auf).

Es besteht weitestgehend Konsens darüber, dass für erfolgreiches Unterrichten Wissen nötig ist, das über das in einer bestimmten Schulstufe vermittelte Wissen hinaus geht (Tepner et al., 2012, S. 10). In ProwiN wird *Schulwissen*, also physikalisches Wissen, das im Unterricht in der Sekundarstufe I oder der Oberstufe vermittelt wird und über das daher auch leistungsstarke Schülerinnen und Schüler verfügen können, und *vertieftes Schulwissen* erhoben (Kirschner, 2013, S. 27). Universitäres Wissen, also physikalisches Wissen, das an der Universität gelehrt wird und keinen expliziten Schulbezug hat, wird nicht erhoben. In bestimmten Fällen (z. B. für Hauptschullehrkräfte, die lediglich in der Sekundarstufe 1 unterrichten) können die Aufgaben des CK-Tests allerdings Wissen abfragen, das aus Perspektive dieser Lehrkräfte dem universitären Wissen zugeordnet werden könnte. Im ProwiN-Projekt wird keine präzisere Definition des vertieften Schulwissens vorgenommen, Kirschner (2013) beschreibt allerdings Merkmale von Aufgaben, mit denen vertieftes Schulwissen in ProwiN erfasst werden soll:

[3] Es ist kein erklärtes Ziel des ProwiN-Projekts, diese Bereiche statistisch voneinander zu trennen. Die Differenzierung diente lediglich dazu, sicherzustellen, dass alle Wissensbereiche durch die Testaufgaben abgedeckt wurden.

- „Unbekannt: Keine Standardaufgaben für die Schule; Aufgaben, die an der Universität nicht explizit gelehrt werden
- Vollständiger Verzicht auf Oberstufen- und universitäre Mathematik (insbesondere Analysis)
- Benötigtes deklaratives Wissen geht im Wesentlichen nicht über die Sekundarstufe II hinaus
- Wissen muss flexibel eingesetzt werden
- Erste Lösungsansätze können in die Irre führen“ (S.27)

Da eine Aufgabe für einige Versuchspersonen bekannt, für andere aber unbekannt sein kann, sind vertieftes Schulwissen und Schulwissen im ProwiN-Modell nicht eindeutig voneinander trennbar (Kirschner, 2013, S. 27).

#### 5.1.1.2. Fachdidaktisches Wissen

Das PCK von Physiklehrkräften wird in ProwiN über die als zentral angesehenen Facetten *Wissen über Schülervorstellungen* und Wissen über Instruktionsstrategien und Repräsentationen modelliert. Letzteres wird als *Wissen über Konzepte* und *Wissen über Experimente* für den naturwissenschaftlichen Kontext spezifiziert (Kirschner, 2013, S. 33; Tepner et al., 2012, S. 15). Hiermit wird dem Umstand Rechnung getragen, dass Experimente zum einen eine zentrale Rolle für die Erkenntnisgewinnung in der Physik spielen und zum anderen als eine der wichtigsten Methoden für den Physikunterricht betrachtet werden (Tesch, 2011, S. 191). Die zentrale Stellung des Experiments im Physikunterricht wird zudem explizit in den Kernlehrplänen der gymnasialen Mittelstufe erwähnt (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen [MSW], 2008, 2011).

Die Aufgaben im PCK-Test wurden facettenübergreifend konstruiert: Zum einen wäre für eine getrennte Analyse der Facetten eine Anzahl an Aufgaben nötig gewesen, die den vorgesehenen Testumfang weit überschritten hätte, zum anderen ist bisher nicht geklärt, ob sich das PCK einer Lehrkraft als Summe über das Wissen in verschiedenen Facetten beschreiben lässt oder sich vielmehr aus der Integration verschiedener Facetten ergibt und PCK damit mehr ist, als lediglich die Summe seiner Teile (vergl. hierzu Kirschner, 2013, S. 30/103).

Das Wissen über Schülervorstellungen umfasst in ProwiN das Wissen über korrekte und inkorrekte Vorstellungen der Lernenden und Wissen darüber, welche Darstellungsformen und Repräsentationen die Verfestigung inkorrekter Vorstellungen von Schülerinnen und Schülern noch begünstigen. Aufgaben zu dieser Facette können daher zum Teil auch dem Wissen über Konzepte und deren fachdidaktischer Aufbereitung zugeordnet werden. Die Aufgaben zum Wissen über Konzepte erfassen Wissen darüber, wie physikalische Konzepte aufbereitet werden können, um Lernende in ihren Lernprozessen zu unterstützen, und Wissen über physikalisch angemessene Kriterien für die Leistungsbeurteilung von Unterrichtsprodukten. Das Wissen über Experimente umfasst Wissen über die fachdidaktisch angemessene Gestaltung von Experimenten sowie über verschiedene Funktionen von Experimenten im Unterricht. Auch hier existieren Aufgaben, die der Facette Wissen über

Schülervorstellungen oder der Facette Wissen über Konzepte zugeordnet werden können.

Ähnlich wie in der Konzeptualisierung von Gess-Newsome et al. (2010), die innerhalb von PCK die Bereiche PCK-CK und PCK-PK unterscheiden, gibt es Aufgaben im PCK-Test, deren Fachbezug besonders ausgeprägt ist. Dies gilt beispielsweise für Aufgaben, die das Wissen über Konzepte erfassen. Diese Aufgaben liegen daher am Übergang zwischen PCK und CK.

Da das PCK von Lehrkräften als themenspezifisches Wissen angesehen wird (vergl. z. B. Sadler et al., 2013, S. 1041), fokussiert auch der PCK-Test im Wesentlichen auf ein Thema. Erfasst wird Wissen im Inhaltsbereich *Mechanik*. Im Inhaltsbereich *Physik* wird zudem Wissen erhoben, das sich themenunabhängig auf den Physikunterricht bezieht (z. B. „Was spricht für die Verwendung von Einheiten bei Rechnungen im Physikunterricht?“) (Kirschner, 2013, S. 29-30). Ebenso wie der CK-Test umfasst auch der PCK-Test eine Aufgabe zum Inhaltsbereich Elektrizitätslehre.

#### 5.1.1.3. Pädagogisches Wissen

Die Modellierung von PK berücksichtigt die Facetten *Klassenführung*, *Unterrichtsmethoden*, *individuelle Lernprozesse* und *Leistungsbeurteilung*. Dabei handelt es sich um Facetten, die sich in den Standards für die Lehrerbildung wiederfinden und auf Grundlage empirischer Forschungsergebnisse als bedeutsam erachtet werden (Tepner et al., 2012, S. 20). Das Wissen über Klassenführung umfasst Wissen über Regeln und Rituale, die die Aufrechterhaltung des Unterrichtsflusses unterstützen sowie Wissen über Störungsprävention und den Umgang mit Disziplinproblemen. Die Aufgaben zum Wissen über Unterrichtsmethoden beziehen sich insbesondere auf die Passung zwischen angewendeten Unterrichtmethoden und der übergeordneten Zielsetzung einer Unterrichtseinheit und darauf, ob ausgewählte Unterrichtsmethoden adäquat umgesetzt werden. Bezüglich der Facette zu individuellen Lernprozessen wird Wissen über Maßnahmen zur Förderung des selbstregulierten Lernen erfasst. Das Wissen über Leistungsbeurteilung bezieht sich auf die Gestaltung von motivational und kognitiv förderlichem Feedback (vergl. Lenske et al., 2015; Tepner et al., 2012). Da mit dem PK-Test fachunspezifisches Wissen von Lehrkräften erhoben wird, wird in den PK-Aufgaben kein Bezug zum Unterrichtfach Physik hergestellt. Deklaratives und konditional-prozedurales Wissen wird in zwei separaten Testteilen des PK-Tests erfasst ($PK_D$ bzw. $PK_{KP}$).

### 5.1.2. Validierung der „ProwiN“-Testinstrumente

Die Validierung der Testinstrumente zum Professionswissen von Physiklehrkräften erfolgte im Rahmen der ersten Phase des ProwiN-Projekts auf Basis von Expertenbefragungen, einer Modellprüfung zur angenommenen dreidimensionalen Struktur des Professionswissens, Korrelationsanalysen zwischen den Dimensionen und durch den Vergleich bekannter Gruppen mit zu erwartenden Fähigkeitsunterschieden (Kirschner, 2013; Lenske et al., 2015). Um Hinweise auf die Praxisrelevanz des erfassten Wissens zu bekommen, wurden außerdem Unterschiede zwischen Lehr-

kräften bzw. Lehrkräften im Vorbereitungsdienst und Studierenden untersucht. Tabelle 5.1 auf Seite 61 zeigt die von Kirschner (2013) und Lenske et al. (2015) geprüften Hypothesen zur Überprüfung der Validität der Instrumente zur Erfassung des Professionswissens von *Physik*lehrkräften. Die Ergebnisse von Kirschner wurden auf Basis von Rasch-Personenfähigkeiten gerechnet, während die Ergebnisse von Lenske et al. auf klassisch berechneten Summenscores basieren. In der Auswertung des $PK_D$-Tests wurden von Kirschner (2013, S. 137) einige Aufgaben, die keine gute Passung ins Rasch-Modell zeigten, ausgeschlossen. In Einzelfällen kommt es daher zu leicht unterschiedlichen Validierungsergebnissen.

Die Inhaltsvalidität der fachspezifischen Professionswissenstests wurde über den Abgleich mit Curricula und Fachliteratur, Expertenbefragungen und der Testentwicklung anhand des Modells sichergestellt (Kirschner, 2013, S. 77). Die Testinstrumente zur Messung des deklarativen und konditional-prozeduralen pädagogischen Wissens basieren auf einem theoriegeleitet entwickelten und mittlerweile als validiert geltenden Testinstrument aus der COACTIV-R Studie. Die Inhaltsvalidität der dort verwendeten Aufgaben wurde von Voss, Kunter und Baumert (2011b, S. 6) über die Einschätzung der Unterrichtsrelevanz, der Fachunabhängigkeit und der Authentizität der im Test beschriebenen Unterrichtssituationen durch 20 Lehrkräfte (im Mittel acht Lehrkräfte pro Aufgabe) sichergestellt. Der $PK_{KP}$-Test wurde zusätzlich durch die Befragung von acht Experten mit fachdidaktischem oder pädagogisch-psychologischem Hintergrund (Professoren/Professorinnen oder Postdoktorierende) inhaltlich validiert (Lenske et al., 2015).

Die Dimensionalitätsprüfungen und Korrelationsanalysen zur Überprüfung der Konstruktvalidität bestätigen, dass CK, PCK und PK zusammenhängende, aber trennbare Dimensionen darstellen. Erwartungsgemäß korrelieren CK und PCK stärker miteinander als CK und PK.

Die Ergebnisse zur Kriteriumsvalidität weisen darauf hin, dass die fachspezifischen Professionswissenstests das fachspezifische Wissen messen, das am Gymnasium unterrichtende Physiklehrkräfte in Abgrenzung zu Lehrkräften anderer Fächer und anderer Schulformen auszeichnet. Dabei handelt es sich allerdings um Wissen, über das auch Diplomphysiker mit universitärer Lehrerfahrung verfügen können. Physiklehrkräfte und Physiklehrkräfte im Vorbereitungsdienst, die ein gymnasiales Lehramt studiert hatten, verfügten über ein höheres PCK und CK als Studierende des gymnasialen Lehramts. Dieses Ergebnis kann als Hinweis darauf gedeutet werden, dass die Tests Wissen erfassen, das Lehrkräfte mit Praxiserfahrung auszeichnet und das daher relevant für das Unterrichten von Physik am Gymnasium sein könnte. Da es sich allerdings nur um quasi-längsschnittlich erhobene Daten handelt, kann keine Aussage darüber getroffen werden, ob dieses Wissen durch Praxiserfahrung erworben wird. Die Tests zum pädagogischen Wissen messen fachunspezifisches Wissen, über das Lehrkräfte unterschiedlicher Fachrichtungen gleichermaßen verfügen. Dozenten der universitären Lehrerbildung schnitten im Test zum deklarativen Wissen im Mittel besser ab als Lehrkräfte, was dafür spricht, dass dieses Wissen im Rahmen der universitären Lehrerausbildung von ihnen gelehrt wird. Im Test zum konditional-prozeduralen Wissen schnitten Lehrkräfte und Dozenten gleichermaßen gut ab, was dafür spricht, dass der Test eher praxisnahes Wissen erfasst. Die Ergebnisse zum Vergleich zwischen Lehrkräften und Studieren-

den, die weder Erziehungswissenschaften, Lehramt noch Psychologie studierten, und in beiden Tests ähnlich gut abschnitten wie Lehrkräfte, werfen allerdings die Frage auf, ob die Tests eher eine Art pädagogisches Allgemeinwissen abfragen, über das man nach Durchlaufen der eigenen Schullaufbahn verfügt. Auch die nicht vorhandenen Unterschiede im deklarativen PK zwischen Gymnasiallehrkräften der Physik und Diplomphysikern, die weder an der Universität tätig waren noch lehrten, könnten in diese Richtung deuten.

Zusammengenommen deuten die Ergebnisse der ersten Phase des ProwiN-Projekts (mit den genannten Einschränkungen) auf eine valide Erfassung des Professionswissens von Physiklehrkräften hin. Die prädiktive Validität der Testinstrumente wurde allerdings nicht untersucht. Dieser Aufgabe widmet sich die zweite Phase des ProwiN-Projekts.

### 5.1.3. Ziele der ProwiN-Videostudie

Das Ziel der zweiten Phase des ProwiN-Projekts ist die Analyse des Zusammenhangs zwischen dem mit den ProwiN-Testinstrumenten erfassten Professionswissens von Lehrkräften, verschiedenen Merkmalen guten Unterrichts und Unterrichtserfolg in den naturwissenschaftlichen Fächern Physik, Biologie und Chemie. Mit primärem Bezug zum pädagogischen Wissen der Lehrkräfte sollen die Merkmale Klassenführung, Einsatz variabler Unterrichtsmethoden, Art der Leistungsbeurteilung und Förderung individueller Lernprozesse untersucht werden. Entsprechende Analysen werden in der Lehr-Lern-Psychologie durchgeführt. Mit Bezug zum fachspezifischen Professionswissen der Lehrkräfte soll die Sachstruktur der vermittelten Unterrichtsinhalte, die Nutzung der Fachsprache, der Umgang mit Schülervorstellungen und Schülerfehlern und der Umgang mit Experimenten und Modellen in den beteiligten Fachdidaktiken untersucht werden. Aufgrund von Unterschieden in der Umsetzung des ProwiN-Modells in die entsprechenden Testinstrumente zur Erfassung des fachspezifischen Professionswissens in den verschiedenen Fächern werden hier allerdings unterschiedliche Schwerpunkte gesetzt. Über den Vergleich der Ergebnisse in den einzelnen Fächern soll deren Generalisierbarkeit überprüft werden.

Das ProwiN-Projekt wählt für die Untersuchung der Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg einen etwas anderen Zugang als beispielsweise die COACTIV-Studie. COACTIV war in die nationale PISA 2003/2004-Studie integriert und konnte daher auf die PISA-Leistungstestergebnisse der Lernenden am Ende des 9. und 10. Schuljahrs zugreifen, um die prädiktive Validität der COACTIV-Professionswissenstests zu überprüfen. Der Unterricht der teilnehmenden Lehrkräfte wurde nicht direkt beobachtet, sondern lediglich auf Grundlage der im Unterricht oder im Rahmen von Klassenarbeiten oder Hausaufgaben eingesetzten Aufgaben rekonstruiert. Zusätzlich wurden Schülerbefragungen zum Unterricht durchgeführt. Ein solches Vorgehen ermöglicht zwar die Untersuchung von sehr großen Stichproben (in COACTIV $N = 181$ Lehrkräfte und ihre Klassen) und damit auch den statistischen Nachweis kleiner Zusammenhänge, es hat aber auch Nachteile. Zum einen beziehen sich die PISA-Leistungstests nicht unmittelbar auf den in den Klassen stattgefundenen Unterricht, die Testinstrumente

Tabelle 5.1.
*Übersicht über die Ergebnisse von Kirschner (2013) und Lenske et al. (2015) zur Validierung der ProwiN-Professionswissenstests (Für Korrelationen oder Effektstärken gilt ein Mindestsignifikanzniveau von $p < 0.5$) (Fortsetzung auf der nächsten Seite)*

| Aspekte | Maßnahmen/Hypothesen | CK | PCK | $PK_D$ | $PK_{KP}$ |
|---|---|---|---|---|---|
| Inhalts-validität | Expertenbefragung | x | x | x[1] | x |
| | Modellbasierte Entwicklung | x | x | x | x |
| | Abgleich mit Curricular/Fachliteratur | x | x | | |
| Konstrukt-validität | Modell: 3D besser als 1D/2D | bestätigt für CK, PCK, $PK_D$ | | | n.u. |
| | Zusammenhang zw. den Dimensionen[2] | $r_{CK\text{-}PCK}$ | $r_{PCK\text{-}PK_D/PK_{KP}}$ | $r_{CK\text{-}PK_D/PK_{KP}}$ | $r_{PK_D\text{-}PK_{KP}}$ |
| | Kirschner (2013): | .45 | .27/ n.u. | .17/ n.u. | n.u. |
| | Lenske et al. (2015): | n.u. | .31/.19 | .19/ n.s. | .50[3] |
| | Korrelationen: $r_{CK\text{-}PCK} > r_{CK\text{-}PK}$ | bestätigt für $PK_D$ ($p_{1\text{-seitig}} < .001$) | | | n.u. |
| | Korrelationen: $r_{PK\text{-}PCK} > r_{CK\text{-}PK}$ | n.s. bei Kirschner (2013), $p < .05$ bei Lenske et al. (2015) | | | |

*Legende:* LK=Lehrkraft; LiV=Lehrkräfte im Vorbereitungsdienst; Gym=unterrichten am Gymnasium; GL=gymnasiales Lehramt studiert; HS= unterrichten an der Hauptschule; NW=Naturwissenschaften; Ph=Physik; +/- =Hypothese bestätigt/abgelehnt; n.u.=Hypothese nicht untersucht; n.s.=nicht signifikant

[1] Voss, Kunter & Baumert, 2011b

[2] Stichprobe Kirschner (2013): $N_{CK,PCK} = 279$, $N_{PK_D} = 186$, Stichprobe Lenske et al. (2015): $N = 171$

[3] Latente Korrelation in 2-dimensionaler konfirmatorischer Faktorenanalyse

Tabelle 5.1.
*(Fortsetzung) Übersicht über die Ergebnisse von Kirschner (2013) und Lenske et al. (2015) zur Validierung der ProwiN-Professionswissenstests (Für Korrelationen oder Effektstärken gilt ein Mindestsignifikanzniveau von $p < 0.5$)*

| Aspekte | Maßnahmen/Hypothesen | CK | PCK | $PK_D$ | $PK_{KP}$ |
|---|---|---|---|---|---|
| Kriteriums-validität | $LK_{Ph,Gym,N=216} > LK_{Ph,Nicht\text{-}Gym,N=62}$ | $d = 1.1$ | $d = 1.2$ | n.u. | n.u. |
| | $LK_{NW,Gym,N=391} < LK_{NW,HS,N=62}$[4] | | | - | - |
| | $LK_{Ph,Gym,N=216} > LK_{Andere\ Fächer,Gym,N=31}$ | $d = 2.1$ | $d = 0.9$ | | |
| | $LK_{Ph,Gym,N=149} = LK_{Andere\ Fächer,Gym,N=21}$ | | | + | + |
| | $LK_{Ph,Gym,N=216} = \text{Diplom-Physiker}_{N=22}$ | + | | | |
| | $LK_{Ph,Gym,N_{PCK/PK}=216/148} > \text{Dipl.-Phys.}_{N=22}$ | | n.s.[5] | - | n.u. |
| | $LK_{N=21} < \text{Dozenten}_{N=23}$ | | | $d = 0.77$ | |
| | $LK_{N=21} = \text{Dozenten}_{N=23}$ | | | | + |
| | $LK/LiV_{Ph,GL,N=71/37} > \text{Studierende}_{Ph,GL,N=43}$ | $d = 0.99/0.70$[6] | $d = 0.68/0.60$[7] | n.u. | n.u. |
| | $LK_{N=21} > \text{Studierende}_{\text{nicht päd.-psych. Fächer},N=31}$ | | | - | - |

*Legende:* LK=Lehrkraft; LiV=Lehrkräfte im Vorbereitungsdienst; Gym=unterrichten am Gymnasium; GL=gymnasiales Lehramt studiert; HS= unterrichten an der Hauptschule; NW=Naturwissenschaften; Ph=Physik; +/- =Hypothese bestätigt/abgelehnt; n.u.=Hypothese nicht untersucht; n.s.=nicht signifikant

[4] Kirschner et al., in Druck; Für Teilstichprobe der Physiklehrkräfte werden Unterschiede nicht signifikant ($N_{Gym/HS} = 149/23$)

[5] Unterschied wird nur für die Teilstichprobe der Diplomphysiker, die weder an einer Universität tätig sind noch lehren, signifikant (N=7)

[6] Ergebnisse zeigen sich deskriptiv (nicht-signifikant!) auch in kleiner Stichprobe von LK/LiV und Studierenden des nicht-gymnasialen Lehramts

[7] Ergebnisse gelten nicht in kleiner Stichprobe von LK/LiV und Studierenden des nicht-gymnasialen Lehramts

messen also nicht zwingend den durch den Unterricht bedingten Leistungszuwachs der Lernenden.[4] Zum anderen können keine Aussagen zur Qualität des Unterrichts gemacht werden, da dieser nicht direkt beobachtet wird. In ProwiN werden die Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg untersucht, indem pro Lehrkraft zwei Unterrichtsstunden videographiert und durch externe Beobachter analysiert werden und der Leistungszuwachs der Lernenden über eine Unterrichtseinheit zu einem bestimmten Fachthema mit eigens hierfür entwickelten Testinstrumenten erhoben wird. Aufgrund des Aufwands, der mit der Videographie des Unterrichts verbundenen ist, und aufgrund der Umstands, dass ProwiN nicht in PISA-Erhebungen integriert werden konnte und daher auf die freiwillige Teilnahme der Lehrkräfte angewiesen ist, werden allerdings kleinere Stichproben als in COACTIV untersucht (pro Fach ca. $20 - 40$ Lehrkräfte mit ihren Klassen). Mit Blick auf die in COACTIV gefundenen Effektstärken (vergl. Baumert et al., 2010, S. 161) sollten diese Stichprobengrößen dennoch ausreichend sein, um die interessierenden Zusammenhänge nachzuweisen.

Die Einordnung der vorliegenden Studie in das ProwiN-Projekt erfolgt in Abschnitt 5.4 auf Seite 74. Im Folgenden soll zunächst dargelegt werden, welchen Ansatz die vorliegende Arbeit für die Überprüfung der prädiktiven Validität der ProwiN-Testinstrumente für gutes und erfolgreiches Unterrichten wählt. Hierfür wird zunächst begründet welche Zielkriterien erfolgreichen Unterrichts betrachtet werden. Anschließend wird ausgehend von diesen Zielkriterien und ausgehend von den in der Physik entwickelten Professionswissenstests abgeleitet, welches Merkmal guten Unterrichts in dieser Arbeit untersucht werden soll.

## 5.2. Auswahl der Kriterien für erfolgreiches Unterrichten: Fachwissenserwerb und situationales Interesse

In Abschnitt 3.2 auf Seite 24 wurden bereits mögliche Zielkriterien für erfolgreiches Unterrichten vorgestellt. Nach Helmke (2009, S. 41) ist „[die] wichtigste Voraussetzung für kumulative und anspruchsvolle Lernprozesse [...] eine solide und gut organisierte Wissensbasis, das heißt ein in sich vernetztes, in verschiedenen Situationen erprobtes und flexibel anpassbares Wissen (‚intelligentes Wissen'), das Fakten, Konzepte, Theorien und Methoden gleichermaßen umfasst." Als ein Kriterium für erfolgreiches Unterrichten werden in dieser Studie daher Schülerleistungen im physikalischen Fachwissen betrachtet. Ein weiterer Grund für die Auswahl der Schülerleistung als Zielkriterium für erfolgreiches Unterrichten liegt in der Anschlussfähigkeit an die in Abschnitt 4.3 auf Seite 41 beschriebenen Studien zur Überprüfung der prädiktiven Validität von Professionswissenstests.

Zu beachten ist an dieser Stelle, dass nach dem Angebots-Nutzungsmodell von Helmke (2009) die individuellen Voraussetzungen der Lernenden, wie z. B.

[4] Lediglich die Hälfte der Testaufgaben bezieht sich auf Inhalte der Jahrgangsstufe 10 (vergl. Ehmke et al., 2006, S. 69).

Vorwissen, Sprache, Intelligenz, Lernmotivation oder Anstrengungsbereitschaft, sowie die Unterrichtszeit oder die aktive Lernzeit im Unterricht (Time on Task) einen Einfluss auf Schülerleistungen haben (vergl. Abbildung 2.1 auf Seite 8). In der vorliegenden Arbeit sollen Variablen, die auf die Schülerleistung wirken, aber ihrerseits nicht durch das professionellen Wissen der Lehrkraft beeinflusst werden können, kontrolliert werden. Während die Lehrkraft die Lernmotivation oder die Anstrengungsbereitschaft der Lernenden durchaus beeinflussen könnte oder beispielsweise auf Basis ihres pädagogischen Wissens versuchen könnte über eine effektive Klassenführung die aktive Lernzeit im Unterricht zu erhöhen, gilt dies nicht für das Vorwissen, die kognitiven Fähigkeiten oder den sprachlichen Hintergrund der Lernenden und nur im begrenztem Ausmaß für die tatsächliche Unterrichtszeit (also die Anzahl tatsächlich stattgefundener Unterrichtsstunden). Letztere kann z. B. durch schulinterne Curricula vorgeschrieben werden oder durch Unterrichtsausfall bedingt sein. Das Fachwissen der Schülerinnen und Schüler am Ende einer Unterrichtseinheit hängt davon ab, über welches Wissen bezüglich des Lerngegenstands die Lernenden schon vor Beginn des Unterrichts verfügten. Einen bedeutsamen Einfluss auf Schülerleistung haben auch die kognitiven Fähigkeiten der Lernenden (Fischer et al., 2014b, S. 19; Schroeders et al., 2013, S. 341-342) oder die von den Lernenden zuhause gesprochene Sprache (Pöhlmann, Haag & Stanat, 2013, S. 324).[5] Auch konnten in Schulleistungsstudien Geschlechterdifferenzen im physikalischen Fachwissen festgestellt werden (vergl. z. B. Schroeders, Penk, Jansen & Pant, 2013, S. 264) – Jungen schnitten (nach Kontrolle der Schulform) signifikant besser ab als Mädchen. Darüber hinaus hängt das Fachwissen von Schülerinnen und Schülern am Ende einer Unterrichtseinheit von dem Zeitraum für Lerngelegenheiten und damit von der tatsächlichen Unterrichtszeit ab (vergl. Helmke, 2009, S. 81). Als Kontrollvariablen werden daher das Vorwissen, die kognitiven Fähigkeiten, das Geschlecht und die von den Lernenden zuhause gesprochene Sprache sowie die Unterrichtszeit ausgewählt.

Unterricht kann nur dann erfolgreich sein, wenn die Lernenden motiviert sind das durch die Lehrkraft bereitgestellte Lehrangebot auch zu nutzen. Grundsätzlich unterscheidet man zwischen *intrinsischer* und *extrinsischer* Lernmotivation. Die intrinsische Motivation bezeichnet „die Absicht, eine bestimmte Lernhandlung um ihrer selbst willen durchzuführen, weil diese z.B. als interessant, spannend, herausfordernd usw. erscheint“ (Schiefele & Schreyer, 1994, S. 1-2). Dagegen bezeichnet die extrinsische Motivation die Absicht mit der Lernhandlung bestimmte Ziele zu verfolgen, die außerhalb der Lernhandlung als solcher liegen. Im Vordergrund steht dabei das Herbeiführen positiver oder die Vermeidung negativer Folgen, also z. B. eine Anerkennung durch die Lehrkraft erfahren, eigenen oder fremden Leistungsansprüchen genügen oder Ärger mit den Eltern vermeiden (Schiefele & Schreyer, 1994, S. 2). Krapp (2003, S. 97) weist allerdings darauf hin, dass intrinsische und extrinsische Motivation manchmal nicht eindeutig voneinander trennbar sind. Insbesondere die intrinsische Motivation gilt als „unerlässliche Voraussetzung des

[5] Die zuhause gesprochene Sprache stellt auch einen Indikator für den Migrationshintergrund der Lernenden dar (Quesel, Möser & Husfeldt, 2014, S. 296).

Wissenserwerbs“ (Edelmann, 2003, S. 32). Der Einfluss der intrinsischen Motivation auf Lernerfolg ist zudem auch empirisch belegt (vergl. Schiefele & Schreyer, 1994).

Als „entscheidende Grundlage für das Auftreten intrinsischer Motivation“ wird das Interesse angesehen, also „die besondere Beziehung einer Person zu einem Lerngegenstand“ (Krapp, 2003, S. 96). Interessiert man sich für einen Lerngegenstand, kann das Lernen an sich als persönlicher Gewinn empfunden werden, was einer intrinsischen Lernmotivation entsprechen würde (Krapp, 2003, S. 97). So lassen sich auch deutliche Zusammenhänge zwischen den schulfachbezogenen Interessen und Leistungen von Schülerinnen und Schülern nachweisen (vergl. hierzu die Meta-Analyse von Schiefele, Krapp & Schreyer, 1993).

Das *individuelle* Fachinteresse einer Person stellt ein relativ stabiles Merkmal dar (Schiefele, 2008, S. 46). Es wird allerdings angenommen, dass individuelles Fachinteresse durch wiederholtes Auftreten von *situationalem* Interesse bezüglich eines Fachgegenstands manifestiert werden kann (Spoden & Geller, 2014, S. 50). Auch im Rahmenmodell der Interessengenese von Krapp (1998, S. 191) wird situationales Interesse als Ausgangspunkt für individuelles Interesse modelliert.[6] Das *situationale* Interesse bezeichnet „den durch äußere Umstände (z. B. einen spannenden Vortrag) hervorgerufenen Zustand des Interessiertseins, der u. a. durch eine erhöhte Aufmerksamkeit gekennzeichnet ist“ (Schiefele, 2008, S. 46). Nach Baumert und Kunter (2006, S. 476) „tritt die Lehrkraft als didaktischer Mittler zwischen Sachverhalt und Schüler auf, wenn es ihr gelingt, situationales Interesse und Aufmerksamkeit zu wecken“. So kann das Erzeugen situationalen Interesses im Unterricht das Auftreten intrinsischer Lernmotivation bei den Lernenden begünstigen (Schiefele, 2008, S. 46). Situationales Interesse kann also einerseits die Nutzung des Lehrangebots durch die Lernenden und anderseits die Ausbildung von Fachinteresse begünstigen. Krapp (1998, S. 196) vermutet zudem, dass „auch ein zeitlich begrenztes situationales Interesse eine dauerhafte Bildungswirkung“ haben kann, da vieles dafür spreche, dass „ein mit positiven Erlebensqualitäten erworbenes Wissen nach einer längeren Latenzphase nicht nur besser erinnert, sondern auch mit höherer Wahrscheinlichkeit erneut aktiviert und gegebenenfalls selbständig erweitert wird“. Als weiteres Kriterium für erfolgreichen Unterricht soll daher das situationale Interesse der Lernenden im Unterricht betrachtet werden.

## 5.3. Auswahl eines Merkmals guten Unterrichts: Kognitive Aktivierung

Diese Studie soll sich nicht auf die Untersuchung des Zusammenhangs zwischen Professionswissen der Lehrkraft und Unterrichtserfolg beschränken, da dieser Zusammenhang durch zahlreiche Variablen und nicht zuletzt durch das Unterrichtsgeschehen mediiert oder moderiert werden kann. Zudem können nicht gefundene Zusammenhänge zwischen Professionswissen und den betrachteten Zielkriterien von Unterricht keine Aufschlüsse darüber liefern, ob ein Zusammenhang zwischen dem Wissen und Handeln der Lehrkraft im Unterricht besteht, sich dieser aber nicht bis

[6] Hierbei handelt es sich allerdings nicht um empirisch abgesicherte Erkenntnisse.

auf die Zielkriterien auswirkt, oder aber ein Zusammenhang zwischen dem in einem Testinstrument abfragbaren und damit explizierbaren (und nicht rein implizitem) Wissen einer Lehrkraft und ihrem Handeln als solcher schon nicht nachweisbar ist. Auch erscheint eine reine Fokussierung auf Variablen des Unterrichtserfolgs nicht angebracht, da dieser nach dem Angebots-Nutzungsmodell nur bis zu einem gewissen Grad durch die Lehrkraft realisiert werden kann. Darüber hinaus ist erfolgreiches Unterrichten nicht mit qualitätvollem Unterrichten gleichzusetzen, wie bereits in Abschnitt 3.3 auf Seite 25 erläutert wurde.

Welches Merkmal „guten“ Unterrichts soll nun aber betrachtet werden? Dieser Frage kann man sich von zwei Seiten her nähern: ausgehend von den betrachteten Zielkriterien erfolgreichen Unterrichtens oder ausgehend von dem mit dem ProwiN-Testinstrumenten abgefragten Wissen, zu dem ein Zusammenhang hergestellt werden soll. Unter Berücksichtigung der Zielkriterien sollte das betrachtete Merkmal der Unterrichtsqualität einen Einfluss auf den Fachwissenserwerb der Lernenden haben und zudem das Auftreten situationalen Interesses bei den Lernenden begünstigen. Ausgehend von den Testinstrumenten zum Professionswissen sollte zum einen theoretisch abgeleitet werden können, warum Zusammenhänge zwischen dem betrachteten Merkmal und dem Fachwissen der Lehrkräfte zu erwarten wären. Zum anderen sollte ein Merkmal ausgewählt werden, für dessen Realisierung insbesondere die im ProwiN-Projekt betrachteten PCK-Facetten (Wissen über Schülervorstellungen, Experimente und Konzepte) eine Rolle spielen. Zusammenhänge zum pädagogischen Wissen (hier: Wissen über Klassenführung, Unterrichtsmethoden, individuelle Lernprozesse und Leistungsbeurteilung) stehen nicht im Fokus dieser Arbeit, sollen aber mit untersucht werden.

Um diesen Anforderungen gerecht zu werden, wird als Merkmal guten Unterrichts die kognitive Aktivierung betrachtet, was im Folgenden ausführlich begründet werden soll. Ein kognitiv aktivierender Unterricht wird in dieser Arbeit in Anlehnung an Hugener (2008), Kunter (2005), Rakoczy und Pauli (2006), Vogelsang (2014) und Widodo und Duit (2004) durch folgende Lehrerhandlungen beschrieben (vergl. auch Abschnitt 3.3.3.1 auf Seite 29):

- **Bewusstmachen des Lernstatus** z. B. Ausblick/Rückblick geben auf Inhalte, die in einer Unterrichtsstunde thematisiert werden bzw. wurden, Verbindungen zu früher Gelerntem aufzeigen
- **Exploration des Vorwissens und der Vorstellungen** z. B. Anregung das Unterrichtsthema nach ihrem Verständnis zu erläutern, nach Vorwissen fragen, ohne auf bestimmte Antwort abzuzielen oder Wertungen vorzunehmen
- **Exploration der Denkweisen** z. B. Einfordern von Begründungen, Nachfragen wie Lernende zu ihren Antworten gelangen, Anregung Sachverhalte in eigenen Worten zu erläutern
- **Evolutionärer Umgang mit Schülervorstellungen** z. B. Aufgreifen und Weiterverwenden von Vorstellungen der Lernenden, Erzeugen kognitiver Konflikte, genetisch-sokratisches Vorgehen, indem Lernende auch mal in

die Irre gelaufen werden lassen, Aufforderungen auf aktuellem Wissenstand aufbauend zu argumentieren und Schlussfolgerungen zu ziehen

- **Einnehmen einer Mediatorfunktion** z. B. Beiträge der Lernenden aufeinander beziehen, Nachfragen bei missverständlichen oder unklaren Äußerungen, Einfordern von Begründungen, Unterstützung bei der Ausformulierung von Ideen, Lernende durch eigene Beiträge aktiv am Unterricht beteiligen
- **Unterlassen von Handlungen, die auf ein rezeptives Lernverständnis der Lehrperson hinweisen** z. B. kleinschrittiges Frageverhalten, rezeptartige Aufgabenstellungen, Betonen von genauem Auswendiglernen fachlicher Inhalte
- **Schaffung herausfordernder Lerngelegenheiten** z. B. Aufgaben- und Fragestellungen, die zum Nachdenken anregen, mehr als Ja- oder Nein-Antworten bedürfen, nicht nur auswendiggelerntes Wissen abfragen und Vergleichs- und Analyseprozesse erfordern, Erfragen von Hypothesen in Experimentiersituationen

An dieser Stelle sei darauf hingewiesen, dass eine derartige Beschreibung von kognitiv aktivierendem Unterricht mit gewissen Einschränkungen verbunden ist: So scheinen sich einige der genannten Merkmale lediglich zur Beschreibung kognitiv aktivierenden Unterrichts in *Einführungs*stunden zu eignen (Praetorius, Pauli, Reusser, Rakoczy & Klieme, 2014, S. 9). Auf Grundlage von fünf videographierten Unterrichtsstunden von $N = 38$ Mathematiklehrkräften aus der Pythagoras-Studie konnten Praetorius et al. (2014) zeigen, dass die kognitive Aktivierung (operationalisiert über die Exploration der Denkweisen, das rezeptive Lernverständnis der Lehrkraft und herausfordernde Lerngelegenheiten) zwischen unterschiedlichen Stundenarten und in Unterrichtsstunden zu verschiedenen Themen erheblich variierte.[7]

### 5.3.1. Kognitive Aktivierung und Fachwissen der Lernenden

Ein Fachwissenserwerb auf Schülerseite setzt eine Nutzung des Lehrangebotes durch die Schülerinnen und Schüler und damit, im Sinne eines konstruktivistischen Lernverständnisses, eine aktive Auseinandersetzung der Lernenden mit dem Lerngegenstand voraus. Die Unterstützung solch einer kognitiv aktiven Auseinandersetzung der Lernenden mit dem Lerngegenstand durch die Lehrkraft wird durch Merkmale einer kognitiv aktivierenden Unterrichtsgestaltung beschrieben. So adressiert die kognitive Aktivierung viele Aspekte des aktuellen Lernverständnisses. Dieses wird von Hugener (2008) wie folgt zusammengefasst:

[7] Unter der Annahme, dass für kognitiv aktivierendes Unterrichten fachspezifisches Professionswissen notwendig ist, ist die Themenabhängigkeit der kognitiven Aktivierung wenig überraschend – schließlich kann auch das CK und PCK von Lehrkräften bezüglich unterschiedlicher Themen variieren.

> Der Aufbau von Wissensstrukturen erfolgt an einen Inhalt gebunden, in einem bestimmten Kontext (situativ) und in der handelnden Auseinandersetzung mit einem Lerngegenstand (aktiv), während der Strukturen aufgebaut werden (konstruktiv). Dem Vorwissen kommt dabei eine vorrangige Bedeutung zu, da neue Wissensstrukturen verknüpfend auf bestehende aufgebaut werden (kumulativ). Lernprozesse sind am erfolgreichsten, wenn der Schüler oder die Schülerin das Ziel kennt und darauf hinarbeitet (zielgerichtet), das eigene Vorgehen überwacht und steuert (selbstreguliert), sich mit anderen austauschen kann (kooperativ und interaktiv) und von Experten unterstützt und begleitet wird (fremdgesteuert). Lernvoraussetzungen, Lernprozesse und Lernwirkungen sind bei jeder und jedem Lernenden individuell verschieden (individuell). Nach diesem Verständnis von Wissenserwerb ist Lernen ein aktiver, konstruktiver, kumulativer und zielorientierter Prozess, der ko-konstruktiv in Lerngemeinschaften und in bestimmten Kontexten abläuft und metakognitiv gesteuert wird. (S. 21-22)

Eine kognitiv aktivierende Unterrichtsgestaltung zeichnet sich durch die Schaffung herausfordernder Lerngelegenheiten und die Vermeidung rezeptartiger Aufgabenstellungen aus, was die kognitive Aktivität der Lernenden fördern kann. Die besondere Berücksichtigung des Vorwissens der Lernenden und dessen Aktivierung schafft zudem eine Basis für kumulative Lernprozesse und berücksichtigt die Individualität der Lernenden. Das Bewusstmachen des Lernstatus gegenüber den Lernenden ermöglicht ein zielgerichtetes Lernen und die Vernetzung von Wissen. Indem die Lehrkraft die Rolle eines Mediators im Unterricht einnimmt, wird außerdem die soziale Ko-Konstruktion von Wissen möglich.

### 5.3.2. Kognitive Aktivierung und situationales Interesse der Lernenden

Merkmale eines kognitiv aktivierenden Unterrichts, wie z. B. die Schaffung herausfordernder Lerngelegenheiten oder ein genetisch-sokratisches Vorgehen der Lehrkraft (als Aspekt des evolutionären Umgangs mit Schülervorstellungen) begünstigen das Autonomie- und Kompetenzerleben der Lernenden (Kunter, 2005, S. 140). Nach der Selbstbestimmungstheorie von Deci und Ryan (1993) ist die Befriedigung des angeborenen psychologischen Bedürfnisses nach Autonomie, Kompetenz und sozialer Eingebundenheit ausschlaggebend für die Entwicklung intrinsischer Motivation. Da situationales Interesse und intrinsische Motivation in Beziehung zueinander stehen (vergl. Abschnitt 3.2 auf Seite 24), kann ein Zusammenhang von kognitiv aktivierendem Unterricht und situationalem Interesse der Lernenden angenommen werden. Zu vermuten ist auch, dass neben der Schaffung herausfordernder Lerngelegenheiten und einem evolutionären Umgang mit Schülervorstellungen auch weitere Merkmale eines kognitiv aktivierenden Unterrichts Einfluss auf das situationale Interesse der Lernenden haben. Ein Unterricht, in dem die Lernenden zu einem aktiven Diskurs im Klassenraum ermuntert werden, in dem Diskussionen angeleitet werden und die Vorstellungen der Lernenden nicht ignoriert, sondern einbezogen

werden, wird wahrscheinlich als interessanter empfunden als ein Unterricht, in dem die Lernenden lediglich kleinschrittige Arbeitsaufträge erledigen müssen oder sich mit Fragestellungen beschäftigen, die kein wirkliches Mitdenken erfordern.

In der Tat erwies sich kognitive Aktivierung, so wie sie in dieser Arbeit operationalisiert wurde, unter Kontrolle von intrinsischer Motivation und Fachinteresse als signifikanter Prädiktor für das situationale Interesse von Lernenden im Biologieunterricht (Förtsch, Werner, Dorfner, von Kotzebue & Neuhaus, 2015). Weitere Studien, die diesen Zusammenhang untersuchen, gibt es bisher aber nicht. Seidel, Rimmele und Prenzel (2003, S. 158) konnten allerdings einen negativen Einfluss eines enggeführten Klassengesprächs auf intrinsische Motivation und Interessiertheit nachweisen (die Operationalisierung der Autoren beinhaltet Aspekte der Merkmale herausfordernde Lerngelegenheiten, Lehrkraft als Mediator und rezeptives Lernverständnis der Lehrkraft).[8]

Zusammenfassend kann angenommen werden, dass eine kognitiv aktivierende Unterrichtsgestaltung sowohl einen Einfluss auf den Fachwissenserwerb der Lernenden als auch auf deren situationales Interesse hat. Ausgehend von den für diese Studie ausgewählten Zielkriterien von Unterricht scheint die Betrachtung der kognitiven Aktivierung als Merkmal guten Unterrichts also zielführend zu sein. Über einen Zusammenhang zwischen kognitiver Aktivierung und den betrachteten Zielkriterien von Unterricht kann allerdings lediglich sichergestellt werden, dass nicht nur ein Merkmal guten Unterrichts betrachtet wird, sondern dass dieses Merkmal im Sinne von Fenstermacher und Richardson (2005) auch als Merkmal der Unterrichtsqualität angesehen werden kann. Das Ziel dieser Arbeit ist die Untersuchung der prädiktiven Validität der ProwiN-Professionswissenstests für Physiklehrkräfte und damit auch die Untersuchung des Zusammenhangs zwischen dem Professionswissen der Lehrkräfte und der Qualität ihres Unterrichtens. Weitaus wichtiger ist daher, dass das mit den ProwiN-Instrumenten erfasste Wissen als notwendige Wissensbasis angenommen werden kann, um eine kognitiv aktivierende Unterrichtsgestaltung zu realisieren.

### 5.3.3. CK und kognitive Aktivierung

Um einen kognitiv aktivierenden Physikunterricht durchzuführen, muss eine Lehrkraft zweifelsohne über physikalisches Fachwissen verfügen. Ohne Fachwissen kann die Lehrkraft keine herausfordernden Lerngelegenheiten schaffen: Das Abfragen von auswendig gelerntem Wissen oder die Beantwortung von Fragestellungen, die lediglich Ja- oder Nein-Antworten erfordern, könnte wahrscheinlich auch eine Lehrkraft mit Lücken im Fachwissen realisieren. Um jedoch Fragen und Aufgaben zu stellen, die zum Nachdenken anregen, muss die Lehrkraft selbst über ein gewisses Verständnis der Fachinhalte verfügen. Auch die Exploration der Denkweisen setzt ein gewisses Fachwissen der Lehrkraft voraus. Fordert die Lehrkraft die Lernenden dazu auf, ihre Antworten zu begründen oder Sachverhalte in eigenen

[8] Untersucht wurden hier allerdings nur der Unterricht und die Schulklassen von einer kleinen Stichprobe von 13 Lehrkräften.

Worten zu erläutern, kann sie nur dann Rückschlüsse auf die Denkweisen der Lernenden ziehen, wenn sie Unterschiede zu den wissenschaftlichen Denkweisen erkennt, mit denen sie demnach selbst vertraut sein muss. Eine Lehrkraft mit unzureichendem Fachwissen würde im Unterricht daher vermutlich wenig Anstrengung unternehmen, die Denkweisen der Lernenden nachzuvollziehen, da sie mit den gewonnen Informationen nicht weiterarbeiten kann. Auch ein evolutionärer Umgang mit Schülervorstellungen erfordert Fachwissen aufseiten der Lehrkraft. Beispielsweise können nur dann kognitive Konflikte erzeugt werden, wenn die Lehrkraft eine inkorrekte Schülervorstellung als solche erkennt und zudem in der Lage ist, den Lernenden Ungereimtheiten in ihren Vorstellungen aufzuzeigen. Auch ein Bewusstmachen des Lernstatus erfordert in einem gewissen Maße Fachwissen der Lehrkraft, beispielsweise um Ausblick auf Inhalte zu geben, die sich aus dem aktuellen Unterricht ergeben oder Verbindungen zu bereits Gelerntem aufzeigen zu können. Lehrkräfte mit niedrigem Fachwissen könnten zudem dazu neigen, Unterricht eher rezeptiv orientiert zu gestalten: Rezeptartige Aufgabenstellungen (und damit enge Vorgaben wie Aufgaben zu bearbeiten sind), das Betonen des Auswendiglernens von Fachinhalten und kleinschrittige Fragestellungen können das Risiko einer Lehrkraft reduzieren sich in Gesprächssituationen wiederzufinden, denen sie sich fachlich nicht gewachsen fühlt. Ein negativer Zusammenhang zwischen dem Fachwissen von Mathematiklehrkräften und einem rezeptiven Lehr-Lernverständnis (erhoben durch Lehrerfragebögen) konnte sowohl von Krauss, Neubrand et al. (2008, S. 247) im Rahmen der COACTIV-Studie als auch von Kessler (2011, S. 133) nachgewiesen werden. Aussagen über den kausalen Zusammenhang zwischen den beiden Variablen konnten allerdings nicht getroffen werden. Krauss, Neubrand et al. (2008, S. 247) konnten außerdem auch einen negativen Zusammenhang zwischen dem Fachwissen der Lehrkräfte und deren Selbstberichten über die Verwendung kleinschrittiger Anleitungen im Unterricht nachweisen. Positiv korrelierte das Fachwissen hingegen mit einer konstruktivistischen Sichtweise von Unterricht und dem Insistieren auf Begründungen und Erklärungen im Unterricht.

Der Einfluss des Fachwissens auf die in dieser Studie betrachteten Merkmale eines kognitiv aktivierenden Unterrichts wurden lediglich in der Studie von Vogelsang (2014) untersucht. Hier fanden sich keine Zusammenhänge. Da das Fachwissen in dieser Studie allerdings mit allen weiteren betrachteten Merkmalen guten Unterrichts ebenfalls gar nicht oder sogar signifikant negativ korrelierte, geht Vogelsang (2014, S. 511) davon aus, dass „die im Paderborner Instrument erfassten Konstrukte keine Handlungsressourcen zur Gestaltung ‚angemessenen' Physikunterrichts bilden". In COACTIV wurde der Einfluss von Fachwissen auf den Einsatz kognitiv herausfordernder Aufgaben im Unterricht untersucht. Der Unterricht wurde allerdings nicht direkt beobachtet. Stattdessen wurde das kognitive Potenzial aller im Unterricht eingesetzter oder im Rahmen von Klassenarbeiten oder Hausaufgaben gestellter Aufgaben hinsichtlich des Typus mathematischen Arbeitens, des Niveaus der verlangten mathematischen Argumentation und hinsichtlich der innermathematischen Übersetzungsleistungen sowie deren Passung zum curricularen Niveau der untersuchten Jahrgangsstufe betrachtet (Baumert &

Kunter, 2011, S. 173). Es zeigte sich ein Einfluss des Fachwissens auf die curriculare Passung der Aufgaben, nicht aber auf das kognitive Potenzial der Aufgaben.[9]

Auf Grundlage der angeführten Überlegungen wird davon ausgegangen, dass grundsätzlich ein Zusammenhang zwischen dem Fachwissen von Physiklehrkräften und kognitiver Aktivierung bestehen sollte. Lehrkräfte, die über ein hohes Fachwissen verfügen, sollten sich daher durch eine kognitiv aktivierende Unterrichtsgestaltung auszeichnen.

### 5.3.4. PCK und kognitive Aktivierung

Der ProwiN-PCK-Test erfasst hauptsächlich Wissen über Schülervorstellungen. Physiklehrkräfte, die populäre Schülerfehlvorstellungen kennen, können gezielt versuchen, herauszufinden, ob auch ihre eigenen Schülerinnen und Schüler derartige Fehlvorstellungen haben. Auch kognitive Konflikte können nur dann erzeugt werden, wenn die Lehrkraft potentielle Schülerfehlvorstellungen erkennt. Darüber hinaus ist anzunehmen, dass Lehrkräfte, die über Wissen über Schülervorstellungen verfügen, zum einen eher versuchen im Unterricht an diese anzuknüpfen und zum anderen versuchen diesen auch auf den Grund zu gehen, indem sie die Denkweisen der Lernenden explorieren. Auch für das Einnehmen einer Mediatorfunktion im Unterricht kann das Wissen über Schülervorstellungen hilfreich sein, beispielsweise um Schüleräußerungen zu identifizieren, die missverständlich für andere Lernende sein könnten. Die Facette Wissen über Schülervorstellungen umfasst auch Wissen darüber, welche Darstellungsformen und Repräsentationen die Verfestigung inkorrekter Vorstellungen von Schülerinnen und Schülern noch begünstigen. Die Facette Wissen über Experimente umfasst Wissen über die fachdidaktisch angemessene Gestaltung von Experimenten. Beides kann die Lehrkraft beispielsweise anwenden, wenn sie im Anschluss an eine Exploration der Denkweisen der Lernenden versucht angemessene Aktivitäten oder besser an das Verständnis der Lernenden anknüpfende Repräsentationsformen für den weiteren Unterrichtsverlauf auszuwählen. Anzunehmen ist außerdem, dass Lehrkräfte mit Wissen über Schülervorstellungen um die Bedeutung dieser Vorstellungen für verständnisvolles Lernen wissen und daher weniger dazu neigen ihre Schülerinnen und Schüler physikalische Konzepte lediglich auswendig lernen zu lassen oder rezeptartige Aufgabenstellungen zu formulieren, die nicht auf eine aktive Auseinandersetzung der Lernenden mit dem Lerngegenstand abzielen. In COACTIV konnten Zusammenhänge zwischen dem fachdidaktischen Wissen von Mathematiklehrkräften und einer rezeptiven Ansicht von Lehren und Lernen sowie deren Selbstberichten über die Verwendung kleinschrittiger Anleitungen im Unterricht empirisch nachgewiesen werden (Krauss,

[9] Auch in der Arbeit von Kessler (2011) wurde der Zusammenhang zwischen Fachwissen und kognitiver Aktivierung untersucht. Als Merkmal eines kognitiv aktivierenden Unterrichts wurde hier allerdings ausschließlich der von den Lernenden wahrgenommene konstruktive Umgang mit Schülerfehlern betrachtet, der eigentlich eher dem Konstrukt der konstruktiven Unterstützung zugeordnet werden kann. Die Ergebnisse zeigen negative Zusammenhänge zwischen Fachwissen und dem Umgang mit Schülerfehlern.

Neubrand et al., 2008, S. 247).[10] Des Weiteren wird in ProwiN Wissen über physikalisch angemessene Kriterien für die Leistungsbeurteilung und Wissen darüber, welche verschiedenen Funktionen Experimente im Unterricht einnehmen können, erfasst. Von diesem Wissen kann kein direkter Bezug zu Merkmalen eines kognitiv aktivierenden Unterrichts abgeleitet werden.[11]

Es ist allerdings nicht ausreichend, lediglich auf Basis des konkret im ProwiN-Testinstrument erfassten Wissens für einen Zusammenhang zwischen dem fachdidaktischen Wissen von Lehrkräften und kognitiver Aktivierung zu argumentieren. Schließlich geht man davon aus, dass die Testergebnisse valide Indikatoren für das *Konstrukt* PCK bilden. Grundsätzlich sollte das fachdidaktische Wissen einer Lehrkraft notwendig dafür sein, herausfordernde Lerngelegenheiten zu gestalten. Fachwissen allein ist hierfür nicht ausreichend. Die Lehrkraft muss wissen, wie sie Fachinhalte didaktisch aufbereiten muss und in welcher Tiefe sie physikalische Konzepte mit ihren Schülerinnen und Schülern erarbeiten kann, um diese weder zu unter- noch zu überfordern. Sofern das fachdidaktische Wissen von Lehrkräften durch das in ProwiN abgefragte Wissen hinreichend gut repräsentiert wird, sollten also auch hier Zusammenhänge bestehen. Ähnliches gilt für das Bewusstmachen des Lernstatus. Zeigt die Lehrkraft beispielsweise Verbindungen zu früher Gelerntem oder neu zu Lernendem auf, muss sie einerseits über curriculares Wissen verfügen und andererseits einschätzen können, welche Verknüpfungen von den Lernenden überhaupt nachvollzogen werden können und damit die Vernetzung von Wissen erst möglich machen.

In COACTIV konnte gezeigt werden, dass Zusammenhänge zwischen dem fachdidaktischen Wissen von Mathematiklehrkräften und den Lernleistungen von Schülerinnen und Schülern vollständig über das kognitive Potenzial der im Unterricht eingesetzten Aufgaben und deren curriculares Niveau mediiert wurden (Baumert & Kunter, 2011, S. 180). Die kognitive Aktivierung im Unterricht war wesentlich stärker durch das PCK der Lehrkräfte beeinflusst als durch deren CK (vergl. Baumert & Kunter, 2011, S. 182-183). Auch im QuiP-Projekt konnte ein Zusammenhang zwischen dem PCK von Physiklehrkräften und der Passung zwischen den Komplexitätsniveaus von Lehrerfragen und Schülerantworten nachgewiesen werden. Letzteres kann als weiterer Hinweis für die Bedeutung von fachdidaktischem Wissen für die Schaffung herausfordernder Lerngelegenheiten angesehen werden (Ergönenç et al., 2014, S. 155).[12] In der Studie von Vogelsang (2014) konnte für das fachdidaktische Wissen kein Einfluss auf die kognitive Aktivierung im Unterricht nachgewiesen werden. Wie das Fachwissen korrelierte das fachdidaktische Wissen in dieser Studie allerdings, sofern es überhaupt korrelierte, signifikant negativ mit weiteren betrachteten Merkmalen guten Unterrichts, was gegen die Handlungsrelevanz des mit dem Paderborner Instrument erhobenen Wissens spricht.

[10] In COACTIV wurde allerdings das Wissen über Erklären und Repräsentieren, Wissen über typische Schülerfehler und -schwierigkeiten und Wissen über das Potenzial für multiple Lösungsansätze von Mathematikaufgaben erfasst (Krauss, Neubrand et al., 2008, S. 234-237).

[11] Hierbei handelt es sich allerdings auch lediglich um jeweils eine Aufgabe im Testinstrument.

[12] Die Passung zwischen den Komplexitätsniveaus, über die kognitive Aktivierung in QuiP operationalisiert wurde, hing allerdings nicht mit Schülerleistung zusammen.

Zusammenfassend ist anzunehmen, dass sich insbesondere das fachdidaktische Wissen einer Lehrkraft in einer kognitiv aktivierenden Unterrichtsgestaltung widerspiegeln sollte. Ausgehend von den Ergebnissen der COACTIV-Studie ist anzunehmen, dass fachdidaktisches Wissen und kognitive Aktivierung stärker zusammenhängen als Fachwissen und kognitive Aktivierung.

### 5.3.5. PK und kognitive Aktivierung

Das pädagogische Wissen spielt sicherlich eine größere Rolle für die allgemein-pädagogischen Merkmale guten Unterrichts (z. B. Klassenführung) als für die kognitive Aktivierung. So lässt sich das in ProwiN erhobene Wissen über Klassenführung, Unterrichtsmethoden, individuelle Lernprozesse und Leistungsbeurteilung nicht direkt in Bezug zu den Merkmalen eines kognitiv aktivierenden Unterrichts setzen. Wie bereits erwähnt, scheint eine gute Klassenführung allerdings eine notwendige Voraussetzung für kognitive Aktivierung zu sein (Klieme et al., 2001, S. 53). Das pädagogische Wissen kann daher als notwendig dafür angenommen werden, um die Grundvoraussetzungen für einen kognitiv aktivierenden Unterricht zu schaffen. Löst man sich auch an dieser Stelle von dem konkret in den ProwiN-Testinstrumenten erhobenem Wissen, kann grundsätzlich ein Zusammenhang zwischen pädagogischem Wissen und einer kognitiv aktivierenden Unterrichtsgestaltung angenommen werden. Lehrkräfte, die über allgemein-pädagogisches Wissen über Lehren und Lernen verfügen, sollten sich der Bedeutung aller hier betrachteten Merkmale einer kognitiv aktivierenden Unterrichtsgestaltung für das initiieren und unterstützen von Lernprozessen bewusst sein. Um den Unterricht wirklich kognitiv aktivierend zu gestalten, benötigt eine Lehrkraft allerdings fachspezifisches Professionswissen.

Vogelsang (2014, S. xlv im Anhang) konnte signifikante Zusammenhänge zwischen dem pädagogischen Wissen von angehenden Physiklehrkräften und dem Bewusstmachen des Lernstatus, der Exploration der Denkweisen und dem Einnehmen einer Mediatorfunktion nachweisen. Im Rahmen von COACTIV-R konnten keine signifikanten Zusammenhänge zwischen dem PK von angehenden Lehrkräften und kognitiver Aktivierung gezeigt werden (Voss et al., 2014). Die kognitive Aktivierung wurde in dieser Studie allerdings darüber operationalisiert, inwieweit Aufgaben selbstständig bearbeitet und Lösungswege begründet werden sollten und wie sehr unterschiedliche Schülerlösungen diskutiert wurden. Ein Einfluss des pädagogischen Wissens ist hier nicht unbedingt zu erwarten. Außerdem wurde das PK der Probanden zu Beginn des Referendariats erhoben, ihr Unterricht aber durch Schülerinnen und Schüler eingeschätzt, die diese zwei Jahre später unterrichteten.

Auf Grundlage dieser Überlegungen und der vorhandenen Befunde kann zwar vermutet werden, dass das pädagogische Wissen von Lehrkräften Einfluss auf die kognitive Aktivierung im Unterricht hat, dieser sollte aber (falls überhaupt vorhanden) geringer als der Einfluss des fachdidaktischen Wissens oder des Fachwissens sein.

## 5.4. Einordnung der vorliegenden Studie in das ProwiN-Projekt

Die vorliegende Studie ist eins von zwei Dissertationsprojekten, die die prädiktive Validität der ProwiN-Professionswissenstests für Physiklehrkräfte in Bezug auf verschiedene Merkmale guten Unterrichts sowie Schülerleistung und Schülermotivation untersuchen sollen. Bezüglich der Schülerleistung fokussiert diese Studie auf den Fachwissenserwerb im Physikunterricht. Als motivationales Zielkriterium von Unterricht wird das situationale Interesse der Lernenden im Unterricht untersucht. Die Betrachtung der kognitiven Aktivierung als Merkmal guten Unterrichts beinhaltet die Analyse des Umgang mit Schülervorstellungen (vergl. Abschnitt 5.1.3 auf Seite 60), geht aber noch darüber hinaus und ermöglicht zudem Anschlussfähigkeit an Studien wie COACTIV oder Vogelsang (2014). Die Zusammenhänge zwischen dem Professionswissen der Lehrkräfte und kognitiver Aktivierung im Unterricht werden daher im Rahmen von ProwiN auch für den Chemie- und Biologieunterricht untersucht.

# 6. Forschungsfragen und Hypothesen

Im Rahmen der vorliegenden Arbeit soll untersucht werden, ob das mit den ProwiN-Testinstrumenten erfasste Professionswissen von Physiklehrkräften prädiktiv für deren Unterrichtsqualität und Unterrichtserfolg ist. Auf Grundlage der Überlegungen im letzten Kapitel (Abschnitt 5.2 auf Seite 63 und Abschnitt 5.3 auf Seite 65) werden zwei Forschungsfragen (F) und die zugehörigen Hypothesen (H) formuliert. Die erste Fragestellung bezieht sich auf den Zusammenhang zwischen Professionswissen und Unterrichtserfolg, der über den Zusammenhang zwischen dem Fachwissen, fachdidaktischem Wissen bzw. pädagogischem Wissen von Physiklehrkräften und der Fachwissensleistung und dem situationalen Interesse der Lernenden modelliert wird. Wie bereits erläutert, erscheint es allerdings nicht angemessen lediglich Zusammenhänge zu Schüleroutputvariablen zu betrachten. Die zweite Fragestellung bezieht sich daher auf den Zusammenhang zwischen Professionswissen und Unterrichtsqualität (also gutem *und* erfolgreichem Unterricht), der über Zusammenhänge zwischen den Professionswissensdimensionen, kognitiver Aktivierung und den hier betrachteten Zielkriterien erfolgreichen Unterrichts modelliert wird.

In Abschnitt 2.3.3 auf Seite 17 und Abschnitt 5.3.5 auf Seite 73 wurde die Bedeutung der Klassenführung als wichtige Voraussetzung für die Sicherung anspruchsvollen und kognitiv aktivierenden Unterrichts thematisiert. Eine Kontrolle der Klassenführung wäre vor diesem Hintergrund wünschenswert, übersteigt allerdings den in dieser Arbeit leistbaren Forschungsaufwand. Der Einfluss der Professionswissensdimensionen auf Unterrichtserfolg und kognitive Aktivierung wird außerdem getrennt betrachtet – eine ebenfalls wünschenswerte Kontrolle des pädagogischen Wissens bei der Betrachtung der fachspezifischen Professionswissensdimensionen ist aufgrund des begrenzten Umfangs der in der vorliegenden Arbeit untersuchten Stichprobe nicht möglich.

An dieser Stelle sei darauf hingewiesen, dass nur Belege *für* die Gültigkeit der Annahmen über Zusammenhänge zwischen Professionswissen und gutem und erfolgreichem Unterrichten gesammelt werden können. Nicht gefundene Zusammenhänge können, wie bereits erläutert, nicht eindeutig interpretiert werden – eine eindeutige Falsifizierung der Annahmen ist also nicht möglich. Um die Forschungsfragen eindeutig beantworten zu können, beziehen sich diese daher auf das mit den ProwiN-Testinstrumenten erfasste Wissen und nicht auf Professionswissen von Lehrkräften im Allgemeinen.

## 6.1. Forschungsfrage 1: Professionswissen und Unterrichtserfolg

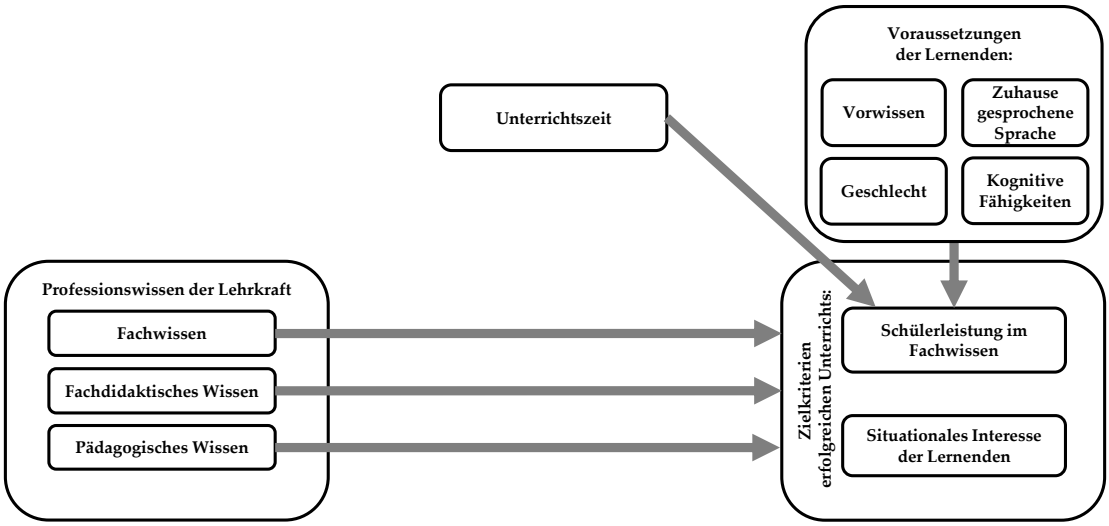



Abbildung 6.1.
*Zusammenhang zwischen Professionswissen und Unterrichtserfolg.*

Die erste Fragestellung, mit der sich diese Arbeit auseinandersetzt, bezieht sich auf den Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und ihrem Unterrichtserfolg. Der Unterrichtserfolg wird über die Schülerleistung im Fachwissen und das situationale Interesse der Lernenden modelliert. Es wird davon ausgegangen, dass das Fachwissen, fachdidaktische Wissen und pädagogische Wissen einer Lehrkraft einen Einfluss auf die Fachwissensleistung und das situationale Interesse der Lernenden hat. Die Schülerleistungen im Fachwissen werden zudem durch das Vorwissen der Lernenden, deren kognitive Fähigkeiten, Geschlecht und die von den Lernenden zuhause gesprochene Sprache sowie die tatsächliche Unterrichtszeit beeinflusst. Abbildung 6.1 zeigt die grafische Darstellung der angenommenen Zusammenhänge.

**F1:** **Welche Zusammenhänge existieren zwischen dem mit den ProwiN-Tests gemessenen Fachwissen, fachdidaktischen Wissen und pädagogischen Wissen von Physiklehrkräften und Unterrichtserfolg?**

H1a-c: *Unterschiede in den Fachwissensleistungen der Lernenden werden (nach Kontrolle des Vorwissens, der kognitiven Fähigkeiten, des Geschlechts, der zuhause gesprochenen Sprache und der Unterrichtszeit) durch Unterschiede im a) CK, b) PCK und c) PK der Lehrkräfte erklärt. Höhere Testergebnisse der Lehrkräfte im CK, PCK bzw. PK hängen mit höheren Fachwissensleistungen der Lernenden zusammen.*

H1d-f: *Unterschiede im situationalen Interesse der Lernenden werden durch Unterschiede im d) CK, e) PCK und f) PK der Lehrkräfte erklärt.*

*Höhere Testergebnisse im CK, PCK bzw. PK hängen mit höheren Ausprägungen des situationalen Interesses der Lernenden im Unterricht zusammen.*

**Methodik:**
Zur Überprüfung der Hypothesen werden Mehrebenenanalysen gerechnet. Dadurch wird dem Umstand Rechnung getragen, dass die zueinander in Bezug zu setzenden Variablen auf unterschiedlichen Ebenen liegen (Klassen- vs. Schülerebene) und die Lernenden in Schulklassen gruppiert sind.[1] In den Modellen 1a-c wird die Schülerleistung im Fachwissen am Ende einer Unterrichtseinheit als abhängige Variable betrachtet. Auf Individualebene werden das Vorwissen zu Beginn der Unterrichtseinheit, die kognitiven Fähigkeiten, das Geschlecht und die von den Lernenden zuhause gesprochene Sprache als Prädiktoren in die Modelle aufgenommen. Auf Klassenebene wird die Unterrichtszeit als Prädiktor in die Modelle aufgenommen. In den Modellen 1d-f wird das situationale Interesse der Lernenden als abhängige Variable betrachtet. Die Hypothesen werden angenommen, wenn das CK, PCK bzw. PK der Lehrkräfte signifikante Prädiktoren für die Fachwissensleistungen (H1a-c) bzw. für das situationale Interesse der Lernenden (H1d-f) sind ($\gamma_{\mathrm{W}}^{\mathrm{StdYX}} > 0$, $p_{\text{1-seitig}} < 0.05$).

## 6.2. Forschungsfrage 2: Professionswissen und Unterrichtsqualität

Die zweite Fragestellung bezieht sich auf den Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und der Qualität ihres Unterrichts. Als Merkmal guten Unterrichts wird die kognitive Aktivierung betrachtet. Um sicherzustellen, dass es sich hierbei um ein Merkmal der Unterrichtsqualität im Sinne von Fenstermacher und Richardson (2005) handelt, muss auch der Zusammenhang zwischen kognitiver Aktivierung und den Zielkriterien erfolgreichen Unterrichts untersucht werden. Die Forschungsfrage 2 teilt sich demnach in zwei Teilfragestellungen auf. Auf Grundlage der Ausführungen in Abschnitt 5.3 auf Seite 65 wird angenommen, dass insbesondere das fachdidaktische Wissen, aber auch das Fachwissen einer Lehrkraft Einfluss darauf hat, inwieweit diese ihren Unterricht kognitiv aktivierend gestaltet. Bezüglich des pädagogischen Wissens ist hingegen unklar, ob von einem Zusammenhang zur kognitiven Aktivierung ausgegangen werden kann – negative Zusammenhänge sind allerdings nicht zu erwarten. Abbildung 6.2 auf der nächsten Seite zeigt die grafische Darstellung der angenommenen Zusammenhänge.

**F2.1: Besteht ein Zusammenhang zwischen der kognitiv aktivierenden Gestaltung von Unterricht und Unterrichtserfolg?**

[1] In Abschnitt 7.4.5 auf Seite 99 im Methodenteil dieser Arbeit wird ausführlich auf die Problematik hierarchischer Datenstrukturen und einen methodisch angemessenen Umgang mit diesen eingegangen.

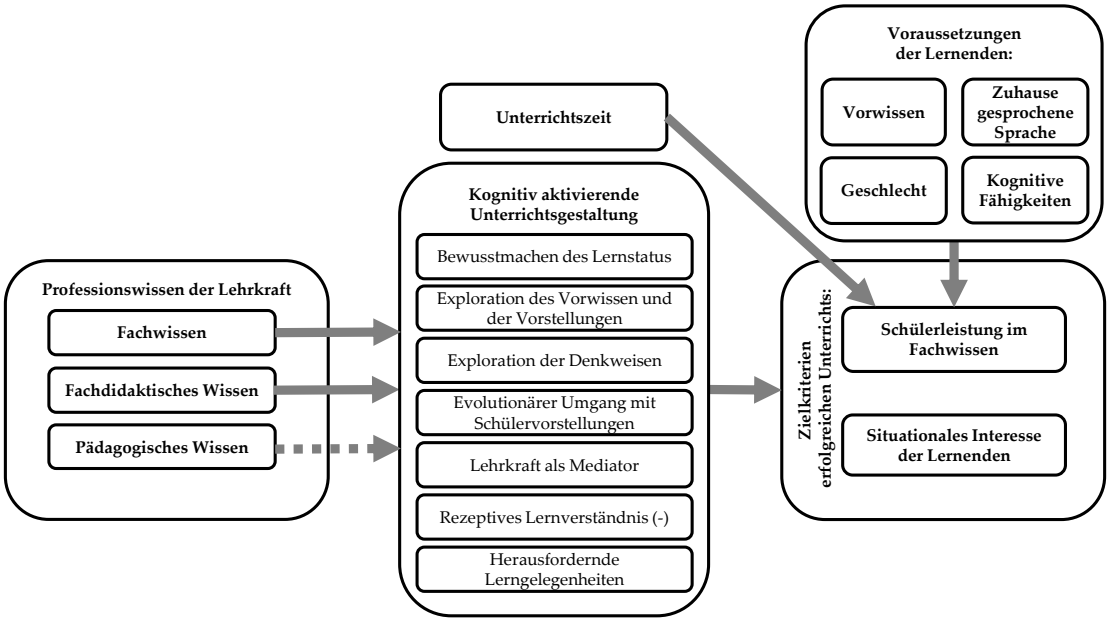



Abbildung 6.2.
*Zusammenhang zwischen Professionswissen und Unterrichtsqualität.*

H2.1a: *Unterschiede in den Fachwissensleistungen der Lernenden werden (nach Kontrolle des Vorwissens, der kognitiven Fähigkeiten, des Geschlechts, der zuhause gesprochenen Sprache und der Unterrichtszeit) durch Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts erklärt. Höhere Ausprägungen in der kognitiven Aktivierung hängen mit höheren Fachwissensleistungen der Lernenden zusammen.*

H2.1b: *Unterschiede im situationalen Interesse der Lernenden werden durch Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts erklärt. Höhere Ausprägungen in der kognitiven Aktivierung hängen mit höheren Ausprägungen des situationalen Interesses der Lernenden im Unterricht zusammen.*

**Methodik:**
Zur Überprüfung der Hypothesen werden auch hier Mehrebenenanalysen gerechnet. In Modell 2.1a wird die Schülerleistung im Fachwissen am Ende einer Unterrichtseinheit als abhängige Variable betrachtet. Auf Individualebene werden das Vorwissen zu Beginn der Unterrichtseinheit, die kognitiven Fähigkeiten, das Geschlecht und die von den Lernenden zuhause gesprochene Sprache als Prädiktoren in das Modell aufgenommen. Auf Klassenebene wird die Unterrichtszeit als Prädiktor in das Modell aufgenommen. In Modell 2.1b wird das situationale Interesse der Lernenden als abhängige Variable betrachtet. Die Hypothesen H1a bzw. H1b werden angenommen, wenn die kognitive Aktivierung ein signifikanter Prädiktor für die Fachwissensleistungen bzw. für das situationale Interesse der Lernenden ist ($\gamma_{\mathrm{W}}^{\mathrm{StdYX}} > 0$, $p_{\text{1-seitig}} < 0.05$).

**F2.2:** **Welche Zusammenhänge existieren zwischen dem mit den ProwiN-Tests gemessenen Fachwissen, fachdidaktischen Wissen und pädagogischen Wissen von Physiklehrkräften und der kognitiv aktivierenden Gestaltung ihres Unterrichts?**

H2.2a-b: *Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts werden durch Unterschiede im a) CK und b) PCK der Lehrkräfte erklärt. Höhere Ausprägungen im CK bzw. PCK hängen mit höheren Ausprägungen in der kognitiven Aktivierung zusammen.*

H2.2c: *Kognitive Aktivierung hängt stärker mit PCK als mit CK zusammen.*

H2.2d: *Falls ein Zusammenhang zwischen PK und kognitiver Aktivierung existiert, ist dieser schwächer als die Zusammenhänge zwischen CK bzw. PCK und kognitiver Aktivierung.*

**Methodik:**
Zur Überprüfung der Hypothesen werden Korrelationen berechnet, da es sich ausschließlich um Zusammenhänge auf Klassenebene handelt. Die Hypothesen H2.2a-b werden angenommen, wenn das CK bzw. PCK der Lehrkräfte signifikant positiv mit der kognitiven Aktivierung korreliert ($r > 0$, $p_{\text{1-seitig}} < 0.05$). Die Hypothesen H2.2c-d werden angenommen, wenn die folgende Reihung für die Korrelationen zwischen den Professionswissensdimensionen und kognitiver Aktivierung gilt und die Unterschiede zwischen den Korrelationen signifikant werden: $r_{\text{PCK-KA}} > r_{\text{CK-KA}} > r_{\text{PK-KA}}$ ($p_{\text{1-seitig}} < 0.05$).

# 7. Methoden und Anlage der Studie

Die vorliegende Studie wurde im Rahmen der zweiten Phase des ProwiN-Projektes durchgeführt. Im Rahmen der ProwiN-Videostudie wurden in jedem Fach zwei Dissertationsprojekte finanziert. Das Studiendesign war weitestgehend durch das Rahmenprojekt festgelegt. Während die Entwicklung und Anpassung der schriftlichen Erhebungsinstrumente durch die Autorin dieser Arbeit erfolgte, wurde die Datenerhebung ab dem zweiten Erhebungsjahr von beiden Doktorierenden gemeinsam durchgeführt. Nach Ende der Datenerhebung für die vorliegende Studie wurden die Erhebungen im Zuge des zweiten Dissertationsprojekts ein weiteres Jahr fortgeführt.

In diesem Kapitel wird zunächst das Studiendesign vorgestellt. Anschließend erfolgt eine Beschreibung der Studiendurchführung und des Ablaufs der Erhebungen (Abschnitt 7.2 auf der nächsten Seite). Die Stichprobe wird in Abschnitt 7.3 auf Seite 89 beschrieben. In Abschnitt 7.4 auf Seite 90 werden die in der vorliegenden Arbeit angewendeten statistischen Methoden beschrieben. Die Datenerhebung für das Rahmenprojekt beinhaltete die Erhebung zahlreicher Kontrollvariablen. Diese werden zwar aufgeführt, aber im Rahmen dieser Arbeit nur zum Teil in die Analysen einbezogen. In Abschnitt 7.5 auf Seite 102 werden daher nur die für diese Arbeit relevanten schriftlichen Erhebungsinstrumente vorgestellt. Eine Beschreibung des videobasierten Ratinginstruments zur kognitiven Aktivierung findet sich in Abschnitt 7.6 auf Seite 134.

## 7.1. Untersuchungsdesign

Um zu untersuchen, ob das mit den ProwiN-Testinstrumenten erfasste Professionswissen von Physiklehrkräften prädiktiv für deren Unterrichtsqualität und Unterrichtserfolg ist, wurde eine Videostudie durchgeführt. Das Design der Studie folgt dem Ziel, Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts, im physikalischen Fachwissen und im situationalen Interesse von Schulklassen auf Unterschiede im Professionswissen der unterrichtenden Lehrkräfte zurückzuführen. Hierfür wurde zunächst das Professionswissen der Lehrkräfte erfasst. Die ProwiN-Testinstrumente zur Erfassung des fachspezifischen Professionswissens fokussieren auf den physikalischen Fachinhalt Mechanik. Daher wurde als Intervention die von den Lehrkräften regulär geplante Unterrichtseinheit zur Mechanik gewählt. Das Fachwissen der Lernenden in Mechanik wurde im Rahmen eines Prä-Post-Designs vor und nach der Unterrichtseinheit erhoben. Die Dauer der Unterrichtseinheit und damit der Zeitraum zwischen Prä- und Post-Test hing von der individuellen Planung der Lehrkräfte und von den internen Curricula der Schulen ab, an denen die Lehrkräfte unterrichteten. Um einen Einblick in die

Unterrichtspraxis der Lehrkräfte zu gewinnen, wurden zwei aufeinanderfolgende Unterrichtsstunden innerhalb der Mechanikeinheit videographiert. Am Ende jeder Unterrichtsstunde wurde das situationale Interesse der Lernenden erfasst.

Die Durchführung der Studie sollte die reguläre Unterrichtsplanung und -durchführung der Lehrkräfte so wenig wie möglich beeinflussen. Gleichzeitig musste aber eine möglichst gute Vergleichbarkeit der aufgezeichneten Unterrichtsstunden sichergestellt werden. Eine Lehrplananalyse für die Mittelstufe ergab, dass ein zentrales Thema innerhalb der Unterrichtseinheit Mechanik die Behandlung des Kraftbegriffes ist. Die Lehrkräfte hatten daher die Vorgabe, in der ersten aufgezeichneten Unterrichtsstunde den physikalischen Kraftbegriff einzuführen. Außerdem sollte die Unterrichtstunde ein Lehrerexperiment beinhalten und das primäre Lehrziel sollte im Kompetenzbereich Fachwissen liegen. Das Vorführen eines Lehrerexperimentes sollte eine sichtbare Aktivität der Lehrkraft in der Unterrichtsstunde sicherstellen, schloss aber die Durchführung von Schülerexperimenten oder Gruppenarbeit im Rest der Stunde nicht aus. Die Interpretation dieser Vorgaben wurde den Lehrkräften selbst überlassen. Während einige Lehrkräfte aufwendigere Experimente durchführten, führten andere Lehrkräfte lediglich kleine Demonstrationsexperimente vor (z. B. das Zusammendrücken einer Knetkugel oder die Beschleunigung eines Spielzeugautos zur Demonstration der Kraftwirkungen). Für die zweite Unterrichtsstunde wurden keine einschränkenden Vorgaben gemacht. Den Zeitpunkt der Videoaufnahmen konnten die Lehrkräfte entsprechend ihrer individuellen Unterrichtsplanung frei wählen. Auch die Länge der aufgezeichneten Unterrichtsstunden orientierte sich an den Realbedingungen vor Ort. Abhängig von der Stundenplanung der jeweiligen Schule variierte die Länge der aufgezeichneten Unterrichtsstunden zwischen 45 und 90 Minuten.

Als Kontrollvariablen auf Schülerebene wurden die kognitiven Fähigkeiten der Lernenden und die von den Lernenden zuhause gesprochene Sprache sowie deren Geschlecht erhoben. Da die Länge der Unterrichtseinheit Mechanik nicht vorgeschrieben wurde, wurde als Kontrollvariable auf Klassenebene die Anzahl der Unterrichtsstunden erfasst, die die Lehrkräfte in Mechanik unterrichtet hatten. Abbildung 7.1 auf der nächsten Seite zeigt das Untersuchungsdesign der ProwiN-Videostudie.

## 7.2. Durchführung der Studie

Die Datenerhebung erfolgte im Bundesland Nordrhein-Westfalen (NRW) in den Schuljahren 2011/2012 und 2012/2013. Die Datenerhebung im Rahmenprojekt wurde durch den zweiten im ProwiN-Projekt arbeitenden Physikdoktoranden im Schuljahr 2013/2014 fortgeführt. Ursprünglich wurden Daten sowohl an Gymnasien als auch an Gesamtschulen erhoben, um die Varianz im Professionswissen der Lehrkräfte und im Schülerleistungszuwachs zu erhöhen. Im Erhebungszeitraum 2011-2013 nahmen allerdings nur zwei Gesamtschullehrkräfte mit ihren Klassen an der Studie teil. Zwischen Gymnasien und Gesamtschulen sind sowohl bezüglich des Leistungsniveaus der Lernenden als auch bezüglich des Professionswissens der Lehrkräfte Unterschiede zu erwarten: Zum einen stellten Pant et al. (2013,

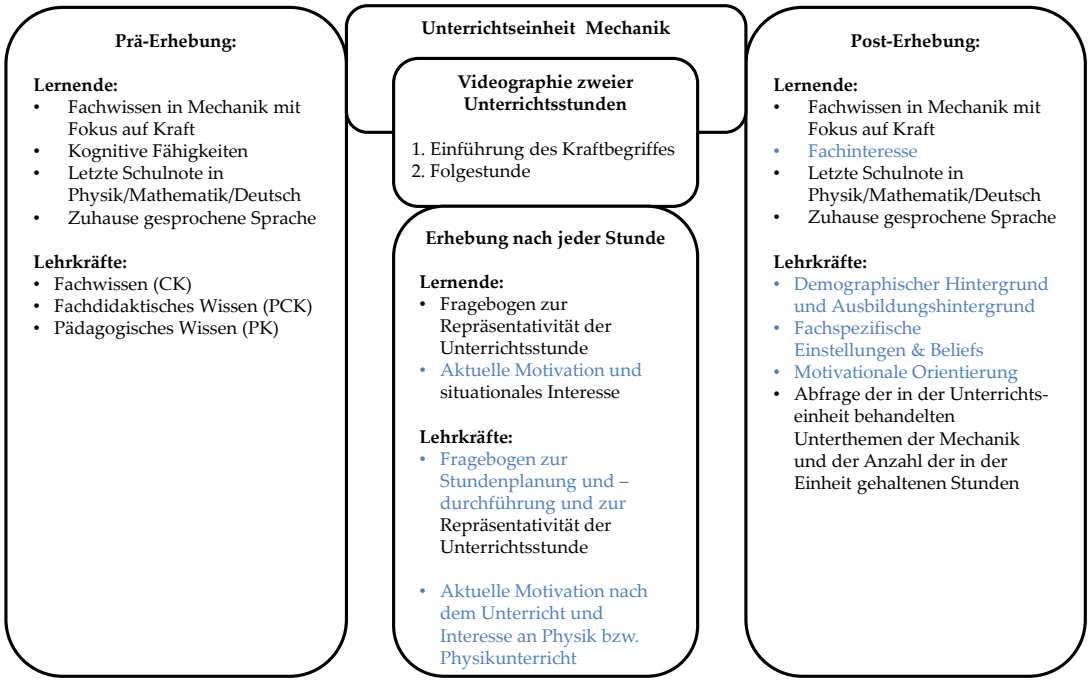



Abbildung 7.1.
*Grafische Darstellung des Untersuchungsdesigns mit allen im Rahmen der ProwiN-Videostudie erhobenen Variablen (Blau dargestellte Variablen wurden in dieser Arbeit nicht berücksichtigt).*

S. 215) fest, dass das Leistungsniveau an Gymnasien in NRW höher ist als an nicht-gymnasiale Schulformen, zum anderen konnte Kirschner (2013, S. 88) zeigen, dass Gymnasiallehrkräfte über ein höheres fachspezifisches Professionswissen verfügen. Ob das höhere Leistungsniveau an Gymnasien auch durch das höhere Professionswissen der Lehrkräfte erklärt werden kann, kann aufgrund der geringen Anzahl der Gesamtschulklassen in der erhobenen Stichprobe im Rahmen dieser Arbeit nicht untersucht werden. Falls die beobachteten Leistungsunterschiede allerdings auf andere, hier nicht untersuchte Variablen zurückzuführen sind, könnte der Einbezug der Daten der Gesamtschulklassen in die Analysen zu einer Verzerrung der Ergebnisse und zu falschen Interpretationen führen. Die Daten dieser Klassen werden daher nicht in die Auswertung aufgenommen. Folglich werden die Gesamtschulklassen nicht in der Stichprobenbeschreibung in Abschnitt 7.3 auf Seite 89 aufgeführt.

### 7.2.1. Auswahl der Jahrgangsstufe

Wie bereits erwähnt, wurde als Intervention die Unterrichtseinheit Mechanik und als Thema der ersten videographierten Unterrichtsstunde die Einführung des Kraftbegriffes gewählt. Um zu entscheiden, in welcher Jahrgangsstufe die Studie durchgeführt werden sollte, wurde im Rahmen einer Lehrplanumfrage an 80 Gymnasien und 66 Gesamtschulen vor Beginn der Studie erfragt, in welcher Jahrgangsstufe das Thema Kraft an den Schulen unterrichtet wurde. Aus dem

Diagramm in Abbildung 7.2 geht hervor, dass das Thema Kraft an Gymnasien fast ausschließlich in den Jahrgangsstufen 8 und 9 behandelt wurde. Dementsprechend wurden diese beiden Jahrgangsstufen für die Datenerhebung ausgewählt.

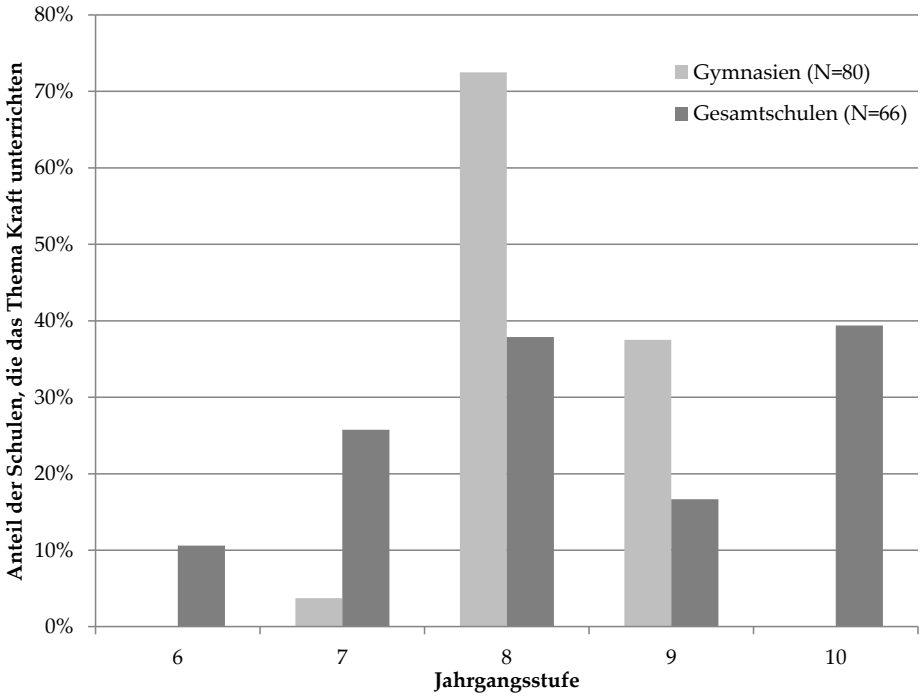



Abbildung 7.2.
*Anteil der Schulen, die das Thema Kraft in einer bestimmten Jahrgangsstufe behandeln (Da Mehrfachangaben möglich waren, addieren sich die gezeigten Anteile nicht zu* 100%*).*

## 7.2.2. Teilnehmerakquise und Teilnahmeanreize

Die Akquise der Lehrkräfte erfolgte über verschiedene Wege und gestaltete sich insgesamt schwierig, da nur sehr wenige Lehrkräfte Interesse an einer Studienteilnahme hatten. NRW-weit wurden fast 600 Schulen angeschrieben und zusätzlich telefonisch über die Studie und die Teilnahmemöglichkeiten informiert. Interessierte Lehrkräfte und persönliche Kontakte wurden außerdem gezielt angerufen. Darüber hinaus wurde die Studie auf Veranstaltungen und Lehrerfortbildungen vorgestellt und beworben.

Im ersten Jahr der Erhebung erhielten die teilnehmenden Lehrkräfte als Anreiz zur Teilnahme 100 €, die teilnehmenden Klassen 50 € für die Klassenkasse und die Lernenden in den Klassen jeweils 5 €. Die Schulen wurden unmittelbar nach den Sommerferien angeschrieben, da die Lehrkräfte an vielen Schulen erst zu diesem Zeitpunkt endgültig erfuhren, ob sie eine 8. oder 9. Klasse in Physik unterrichten würden und ob eine Teilnahme somit überhaupt möglich war. Dieses Vorgehen erwies sich allerdings als problematisch, da viele Lehrkräfte direkt nach den Sommerferien mit der Unterrichtseinheit Mechanik oder der Behandlung des

Kraftbegriffes anfingen und es daher für die Durchführung des Prä-Tests oder die Aufzeichnung der Einführungsstunde zum Kraftbegriff in vielen Fällen bereits zu spät war.

Im zweiten Erhebungsjahr wurden die Lehrkräfte daher bereits nach den Osterferien erstmals kontaktiert. Bis die Klassenverteilung an den Schulen festgelegt war, wurde ein regelmäßiger Kontakt zu den Lehrkräften gepflegt. Außerdem wurde das Anreizsystem dahingehend geändert, dass die Lernenden nur noch den Geldbetrag für die Klassenkasse erhielten und stattdessen 150 € an die Physiksammlungen der teilnehmenden Schulen ausgezahlt wurden. Dieses Vorgehen und das neue Anreizsystem erwiesen sich als wesentlich erfolgreicher bei der Teilnehmerakquise.

## 7.2.3. Ablauf der Erhebungen

Vor Beginn der Erhebung wurden die Lehrkräfte gebeten, eine vollständige Klassenliste und (soweit bereits vorhanden) einen Sitzplan der Klasse, mit der sie an der Studie teilnahmen, zur Verfügung zu stellen. Außerdem musste dem Forscherteam spätestens bis zur Videographie der ersten Unterrichtsstunde von jedem Kind eine Einverständniserklärung der Erziehungsberechtigten vorliegen. Lag diese Einverständniserklärung nicht vor oder wurde sie nicht erteilt, wurden die entsprechenden Schülerinnen und Schüler bei den Videoaufnahmen in den toten Winkel der Kameras gesetzt. Die Erziehungsberechtigten konnten außerdem darüber entscheiden, ob die Lehrkraft eine Kopie der Unterrichtsvideos erhalten durfte. Die Lehrkräfte bekamen nur dann eine Kopie der Videos, wenn alle Erziehungsberechtigten einer Klasse der Aushändigung zugestimmt hatten.

Jedem Lernenden wurde vor der Prä-Erhebung eine Identifikationsnummer (ID) zugeteilt, die aus einer Klassen-ID, einer Schüler-ID und einer ID für die Schulart (Gymnasium oder Gesamtschule) bestand. Diese Nummer wurde zusammen mit einer Testheft-ID für die verschiedenen Erhebungsinstrumente vor den Erhebungen auf alle Testhefte geklebt. Um während der Erhebung eine Zuordnung der Testhefte zu ermöglichen, wurden zusätzlich die Namen der Lernenden auf einem Klebezettel auf die Testhefte geklebt. Vor der Abgabe der Testhefte wurden die Schülerinnen und Schüler aufgefordert die Klebezettel zu entfernen. Alle Erhebungen wurden bis auf wenige Ausnahmen bei der Prä- oder Post-Erhebung von zwei Testleitern durchgeführt. Dabei handelte es sich um die Autorin, den zweiten im Projekt arbeitenden Doktoranden und geschulte studentische Hilfskräfte.

### 7.2.3.1. Prä-Erhebung

Die Prä-Erhebung fand in der letzten Stunde vor Beginn der Unterrichtseinheit Mechanik statt und war auf 45 Minuten ausgelegt. Um ein standardisiertes Vorgehen bei der Erhebung sicherzustellen, wurde ein Testleitermanual (siehe Anhang A.1.1 auf Seite 218) genutzt. Alle Informationen zur Studie, zum Ablauf der Testung und zum Ausfüllen der Testhefte wurden vorgelesen. Nach einer kurzen Einleitung wurde zunächst die Testung der kognitiven Fähigkeiten der Lernenden durchgeführt. Hierfür wurde, dem Manual des standardisierten Testinstrumentes folgend, ein kurzes Aufgabenbeispiel gemeinsam besprochen. Anschließend hatten die Ler-

nenden acht Minuten Zeit für die Bearbeitung des Testheftes. Danach erfolgte die Bearbeitung des Schülerfachwissenstests. Um gegenseitigem Abschreiben unter den Lernenden vorzubeugen, gab es zwei verschiedene Testhefte. Sitznachbarn erhielten unterschiedliche Testhefte. Die Bearbeitungszeit für den Prä-Test betrug 30 Minuten. Die meisten Schülerinnen und Schüler bearbeiteten das Testheft in $15-20$ Minuten. Um auch für die letzten Lernenden eine ruhige Testatmosphäre sicherzustellen, erhielten die Schülerinnen und Schüler bei Abgabe des Testhefts ein Sudoku und ein Mandala zum Ausmalen.

Die Lehrkraft füllte während der 30-minütigen Bearbeitungszeit für den Schülerfachwissenstest den Test zum pädagogischen Wissen aus. Abhängig vom Stundenplan der Lehrkraft, wurde am selben Tag außerdem vor oder nach der Prä-Erhebung die Testung zum Fachwissen und fachdidaktischen Wissen durchgeführt. Auch hier wurde ein Testleitermanual genutzt (siehe Anhang A.1.3 auf Seite 220). Bis auf wenige Ausnahmen wurden beide Tests ohne Unterbrechung in einer vorgegebenen Bearbeitungszeit von 85 Minuten (45 Minuten für den PCK-Test, 40 Minuten für den CK-Test) direkt nacheinander durchgeführt. Bei einigen Lehrkräften mussten die Testungen aus organisatorischen Gründen in separaten Freistunden durchgeführt werden.

#### 7.2.3.2. Post-Erhebung

Der Ablauf der Post-Erhebung verlief ähnlich zur Prä-Erhebung. Erneut wurde ein Testleitermanual (siehe Anhang A.1.2 auf Seite 219) genutzt. Die Lernenden bearbeiteten zunächst den Schülerfachwissenstest. Es wurden die gleichen Testhefte wie bei der Prä-Erhebung eingesetzt. Dabei erhielten die Lernenden das Testheft, das sie bei der Prä-Erhebung nicht bearbeitet hatten. Sitznachbarn erhielten wieder unterschiedliche Testhefte. Die Bearbeitungszeit betrug wie im Prä-Test 30 Minuten. Die meisten Lernenden bearbeiteten das Testheft in $20-25$ Minuten. Die im Vergleich zum Prä-Test längere tatsächliche Bearbeitungszeit kann als Indiz für die intensivere Auseinandersetzung mit den Aufgaben gewertet werden. Da zum Zeitpunkt der Prä-Erhebung die im Schülerfachwissenstest adressierten Fachinhalte noch nicht vermittelt worden waren, wurde bei der Bearbeitung der Aufgaben vermutlich häufiger geraten und weniger intensiv nachgedacht. Bei Abgabe des Schülerfachwissenstests erhielten die Schülerinnen und Schüler den Fragebogen zum Fachinteresse, der etwa fünf Minuten Bearbeitungszeit in Anspruch nahm und im Anschluss erneut eine Beschäftigung.

Die Lehrkräfte erhielten während der Post-Erhebung einen Lehrerfragebogen, in dem Angaben zu ihrem demographischen Hintergrund und zum Ausbildungshintergrund erbeten und ihre fachspezifischen Einstellungen und Beliefs sowie ihre motivationale Orientierung erhoben wurden. Außerdem wurden die Anzahl der in der Unterrichtseinheit Mechanik unterrichteten Stunden und die dort behandelten Unterthemen erfragt. Darüber hinaus wurde von den Lehrkräften ein Fragebogen zur Selbstwirksamkeitserwartung eingesammelt, der ihnen bei einem der vorherigen Erhebungszeitpunkte ausgehändigt worden war. Freiwillig konnten die Lehrkräfte außerdem an einem Expertenrating zur Zuordnung der Schülerfachwissenstestaufgaben zu den abgefragten Unterthemen der Mechanik teilnehmen. Hier konnten

die Lehrkräfte auch für jede Aufgabe einschätzen, ob ihre Schülerinnen und Schüler die Aufgabe lösen können sollten.

#### 7.2.3.3. Video-Erhebung

Im folgenden Abschnitt wird zunächst kurz auf technische Details zur Videographie des Unterrichts eingegangen. Anschließend wird der Ablauf der Videoerhebungen beschrieben.

**Videographie** Die im Projekt arbeitenden Doktoranden und die studentischen Hilfskräfte erhielten zunächst eine halbtägige Schulung im Umgang mit der Videoausrüstung. Auf Basis des Video-Manuals aus dem QuiP Projekt (Keller, 2011) wurden Absprachen zur Videographie des Unterrichts getroffen. Im ersten Erhebungsjahr bestand der Videoaufbau aus zwei Kameras (vergl. Abbildung 7.3). Die handgeführte Aktionskamera verfolgte die gerade handelnden Akteure und insbesondere das Handeln der Lehrkraft. Die Totalenkamera filmte hingegen von der Pultseite aus in die Klasse hinein und wurde so ausgerichtet, dass Schülerinnen und Schüler ohne Einverständniserklärung der Erziehungsberechtigten im toten Winkel der Kamera saßen. Die Lehrkraft erhielt ein Diktiergerät mit Mikrofon, zwei weitere Diktiergeräte wurden im Raum verteilt. Im zweiten Jahr der Erhebung wurde der Videoaufbau durch eine geführte Lehrerkamera ergänzt, die ausschließlich auf Handeln, Mimik und Gestik der Lehrkraft fokussierte.

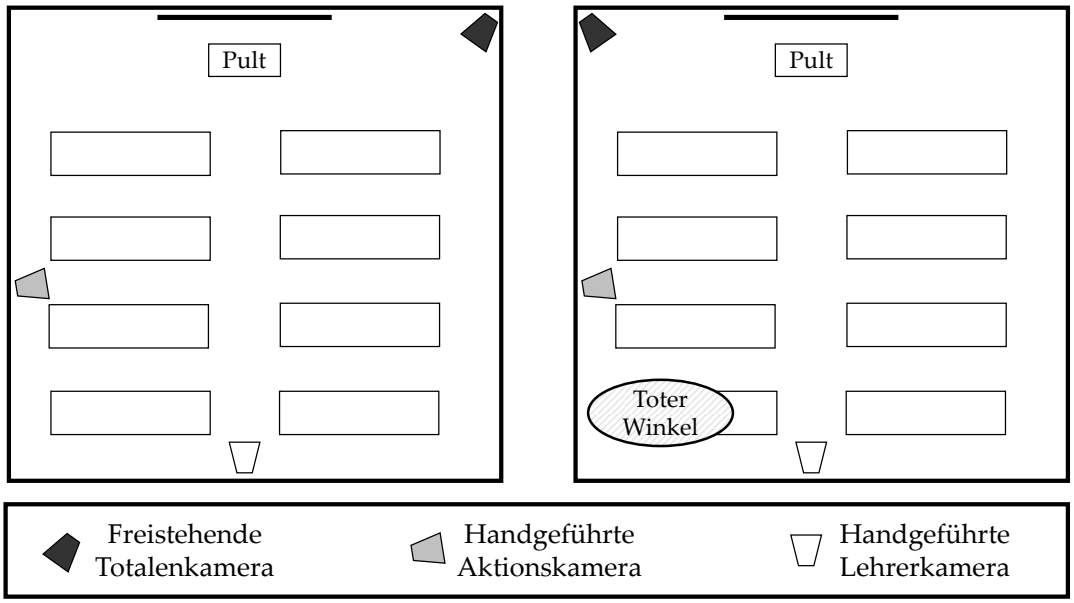



Abbildung 7.3.
*Beispielhafte Anordnung der Videokameras im Klassenraum.*

**Ablauf der Videoerhebung** Nach dem Aufbau der Kameras wurde die gesamte Unterrichtsstunde gefilmt. Wie bereits erwähnt, variierte die Länge der aufge-

zeichneten Unterrichtsstunden zwischen 45 und 90 Minuten. In den letzten fünf Minuten der Unterrichtsstunden füllten die Lernenden einen Fragebogen zur Repräsentativität der Unterrichtsstunden und zu ihrer aktuellen Motivation aus, in dem auch ihr situationales Interesse am Unterricht erfasst wurde. Die Lehrkraft beantwortete im Anschluss an den Unterricht einen ca. zehnminütigen Fragebogen zur Stundenplanung- und -durchführung, zur Repräsentativität der Unterrichtsstunde und zu ihrer aktuellen Motivation und ihrem Interesse an dem unterrichteten Inhalt bzw. ihrem Interesse am Unterrichten des Inhalts.

#### 7.2.3.4. Zeitraum zwischen den Erhebungen

Im Mittel lagen zwischen Prä- und Post-Testung 25.0 Wochen, was in etwa einem Schulhalbjahr entspricht. Dieser Zeitraum variierte allerdings stark zwischen den teilnehmenden Klassen ($Min$ = 10.0 Wochen, $Max$ = 44.3 Wochen, $SD$ = 8.5 Wochen). Grund dafür waren unterschiedliche schulinterne Curricula, der krankheitsbedingte Ausfall einzelner Lehrkräfte und innerhalb der Erhebung liegende Ferienzeiten (Herbst-, Winter- und Osterferien). Die Aufzeichnung des ersten Unterrichtsvideos erfolgte im Mittel 26 Tage nach der Prä-Testung ($Min$ = 1 Tag, $Max$ = 91 Tage, $SD$ = 22 Tage). Zwischen der ersten und der zweiten Videostunde lagen in der Regel nicht mehr als sieben Tage, da zwei aufeinanderfolgende Unterrichtsstunden gefilmt wurden. Vier Klassen bildeten hier eine Ausnahme. In zwei Klassen (ID 22, ID 28) lagen aufgrund von Stundenausfall und aufgrund der Herbstferien 28 bzw. 11 Tage zwischen den Videoerhebungen, dennoch handelte es sich um aufeinanderfolgende Unterrichtsstunden. In zwei anderen Klassen (ID 03, ID 05) war eine Aufzeichnung aufeinanderfolgender Unterrichtsstunden nicht möglich. Hier wurde 23 bzw. 8 Wochen nach der ersten Videostunde eine Einführungsstunde in ein anderes, von den Lehrkräften frei gewähltes, Unterthema der Mechanik aufgezeichnet.

### 7.2.4. Maßnahmen zur Sicherung der Datenqualität

Um die Qualität der mit den schriftlichen Testinstrumenten erhobenen Daten sicherzustellen, wurden im ersten Erhebungsjahr alle Testhefte (mit Ausnahme der Testhefte zum fachspezifischen Professionswissen) eingescannt und die Daten wurden mit Hilfe der Software Teleform (Cardiff, 2011) elektronisch in die Statistiksoftware SPSS (IBM Corp., 2012) eingelesen. Aufgrund von Softwareproblemen kam es hierbei allerdings bei einigen Klassensätzen zu systematischen Falscheingaben. Die SPSS-Matrizen aller elektronisch eingelesenen Daten wurden daher mit den Originaltestheften verglichen und vorhandene Falscheingaben korrigiert. Im zweiten Jahr der Erhebung wurden alle Daten zusätzlich per Hand in SPSS eingegeben. Ein Vergleich der SPSS-Matrizen ergab eine Fehlerquote von 0.3% bei der Handeingabe und 0.1% bei der elektronischen Einlesung der Daten (bezogen auf 104 769 Eingaben). Alle identifizierten Fehler wurden korrigiert. Die Datenqualität der hier verwendeten Daten in Bezug auf Falscheingaben ist damit wesentlich höher als in anderen Studien (vergl. z. B. Schoppmeier, 2013, S. 61).

Als weitere Maßnahme zur Sicherung der Datenqualität wurden Testhefte identifiziert, die stichhaltige Hinweise auf eine nicht ordnungsgemäße Bearbeitung durch die Schülerinnen und Schüler lieferten. Die Prä-Testdaten von vier Schülerinnen und Schülern und die Post-Testdaten eines Schülers wurden nicht in die Datenauswertung einbezogen: drei Schülerinnen hatten die Bearbeitung des unmittelbar vor dem Prä-Test durchgeführten Tests zur Messung ihrer kognitiven Fähigkeiten verweigert, ein Schüler hatte eindeutig Muster gekreuzt und ein weiterer Schüler hatte das komplette Testheft bemalt und bei fast allen Aufgaben alle Antwortmöglichkeiten angekreuzt. Diese Schülerinnen und Schüler werden daher auch nicht in der Stichprobenbeschreibung in Abschnitt 7.3 aufgeführt.

## 7.3. Stichprobe

Aus organisatorischen Gründen beschränkte sich die Datenerhebung auf das Bundesland NRW. Zu einer Teilnahme an der Studie berechtigt waren festangestellte Gymnasiallehrkräfte, die in einer 8. oder 9. Klasse Physik unterrichteten und das Thema Kraft im Rahmen der Unterrichtseinheit Mechanik behandelten. Eine Lehrbefähigung für das Fach Physik war nicht zwingend erforderlich. Insgesamt nahmen 23 Physiklehrkräfte mit ihren Klassen an der Studie teil.[1] Zwei Lehrkräfte unterrichteten an einem Mädchengymnasium (ID 16, ID 23). Eine Lehrkraft (ID 5) war zwar festangestellt, stand aber erst kurz vor Abschluss ihres Vorbereitungsdienstes (Referendariats) im Rahmen der Ordnung zur berufsbegleitenden Ausbildung von Seiteneinsteigerinnen und Seiteneinsteigern und der Staatsprüfung (OBAS).

Die Klassengröße variierte zwischen 20 und 34 Schülerinnen und Schülern pro Klasse ($M = 28.7$, $SD = 3.4$). Tabelle 7.1 zeigt die Verteilung der Klassen und der Schülerinnen und Schüler auf die Jahrgangstufen und die Anzahl der Schülerinnen und Schüler, die an den jeweiligen Testzeitpunkten anwesend waren.

Tabelle 7.1.
*Anzahl der Klassen und der Schülerinnen und Schüler (SuS) in den verschiedenen Jahrgangsstufen (JS) und zu den verschiedenen Erhebungszeitpunkten*

| | JS 8 | JS 9 | Prä | Post | Video 1 | Video 2 | Gesamt |
|---|---|---|---|---|---|---|---|
| $N_{\text{Klassen}}$ | 15 | 8 | 23 | 23 | 23 | 23 | 23 |
| $N_{\text{SuS}}$ | 440 | 221 | 640 | 630 | 633 | 625 | 661 |

Da die Teilnahme an der Studie freiwillig war, konnte keine repräsentative Stichprobe, sondern lediglich eine Gelegenheitsstichprobe untersucht werden. Eine

[1] Im Rahmenprojekt wurden zusätzlich zum Erhebungszeitraum dieser Studie im Schuljahr 2013/2014 die Daten von 12 weiteren Lehrkräften und ihren Klassen erhoben. Die Daten der Gesamtstichprobe von 35 Lehrkräften werden im Rahmen der Dissertation von Liepertz (2016) ausgewertet. Außerdem nahmen fünf Lehrkräfte mit zwei oder drei Parallelklassen an der Studie teil.

ausführliche Beschreibung der Stichprobe befindet sich in Abschnitt 8.1 auf Seite 159 im Ergebnisteil dieser Arbeit. Um zu untersuchen, ob es sich bei der Stichprobe der Lehrkräfte der zweiten Phase des ProwiN-Projekts um eine starke Positivauswahl handelt, erfolgt hier auch ein Vergleich mit der Stichprobe der im Rahmen der ersten Projektphase untersuchten Gymnasiallehrkräfte.

## 7.4. Statistische Methoden

In diesem Abschnitt werden die im Rahmen dieser Arbeit angewendeten statistischen Methoden beschrieben. Die fachspezifischen Professionswissenstests, der Schülerfachwissenstest und der Kognitive Fähigkeitentest werden auf Basis der probabilistischen Testtheorie im Rahmen einer Rasch-Analyse ausgewertet. In Abschnitt 7.4.2 auf Seite 93 werden daher die Grundlagen der Rasch-Analyse erläutert. Abschnitt 7.4.3 auf Seite 96 geht auf die Berechnung von Reliabilitäten im Rasch-Modell und im Rahmen der klassischen Testtheorie ein. Da die Kodierung der fachspezifischen Professionswissenstests und die Ratings zur kognitiven Aktivierung im Unterricht von verschiedenen Personen durchgeführt wurden, wird in Abschnitt 7.4.4 auf Seite 98 die Intraklassenkorrelation als Maß für die Beurteilung von Interrater-Übereinstimmungen eingeführt. Die Fragestellung, inwieweit sich Unterschiede im Fachwissen der Lernenden am Ende der Unterrichtseinheit zur Mechanik und im situationalen Interesse der Lernenden im Unterricht durch das Professionswissen der Lehrkräfte und die kognitiv aktivierende Gestaltung des Unterrichts erklären lassen, wird nicht mit herkömmlichen Regressionsanalysen, sondern im Rahmen von Mehrebenenanalysen untersucht. Die Gründe für dieses Vorgehen sowie die Grundlagen der Mehrebenenanalyse werden in Abschnitt 7.4.5 auf Seite 99 erläutert. Außerdem werden in Abschnitt 7.4.6 auf Seite 101 die im Zuge der Instrumentvalidierung genutzten Verfahren zur Berechnung messfehlerbereinigter Korrelationen beschrieben.

### 7.4.1. Allgemeine Hinweise zur Datenanalyse

Die statistischen Berechnungen in dieser Arbeit werden größtenteils mit der Statistik-Software SPSS Statistics 21 (IBM Corp., 2012) durchgeführt. Hierfür werden die im Rahmen der Rasch-Analysen geschätzten Personenparameter in SPSS importiert. In Fällen, in denen weitere Programme genutzt werden, ist dieses explizit vermerkt.

**Signifikanztestung** Sofern Annahmen bezüglich des Vorzeichens von Zusammenhängen oder von Mittelwertunterschieden vorliegen, werden diese, mit Blick auf die Teststärke, einseitig auf Signifikanz getestet (vergl. z. B. Bortz & Döring, 2006, S. 511; Field, 2009, S. 54) – entsprechende $p$-Werte werden als $p_{\text{1-seitig}}$ gekennzeichnet. Korrelationen, die lediglich bei einseitiger Testung signifikant werden, sind daran zu erkennen, dass ihre 95%-Konfidenzintervalle den Nullpunkt einschließen.

**Standardfehler und Konfidenzintervalle** Korrelationen und Regressionskoeffizienten werden in dieser Arbeit mit Standardfehlern und 95%-Konfidenzintervallen angegeben, außer es handelt sich um aus der Literatur zitierte Werte. Es erfolgt allerdings keine Propagation der im Rasch-Modell geschätzten Standardfehler der Personenparameter im Rahmen einer Fehlerentwicklung. Da in SPSS keine Standardfehler und Konfidenzintervalle für Korrelationskoeffizienten ausgegeben werden, werden diese über die in SPSS implementierten Bootstrappingverfahren bestimmt (Schülerebene/Klassenebene: geschichtete/einfache Ziehung von 1000 Stichproben). Das für die Mehrebenenanalysen verwendeten Programm Mplus (L. K. Muthén & Muthén, 2007) schätzt robuste Standardfehler für alle Modellparameter (L. K. Muthén & Muthén, 2007, S. 8). Für die Ausgabe von Konfidenzintervallen wurde die Option CINTERVAL(Symmetric) benutzt. Bootstrapping ist in Mplus für den gewählten Analysetyp (Type=TWOLEVEL) nicht implementiert (L. K. Muthén & Muthén, 2007, S. 496).

**Angabe signifikanter Stellen** Gerundet wird nach den Rundungsregeln der DIN 1333: Ist die erste signifikante Stelle des Standardfehlers $< 3$, wird der Standardfehler auf zwei signifikante Stellen gerundet, für Werte $\geq 3$ wird auf eine signifikante Stelle gerundet – der Standardfehler wird dabei stets aufgerundet (vergl. Fleischmann, 2013). Ein fehlerbehafteter Wert wird mit der gleichen Anzahl Dezimalstellen berichtet wie sein Standardfehler.

**Prüfung auf Normalverteilung** Die Normalverteilung der Daten wird über den *Shapiro-Wilk-Test* und über Betrachtungen der Schiefe und Kurtosis der Datenverteilung geprüft (Field, 2009, S. 138 und S. 148; zu Vorteilen des Shapiro-Wilk-Tests gegenüber anderen Normalverteilungstest vergl. Ghasemi & Zahediasl, 2012; Razali & Wah, 2011). In Tabelle B.3 auf Seite 244 befindet sich eine Übersicht über die Verteilung der Daten in allen untersuchten Variablen.

**Klassische Analyseverfahren** Für Zusammenhangsmaße zwischen normalverteilten Variablen werden *Pearson-Korrelationen* ($r_{\text{Pearson}}$) berechnet. Weicht die Verteilung in einer Variablen signifikant von der Normalverteilung ab, werden für Stichproben mit $N < 30$ die nicht-parametrische Rangkorrelationen *Spearmanns Rho* ($r_{\text{Spearman}}$) und *Kendalls Tau* ($\tau_{\text{Kendall}}$) berechnet. Für große Stichproben können bereits geringfügige Abweichungen von der Normalverteilung signifikant werden (Field, 2009, S. 148). Für Stichproben mit $N \geq 30$ geht man daher aufgrund des zentralen Grenzwerttheorems in der Regel von einer Normalverteilung der Daten aus und kann parametrische Korrelationen berechnen (Bortz & Döring, 2006, S. 411; Ghasemi & Zahediasl, 2012, S. 486). Bei signifikanter Abweichung von der Normalverteilung für Stichproben mit $N \geq 30$ werden in dieser Arbeit zusätzlich nicht-parametrische Korrelationen angegeben. Die Interpretation bezieht sich in diesen Fällen auf die Pearson-Korrelationskoeffizienten. Entsprechendes gilt bei der Überprüfung von Mittelwertunterschieden auf statistische Signifikanz. Für normalverteilte Variablen werden t-Tests gerechnet. Als Effektstärke wird *Cohens d* angegeben. Die Effektstärken und ihre Bootstrap-

Konfidenzintervalle werden mit Hilfe der Statistik-Software R berechnet (R Core Team, 2015). Hierfür wird das R-Paket „bootES“ (Kirby & Gerlanc, 2013) benutzt. Für nicht normalverteilte Variablen werden Mittelwertunterschiede von unabhängigen Stichproben über *Mann-Whitney-U-Tests* und von abhängigen Stichproben über *Wilkoxon-Vorzeichen-Rang-Tests* auf Signifikanz geprüft. Effektstärken können über $r_{\mathrm{MW}} = z(\text{Mann-Whitney } U)/\sqrt{N}$ bzw. $r_{\mathrm{W}} = z(\text{Wilkoxon } T)/\sqrt{N}$ berechnet werden (vergl. Field, 2009, S.550 bzw. S.558).

**Signifikanztests für Korrelationsunterschiede** Im Rahmen der Validierung der fachspezifischen Professionswissenstests und des Schülerfachwissenstests werden in Abschnitt 7.5.1.7 und 7.5.3.5 auf Seite 109 und auf Seite 126 Korrelationskoeffizienten miteinander verglichen. Transformiert man die Korrelationskoeffizienten mit Hilfe von *Fishers Z-Transformation* in $z$-Werte, kann mit Hilfe der Methode von Steiger die standardisierte Differenz $Z$ zwischen den Korrelationen auf Signifikanz geprüft werden (vergl. Bortz, 2005, S. 222). Die Signifikanz-Berechnungen in dieser Arbeit werden mit den Online-Tools von Lee und Preacher (2013a und 2013b) durchgeführt.

**Umgang mit Ordinalskalen** Werden Messgrößen auf Grundlage von mehrstufigen Rating-Skalen erhoben, kann nicht sichergestellt werden, dass es sich dabei um intervallskalierte und nicht lediglich um ordinalskalierte Messgrößen handelt (vergl. z. B. Baur, 2008, S. 279-289; Bortz & Döring, 2006, S. 176-188). In der vorliegenden Arbeit gilt dies für die Aufgaben des PCK- und des CK-Tests, die auf einer dreistufigen Punkteskala bewertet werden, für die kognitive Aktivierung im Unterricht, die auf einer dreistufigen Ratingskala beurteilt wird, und für das situationale Interesse der Lernenden, das auf einer siebenstufigen Ratingskala eingeschätzt wird. Nach Baur (2008, S. 286-287) sinkt das Fehlerrisiko, das entsteht, wenn ordinalskalierte Variablen fälschlicherweise als intervallskaliert angenommen werden, mit der Anzahl möglicher Ausprägungen der Ratingskala: Um Intervallskalenniveau annehmen zu können, sind Ratingskalen mit mindestens fünf Ausprägungen wünschenswert. Für das situationale Interesse der Lernenden wird auf Grund der siebenstufigen Ratingskala Intervallskalenniveau angenommen. Eine Möglichkeit mit ordinal skalierten Daten umzugehen bietet das Rasch-Modell (siehe nächster Abschnitt), mit dessen Hilfe intervallskalierte Personenfähigkeiten geschätzt werden können (vergl. z. B. Bühner, 2006, S. 300) – hierfür sind allerdings Stichprobengrößen $> 100$ wünschenswert (vergl. z. B. W.-H. Chen et al., 2014). Da die Stichprobe für die Auswertung der fachspezifischen Professionswissenstests mit Lehrkräften aus der ersten Phase des ProwiN-Projekts erweitert werden konnte (Abschnitt 7.5.1.3 auf Seite 104), können mit diesen Tests intervallskalierte Personenfähigkeiten geschätzt werden. Auf Grund der geringen Stichprobengröße ($N = 23$), die für die Beurteilung der Unterrichtsstunden zur Verfügung steht, ist dies für die Qualitätsmaße zur kognitiven Aktivierung nicht möglich. Es kann daher nicht sichergestellt werden, dass die über die Beurteilung von Handlungsindikatoren generierten Qualitätsmaße für die kognitiv aktivierende Gestaltung der Unterrichtsstunden intervallskaliert sind. Sofern Korrelationen zwischen die-

sen Maßen und anderen Variablen berechnet werden, werden daher zusätzlich zu den parametrischen Korrelationen auch nicht-parametrische Rangkorrelationen angegeben. Lediglich im Rahmen der Mehrebenenanalysen ist es nicht möglich, Vergleichsparameter aus Analysen anzugeben, die von einem ordinalen Skalenniveau ausgehen. Hierfür müssten verschiedene Ausprägungen in der kognitiven Aktivierung in Form von Dummy-Variablen als zusätzliche Prädiktoren in die Modelle mit aufgenommen werden, was auf Grund der geringen Stichprobengröße nicht möglich ist (vergl. Abschnitt 7.4.5 auf Seite 99).

### 7.4.2. Die Rasch-Analyse

Die Testinstrumente zum fachspezifischen Professionswissen der Lehrkräfte, zum Fachwissen und zu den kognitiven Fähigkeiten der Lernenden wurden in dieser Arbeit mit Hilfe der Item-Response-Theorie (IRT) ausgewertet. Dabei handelt es sich um eine probabilistische Testtheorie, die im Gegensatz zur klassischen Testtheorie auf der Annahme basiert, dass Testergebnisse nicht allein ein Produkt der latenten (und damit nicht direkt messbaren) Fähigkeiten der Versuchspersonen sind, sondern immer auch zufällige Kriterien in das Testergebnis einer Person einfließen (wie z.B. Flüchtigkeitsfehler oder Glück beim Raten) (Strobl, 2012, S. 6). Das Testergebnis einer Person hängt also nur mit einer bestimmten Wahrscheinlichkeit von ihrer latenten Fähigkeit zur Lösung der Testaufgaben ab (Strobl, 2012, S. 7).

Im Folgenden wird zunächst auf das hier verwendete IRT-Modell, das sogenannte Rasch-Modell, und die grundlegenden Annahmen, auf denen dieses Modell basiert, eingegangen. Maßnahmen zur Überprüfung der Modellpassung werden erklärt und die Reliabilitätsschätzung im Rasch-Modell wird erläutert. Abschließend wird das Schätzverfahren, das in dem für die Rasch-Analyse verwendeten Programm Winsteps (Linacre, 2011) implementiert ist, vorgestellt.

**Das Rasch-Modell** Das Rasch-Modell basiert auf der Annahme, dass die Lösungswahrscheinlichkeit einer Aufgabe im mittleren Schwierigkeitsbereich nahezu linear mit Zunahme der latenten Personenfähigkeit ansteigt, während sich die Wahrscheinlichkeit eine sehr leichte oder sehr schwere Aufgabe zu lösen nur gering ändert (Rost, 2004, S. 115). Die Lösungswahrscheinlichkeit einer bestimmten Aufgabe und die Wahrscheinlichkeit, dass eine bestimmte Person eine Aufgabe lösen kann, hängen davon ab, wie sich Personenfähigkeit und Aufgabenschwierigkeit zueinander verhalten (Strobl, 2012, S. 8). Über die Beschreibung dieser Wahrscheinlichkeiten können Personenfähigkeiten und Aufgabenschwierigkeiten auf einer gemeinsamen Skala geschätzt werden. Bearbeitet eine Person eine Aufgabe, deren Schwierigkeit auf der gemeinsamen Skala auf Höhe ihrer Personenfähigkeit liegt, bedeutet das, dass die Wahrscheinlichkeit dafür, dass die Person die Aufgabe lösen kann, bei 50% liegt (Strobl, 2012, S. 10). Liegt die Personenfähigkeit über der Aufgabenschwierigkeit, erhöht sich die Lösungswahrscheinlichkeit der Aufgabe für diese Person.

Die Personenfähigkeiten im Rasch-Modell können so geschätzt werden, dass die Wahrscheinlichkeit für die Beobachtung der individuell erreichten Anzahl gelöster

Aufgaben maximiert wird (Bortz & Döring, 2006, S. 226).[2] Damit die Anzahl gelöster Aufgaben alle für die Schätzung der Personenfähigkeit relevanten Informationen enthält und durch dieses Vorgehen keine Information über die Personenfähigkeit verloren geht (z.B. weil nicht berücksichtigt wird, dass eine Person zwar nur wenige, dafür aber besonders schwierige Aufgaben gelöst hat), müssen die Aufgaben eine gute Passung ins Rasch-Modell zeigen (Bortz & Döring, 2006, S. 227). Nur dann kann davon ausgegangen werden, dass weniger fähige Personen tendenziell lediglich leichte Aufgaben richtig beantworten, während Personen mit höherer Fähigkeit zusätzlich schwere Aufgaben beantworten können und damit im Mittel mehr Aufgaben richtig lösen (Strobl, 2012, S. 15). Die Gültigkeit des Rasch-Modells impliziert auch die *spezifische Objektivität* eines Testinstruments. Der Vergleich von Personenfähigkeiten ist demnach unabhängig von den für den Vergleich ausgewählten Aufgaben. Ebenso sollte der Vergleich von Aufgabenschwierigkeiten in verschiedenen Personengruppen der untersuchten Stichprobe zu ähnlichen Aussagen führen (Strobl, 2012, S. 20). Gilt das Rasch-Modell, sollte eine Aufgabe, die zum Beispiel für Mädchen einfacher ist als eine andere, auch für Jungen einfacher zu lösen sein. Wenn eine Aufgabe in verschiedenen Personengruppen unterschiedlich funktioniert, spricht man von *Differential Item Functioning* (DIF). DIF kann auch dann auftauchen, wenn die vom Rasch-Modell geforderte Eindimensionalität der Daten verletzt ist, wenn also die Lösung einzelner Aufgaben z. B. nicht nur von der Personenfähigkeit bzgl. des intendierten Konstrukts, sondern von weiteren, nicht untersuchten Fähigkeitsdimensionen abhängt (Strobl, 2012, S. 23).

Das Rasch-Modell setzt außerdem die *lokale stochastische Unabhängigkeit* der Beobachtungen voraus. Bezogen auf die Aufgaben bedeutet dies, dass sich für eine bestimmte Person die Lösungswahrscheinlichkeit einer Aufgabe nicht durch die Lösungswahrscheinlichkeit einer anderen Aufgabe verändern darf. Dies wäre beispielsweise der Fall, wenn einzelne Aufgaben aufeinander aufbauen, die Lösung einer Aufgabe also von der Lösung einer anderen Aufgabe abhängt. Bezogen auf die Personen bedeutet die lokale stochastische Unabhängigkeit, dass die Lösungswahrscheinlichkeit einer Aufgabe für eine Person nicht von der Lösungswahrscheinlichkeit der gleichen Aufgabe für eine andere Person abhängt. Dies kann beispielsweise dann passieren, wenn Testpersonen abschreiben oder wenn im Rahmen von Veränderungsmessungen Daten von zwei voneinander abhängigen Stichproben (z.B. Prä-Post-Daten derselben Schülerstichprobe) in einer gemeinsamen Rasch-Analyse analysiert werden (Strobl, 2012, S. 18-19).

**Überprüfung der Modellpassung** Um sicherzustellen, dass die aus den Summenscores geschätzten Personenparameter gute Schätzer für die latenten Fähigkeiten der untersuchten Personen darstellen, muss die Passung der mit den Testinstrumenten erhobenen Daten ins Rasch-Modell überprüft werden.

> **Exkurs:** Formal gesehen unterscheidet das Rasch-Modell nicht zwischen Personen und Aufgaben (Rost, 2004, S. 364). Die Modellpassung

[2] Es gibt allerdings auch Methoden zur Schätzung der Personenfähigkeiten, die auf anderen Ansätzen aufbauen (vergl. z. B. Rost, 2004, S. 309-317).

kann also sowohl durch Ausschluss von Aufgaben als auch durch Ausschluss von Personen verbessert werden. Rost (2004) stellt fest:

> Von einem wissenschafts-ethischen Standpunkt aus betrachtet, gibt es jedoch eine Asymmetrie in dieser Frage. Während die Selektion von Items als legitim gilt, schließlich sind sie von Menschenhand gemacht und können mit den Fehlern behaftet sein, die eine Eliminierung rechtfertigen, gilt die Eliminierung unpassender Personen aus der Datenmatrix als illegitim. (S. 365)

Er nennt aber dennoch Gründe warum der Ausschluss von Personen sinnvoll sein kann. Beispielsweise tragen Personen, deren Antwortverhalten nicht modellkonform ist (z.B. aufgrund von mangelnder Testmotivation oder Konzentrationsschwäche), überproportional viel zum Messfehler eines Testinstruments bei.

Die vorliegende Arbeit folgt den wissenschafts-ethischen Argumenten und schließt keine Personen aufgrund von schlechter Modellpassung aus den Analysen aus. Im folgenden wird daher lediglich auf die Passung der Aufgaben ins Rasch-Modell eingegangen. Aufgrund der bereits erwähnten Symmetrie zwischen Aufgaben und Personen, können die Statistiken zur Überprüfung der Modellpassung von Personen allerdings analog formuliert werden.

Wie gut eine Aufgabe ins Rasch-Modell passt, kann mit Hilfe der *Mean-Square-Statistik* (*MnSq*-Statistik) bewertet werden, die aus der mittleren quadratischen Abweichung zwischen dem im Modell erwarteten Antwortverhalten bezüglich einer Aufgabe und dem tatsächlich beobachteten Antwortverhalten berechnet wird. Der Erwartungswert des *MnSq* liegt bei 1, was einer optimale Passung der Daten ins Modell entsprechen würde. Ein *MnSq* von 1.3 bedeutet beispielsweise, dass in den beobachteten Antworten für eine Aufgabe 30% mehr Varianz vorhanden ist, als das Rasch-Modell auf Grundlage der bestimmten Personenfähigkeiten für diese Aufgabe erwarten würde (Bond & Fox, 2007, S. 239). Dieser sogenannte *Underfit* resultiert z.B. daraus, dass sehr fähige Personen entgegen der Erwartung eine leichte Aufgabe nicht lösen können, oder dass weniger fähige Personen plötzlich eine sehr schwere Aufgabe lösen und damit eine zu geringe Abhängigkeit des Antwortverhaltens von der Personenfähigkeit existiert (Rost, 2004, S. 374). Ein $MnSq < 1$ bedeutet hingegen, dass weniger Varianz in den Daten vorhanden ist, als das Rasch-Modell vorhersagen würde. Die Daten passen also „zu gut“ ins Modell. Der *Overfit* kann ein Hinweis auf Verletzung der lokalen stochastischen Unabhängigkeit sein, da das Antwortverhalten für eine Aufgabe unerwartet präzise durch das Antwortverhalten im Rest der Aufgaben beschrieben werden kann (Bond & Fox, 2007, S. 241).

Man unterscheidet zwei verschiedene *MnSq*-Statistiken, den *Outfit* und den *Infit*. Bei der Berechnung des Outfits wird ungewichtet über die Residuen, d.h. über die Abweichung zwischen vorhergesagten und beobachteten Daten, summiert (Bond & Fox, 2007, S. 238). Der Outfit reagiert daher sensibler auf Ausreißer und ist vor allem durch unerwartetes Antwortverhalten von Personen an den Rändern der

Fähigkeitsskala beeinflusst (Linacre, 2011, S. 594). Beim Infit handelt es sich um ein gewichtetes Maß, dass sowohl die Residualvarianz als auch die Modellvarianz in der Berechnung mit berücksichtigt (Bond & Fox, 2007, S. 238). Der Infit zeigt damit unerwartetes Antwortverhalten von Personen an, deren Fähigkeit nahe an der Schwierigkeit der betrachteten Aufgabe liegt, also genau bei den Personen, bei denen eine Aufgabe besonders präzise messen sollte (Linacre, 2011, S. 596).
Um die statistische Bedeutsamkeit einer Modellabweichung zu ermitteln, können die *MnSq*-Werte in standardisierte t-Werte transformiert werden. Der Erwartungswert der t-Werte bei perfekter Modellpassung liegt bei null. Werte mit $|t| > 2$ indizieren eine statistisch signifikante Modellabweichung, den sogenannten *Misfit* einer Aufgabe (Bond & Fox, 2007, S. 239).

Über DIF-Analysen in verschiedenen zur Stichprobe gehörenden Personengruppen können außerdem Aufgaben identifiziert werden, die die Annahme der spezifischen Objektivität verletzen. Ob eine Aufgabe DIF zeigt, kann auf Grundlage der *DIF.Contrast-Statistik*, der Differenz der in zwei unterschiedlichen Personengruppen (z.B. Gymnasial-und Gesamtschüler) bestimmten Aufgabenschwierigkeiten, entschieden werden. Ob der DIF statistisch bedeutsam ist, kann durch Transformation des DIF.Contrast in Welchs $t$-Werte ermittelt werden (Linacre, 2011, S. 416). Tabelle 7.2 zeigt die in dieser Arbeit verwendeten Kriterien zur Überprüfung der Modellpassung.

Tabelle 7.2.
*Kriterien zur Prüfung der Modelpassung der eingesetzten Aufgaben*

| Kriterium | Grenzwert | Signifikanz | Quellenangabe |
|---|---|---|---|
| Underfit | $MnSq_{\text{In/Out}} > 1.2$ | $\|t\| > 2$ | Bond & Fox, 2007, S. 243 |
| Overfit | $MnSq_{\text{In/Out}} < 0.8$ | $\|t\| > 2$ | Bond & Fox, 2007, S. 243 |
| DIF | $\|DIF.Contrast\| > 0.64$ | $\|t\| > 2$ | Linacre, 2011, S. 417 |

**Verwendetes Schätzverfahren** Die im Rahmen dieser Arbeit durchgeführten Rasch-Analysen werden mit dem Programm Winsteps 3.72.3 durchgeführt (Linacre, 2011). Winsteps verwendet zur Bestimmung der Personen-und Aufgabenparameter sogenannte *Unbedingte Maximum-Likelihood-Schätzer* (UML-Schätzer), d.h. die Likelihoodfunktion, die die Wahrscheinlichkeit der beobachteten Daten beschreibt, hängt sowohl von den Personen- als auch von den Aufgabenparametern ab und beide Parameter werden im Rahmen der Analyse gemeinsam geschätzt (Rost, 2004, S. 309). Vor-und Nachteile dieses Verfahrens gegenüber anderen Schätzverfahren sind in Linacre (2011, S. 553) und Rost (2004, 309ff.) beschrieben.

### 7.4.3. Reliabilitätsberechnungen

Die Reliabilität ist ein Maß für die Zuverlässigkeit eines Testinstruments. Sie gibt an, wie präzise das Testinstrument misst und ist über den Anteil der wahren (also

nicht messfehlerbehafteten) Varianz an der beobachteten Varianz definiert (Bortz & Döring, 2006, S. 196). In dieser Arbeit werden der PK-Test, der Fragebogen zum situationalen Interesse der Lernenden und das Rating zur kognitiven Aktivierung klassisch ausgewertet, während die Aufgaben des PCK- und CK-Tests und des Kognitive Fähigkeiten Tests im Rahmen von Rasch-Modellen analysiert werden. In diesem Abschnitt werden daher die verschiedenen Verfahren zur Reliabilitätsbestimmung in der klassischen und der probabilistischen Testtheorie erläutert und verglichen.

In der klassischen Testtheorie wird als Maß für die Reliabilität die *interne Konsistenz* betrachtet, die über *Cronbachs Alpha* ($\alpha_C$) beschrieben werden kann. Cronbachs Alpha schätzt die wahre Varianz in den Personenfähigkeiten auf Grundlage der Korrelationen zwischen den Testaufgaben. Der Alphakoeffizient entspricht der Korrelation zwischen zwei Testhälften, gemittelt über alle möglichen Testhalbierungen (Bortz & Döring, 2006, S. 198). Eine Voraussetzung für eine korrekte Schätzung der Reliabilität durch Cronbachs Alpha ist die wechselseitige Unkorreliertheit der Messfehler der Testaufgaben. In der Praxis ist diese meist nicht gegeben, da zum Messfehler beitragende Störfaktoren, wie z.B. Motivation, Prüfungsstress oder Tagesform, Einfluss auf die Bearbeitung aller Testaufgaben haben. Korrelierte Fehlerterme führen daher in der Regel zu einer Überschätzung der Reliabilität durch Cronbachs Alpha (Bortz & Döring, 2006, S. 199). Klassische Reliabilitätsanalysen werden in dieser Arbeit mit SPSS durchgeführt. Da die Bestimmung von Konfidenzintervallen für Cronbachs Alpha in SPSS nicht möglich ist, werden diese mit Hilfe der Statistik-Software R berechnet (R Core Team, 2015). Hierfür wird das R-Paket „psych“ (Revelle, 2015) benutzt.[3]

In der IRT unterscheidet man zwischen der *Personenreliabilität* und der *Itemreliabilität*. Die Personenreliabilität beschreibt, wie in der klassischen Testtheorie, den Anteil der beobachteten Varianz in den Personenparametern, der auf wirkliche Personenunterschiede zurückgeht (Rost, 2004, S. 39). Sie ist damit auch ein Maß dafür, wie wahrscheinlich es ist, dass die Rangfolge von Personen auf der latenten Personenfähigkeitsskala über ihre Messergebnisse beschrieben werden kann (Linacre, 2011, S. 618). Im Rahmen von IRT-Modellen wird die wahre Varianz über die Erwartungswerte der Standardschätzfehler der Personenparameter direkt geschätzt (Linacre, 2011, S. 618, Rost, 2004, S. 380). Die Rasch-Personenreliabilitäten sind in der Regel niedriger als Cronbachs Alpha, da die wahre Reliabilität im Rasch-Modell eher unterschätzt wird (Linacre, 2011, S. 619). Im Gegensatz zu Cronbachs Alpha kann die Rasch-Reliabilität auch für unvollständige Datensätze berechnet werden. Fehlende Werte können die Reliabilität eines Testinstrumentes allerdings verringern (Linacre, 2011, S. 618). Die Personenreliabilität hängt vor allem von der Varianz der Personenfähigkeiten und von der Anzahl der Testaufgaben ab (Linacre, 2011, S. 618). Im Rasch-Modell kann außerdem die Itemreliabilität geschätzt werden, die den Anteil der wahren Varianz an der beobachteten Varianz in den Aufgabenschwierigkeiten beschreibt (Linacre, 2011, S. 619). Die Itemreliabilität erhöht sich

[3]Standardfehler für Cronbachs Alpha werden von „psych“ nicht ausgegeben. Die berichteten Werte für Cronbachs Alpha werden daher, wie in anderen Arbeiten üblich, stets mit zwei signifikanten Stellen angegeben.

mit steigender Varianz in den Aufgabenschwierigkeiten und mit steigender Anzahl untersuchter Personen (Linacre, 2011, S. 618).

Winsteps berechnet obere (Model) und untere (Real) Grenzwerte für die wahren Personen- und Itemreliabilitäten. Da die Reliabilität der UML-Schätzer im Rasch-Modell eher unterschätzt wird (Linacre, 2011, S. 619) und die Personenreliabilitäten oft mit Cronbachs Alpha verglichen werden, werden im Rahmen dieser Arbeit zwar beide Grenzwerte angegeben, die Interpretation bezieht sich aber stets auf die Model-Reliabilitäten.

Bezüglich der Bewertung von Reliabilitäten gibt es unterschiedliche Ansichten. Nach Bortz und Döring (2006, S. 199) gelten Reliabilitäten ab .8 als mittelmäßig und ab .9 als hoch. Nach Lamberti (2001, S. 31) sind Reliabilitäten von über .5 gerade noch als ausreichend und Reliabilitäten von .75 bereits als gut zu bezeichnen.

### 7.4.4. Beurteilung von Interrater-Übereinstimmungen

In Abschnitt 7.5.1.5 auf Seite 106 und Abschnitt 7.6.6 auf Seite 142 zur Objektivität der Testinstrumente zum fachspezifischen Professionswissen und des Videoratings zur kognitiven Aktivierung wird über die Übereinstimmung zwischen verschiedenen Kodierern bzw. Ratern berichtet. Die Bepunktung der PCK- und CK-Aufgaben erfolgt auf einer dreistufigen Skala von null bis zwei Punkten, die Ratingskala zur kognitiven Aktivierung ist ebenfalls dreistufig. Beide Skalen werden als intervallskaliert angenommen, obwohl nicht eindeutig entschieden werden kann, ob es sich um ordinal- oder intervallskalierte Skalen handelt. Für intervallskalierte Daten kann die Interrater-Übereinstimmung über die sogenannte *Intraklassenkorrelation* (ICC) bestimmt werden. Liegen entgegen der Annahme lediglich ordinalskalierte Daten vor, kann die Interrater-Übereinstimmung durch die ICC allerdings deutlich unterschätzt werden (vergl. Wirtz & Caspar, 2002, S. 126). Sofern im Rahmen der weiteren Auswertung parametrische Methoden zum Einsatz kommen, empfehlen Wirtz und Caspar (2002, S. 127) allerdings auch für die Beurteilung der Interrater-Übereinstimmung die Anwendung parametrischer Verfahren. Daher wird zur Beurteilung der Übereinstimmung zwischen den Kodierern der PCK- und CK-Aufgaben und zwischen den Ratern der kognitiven Aktivierung in den Unterrichtsvideos die ICC verwendet.

Die ICC ist ein Maß für den Anteil der Varianz in den Raterurteilen, der durch Unterschiede in den wahren Werten der beurteilten Objekte erklärt werden kann (vergl. Wirtz & Caspar, 2002, S. 190). Die Berechnung der ICC setzt Varianzhomogenität und angenähert normalverteilte Daten voraus (vergl. Wirtz & Caspar, 2002, 160ff.). Shrout und Fleiss (1979) unterscheiden sechs verschiedene ICCs. Die Wahl des ICCs hängt zum einen davon ab, ob die Beurteilungseinheiten durch verschiedene Rater geratet wurden (einfaktorielles Modell) oder ob alle Beurteilungseinheiten von allen Ratern geratet wurden (zweifaktorielles Modell). Im zweifaktoriellen Modell kann zusätzlich spezifiziert werden, ob die Übereinstimmung einer Grundgesamtheit von Ratern (Rater-fixed) oder einer zufällig gezogenen Stichprobe aus der Grundgesamtheit der Rater (Rater-random) betrachtet werden soll. Außerdem kann mit der ICC sowohl die Reliabilität der Skalenwerte eines einzelnen Raters beschrieben werden als auch die Reliabilität eines über alle Rater gemittelten

Ratings (vergl. Shrout & Fleiss, 1979; Wirtz & Caspar, 2002). Schlussendlich kann die absolute Übereinstimmung in den Ratings (unjustierte ICC) oder lediglich die Konsistenz der Ratings (justierte ICC) bestimmt werden, je nachdem, ob die Skalenwerte unabhängig vom jeweiligen Rater oder lediglich relativ zu anderen, durch den jeweiligen Rater vergebenen, Skalenwerten interpretiert werden sollen (vergl. Wirtz & Caspar, 2002, S. 190).

In dieser Studie wurden alle Testhefte und alle Videos von allen Ratern beurteilt. Die Rater stellen lediglich eine zufällige Auswahl für die Grundgesamtheit der Rater dar, zu der theoretisch jeder Forscher gehört, der die verwendeten Instrumente einsetzen will. Die generierten Daten sollen unabhängig vom jeweiligen Rater beurteilt werden. Die Beurteilung der Interrater-Übereinstimmung erfolgt daher über die $ICC_{\text{2-fakt.,unjust}}$ (Shrout & Fleiss, 1979) für zufällig ausgewählte Rater und bezieht sich auf die Skalenwerte der einzelnen Rater und nicht auf die über alle Rater gemittelten Skalenwerte.

ICCs ab $> .7$ können als gut bezeichnet werde und lassen einen Gruppenvergleich auf Basis von Ratingdaten zu (Wirtz & Caspar, 2002, S. 25, 234). Der wahre Wert von Personen kann allerdings erst ab ICCs $> .85$ ausreichend präzise durch ein Rating beschrieben werden (Wirtz & Caspar, 2002, S. 234). Eine Differenzierung zwischen Personen auf individueller Ebene sollte erst ab ICCs $> .9$ erfolgen (Wirtz & Caspar, 2002, S. 199). Wirtz und Caspar (2002, S. 234) merken allerdings an, dass im Falle geringer Übereinstimmungen (trotz intensiven Ratertrainings) die Anwendung eines Ratings dennoch sinnvoll sein kann, sofern kein reliableres Bewertungsinstrument für das interessierende Merkmal zur Verfügung steht.

### 7.4.5. Mehrebenenanalysen

In dieser Arbeit wird untersucht, inwieweit Unterschiede im Fachwissen der Lernenden am Ende der Unterrichtseinheit Mechanik und im situationalen Interesse der Lernenden am Unterricht durch das Professionswissen der Lehrkräfte und die kognitiv aktivierende Gestaltung des Unterrichts erklärt werden können. Das Schülerfachwissen wird sowohl als Funktion von Variablen auf Schülerebene als auch als Funktion von Variablen auf Klassenebene modelliert. Den im Rahmen dieser Studie erhobenen Daten liegt allerdings eine hierarchische Datenstruktur zugrunde, da die Lernenden in Schulklassen gruppiert sind. Die Stichprobe kann daher lediglich auf Klassenebene (und auch hier nur eingeschränkt, da die Klassen nicht zufällig gezogen wurden, sondern freiwillig an der Studie teilnehmen konnten) als Zufallsstichprobe betrachtet werden. Auf Schülerebene ist die Annahme, dass es sich um eine Zufallsstichprobe unabhängiger Beobachtungseinheiten handelt, nicht gerechtfertigt. Lernende innerhalb einer Klasse sind einander hinsichtlich schulleistungsrelevanter Merkmale oftmals ähnlicher als Schülerinnen und Schüler verschiedener Klassen, schließlich werden sie von derselben Lehrkraft unterrichtet, interagieren untereinander und können sich auch bezüglich weiterer Merkmale, wie z. B. bzgl. ihres sozialen Hintergrunds ähneln.

Diese Ähnlichkeit kann über die in Abschnitt 7.4.4 auf Seite 98 eingeführte ICC (einfaktorielles unjustiertes Modell für die einzelnen Skalenwerte) beschrieben werden (J. Cohen, Cohen, West & Aiken, 2003, S. 537). Die Leistungsdaten der

Lernenden werden dabei als Beurteilung für das mittlere Leistungsniveau der Klasse interpretiert. Hohe ICCs indizieren daher hohe Ähnlichkeit innerhalb der Klassen und damit große Unterschiede zwischen den Klassen. Die ICC misst hier den Anteil der Gesamtvarianz in den Leistungsdaten, der durch die Klassenzugehörigkeit erklärt werden kann (J. Cohen et al., 2003, S. 537). Schon ICCs von .05 oder .01 können dazu führen, dass die auf Schülerebene erhobenen Daten nicht mehr als unabhängig voneinander betrachtet werden können (vergl. J. Cohen et al., 2003, S. 537). Die Unabhängigkeit der Beobachtungseinheiten ist allerdings eine wichtige Voraussetzung für die Durchführbarkeit von herkömmlichen Regressionsanalysen (Geiser, 2011, S. 199). Bei der Schätzung der Standardfehler auf die Regressionskoeffizienten wird außerdem davon ausgegangen, dass die Daten in einer Zufallsstichprobe erhoben wurden (Hartig, Jude & Wagner, 2008, S. 45).

Wird die hierarchische Struktur der Daten ignoriert, führt dies zum einen zu einer Unterschätzung der Standardfehler der Regressionskoeffizienten, zu inkorrekt geschätzten Konfidenzintervallen und zu einer Überschätzung der Signifikanz von Regressionskoeffizienten (Geiser, 2011, S. 200). Zum anderen kann die Vernachlässigung einer hierarchischen Datenstruktur zu Fehlschlüssen in der Interpretation von Ergebnissen führen (vergl. Langer, 2009, 21ff.). Ein stark vereinfachtes Beispiel hierfür ist die Untersuchung von Leistungsunterschieden zwischen Jungen und Mädchen in zwei Schulklassen, die von einer „guten“ und einer „schlechten“ Lehrkraft unterrichtet werden. Angenommen, der Anteil der Mädchen ist in der Klasse der „guten“ Lehrkraft ist wesentlich höher als in der Klasse der „schlechten“ Lehrkraft. Würde man bei der Auswertung der Schülerdaten die Klassenzugehörigkeit nicht berücksichtigen, käme man zu dem Fehlschluss, dass Mädchen wesentlich bessere Leistungen zeigen als Jungen. Ursache für die beobachteten Leistungsunterschiede sind aber nicht die Geschlechterunterschiede auf Individualebene, sondern ein Merkmal auf Klassenebene, nämlich die Qualität des Lehrangebots.

Um die hierarchische Struktur der Daten zu berücksichtigen, können Regressionsanalysen im Rahmen von hierarchischen linearen Modellen, sogenannten *Mehrebenenmodellen*, durchgeführt werden (Geiser, 2011). Prinzipiell kann man sich eine Mehrebenenanalyse als eine Reihe geschachtelter Regressionsanalysen vorstellen, in denen die Regressionskoeffizienten auf Schülerebene als abhängige Variablen in die Analysen auf Klassenebene eingehen (Nezlek, Schröder-Abé & Schütz, 2006). Die eigentliche Regressionsgleichung enthält sowohl die Variablen auf Schülerebene als auch die Variablen auf Klassenebene. Die Regressionskoeffizienten werden nicht für alle Schülerinnen und Schüler gemeinsam geschätzt, sondern können zwischen den Klassen variieren (Nezlek et al., 2006). In *Random-Coefficients-Regressionsmodellen* können außerdem, zusätzlich zu den Regressionskoeffizienten, den festen Effekten, Zufallseffekte geschätzt werden, die den Zufallsfehler der Variation der Regressionskoeffizienten zwischen den Klassen beschreiben (vergl. J. Cohen et al., 2003, S. 550, Nezlek et al., 2006). Die Modellierung mit Zufallseffekten ist in den meisten Fällen vorzuziehen, da hier dem Umstand Rechnung getragen wird, dass es sich auch bei den Klassen lediglich um eine Zufallsstichprobe handelt. Die beobachtete Varianz der Regressionskoeffizienten zwischen den Klassen kann nur mit einer gewissen Wahrscheinlichkeit in der Grundgesamtheit der Klassen beobachtet werden und ist daher mit einem Zufallsfehler behaftet (vergl. Nezlek et al., 2006).

Die korrekte Schätzung der Zufallseffekte auf Klassenebene ist allerdings erst für große Stichproben (ab $N = 50$ Klassen) möglich, in kleineren Stichproben werden die Fehler meist unterschätzt. Die Regressionskoeffizienten und deren Standardfehler auf Schülerebene können bereits ab einer Stichprobengröße von 10 Klassen akkurat geschätzt werden. Für die Schätzung fester Effekte auf Klassenebene mit zufriedenstellender Genauigkeit werden allerdings Stichprobengrößen von 30 Klassen empfohlen (Maas & Hox, 2004; Maas & Hox, 2005).

Die im Rahmen dieser Arbeit erhobenen Daten werden unter Berücksichtigung der hierarchischen Datenstruktur ausgewertet, da 10% der Gesamtvarianz in den Schülerposttestdaten und je nach betrachteter Unterrichtsstunde zwischen $17 - 20\%$ der Gesamtvarianz im situationalen Interesse der Lernenden zwischen den Klassen liegt (vergl. Abschnitt 8.3.1 und Abschnitt 8.3 auf Seite 170 und auf Seite 171). Da die untersuchte Stichprobe von 23 Klassen sehr klein ist, werden die Standardfehler auf die Regressionskoeffizienten wahrscheinlich unterschätzt. Die im Rahmen der Mehrebenenanalyse generierten Ergebnisse sollten daher mit Vorsicht interpretiert werden. Für herkömmliche Regressionsanalysen sollten nach Field (2009, S. 222) für jede ins Modell aufgenommene erklärende Variable mindestens $10 - 15$ Beobachtungseinheiten vorhanden sein, wobei die Anzahl der benötigten Fälle von der erwarteten Varianzaufklärung durch die jeweiligen Variablen abhängig ist. Überträgt man diese Empfehlungen auf die mehrebenenanalytische Auswertung der $N = 23$ Klassen, sollten auf Klassenebene keinesfalls mehr als zwei erklärende Variablen in die Mehrebenenmodelle aufgenommen werden. Für die Durchführung der Mehrebenenanalysen wird das Programm Mplus (L. K. Muthén & Muthén, 2007) genutzt (Type=TWOLEVEL, Maximum-Likelihood-Schätzung mit robusten Standardfehlern (MLR)). Eine Beispielsyntax für jeweils ein Mehrebenenmodell für die Post-Testwerte bzw. die Maße zum situationalen Interesse der Lernenden findet sich in Abbildung B.5 auf Seite 253 im Anhang.

### 7.4.6. Messfehlerbereinigte Korrelationen

Um Aussagen über die Validität der fachspezifischen Professionswissenstests zu treffen, werden in Abschnitt 7.5.1.7 auf Seite 109, im Rahmen einer konvergenten Validierung, Korrelationen zwischen den in der ersten und zweiten Phase des ProwiN-Projekts gemessenen PCK- und CK-Testwerten der Lehrkräfte berechnet. Über die Berechnung von Korrelationen erfolgt im Zuge der Validierung des Schülerfachwissenstests in Abschnitt 7.5.3.5 auf Seite 126 eine Abgrenzung des mit dem Schülerfachwissenstest erfassten Konstrukts zum Konstrukt der Intelligenz. In Abschnitt 7.5.3.5 auf Seite 126 werden aus den Korrelationen zu den Schulnoten der Lernenden Rückschlüsse auf die Kriteriumsvalidität des Schülerfachwissenstests gezogen.

Die „wahren“ Korrelationen werden in messfehlerbehafteten Messungen allerdings stets unterschätzt. Das ist plausibel, schließlich sollte ein Messwert mit keinem anderen Wert höher korrelieren als mit seinem eigenen wahren (messfehlerfreien) Wert. Die Quadratwurzel der Reliabilität eines Testinstruments entspricht gerade der Korrelation zwischen dem fehlerbehafteten Messwert und seinem wahren Wert. Die Höhe der Korrelation zwischen zwei Testinstrumenten ist damit theoretisch

durch die Quadratwurzel der Reliabilität des weniger reliablen Testinstruments begrenzt (Rost, 2004, 389f).

In den genannten Abschnitten, in denen über die Betrachtung von Korrelationen Rückschlüsse auf die Validität der Erfassung der untersuchten Konstrukte gezogen werden, ist es von Interesse, wie hoch die Korrelation im Falle einer fehlerfreien Messung ausgefallen wären. Über eine sogenannte Minderungskorrektur kann die Höhe der Korrelation für den Fall einer fehlerfreien Messung geschätzt werden, indem man die Korrelation durch die Quadratwurzeln der Reliabilitäten der Testinstrumente teilt (Rost, 2004, S. 390). Zusätzlich zu den Korrelationen zwischen messfehlerbehafteten Werten werden in Abschnitt 7.5.1.7 auf Seite 109 und Abschnitt 7.5.3.5 auf Seite 123 daher an den erforderlichen Stellen die bereinigten Korrelationen angegeben.

## 7.5. Beschreibung der schriftlichen Erhebungsinstrumente

In diesem Abschnitt werden die im Rahmen der zweiten Phase des ProwiN-Projekts eingesetzten Erhebungsinstrumente beschrieben. In Abschnitt 7.5.1 und Abschnitt 7.5.2 auf Seite 112 werden die Testinstrumente zur Erfassung des fachspezifischen und pädagogischen Professionswissens von Physiklehrkräften vorgestellt. In Abschnitt 7.5.3 auf Seite 116 wird der Schülerfachwissenstest und in Abschnitt 7.5.4 auf Seite 128 der Fragebogen zum situationalen Interesse der Lernenden am Unterricht vorgestellt. Die Abschnitte sind folgendermaßen gegliedert: Zunächst wird das vorgestellte Testinstrument kurz beschrieben, anschließend werden technische Details zur Auswertung der erhobenen Daten erläutert. Am Ende jedes Abschnitts erfolgt eine Diskussion der Testgüte. In Abschnitt 7.5.5 auf Seite 130 wird außerdem auf die Erfassung der Kontrollvariablen eingegangen und das Testinstrument zur Messung der kognitiven Fähigkeiten der Lernenden, der Kognitive Fähigkeitentest (KFT), kurz beschrieben.

### 7.5.1. Tests zur Messung des fachspezifischen Professionswissens

Das fachspezifische Professionswissen der Physiklehrkräfte wurde mit einem in der ersten Projektphase des ProwiN-Projekts entwickelten Testinstrument erfasst. Das Testinstrument besteht aus zwei unabhängigen Papier-und-Bleistift-Tests zum PCK und CK der Lehrkräfte und wurde von Kirschner (2013) im Rahmen ihres Dissertationsprojekts entwickelt und validiert. Aus zeitökonomischen Gründen wurden in dieser Studie gekürzte Versionen des PCK- und CK-Tests eingesetzt.

Im Folgenden werden zunächst die in dieser Studie eingesetzten gekürzten PCK- und des CK-Tests beschrieben. Anschließend werden technische Details zur Auswertung der Tests erläutert. Da das PCK und CK der Physiklehrkräfte als Personenfähigkeiten im Rasch-Modell geschätzt werden, wird hier auch die Rasch-Analyse der Daten beschrieben. Außerdem wird auf Unterschiede im

Testinstrument und in der Auswertung der Tests im Vergleich zu dem in der ersten Projektphase validierten Testinstrument eingegangen. Abschließend wird die Objektivität, Reliabilität und die Validität des Testinstruments diskutiert.

#### 7.5.1.1. PCK-Test

Mit dem Test zur Erhebung des fachdidaktischen Wissens der Lehrkräfte wurde deklaratives, prozedurales und konditionales Wissen über Schülervorstellungen, Experimente und Konzepte erfasst. Bezüglich des Fachinhalts liegt der Schwerpunkt des PCK-Tests auf dem Fachinhalt Mechanik. Der Test beinhaltet allerdings auch eine Aufgabe zur Elektrizitätslehre, sowie vier Aufgaben, die fachspezifische, aber nicht themenspezifische Sachverhalte zum Inhalt haben. Der PCK-Test umfasst insgesamt 11 Aufgaben: eine Multiple-Choice-Aufgabe (Multiple-Select, 6 Antwortmöglichkeiten), sieben offene Aufgaben, davon zwei Speed-Aufgaben mit einer Minute Bearbeitungszeit, und drei offene Aufgaben, die aus zwei Aufgabenteilen bestehen. In Letzteren umfasst der erste Aufgabenteil beispielsweise die Aufzählung gängiger Schülerantworten auf eine Fragestellung und der zweite Aufgabenteil die Angabe typischer Begründungen für diese Antworten. Abbildung 7.4 zeigt eine Beispielaufgabe aus dem PCK-Test: eine offene, fachspezifische, aber nicht themenspezifische Speedaufgabe zum Wissen über Konzepte. Für den PCK-Test war eine feste Bearbeitungszeit von 45 Minuten vorgegeben.

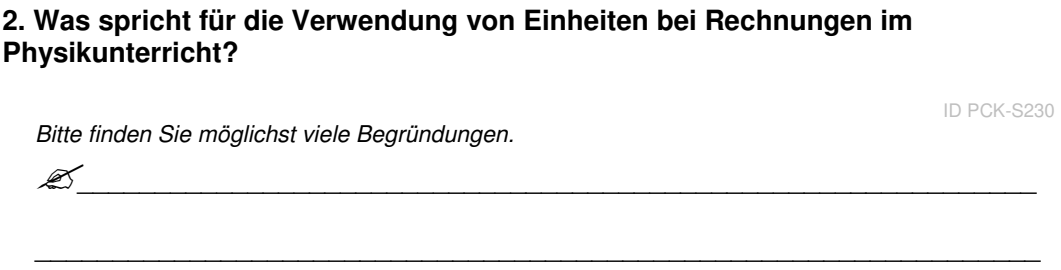



Abbildung 7.4.
*Beispielaufgabe aus dem PCK-Test. „Korrekte Antworten beziehen sich auf die Wissenschaftspropädeutik, die Vermeidung und das Finden von Fehlern und das vertiefte Verständnis von Zusammenhängen. Inkorrekte Antworten beziehen sich beispielsweise auf die reine Übung“ (Kirschner, 2013, S. 45).*

#### 7.5.1.2. CK-Test

Mit dem Testinstrument zur Erhebung des Fachwissens der Lehrkräfte wurde deklaratives, prozedurales und konditionales Schulwissen und vertieftes Schulwissen in Mechanik gemessen. Das Testinstrument umfasst insgesamt 12 Aufgaben: vier Multiple-Choice-Aufgaben (Multiple-Select, 4-5 Antwortmöglichkeiten), eine Aufgabe, bei der fünf physikalische Aussagen als richtig oder falsch bewertet werden müssen, drei Multiple-Choice-Aufgaben (Single-Select, 2-3 Antwortmöglichkeiten), bei denen die Entscheidung für eine Antwortmöglichkeit im Anschluss begründet

werden muss und vier offene Aufgaben, deren Lösungen die Herleitung von Formeln, Berechnungen oder Begründungen erfordern. Für den CK-Test war eine feste Bearbeitungszeit von 40 Minuten vorgegeben.

#### 7.5.1.3. Technische Details zur Auswertung

In diesem Abschnitt wird zunächst die Punktevergabe und der Umgang mit fehlenden Werten in den fachspezifischen Professionswissenstests erläutert. Anschließend wird auf die Rasch-Analyse des CK- und PCK-Tests und die hierfür notwendige Erweiterung der Stichprobe mit Daten aus der ersten Projektphase des ProwiN-Projekts eingegangen.

**Punktevergabe** Die CK- und PCK-Tests wurden mit Hilfe eines Kodiermanuals ausgewertet. Hierfür wurde das von Kirschner im Anschluss an ihre Studie optimierte Kodiermanual im Rahmen einer zweimonatigen Raterschulung überarbeitet und leicht modifiziert (vergl. Liepertz, 2016).
Für jede Aufgabe im CK- und PCK-Test wurden null bis zwei Punkte vergeben. Für die zweiteiligen Aufgaben im PCK-Test galt dieses Bepunktungsschema für jeden Aufgabenteil. Die Punkte der Teilaufgaben wurden addiert, so dass insgesamt null bis vier Punkte in den zweiteiligen Aufgaben vergeben wurden.

**Umgang mit fehlenden Werten** Hohensinn und Kubinger (2011) konnten im Rahmen einer Simulationsstudie zeigen, dass das Behandeln von nicht bearbeiteten Aufgaben als fehlende Werte zu weniger verzerrten Ergebnissen führt als das Bewerten einer solchen Aufgabe als falsch. Für den PCK- und CK-Test waren allerdings jeweils feste Bearbeitungszeiten vorgegeben. Ausgehend von der Annahme, dass Lehrkräfte mit höherem fachspezifischen Professionswissen im Rahmen der Bearbeitungszeit tendenziell mehr Aufgaben bearbeiten können, können nicht bearbeitete Aufgaben in diesem Fall als Ausdruck geringerer Fähigkeit betrachtet werden. Schließlich kann eine Lehrkraft auch in kritischen Unterrichtssituationen nur das Wissen nutzen, das ihr unmittelbar und ohne lange darüber nachdenken zur Verfügung steht. Nicht bearbeitete Aufgaben wurden daher nicht als fehlende Werte behandelt, sondern stattdessen mit null Punkten bewertet.

**Erweiterung der Stichprobe** Die Stichprobe der Lehrkräfte ist mit $N = 23$ zu klein, um im Rahmen einer Rasch-Analyse robuste Schätzungen für die Personenfähigkeiten zu erhalten (vergl. z. B. W.-H. Chen et al., 2014). Die Stichprobe wurde daher mit Daten aus der ersten Projektphase (Kirschner, 2013) erweitert. Hierfür wurde die vergleichbare Teilstichprobe der ebenfalls im Bundesland Nordrhein-Westfalen getesteten Gymnasiallehrkräfte ausgewählt. Die Antworten der $N = 79$ Lehrkräften wurden auf Basis des überarbeiteten Kodiermanuals rekodiert. Die Schätzung der Personenfähigkeiten konnte somit auf der Datengrundlage von insgesamt $N = 102$ Lehrkräften durchgeführt werden.

**Rasch-Analyse** Das PCK und CK der Lehrkräfte wurde jeweils als Personenfähigkeit im Rahmen eines eindimensionalen Partial-Credit-Modells geschätzt. Das Partial-Credit-Modell stellt eine Verallgemeinerung des dichotomen Rasch-Modells (vergl. Abschnitt 7.4.2 auf Seite 93) auf ordinale Antwortkategorien dar (Carstensen, 2000, S. 47). Die Aufgaben im PCK- und CK-Test wurden nicht dichotom ausgewertet (richtig/falsch), sondern mit null bis maximal vier Punkten bewertet. Die Idee des Partial-Credit-Modells besteht darin, die Wahrscheinlichkeit des Übergangs von einer Antwortkategorie zur nächsthöheren Kategorie (z.B. von null auf einen Punkt oder von einem auf zwei Punkte) mit Hilfe des Rasch-Modells zu beschreiben (Strobl, 2012, S. 55).

Für den PCK- und den CK-Test wurden zwei separate Rasch-Analysen durchgeführt. Hierfür wurde die erweiterte Stichprobe der $N = 102$ Gymnasiallehrkräfte genutzt. In den Analysen zeigte sowohl im PCK- als auch im CK-Test jeweils eine Aufgabe eine schlechte Passung ins Rasch-Modell (vergl. Abschnitt 7.4.2 auf Seite 93). Diese Aufgaben (PCK_0040, CK_1450) wurden daher aus den Rasch-Analysen zur Schätzung der Personenfähigkeiten ausgeschlossen. In Abbildungen B.1 und B.2 auf Seite 250 und auf Seite 251 im Anhang finden sich die Wright-Maps für die Aufgaben des PCK- und CK-Tests.

#### 7.5.1.4. Unterschiede zum Testinstrument aus ProwiN I

In diesem Abschnitt wird erläutert, welche Änderungen im Rahmen der zweiten Phase des ProwiN-Projekts (ProwiN II) an dem in der ersten Projektphase (ProwiN I) entwickelten und eingesetzten Testinstrument vorgenommen werden mussten. Außerdem wird auf Unterschiede in der Auswertung der Testinstrumente eingegangen.

**Kürzung des Testinstruments** Die im Rahmen der ersten Phase des ProwiN-Projekts angesetzten Bearbeitungszeiten für den PCK- und CK-Test waren zum Teil knapp bemessen. Aus zeitökonomischen Gründen kam eine Verlängerung der Bearbeitungszeit in den Erhebungen im Rahmen der zweiten Projektphase allerdings nicht in Frage. Die Bearbeitungszeit für den PCK-Test musste um fünf Minuten verkürzt werden, während die Bearbeitungszeit für den CK-Test beibehalten wurde. In dieser Studie wurde daher eine gekürzte Version des Testinstruments zur Erfassung des fachspezifischen Professionswissens eingesetzt.

Die Kürzung der Tests erfolgte auf Basis der zum Zeitpunkt der Testhefterstellung aus ProwiN I vorliegenden Zwischenergebnisse. Die PCK-und der CK-Tests wurden jeweils um drei Aufgaben gekürzt. Im PCK-Test wurden zwei Aufgaben wegen nicht zufriedenstellender Interrater-Übereinstimmung entfernt, im CK-Test eine Aufgabe. Wegen schlechter Passung ins Rasch-Modell wurden im PCK-Test eine Aufgabe und im CK-Test zwei Aufgaben entfernt. Letztere zeigten allerdings in den Analysen des Gesamtdatensatzes aus der ersten Projektphase keine auffällige Fit-Statistik mehr, so dass diese Aufgaben bei Kirschner (2013) für die Berechnung der Lehrerfähigkeiten im fachspezifischen Professionswissen genutzt wurden.

**Unterschiede in der Auswertung** Die Verwendung des in ProwiN I eingesetzten Kodiermanuals führte nicht bei allen Aufgaben zu einer zufriedenstellenden Interrater-Übereinstimmung. Alle Testhefte wurden doppelt kodiert und der Einfluss der Kodiererinnen wurde im Rahmen eines mehrdimensionalen Rasch-Modells mit modelliert (vergl. Kirschner, 2013). Das Kodiermanual wurde daher von Kirschner im Anschluss an ihre Studie optimiert. Im Rahmen einer zweimonatigen Raterschulung wurde das Kodiermanual in der zweiten Projektphase erneut überarbeitet (vergl. Liepertz, 2016). Während in den meisten Fällen lediglich eine Ausschärfung des Erwartungshorizonts erfolgte, musste bei drei Aufgaben im PCK-Test die Bepunktung angepasst werden und bei einer Aufgabe im CK-Test ein fachlicher Fehler korrigiert werden (vergl. Tabelle 7.3 auf der nächsten Seite). Die mit Hilfe des überarbeiteten Manuals erzielte Interrater-Übereinstimmung war für alle Aufgaben zufriedenstellend (vergl. Abschnitt 7.5.1.5). Die Rater-Effekte wurden in der vorliegenden Studie daher nicht mit modelliert, stattdessen wurde die Kodierung eines Kodierers für die Auswertung verwendet.

In der Auswertung des PCK-Tests in ProwiN I (vergl. Kirschner, 2013) wurden die Aufgabenteile der zweiteiligen Aufgaben als separate Aufgaben in die Analysen einbezogen. Da die Aufgabenteile aufeinander aufbauen und demnach nicht unabhängig von einander sind, kommt es bei diesem Vorgehen allerdings zu einer Verletzung der lokalen stochastischen Unabhängigkeit (vergl. Abschnitt 7.4.2 auf Seite 93). In der vorliegenden Studie wurden die Punkte der Teilaufgaben daher addiert, so dass insgesamt null bis vier Punkte in den zweiteiligen Aufgaben erreicht werden konnten.

Im Rahmen der Rasch-Analyse des PCK-Tests wurde die gleiche Aufgabe aufgrund schlechter Modellpassung entfernt wie in der Analyse von Kirschner (2013). In der Rasch-Analyse des CK-Tests zeigten sich allerdings Unterschiede in der Modellpassung. Während in der hier ausgeführten Rasch-Analyse lediglich eine Aufgabe entfernt werden musste, mussten bei Kirschner (2013) drei andere Aufgaben entfernt werden (vergl. Abschnitt 7.5.1.3 auf Seite 104).

Weitere Unterschiede in der Auswertung bestehen in den verwendeten Schätzverfahren im Rahmen der Rasch-Analysen und im Umgang mit fehlenden Werten (vergl.Abschnitt 7.5.1.3 auf Seite 104), die bei Kirschner (2013) als Missings behandelt wurden.

Tabelle 7.3 auf der nächsten Seite zeigt einen Vergleich der Aufgaben, auf deren Basis die Berechnung der Lehrerfähigkeiten im fachspezifischen Professionswissen in der ersten und zweiten Projektphase erfolgte.

#### 7.5.1.5. Objektivität

Damit ein Test als objektives Messinstrument gilt, muss gewährleistet sein, dass bei einer Untersuchung derselben Testpersonen durch verschiedene Testleiter und Auswertende gleiche Ergebnisse erzielt werden. In diesem Abschnitt wird diskutiert, inwieweit das Testinstrument zur Erhebung des fachspezifischen Professionswissens die Objektivitätsanforderungen bezüglich der drei Unterformen der Objektivität, der *Durchführungsobjektivität*, der *Auswertungsobjektivität* und der *Interpretationsobjektivität* erfüllt (vergl. Bortz & Döring, 2006, S. 195).

Tabelle 7.3.
*Übersicht über die in ProwiN I und ProwiN II zur Berechnung der Lehrerfähigkeiten im fachspezifischen Professionswissen hinzugezogenen Aufgaben*

| Aufgabe | Beschreibung | ProwiN I | ProwiN II |
|---|---|---|---|
| PCK_S020 | Warum Experimente | x | x[1] |
| PCK_S230 | Warum Einheiten | x | x |
| PCK_0261 | Lok | x | x |
| PCK_0051 | Diagramm 1 | x | kombiniert |
| PCK_0052 | Diagramm 2 | x | |
| PCK_0151 | Flugbahn 1 | x | kombiniert |
| PCK_0152 | Flugbahn 2 | x | |
| PCK_0071 | Lampe 1 | x | kombiniert |
| PCK_0072 | Lampe 2 | x | |
| PCK_0031 | Anknüpfen an Schülervorstellungen | x | - |
| PCK_0180 | Schülervorstellungen Geschwindigkeit | x | x |
| PCK_0280 | Wirkung von Kraft | x | x |
| PCK_0320 | Zeichnung Kraft | x | x[1] |
| PCK_0080 | Wasserrakete | x | x |
| PCK_0040 | Stundenfortsetzung Experiment | - | - |
| CK_1150 | Rutsche | - | x |
| CK_1512 | Schuss | x | - |
| CK_1460 | Rennstrecke | x | - |
| CK_1160 | Flugzeug Wind | x | x |
| CK_1240 | Puk | - | x |
| CK_1490 | E-Lehre | x | x |
| CK_1450 | Hebel | x | - |
| CK_1410 | Ampel | x | x[2] |
| CK_1300 | Beschleunigung | - | x |
| CK_1470 | Looping | x | x |
| CK_1290 | Schaukel | x | x |
| CK_1180 | Kepler | x | x |
| CK_1140 | Pendel | x | x |
| CK_1220 | LKW | x | x |
| Anzahl Aufgaben PCK | | 14 | 10 |
| Anzahl Aufgaben CK | | 11 | 11 |
| Anzahl identischer Aufgaben PCK | | 7 | |
| davon identisch ausgewertet | | 5 | |
| Anzahl identischer Aufgaben CK | | 8 | |
| davon identisch ausgewertet | | 7 | |

[1] Bepunktung angepasst
[2] fachlicher Fehler in der Musterlösung korrigiert

Die Durchführungsobjektivität ist durch die standardisierte Testdurchführung mit Testleitermanualen sichergestellt (vergl. Anhang A.1.3 auf Seite 220). Eine ausführliche Beschreibung der Testdurchführung findet sich in Abschnitt 7.2.3 auf Seite 85.

Um die Auswertungsobjektivität zu gewährleisten, wurden die Testinstrumente zum fachspezifischen Professionswissen mit Hilfe eines Kodiermanuals von zwei unabhängigen Kodierern ausgewertet (vergl. Liepertz, 2016), dabei handelte es sich um den zweiten im Projekt arbeitenden Physikdoktoranden und eine studentischen Hilfskraft (Lehramt Physik an Gymnasien und Gesamtschulen). Alle Antworten der $N = 102$ Gymnasiallehrkräfte wurden von beiden Kodierern ausgewertet. Mit Hilfe der unjustierten zweifaktoriellen ICC wurde die Interrater-Übereinstimmung bestimmt (vergl. Abschnitt 7.4.4 auf Seite 98 im Kapitel zu statistischen Methoden).

Die ICCs waren für alle Aufgaben gut bis sehr gut (vergl. Tabelle B.4 auf Seite 245 im Anhang). Für die Auswertung konnte daher die Kodierung eines Kodierers genutzt werden. Hierfür wurde die Kodierung der studentischen Hilfskraft ausgewählt, damit diese im Zuge der Fortführung der Datenerhebung für das Rahmenprojekt die Kodierung der Professionswissenstests alleine durchführen konnte. Im CK-Test sind die ICCs hoch genug, um zwischen Personen auf Individualebene differenzieren zu können ($ICC_{\text{2-fakt.,unjust}} \geq .96$). Bis auf drei Aufgaben mit $ICC_{\text{2-fakt.,unjust}} \geq .85$ und einer Aufgabe mit $ICC_{\text{2-fakt.,unjust}} = .77$ ($\text{KI}_{95\,\%} = [0.69, 0.84]$), die streng genommen nur Gruppenvergleiche zulassen, erfüllten dieses Kriterium auch alle Aufgaben im PCK-Test. Durch die Verwendung der Daten eines Kodierers wird vermieden, dass die Aufgaben durch verschiedene Kodierer unterschiedlich streng bewertet werden. Darüber hinaus erfolgt eine Differenzierung auf Individualebene nur indirekt über die Berechnung von Korrelationen. Daher wird die Interrater-Übereinstimmung auch für die PCK-Aufgaben mit $ICC_{\text{2-fakt.,unjust}} \leq .9$ als ausreichend hoch betrachtet und es müssen keine Aufgaben aus den Analysen ausgeschlossen werden.

Das Wissen der Lehrkräfte wird nur innerhalb der Stichprobe miteinander verglichen. Es erfolgt keine Bewertung der absoluten Personenfähigkeiten. Daher kann auch die Interpretationsobjektivität als gewährleistet betrachtet werden (vergl. Bortz & Döring, 2006, S. 195).

#### 7.5.1.6. Reliabilität

Die Reliabilitäten der Testinstrumente wurden im Rahmen einer Rasch-Analyse bestimmt. Das genaue Vorgehen zur Bestimmung der Reliabilitäten sowie die hierfür notwendige Erweiterung der Stichprobe mit Daten aus der ersten Projektphase sind in Abschnitt 7.5.1.3 auf Seite 104 beschrieben. Tabelle 7.4 auf der nächsten Seite zeigt die Reliabilitäten der einzelnen Testinstrumente zur Messung des Professionswissens.

Bei der Bewertung der im Rasch-Modell geschätzten Personenreliabilitäten ist zu beachten, dass diese meist niedrigere Werte annehmen als klassisch berechnete Reliabilitäten wie Cronbachs Alpha (vergl. Abschnitt 7.4.3 auf Seite 96). Die Reliabilität des CK-Tests kann als zufriedenstellend bezeichnet werden, die Reliabilität des PCK-Tests als ausreichend (vergl. Lamberti, 2001, S. 31).

Die im Rahmen der Analysen von Kirschner (2013, S. 75) für die Lehrkräfte der Hauptstudie ($N = 186$) geschätzten Personenreliabilitäten sind sowohl für den PCK-Test (.77) als auch für den den CK-Test (.86) deutlich höher. Die beobachteten Unterschiede in den Reliabilitäten können auf verschiedene Ursachen

Tabelle 7.4.
*Reliabilität der Tests zur Messung des fachspezifischen Professionswissens*

| | | PCK-Test | CK-Test |
|---|---|---|---|
| $N_{\text{Personen}}$ | | 102 | 102 |
| $N_{\text{Aufgaben}}$ | | 10 | 11 |
| Personenreliabilität | Real | .53 | .70 |
| | Model | .59 | .73 |
| Itemreliabilität | Real | .97 | .97 |
| | Model | .97 | .97 |

zurückgeführt werden. Zunächst ist die hier für die Schätzung der Reliabilität betrachtete Stichprobe der Gymnasiallehrkräfte aus NRW wesentlich homogener als die von Kirschner (2013) betrachtete Stichprobe, die in verschiedenen Bundesländern erhoben wurde und neben Gymnasiallehrkräften auch Haupt- und Gesamtschullehrkräfte umfasste. Nach Linacre (2011, S. 618) nimmt die Personenreliabilität mit abnehmender Varianz in den Personenfähigkeiten ebenfalls stark ab. Außerdem könnte die starke Abhängigkeit der Personenreliabilität von der Anzahl der Testaufgaben eine Erklärung für die geringeren Reliabilitäten, insbesondere des PCK-Tests, bieten (vergl. Linacre, 2011, S. 618). Die zweiteiligen Aufgaben im PCK-Test wurden nicht als separate Aufgaben in die Rasch-Analyse einbezogen. Außerdem wurde eine weitere Aufgabe mit schlechter Modellpassung aus den Analysen ausgeschlossen. Im Rahmen der vorliegenden Arbeit wurden daher insgesamt vier Aufgaben weniger zur Schätzung der Personenfähigkeiten im PCK verwendet als bei Kirschner (2013). Darüber hinaus wurden bei Kirschner (2013) EAP/PV-Reliabilitäten berechnet, die in der Regel ähnliche Werte annehmen wie Cronbachs Alpha, und damit höher ausfallen als die hier angegebenen Reliabilitäten der UML-Schätzer (vergl. Abschnitt 7.4.3 auf Seite 96).

#### 7.5.1.7. Validität

Die Validierung des Testinstruments zur Messung des fachspezifischen Professionswissens von Physiklehrkräften im Rahmen der ersten Projektphase (Kirschner, 2013) wurde bereits im Theorieteil dieser Arbeit in Abschnitt 5.1.2 auf Seite 58 ausführlich beschrieben. Da für den Einsatz in der zweiten Projektphase einige Änderungen an dem validierten Testinstrument vorgenommen werden mussten (vergl. Abschnitt 7.5.1.4 auf Seite 105), wird in diesem Abschnitt diskutiert, inwieweit die Ergebnisse der Validierungsstudien aus ProwiN I auf den in dieser Studie eingesetzten PCK- und CK-Test übertragen werden können.

**Inhaltsvalidität** Bezüglich des abgeprüften Inhalts unterscheiden sich die Testinstrumente der ersten und zweiten Projektphase nur geringfügig. In die Auswertung des PCK-Tests wurde lediglich eine bei Kirschner (2013) in die Analysen mit

einbezogene Aufgabe nicht aufgenommen. Im CK-Test traf dieser Umstand auf drei Aufgaben zu. Das unterschiedliche Vorgehen bei der Bewertung der zweiteiligen PCK-Aufgaben reduzierte zwar die Anzahl der zur Schätzung der Personenfähigkeiten genutzten Aufgaben um weitere drei Aufgaben, stellte aber keine Änderung bezüglich des abgeprüften Inhalts dar. Die Inhaltsvalidität des Testinstruments, die in der ersten Projektphase durch den Abgleich mit Curricula und Fachliteratur, Expertenbefragungen und der Testentwicklung anhand des Modells zum Professionswissen sichergestellt worden war, kann demnach auch für das Testinstrument der zweiten Phase als gegeben betrachtet werden.

**Konstruktvalidität** Die Konstruktvalidität des in der zweiten Phase des ProwiN-Projekts eingesetzten Testinstruments zur Erfassung des fachspezifischen Professionswissens kann im Rahmen einer konvergenten Validierung mit den PCK- und CK-Tests der ersten Projektphase untersucht werden.
Hierfür wird auf die Daten der $N = 79$ Physiklehrkräfte aus ProwiN I zurückgegriffen, mit denen die ProwiN II Stichprobe für die Durchführung der Rasch-Analysen erweitert wurde (vergl. Abschnitt 7.5.1.3 auf Seite 104). Für diese Stichprobe liegen sowohl die auf Basis der Kodierung mit dem ursprünglichen Kodiermanual von Kirschner (2013) geschätzten Lehrerfähigkeiten im PCK und CK als auch die auf Basis der Kodierung mit dem überarbeiteten Kodiermanual im Rahmen dieser Arbeit geschätzten Lehrerfähigkeiten vor. Um Aussagen darüber zu machen, ob die PCK- und CK-Tests der ersten und zweiten Projektphase das gleiche Konstrukt messen, wird die Korrelation zwischen den in ProwiN I und II bestimmten Lehrerfähigkeiten berechnet.

Die mit den PCK- und CK-Tests bestimmten Lehrerfähigkeiten korrelieren mit $r_{\text{Pearson,PCK}} = .69 \pm .06$ ($\text{KI}_{95\,\%} = [.60, .81]$, $p_{\text{1-seitig}} < .001$) bzw. $r_{\text{Pearson,CK}} = .799 \pm .028$ ($\text{KI}_{95\,\%} = [.748, .857]$, $p_{\text{1-seitig}} < .001$).[4] Laut Hammann und Jördens (2014) gelten nach einer persönlichen Mitteilung von Rost in konvergenten Validierungen Korrelation über .7 als Beleg für die Konstruktvalidität des untersuchten Testinstruments. Im Normalfall werden für konvergente Validierungen allerdings Testinstrumente eingesetzt, die zwar das gleiche Konstrukt erheben sollen, aber dennoch verschieden sind. Da hier die Messergebnisse zweier unterschiedlicher Versionen des gleichen Testinstruments miteinander korreliert werden, würde man eigentlich höhere Korrelationen erwarten. Allerdings muss bei der Bewertung der Korrelationen jeweils die Reliabilität des weniger reliablen Testinstruments, in diesem Fall also die Reliabilität des PCK- und CK-Tests aus der zweiten Projektphase, berücksichtigt werden (vergl. Abschnitt 7.5.1.6 auf Seite 108). Die Wurzel der Reliabilität gibt an, wie hoch ein messfehlerbehafteter Wert mit seinem „wahren“ Wert korreliert (vergl. Abschnitt 7.4.6 auf Seite 101). Die Korrelation beträgt $\sqrt{.59} = .77$ für den PCK-Test und $\sqrt{.73} = .85$ für den CK-Test. Da nicht

[4] Die Verteilung der ProwiN II CK-Testwerte weicht in der Stichprobe der $N = 79$ Physiklehrkräfte aus ProwiN I signifikant von der Normalverteilung ab, die entsprechenden nicht-parametrischen Korrelationen zwischen den CK-Lehrerfähigkeiten betragen: $r_{\text{Spearman}} = .84 \pm .04$, $\text{KI}_{95\,\%} = [.75, .89]$, $p_{\text{1-seitig}} < .001$; $\tau_{\text{Kendall}} = .66 \pm .04$, $\text{KI}_{95\,\%} = [.58, .73]$, $p_{\text{1-seitig}} < .001$.

zu erwarten ist, dass ein Messergebnis mit einem anderen Messergebnis in gleicher Höhe korreliert wie mit seinem eigenen wahren Wert, stellen diese Werte eine obere Grenze für die zu erwartenden Korrelationen dar. Die beobachteten Korrelationen zwischen den Testergebnissen der ersten und zweiten Projektphase sind daher ausreichend hoch, um eine Übertragung der Validierungsergebnisse von Kirschner (2013) auf das in dieser Studie eingesetzte Testinstrument zu rechtfertigen.

Um diese These zu unterstützen, wird außerdem versucht, die Ergebnisse der Konstruktvalidierung von Kirschner (2013) zu replizieren, die zeigen konnte, dass das PCK der Lehrkräfte mit ihrem CK und PK zusammenhängt und dass diese Zusammenhänge höher sind als der Zusammenhang zwischen CK und PK untereinander. Die Unterschiede in den Zusammenhängen wurden allerdings nur für den Vergleich PCK-CK gegenüber CK-PK statistisch signifikant (Kirschner, 2013, S. 81). Die Zusammenhänge zwischen den Dimensionen des Professionswissens werden in der für die Rasch-Analyse genutzten erweiterten Stichprobe der $N = 102$ Physiklehrkräfte untersucht. Der in der zweiten Projektphase eingesetzte PK-Test wird in Abschnitt 7.5.2 auf der nächsten Seite beschrieben.

Tabelle 7.5.
*Korrelationen zwischen den Dimensionen des Professionswissens in der ersten und zweiten Projektphase*

| | ProwiN II | | | | ProwiN I | | |
|---|---|---|---|---|---|---|---|
| | $N$ | $r_{\text{Pearson}}$ | $\text{KI}_{95\,\%}$ | $p_{\text{1-seitig}}$ | $N$ | $r_{\text{Pearson}}$ | $p$ |
| PCK-CK | 102 | $\mathbf{.39 \pm .09}$ | [.20, .54] | $< .001$ | 216 | **.33** | $< .001$ |
| PCK-PK | 102 | $\mathbf{.27 \pm .11}$ | [.06, .50] | .003 | 149 | **.23** | $< .01$ |
| CK-PK | 102 | $.15 \pm .10$ | $[-.04, .33]$ | .065 | 149 | .15 | .066 |

*Anmerkung.* Signifikante Korrelationen mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Da keine negativen Korrelationen zu erwarten sind, wurde einseitig auf Signifikanz getestet. Der PK-Test wurde bei Kirschner (2013) im Rahmen einer Rasch-Analyse ausgewertet. Im Vergleich zum hier eingesetzten klassisch ausgewerteten Testinstrument wurden 13 Items aus den Analysen entfernt. Da bei Kirschner keine Konfidenzintervalle angegeben wurden und zweiseitig auf Signifikanz getestet wurde, werden für die Korrelationen in ProwiN I nur zweiseitige $p$-Werte berichtet (vergl. Kirschner, 2013, S. 83). Aufgrund signifikanter Abweichungen von der Normalverteilung aller Variablen in der um die ProwiN I-Gymnasiallehrkräfte aus NRW erweiterten Stichprobe der ProwiN II-Lehrkräfte, werden in Tabelle B.5 auf Seite 246 im Anhang zusätzlich nicht-parametrische Korrelationen berichtet.

Im Rahmen der Fehlerabschätzung unterscheiden sich die Korrelationen nicht von denen in ProwiN I bestimmten Korrelationen (vergl. Tabelle 7.5). Auch die Ergebnisse bezüglich des Vergleichs der Korrelationen können repliziert werden (vergl. Kirschner, 2013, S. 81).[5] Die Korrelation zwischen PCK und CK ist signifikant größer als die Korrelation zwischen CK und PK ($Z(102) = 2.11$, $p_{\text{1-seitig}} = .017$).

[5] Angemerkt sei, dass der Test auf signifikante Unterschiede zwischen den Korrelationen in ProwiN I nur für die Gesamtstichprobe aller Lehrkräfte durchgeführt wurde, nicht aber für die hier gezeigten Korrelationen in der Stichprobe der Gymnasiallehrkräfte (Kirschner, 2013, S. 81)

Dies gilt allerdings nicht für die Korrelation zwischen PK und PCK ($Z(102) = 1.12$, $p_{\text{1-seitig}} = .132$).

**Kriteriumsvalidität** Auf Basis der Ergebnisse der konvergenten Validierung mit dem ProwiN I-Testinstrument und der Replikation der Ergebnisse der in der ersten Projektphase durchgeführten Konstruktvalidierung, wird angenommen, dass auch die Ergebnisse zur Untersuchung der Kriteriumsvalidität aus der ersten Projektphase auf das hier eingesetzte Testinstrument übertragen werden können. Hier wurden über die Technik bekannter Gruppen folgende Hypothesen bestätigt (vergl. Kirschner, 2013, S.95 und S.78-79):

- Physiklehrkräfte, die ein gymnasiales Lehramt studiert haben, verfügen über ein höheres PCK und CK als Studierende des gymnasialen Lehramts ($d_{PCK} = 0.68$, $d_{CK} = 0.99$)
- Physiklehrkräfte, die am Gymnasium unterrichten, verfügen über ein höheres PCK und CK als Gymnasiallehrkräfte anderer Fächer ($d_{PCK} = 0.9$, $d_{CK} = 2.1$) und Physiklehrkräfte anderer Schulformen ($d_{PCK} = 1.2$, $d_{CK} = 1.1$)
- Physiklehrkräfte, die am Gymnasium unterrichten, verfügen weder über ein höheres noch über ein niedrigeres CK als nicht an einer Schule unterrichtende Diplomphysiker/innen ($t(236) = 1.55$, $p = .122$)

Zusammengenommen weisen die Betrachtungen zur Inhaltsvalidität, zur Konstruktvalidität und zur Kriteriumsvalidität darauf hin, dass der PCK- und der CK-Test fachspezifisches, voneinander abgrenzbares Wissen erfassen, das Lehrkräfte auszeichnet, die Physik (am Gymnasium) unterrichten. Daher wird von einer validen Erfassung des fachspezifischen Professionswissens ausgegangen.

## 7.5.2. Test zur Messung des pädagogischen Wissens

Das Testinstrument zum pädagogischen Wissen (PK) wurde auf Grundlage von Testaufgaben aus der COACTIV-R Studie (Voss et al., 2011b) entwickelt und validiert (Lenske et al., 2015). In dem PK-Testinstrument werden deklaratives und konditional-prozedurales Wissen in zwei separaten Tests erfasst.

Im Folgenden wird zunächst erläutert, warum in dieser Arbeit nur der Test zum deklarativen Wissen ausgewertet wird. Nach einer kurzen Beschreibung des PK-Tests werden technische Details zu dessen Auswertung erläutert. Abschließend wird die Objektivität, Reliabilität und die Validität des Tests diskutiert.

### 7.5.2.1. Beschränkung der Auswertung auf den Test zum deklarativen Wissen

Der Test zum konditional-prozeduralen Wissen zeigte in der ersten Projektphase in einer Stichprobe aus Gymnasial- und Hauptschullehrkräften der Physik aus Bayern und Nordrhein-Westfalen (NRW) eine sehr geringe Varianz, daher ist fraglich, ob dieser Test das pädagogische Professionswissen differenziert genug erfasst (vergl.

Lenske et al., 2015, S. 239). Dies gilt insbesondere für die in der vorliegenden Studie untersuchte homogenere Stichprobe der Gymnasiallehrkräfte aus NRW. Die Korrelation zwischen dem konditional-prozeduralen Test zum PCK in der in der ersten Projektphase untersuchten Stichprobe der $N = 171$ Physiklehrkräfte war mit $r_{\text{Pearson}} = .19$ ($p < .05$) wesentlich geringer als die entsprechende Korrelation für den deklarativen Test ($r_{\text{Pearson}} = .31$, $p < .01$) und lediglich genauso groß wie die Korrelation des deklarativen Tests zum CK der Physiklehrkräfte (Lenske et al., 2015, S. 240). Da die Ergebnisse für die Konstruktvalidierung für den konditional-prozeduralen Test damit weniger eindeutig ausfallen als für den deklarativen PK-Test, beschränkt sich die Untersuchung des pädagogischen Wissens in dieser Arbeit auf das deklarative Wissen der Lehrkräfte. Wird im Folgendem vom PK-Test gesprochen, ist daher immer der Test zum deklarativen Wissen gemeint.

#### 7.5.2.2. PK-Test

In den Aufgaben des PK-Tests wird deklaratives Wissen über Klassenführung, Unterrichtsmethoden, individuelle Lernprozesse und Leistungsbeurteilung abgefragt. Der PK-Test umfasst neun komplexe Multiple-Choice-Aufgaben (Single-Select, 4-7 Antwortmöglichkeiten). Da die Formulierung von allgemeingültigen Aussagen ohne Einschränkung im pädagogischen Kontext von Praktikern oftmals als schwierig empfunden wird, wurde die Zustimmung zu den jeweiligen Antwortalternativen mit einer vierstufigen Likertskala (1 =„stimmt genau“, 2 =„stimmt eher“, 3 =„stimmt eher nicht“, 4 =„stimmt nicht“) erhoben (vergl. Lenske et al., 2015, S. 234). Abbildung 7.5 zeigt eine Beispielaufgabe des PK-Tests zu individuellen Lernprozessen.

PKD4

**Welche Maßnahmen sind geeignet, das selbstregulierte Lernen zu fördern?**

| | | stimmt genau | stimmt eher | stimmt eher nicht | stimmt nicht |
|---|---|---|---|---|---|
| a) | Lerntagebuch führen lassen | ☐ | ☐ | ☐ | ☐ |
| b) | Concept-Map erstellen lassen | ☐ | ☐ | ☐ | ☐ |
| c) | Gespräch mit den Eltern führen | ☐ | ☐ | ☐ | ☐ |
| d) | Gespräch mit dem Schüler führen | ☐ | ☐ | ☐ | ☐ |
| e) | Lernprozess gut vorstrukturieren | ☐ | ☐ | ☐ | ☐ |

Abbildung 7.5.
*Beispielaufgabe aus dem PK-Test. Korrekte Antworten: a), d), e) stimmt eher oder stimmt genau; b), c) stimmt eher nicht oder stimmt nicht (entnommen aus Kirschner, 2013, S. 49).*

#### 7.5.2.3. Technische Details zur Auswertung

In diesem Abschnitt wird zunächst die Punktevergabe im PK-Test und der Umgang mit fehlenden Werten beschrieben. Anschließend wird erläutert, warum der PK-Test, im Gegensatz zu den anderen in dieser Studie eingesetzten schriftlichen Erhebungsinstrumenten, nicht im Rahmen einer Rasch-Analyse, sondern klassisch ausgewertet wird.

**Punktevergabe und Umgang mit fehlenden Werten** Für die Auswertung des PK-Tests wurde zunächst die Likertskala dichotomisiert (richtig/falsch) und jede Antwortmöglichkeit als eigenständiges Item behandelt (vergl. Lenske et al., 2015, S. 233). Insgesamt ergaben sich 45 Items. Für jede richtige Antwort wurde ein Punkt vergeben. Im Folgenden beziehen sich Aussagen über die Aufgaben des PK-Tests auf die komplexen Multiple-Choice-Aufgaben, während die dichotomisierten Antwortmöglichkeiten als Items bezeichnet werden.

Für den PK-Test war eine feste Bearbeitungszeit vorgegeben. Ausgehend von der Annahme, dass Lehrkräfte mit höherem pädagogischen Wissen im Rahmen der Bearbeitungszeit tendenziell mehr Items bearbeiten können, wurden nicht bearbeitete Items auch hier als Ausdruck geringerer Fähigkeit betrachtet. Sie wurden daher nicht als fehlende Werte behandelt, sondern stattdessen mit null Punkten bewertet (vergl. Lenske et al., 2016).

**Klassische Analyse** Im Zuge der Analyse der PK-Aufgaben in der ersten Phase des ProwiN-Projekts wurde eine der komplexen Multiple-Choice-Aufgaben (5 Items) aus dem PK-Test entfernt. In fünf der komplexen Multiple-Choice Aufgaben wurden zudem einzelne Items (insgesamt zehn) entfernt. Für die Bestimmung der Lehrerfähigkeiten und für die Reliabilitätsanalysen standen daher noch 30 der ursprünglich 45 dichotomisierten Items zur Verfügung (vergl. Lenske et al., 2015, S. 237).

Nach Aussage der Projektpartner aus der Lehr-Lernpsychologie ist das Testinstrument zur Erfassung des pädagogischen Professionswissen nicht für die Analyse im Rahmen eines Rasch-Modells ausgelegt, da das erfasste Konstrukt nicht als eindimensional betrachtet werden kann (Leutner, Persönliche Mitteilung). In der Rasch-Analyse der verbleibenden 30 Items des PK-Tests zeigten in der Tat drei Items eine schlechte Modellpassung. Ein weiteres Item zeigte einen Overfit bzgl. des Rasch-Modells. Das von Lenske et al. (2015) validierte Testinstrument ist also (ohne Ausschluss weiterer Items) nicht raschskalierbar.

Die Auswertung des PK-Tests erfolgte daher im Rahmen einer klassischen Analyse. Obwohl acht Items geringe Trennschärfen (Korrigierte Item-Skala Korrelation $< .1$) zeigten[6], wurden diese Items nicht entfernt, da nach Rücksprache mit den Projektpartner aus der Lehr-Lernpsychologie keine Änderungen an dem in der ersten Projektphase validierten Testinstrument vorgenommen werden sollten.

Die Lehrerfähigkeit im PK wird über den Anteil gelöster Aufgaben bestimmt. Um die Reliabilität des PK-Tests an einer größeren Stichprobe als den $N = 23$

[6] Für drei Items waren die Trennschärfen leicht negativ (Korrigierte Item-Skala Korrelation= $-.11/-.05/-.02$).

Physiklehrkräften untersuchen zu können, wurde die Stichprobe für die Reliabilitätsberechnung analog zu Abschnitt 7.5.1.3 auf Seite 104 mit der Stichprobe der $N = 79$ Gymnasiallehrkäfte aus ProwiN I erweitert.

#### 7.5.2.4. Objektivität

Durch die standardisierte Testdurchführung (vergl. Abschnitt 7.2.3 auf Seite 85), den Verzicht auf offene Aufgaben und den Umstand, dass keine Bewertung der absoluten Personenfähigkeiten, sondern lediglich Vergleiche des Lehrerwissens innerhalb der Stichprobe erfolgen, kann der PK-Test sowohl bezüglich der Durchführung als auch bezüglich der Auswertung und Interpretation als objektives Testinstrument betrachtet werden (vergl. Bortz & Döring, 2006, S. 195).

#### 7.5.2.5. Reliabilität

Die Reliabilität des PK-Tests wurde über die Berechnung von Cronbachs Alpha geschätzt. Hierfür wurde die Stichprobe mit Daten aus der ersten Projektphase erweitert (vergl. Abschnitt 7.5.2.3 auf Seite 114). Die Reliabilität der 30 Items des PK-Tests konnte somit in einer Stichprobe von $N = 102$ Gymnasiallehrkräften bestimmt werden. Sie beträgt $\alpha_{\mathrm{C}} = .67$ ($\mathrm{KI}_{95\,\%} = [.57, .77]$). Lenske et al. (2015, S. 237) geben für eine wesentlich heterogenere Stichprobe von $N = 452$ Lehrkräften der naturwissenschaftlichen Fächer (Physik, Chemie und Biologie) eine geringfügig höhere Reliabilität des deklarativen PK-Tests von $\alpha_{\mathrm{C}} = .70$ an, die innerhalb der 95%-Konfidenzintervalle der in der vorliegenden Arbeit berechneten Reliabilität liegt. Die Reliabilität des PK-Tests kann daher als ausreichend bis zufriedenstellend bezeichnet werden.

#### 7.5.2.6. Validität

Die Validierung des PK-Tests erfolgte im Rahmen der ersten Phase des ProwiN-Projektes und wurde bereits im Theorieteil dieser Arbeit in Abschnitt 5.1.2 auf Seite 58 beschrieben. Die Konstruktvalidität wurde über Korrelationen zu den fachspezifischen Professionswissensdimensionen (CK und PCK) untersucht und bestätigt ($r_{\mathrm{Pearson,PK\text{-}PCK}} = .31$, $p < .01$, $r_{\mathrm{Pearson,PK\text{-}CK}} = .19$, $p < .05$). Über die Validierung mit bekannten Gruppen konnte außerdem gezeigt werden, dass Lehrkräfte unterschiedlicher (naturwissenschaftlicher) Fächer im Mittel über das gleiche deklarative pädagogische Wissen verfügten und dieses erwartungsgemäß niedriger war als das Wissen der in der universitären Lehrerbildung tätigen Probanden ($t(72) = 2.648$, $p < .05$, $d = 0.77$) (Lenske et al., 2015, S. 239 bzw. S. 236).

Da an dem in der ersten Projektphase validierten Test keine Änderungen vorgenommen wurden, kann die Validität des PK-Tests unter den von Lenske et al. (2015) genannten Einschränkungen auch in dieser Studie als gegeben betrachtet werden.

Im Bezug auf die Konstruktvalidität kann außerdem auf Abschnitt 7.5.1.7 auf Seite 110 zur Konstruktvalidität der fachspezifischen Professionswissenstests verwiesen werden, wo gezeigt werden konnte, dass das mit dem PK-Test erhobene Wissen

auch in dieser Studie erwartungsgemäß mit dem CK und PCK der Physiklehrkräfte korreliert (vergl. Tabelle 7.5 auf Seite 111).

Zusammenfassend kann festgehalten werden, dass der PK-Test fachunspezifisches deklaratives pädagogisches Wissen erfasst, welches Lehrkräften im Rahmen der universitären Lehrerbildung vermittelt wird. Auf Basis der Validierungsergebnisse kann allerdings noch keine Aussage darüber getroffen werden, ob das mit dem PK-Test erfasste Wissen als unterrichtsrelevant erachtet werden kann.

## 7.5.3. Schülerfachwissenstest

Mit dem Schülerfachwissenstest wird das Fachwissen der Lernenden in Mechanik mit Fokus auf Kraft erhoben. Der Test besteht aus 39 Multiple-Choice-Aufgaben (Single-Select, 4-5 Antwortmöglichkeiten), die auf zwei Testhefte A und B mit jeweils 24 Aufgaben verteilt sind (siehe Anhang A.2 auf Seite 221). Die Testhefte sind über neun identische Aufgaben im Mittelteil des Tests verankert. Beide Testhefte wurden zu beiden Messzeitpunkten eingesetzt. Lernende, die beim Prä-Test Testheft A bearbeiteten, bearbeiteten beim Post-Test Testheft B (und umgekehrt). Für den Schülerfachwissenstest war eine feste Bearbeitungszeit von 30 Minuten vorgegeben. Tabelle 7.6 auf der nächsten Seite zeigt die Verteilung der Aufgaben auf die im Schülerfachwissenstest bearbeiteten Unterthemen der Mechanik. Eine Kurzbeschreibung aller Aufgaben und eine Übersicht über deren Kennzahlen finden sich bei Cauet (2015).

Im Folgenden wird zunächst die Entwicklung und Pilotierung der Aufgaben für den Schülerfachwissenstest beschrieben. Anschließend werden technische Details zur Auswertung des Testinstruments erläutert. In diesem Abschnitt wird auch auf die Rasch-Analyse der Schülertestdaten eingegangen. Abschließend wird die Objektivität, Reliabilität und die Validität des Testinstruments diskutiert.

### 7.5.3.1. Entwicklung und Pilotierung

In diesem Abschnitt erfolgt eine Beschreibung der Pilotierung des für die Entwicklung des Schülerfachwissenstests zusammengestellten Aufgabenpools. Zunächst wird über die Herkunft der Aufgaben berichtet und der Ablauf sowie die Stichprobe der Pilotierungsstudie beschrieben. Im Anschluss wird über die Erweiterung des Aufgabenpools und die Auswertung der Pilotierungsdaten berichtet. Abschließend wird erklärt, nach welchen Kriterien die Auswahl der Aufgaben für den Schülerfachwissenstest erfolgte.

**Herkunft der Aufgaben** Im Rahmen einer Lehrplananalyse der Kernlehrpläne für Gymnasien und Gesamtschulen in NRW (MSW, 2008, 2011) wurden die in Tabelle 7.6 auf der nächsten Seite aufgeführten Unterthemen als relevant für den Mechanikunterricht in der Jahrgangsstufe 8 und 9 an Gesamtschulen und Gymnasien identifiziert. Ausgehend von diesen Themen wurde ein Aufgabenpool aus 80 Multiple-Choice-Single-Select-Aufgaben erstellt. Bei den Aufgaben handelt es sich um Eigenentwicklungen, um adaptierte Aufgaben aus etablierten Testinstrumenten

Tabelle 7.6.
*Verteilung der Aufgaben auf die Testhefte (TH) A und B und die im Schülerfachwissenstest bearbeiteten Unterthemen der Mechanik. Ankeraufgaben wurden beiden Testheften zugeordnet*

| Unterthemen der Mechanik | Anzahl der Aufgaben TH A | TH B | Anker |
|---|---|---|---|
| Kraftwirkungen/Kraft als Ursache von Bewegungsänderungen | 1 | 2 | 1 |
| Kraft und Gegenkraft | 2 | 1 | 1 |
| Kräftegleichgewicht | 2 | 1 | 1 |
| Addition von Kräften/Kräfteparallelogramm/ Komponentenzerlegung | 2 | 1 | 1 |
| Kraft als Vektor | 0 | 1 | 0 |
| Hebel | 3 | 2 | 1 |
| Gewichtskraft und Masse | 3 | 3 | 0 |
| Gewichtskraft an verschiedenen Orten | 1 | 2 | 0 |
| Gleichförmige Bewegung und ihre Voraussetzung | 5 | 4 | 1 |
| Qualitative Beschreibung beschleunigter Bewegungen | 1 | 2 | 1 |
| Qualitative Beschreibung von Kreisbewegungen | 0 | 1 | 0 |
| Geschwindigkeit | 2 | 2 | 1 |
| Zusammenhang von Geschwindigkeit und Beschleunigung | 1 | 2 | 1 |
| Trägheit | 1 | 1 | 0 |
| Energie/Arbeit/Leistung | 3 | 3 | 2 |
| Berücksichtigung von Reibung oder Luftwiderstand | 2 | 1 | 0 |
| Verwendung und Definition verschiedener Einheiten | 2 | 1 | 1 |
| Differenzierung zwischen Einheiten und Größen | 0 | 1 | 0 |
| Lesen und Interpretieren von Diagrammen | 5 | 6 | 2 |

*Anmerkung.* Die Zuordnung der Aufgaben zu den Unterthemen erfolgte durch das Forscherteam. Einige Aufgaben wurden mehreren Unterthemen zugeordnet. Addiert man die Anzahl an Aufgaben, ergibt sich daher nicht die Anzahl der Aufgaben pro Testheft bzw. die Anzahl an Ankeraufgaben.

(Trends in International Mathematics and Science Study: TIMSS Assessment, 1995, 1999, 2003, 2007, Force Concept Inventory, 1992; Mechanics Baseline Test, 1992), um adaptierte Aufgaben aus dem Internetportal „www.leifiphysik.de“[7] (Leitner & Finckh, o.D.) und um einige Aufgaben aus dem ProwiN CK-Test für Physiklehrkräfte (Kirschner, 2013, adaptiert aus Force Concept Inventory, 1992). Bis auf letztere wurden alle Aufgaben in ein vierstufiges Antwortformat überführt. An den Multiple-Choice-Aufgaben aus dem ProwiN CK-Test wurden keine Veränderungen vorgenommen, um eine parallele Verwendung der Aufgaben im Schüler- und Lehrertest zu ermöglichen. Eine Übersicht über die Aufgaben und die zugehörigen Quellen findet sich bei Cauet (2015).

**Pilotierungsstudie** Die Aufgaben wurden im Frühjahr 2011 im Rahmen einer Pilotierungsstudie in 30 Klassen der Jahrgangsstufe 8 an fünf Gymnasien und vier Gesamtschulen in Nordrhein-Westfalen pilotiert. Die Aufgaben wurden sowohl in Klassen pilotiert, die das Thema Mechanik noch nicht behandelt hatten (Prä-Testung) als auch in Klassen, die die Behandlung des Themas Mechanik bereits abgeschlossen hatten (Post-Testung). Drei Klassen nahmen an beiden Messzeitpunkten teil. Tabelle 7.7 zeigt die Verteilung der Schülerinnen und Schüler auf die Schultypen und Messzeitpunkte.

Tabelle 7.7.
*Verteilung der Pilotierungsstichprobe auf Schultypen und Messzeitpunkte und mittleres Alter der Schülerinnen und Schüler (*50% *weiblich, Altersangabe in Jahren)*

| | Prä-Testung | | Post-Testung | | Summe | | Alter | |
|---|---|---|---|---|---|---|---|---|
| | Klassen | SuS | Klassen | SuS | Klassen | SuS | *M* | *SD* |
| Gymnasium | 6 | 164 | 12 | 280 | 18 | 444 | 13.9 | 0.6 |
| Gesamtschule | 7 | 162 | 8 | 217 | 15 | 379 | 14.1 | 0.6 |
| Summe | 13 | 326 | 20 | 497 | 33 | 823 | 14.0 | 0.6 |

*Anmerkung.* Die Klassen bzw. Schülerinnen und Schüler, die an beiden Messzeitpunkten teilnahmen, sind hier doppelt aufgeführt. Bei einfacher Zählung reduziert sich die Stichprobe auf 30 verschiedene Klassen und 755 Schülerinnen und Schüler.

Die Aufgaben wurden in 6 Aufgabenblöcke (A-F) à 13 bzw. 14 Aufgaben aufgeteilt und in einem rotierten Multi-Matrix-Design auf 6 Testhefte verteilt (AB, CB, CD, ED, EF, AF).

**Auswertung und Erweiterung des Aufgabenpools** Die Pilotierungsdaten wurden im Rahmen einer Rasch-Analyse (vergl. Abschnitt 7.4.2 auf Seite 93) mit Winsteps 3.70.0.5 analysiert. Da lediglich 3 von 30 Klassen zu beiden Messzeitpunkten an der Studie teilgenommen hatten, wird davon ausgegangen, dass es

[7] Ggf. vorhandene Abbildungen wurden neu erstellt.

sich bei der Prä- und Post-Stichprobe weitestgehend um unabhängige Stichproben handelt und die lokale stochastische Unabhängigkeit der Daten somit auch für den gesamten Datensatz als ausreichend betrachtet werden kann (vergl. Abschnitt 7.4.2 auf Seite 93). Prä- und Post-Testdaten wurden daher im Rahmen einer gemeinsamen Rasch-Analyse skaliert.

Die Analyse zeigte, dass die Aufgaben im Mittel zu schwer waren und dass zu wenige Aufgaben im unteren Anforderungsbereich existierten. Daher wurden in den Schülerfachwissenstest für die Hauptstudie zusätzlich Aufgaben aus einer anderen Studie aufgenommen. Hierfür konnte der Umstand genutzt werden, dass zeitgleich zur Pilotierung der Schülertestaufgaben, im Rahmen des Dissertationsprojekts von Zander (2016), 100 Testaufgaben für einen parallel entwickelten Schülertest zum Thema Mechanik in einer Stichprobe von 578 Schülerinnen und Schülern (Jahrgangsstufe 7-9, Gymnasium, NRW, Prä-Testung: $N = 173$, Post-Testung: $N = 405$, $M_{\text{Alter}} = 14.3$, $SD_{\text{Alter}} = 0.9$, 47% weiblich) pilotiert worden waren. Bei den Aufgaben handelte es sich um Eigenentwicklungen und adaptierte Aufgaben aus den in Abschnitt 7.5.3.1 auf Seite 116 bereits erwähnten Testinstrumenten.

Die pilotierten Daten unterschieden sich bezüglich der Stichprobe und bezüglich des Inhalts sowie des Formats der Aufgaben kaum. Da die beiden Stichproben über zwölf Aufgaben, die in beiden Stichproben pilotiert worden waren, verankert werden konnten, war eine gemeinsame Analyse der Daten möglich. In der Rasch-Analyse des Gesamtdatensatzes wurden daher insgesamt 168 Aufgaben und $N = 1401$ Personen analysiert.

Aufgaben mit signifikantem Misfit wurden entfernt. Abweichend von den in Abschnitt 7.4.2 auf Seite 93 angegebenen Misfit-Kriterien wurden im Rahmen der Pilotierung etwas weniger strenge Kriterien verwendet ($MnSq \leq 0.8$ oder $MnSq \geq 1.3$) (Linacre, 2011). Lediglich eine missfittende Aufgabe wurde nicht entfernt, da es sich um eine der Parallelaufgaben zum ProwiN CK-Test für Lehrkräfte handelte. Diese Aufgabe wurde allerdings nicht für den in der Hauptstudie eingesetzten Schülerfachwissenstest ausgewählt. Außerdem wurden Aufgaben entfernt, die DIF bzgl. des Messzeitpunkts oder des Schultyps zeigten (vergl. Abschnitt 7.4.2 auf Seite 93). Der DIF bzgl. des Schultyps konnte allerdings nur für Aufgaben aus dem ProwiN Aufgabenpool bestimmt werden, da die Pilotierung der anderen Aufgaben nur an Gymnasien durchgeführt worden war.

**Auswahl der Aufgaben für den Schülerfachwissenstest** Aus den verbleibenden 137 Aufgaben wurden 39 Aufgaben für die beiden Testhefte A und B des Schülerfachwissenstest für die Hauptstudie ausgewählt. Jedem Testheft wurden 24 Aufgaben zugeordnet. Um die beiden Testhefte zu verankern, wurden neun Aufgaben als Ankeraufgaben beiden Testheften zugeordnet. Bei der Auswahl der Aufgaben und der Verteilung auf die beiden Testhefte wurde darauf geachtet, dass beide Testhefte Aufgaben zu möglichst allen Unterthemen der Mechanik erhielten (vergl. Tabelle 7.6 auf Seite 117) und dass die mittlere Aufgabenschwierigkeit in beiden Testheften etwa gleich groß war. Aufgrund des Multi-Matrix-Designs, das durch die Verankerung mit den Aufgaben aus der Studie von Zander noch zusätzlich

an Komplexität gewonnen hatte, war es im Rahmen der Pilotierungsstudie nicht möglich, eine Reliabilität für den Schülerfachwissenstest zu berechnen.

#### 7.5.3.2. Technische Details zur Auswertung

In diesem Abschnitt wird zunächst die Punktevergabe und der Umgang mit fehlenden Werten im Schülerfachwissenstest erläutert. Im Anschluss wird berichtet, welche Daten aus der Auswertung ausgeschlossen werden mussten. Das Fachwissen der Lernenden wurde im Rahmen eines Rasch-Modells als Personenfähigkeit geschätzt. Abschließend wird die Rasch-Analyse des Schülerfachwissenstest beschrieben.

**Punktevergabe und Umgang mit fehlenden Werten** Für jede korrekt gelöste Aufgabe im Schülerfachwissenstest wurde ein Punkt vergeben. Für den Schülerfachwissenstest war eine feste Bearbeitungszeit vorgegeben. Diese wurde allerdings nur von wenigen Lernenden tatsächlich für die Bearbeitung der Aufgaben benötigt (vergl. Abschnitt 7.2.3.1 und 7.2.3.2 auf Seite 85 und auf Seite 86 zum Ablauf der Prä- bzw. Post-Erhebung). Außerdem wurden die Schülerinnen und Schüler bei der Bearbeitung des Tests darum gebeten, sich bei jeder Aufgabe für eine Antwortmöglichkeit zu entscheiden, auch wenn sie die richtige Antwort nicht wussten. Der Anteil nicht bearbeiteter Aufgaben ist daher insgesamt gering (1% im Prä- und im Post-Test). Anders als im Professionswissenstest für die Lehrkräfte (vergl. Abschnitt 7.5.1.3 auf Seite 104) kann daher nicht davon ausgegangen werden, dass Schülerinnen und Schüler mit höherem Fachwissen im Rahmen der Bearbeitungszeit tendenziell mehr Aufgaben bearbeiten konnten. Nicht bearbeitete Aufgaben sind daher nicht als Ausdruck schlechterer Leistung zu interpretieren. In Bezug auf den Schülerfachwissenstest wurde daher den Argumenten von Hohensinn und Kubinger (2011) gefolgt, die zeigen konnten, dass das Behandeln von nicht bearbeiteten Aufgaben als fehlende Werte zu weniger verzerrten Ergebnissen führt als das Bewerten einer solchen Aufgabe als falsch. Nicht bearbeitete Aufgaben wurden deshalb nicht als falsch, sondern ebenso wie designbedingte Missings, als fehlende Werte behandelt.

**Datenausschluss** Eine Aufgabe musste von Beginn an aus den Analysen ausgeschlossen werden, da die Aufgabenstellung irreführend war.

**Rasch-Analyse** Aus den Schülerfachwissenstestdaten wurden für beide Messzeitpunkte im Rahmen eines eindimensionalen dichotomen Rasch-Modells Personenfähigkeiten geschätzt (vergl. Abschnitt 7.4.2 auf Seite 93).

Um Veränderungen in den Schülerleistungen vom Prä- zum Post-Test beschreiben zu können, müssen die im Rasch-Modell geschätzten Personenfähigkeiten zu den beiden Messzeitpunkten auf einer Skala liegen. Dies lässt sich zum Beispiel über die Bildung sogenannter *virtueller* Personen realisieren, indem die Testdaten einer Person zu zwei unterschiedlichen Messzeitpunkten als Daten zweier unabhängiger Personen behandelt werden. Prä- und Post-Testdaten werden also im Rahmen einer gemeinsamen Analyse skaliert (vergl. Rost, 2004). Im Rahmen einer solchen

Analyse können auch Items identifiziert werden, die DIF bzgl. des Messzeitpunkts zeigen (vergl. Abschnitt 7.4.2 auf Seite 93). Der Nachteil dieses Vorgehens besteht allerdings darin, dass die Abhängigkeit zwischen dem Prä- und Post-Testergebnis einer Person ignoriert wird (Hartig & Kühnbach, 2006) und damit die im Rasch-Modell geforderte lokale stochastische Unabhängigkeit der Beobachtungen verletzt wird (siehe Abschnitt 7.4.2 auf Seite 93).

Um eine Skalierung der Prä- und Post-Test-Personenfähigkeiten auf einer gemeinsamen Skala zu realisieren und gleichzeitig die Stichprobenabhängigkeit zwischen den Messzeitpunkten zu berücksichtigen, wurde ein mehrschrittiges Vorgehen bei der Rasch-Analyse der Schülertestdaten gewählt.

**1.Schritt: Gemeinsame Analyse der Prä- und Post-Testdaten** Um Aufgaben zu identifizieren, die einen DIF bezüglich des Messzeitpunkts zeigten, erfolgte zunächst eine gemeinsame Analyse der Prä- und Post-Testdaten durch Bildung virtueller Personen. Die Analyse wurde mit insgesamt $N = 1270$ Personen (inklusive virtueller Personen) und $N = 38$ Items durchgeführt. Drei Aufgaben zeigten einen signifikantem Prä-Post DIF. Zwei dieser Aufgaben waren für die Schülerinnen und Schüler im Prä-Test wesentlich schwieriger zu lösen, als im Post-Test. Die Lösung der dritten Aufgabe fiel den Lernenden im Prä-Test leichter.

**2.Schritt: Skalierung des Post-Tests** Im nächsten Schritt erfolgte die eigentliche Skalierung der Post-Testdaten. Im Rahmen dieser Analyse wurden vier Aufgaben entfernt, die signifikanten Misfit aufwiesen. Bei einer der entfernten Aufgaben handelte es sich um eine der Ankeraufgaben. Die Anzahl der Ankeraufgaben in den Analysen reduzierte sich damit auf acht. Bei einer anderen Aufgabe handelte es sich um die Aufgabe, die in Schritt 1 einen signifikanten Prä-Post DIF gezeigt hatte und deren Lösung den Lernenden im Prä-Test leichter gefallen war als im Post-Test.

**3.Schritt: Skalierung des Prä-Tests** Im Rahmen der Rasch-Analyse ist es möglich, die Aufgabenschwierigkeit vor Beginn der Analyse auf feste Werte zu fixieren. Dieses Vorgehen kann beispielsweise gewählt werden, wenn die Aufgabenschwierigkeiten bereits in einer anderen Stichprobe bestimmt worden sind.

Um eine Skalierung der Prä- und Post-Testdaten auf einer gemeinsamen Skala zu realisieren, wurden die Aufgabenschwierigkeiten in der Analyse des Prä-Tests auf die Werte der Aufgabenschwierigkeiten im Post-Test fixiert. Lediglich die Schwierigkeit der zwei verbleibenden, in Schritt 1 identifizierten Aufgaben, die einen signifikanten Prä-Post DIF zeigten, wurde nicht fixiert und in der Analyse frei geschätzt.

Bei der Fixierung von Aufgabenschwierigkeiten kann es vorkommen, das die fixierten Werte nicht exakt auf die erhobenen Daten passen. Das kann dazu führen, dass die Passung der Daten ins Modell über- oder unterschätzt wird (Linacre, 2011, S. 596). Obwohl bei der Modellierung des Prä-Tests einige Aufgaben signifikanten Misfit zeigten, wurden daher keine weiteren Aufgaben aus den Analysen zur Schätzung der Personenfähigkeiten im Prä-Test entfernt, da der Misfit lediglich aus

der Fixierung der Aufgabenschwierigkeiten resultieren könnte. In Abbildung B.3 auf Seite 252 im Anhang finden sich die Wright-Maps für die Aufgaben des Prä-Tests und des Post-Tests.

#### 7.5.3.3. Objektivität

Durch die standardisierte Testdurchführung mit Testleitermanualen (vergl. Abschnitt 7.2.3 auf Seite 85) und den Umstand, dass es sich bei dem Schülerfachwissenstest um ein reines Multiple-Choice-Instrument handelt und keine Bewertung der absoluten Personenfähigkeiten, sondern lediglich Vergleiche des Schülerfachwissens innerhalb der Stichprobe auf Klassenebene erfolgen, kann der Schülerfachwissenstest sowohl bezüglich der Durchführung als auch bezüglich der Auswertung und Interpretation als objektives Testinstrument betrachtet werden (vergl. Bortz & Döring, 2006, S. 195).

#### 7.5.3.4. Reliabilität

Die Reliabilitäten des Schülerfachwissenstests wurden im Rahmen der Rasch-Analyse bestimmt (vergl. Abschnitt 7.5.3.2 auf Seite 120). Tabelle 7.8 zeigt die Reliabilitäten für den Prä- und Post-Test.

Tabelle 7.8.
*Reliabilität des Schülerfachwissenstest in der Stichprobe der $N = 23$ Gymnasialklassen*

| | | Prä-Test | Post-Test |
|---|---|---|---|
| $N_{\text{Personen}}$ | | 640 | 630 |
| $N_{\text{Aufgaben}}$ | | 34 | 34 |
| Personenreliabilität | Real | .44 | .58 |
| | Model | .51 | .61 |
| Itemreliabilität | Real | .99 | .99 |
| | Model | .99 | .99 |

Die Reliabilität des Prä-Test kann gerade noch als ausreichend bezeichnet werden, die Reliabilität des Post-Tests als ausreichend (Lamberti, 2001, S. 31). Bei der Bewertung der im Rasch-Modell geschätzten Personenreliabilitäten ist erneut zu beachten, dass diese meist niedrigere Werte annehmen als klassisch berechnete Reliabilitäten wie Cronbachs Alpha (vergl. Abschnitt 7.4.3 auf Seite 96). Darüber hinaus können auch die aus dem Multi-Matrix-Design resultierenden Missings die Reliabilität des Testinstruments verringern (Linacre, 2011, S. 618). Die geringe Reliabilität des Prä-Tests kann zudem daraus resultieren, dass das Fachwissen der Lernenden beim Prä-Test noch unstrukturiert ist und die Lernenden häufiger raten. Dass der Schülerfachwissenstest das relativ heterogene Konstrukt der Mechanik abfragt und darüber hinaus nicht alle Klassen alle im Testinstrument abgefragten Unterthemen auch wirklich im Rahmen der Unterrichtseinheit Mechanik behandelt haben, erklärt auch die vergleichsweise niedrige Reliabilität des Post-Tests.

#### 7.5.3.5. Validität

Der Schülerfachwissenstest wurde eingesetzt, um den Fachwissenszuwachs der Lernenden im Rahmen der Unterrichtseinheit Mechanik zwischen verschiedenen Klassen vergleichen zu können und damit Rückschlüsse auf Unterschiede in der Qualität des Unterrichts zu ziehen. Um einen fairen Vergleich zwischen den Klassen zu ermöglichen, ist es wichtig, dass das im Schülertest abgefragte Wissen das im Rahmen der Unterrichtseinheit Mechanik tatsächlich vermittelte Wissen möglichst gut abbildet. Die *curriculare Validität* des Testinstruments muss also sichergestellt werden. Hierfür ist es wünschenswert, dass die in den Aufgaben bearbeiteten Unterthemen der Mechanik von möglichst allen Lehrkräften im Unterricht behandelt wurden und dass sich die Klassen im Anteil behandelter Unterthemen nicht wesentlich unterscheiden. Ein fairer Vergleich zwischen den Klassen wäre beispielsweise nicht gewährleistet, wenn einige Klassen nur einen Bruchteil der abgefragten Themen im Unterricht behandelt hätten. Daher wird im Folgenden zunächst die curriculare Validität des Schülerfachwissenstests diskutiert.

Um die Konstruktvalidität des Schülerfachwissenstests zu untersuchen, erfolgt anschließend im Rahmen einer diskriminanten Validierung eine Abgrenzung zum Konstrukt der *Intelligenz*. Eine konvergente Validierung des Schülerfachwissenstests war in dieser Studie nicht möglich, da aus testökonomischen Gründen kein weiteres Testinstrument zur Messung des Fachwissens in Mechanik eingesetzt werden konnte. Es kann allerdings darauf verwiesen werden, dass der Schülerfachwissenstest Aufgaben aus anderen Mechaniktests enthält (vergl. Abschnitt 7.5.3.1 auf Seite 116).

Zur Beurteilung der Kriteriumsvalidität werden abschließend die Korrelationen zwischen den Schülertestergebnissen der Lernenden und ihren Schulnoten in den Fächern Physik, Mathematik und Deutsch untersucht.

**Curriculare Validität** Um die curriculare Validität des Schülerfachwissenstests sicherzustellen, wurden im Rahmen einer Lehrplananalyse zunächst Unterthemen der Mechanik identifiziert, die sowohl an Gymnasien als auch an Gesamtschulen potenziell im Physikunterricht der Mittelstufe behandelt werden (vergl. Tabelle 7.6 auf Seite 117). Die Eignung der Aufgaben für den auf Grundlage dieser Unterthemen zusammengestellten Aufgabenpool für die Pilotierung wurde mit drei Fachdidaktikern und einer Lehrkraft diskutiert. Die Zuordnung der Aufgaben zu den Unterthemen erfolgte durch die Autorin, eine weitere Doktorandin und zwei studentische Hilfskräfte, wobei eine Aufgabe auch mehreren Unterthemen zugeordnet werden konnte. Auf Grundlage dieser Zuordnung wurde sicher gestellt, dass die beiden in der Hauptstudie eingesetzten Testhefte die Unterthemen der Mechanik möglichst breit abdeckten. Bei den Aufgaben handelte es sich außerdem zum Großteil um Aufgaben aus etablierten Testinstrumenten.

Da jede Schule ihr eigenes Curriculum hat und meist nur wenig verbindliche Vorgaben für die inhaltliche Gestaltung einer Unterrichtseinheit existieren, ist die Analyse der Kernlehrpläne allerdings nicht ausreichend, um die curriculare Validität des Testinstruments abzusichern. Bei der Post-Erhebung gaben die Lehrkräfte daher

im Lehrerfragebogen an, welche Unterthemen sie im Rahmen der Unterrichtseinheit Mechanik mit ihren Schülerinnen und Schülern behandelt hatten.

Ein Teil der Lehrkräfte ($N = 11$) nahm zusätzlich nach der Post-Erhebung an einem Expertenrating teil und führte selbst eine Zuordnung der Schülerfachwissenstestaufgaben zu den Unterthemen der Mechanik durch. Ursprünglich sollte dieses Vorgehen eine differenzierte Analyse des Schülerfachwissenstests ermöglichen. In jeder Klasse sollten die Aufgaben zu nicht behandelten Unterthemen der Mechanik in der Rasch-Analyse des Schülerfachwissenstest als designbedingte Missings behandelt werden, um so die curriculare Validität des Testinstruments für jede einzelne Klasse sicher zu stellen. Die Lehrkräfte waren sich im Expertenrating bei der Aufgabenzuordnung allerdings sehr uneins und interpretierten die Unterthemen der Mechanik sehr unterschiedlich. Darüber hinaus füllten die Lehrkräfte das Expertenrating in der Regel ziemlich schnell und möglicherweise eher oberflächlich durch und ordneten Aufgaben, die offensichtlich mehreren Unterthemen zugeordnet werden konnten, nur einzelnen Unterthemen zu. Die aus der Themenabfrage und aus dem Expertenrating generierten Daten wurden daher als zu unsicher eingestuft, um auf dieser Grundlage eine individualisierte Auswahl an Aufgaben für jede Klasse zu treffen und werden daher im Folgenden nur ergänzend zur Beurteilung der Validität hinzugezogen.

Aus Tabelle 7.9 auf der nächsten Seite wird ersichtlich, dass die meisten Unterthemen (mehr als 60%) von mehr als 70% der Lehrkräfte im Unterricht behandelt wurden. Lediglich zwei Unterthemen wurden von weniger als 50% der Lehrkräfte behandelt.

Dem Unterthema Zusammenhang von Geschwindigkeit und Beschleunigung, dass immerhin noch 39% der Lehrkräfte behandelten, wurden vom Forscherteam lediglich drei Aufgaben zugeordnet (vergl. Tabelle 7.6 auf Seite 117). Unter Einbezug des Expertenratings wurde diesem Thema lediglich eine Aufgabe zugeordnet[8], die aber gleichzeitig auch der qualitativen Beschreibung beschleunigter Bewegungen (behandelt von 52% der Lehrkräfte) zugeordnet werden konnte. Lediglich 4% der Lehrkräfte gaben an, die qualitative Beschreibung von Kreisbewegungen behandelt zu haben. Diesem Unterthema wurde vom Forscherteam allerdings nur eine Aufgabe zugeordnet, die aber gleichzeitig auch der Kategorie Trägheit zugeordnet werden konnte. Im Expertenrating wurde diese Aufgabe nicht eindeutig zugeordnet. 46% der Lehrkräfte ordnete die Aufgabe der Trägheit zu, lediglich 36% ordneten sie der qualitativen Beschreibung von Kreisbewegungen zu. Außerdem erwarteten 46% der im Expertenrating befragten Lehrkräfte, dass ihre Schülerinnen und Schüler diese Aufgabe lösen können sollten. Die Aufgabe wurde daher nicht aus den Analysen ausgeschlossen.

Im Mittel wurden die einzelnen Unterthemen von 74% der Lehrkräfte behandelt. Aufgeschlüsselt nach Jahrgangsstufen ergab sich ein ähnliches Bild, wobei in der Jahrgangsstufe 8 die einzelnen Unterthemen im Mittel von mehr Lehrkräften behandelt wurden (75%) als in der Jahrgangsstufe 9 (70%). Über alle Klassen

[8]Kriterium für die Zuordnung einer Aufgabe zu einem Unterthema war hierbei, dass sowohl das Forscherteam als auch mindestens 50% der Lehrkräfte diese Einordnung vorgenommen hatten.

Tabelle 7.9.
*Anteil der Lehrkräfte, die die im Schülerfachwissenstest adressierten Unterthemen im Rahmen der Unterrichtseinheit Mechanik behandelt haben*

| Unterthemen der Mechanik | Behandelt von % der LK |
|---|---|
| Kraftwirkungen/Kraft als Ursache von Bewegungsänderungen | 100% |
| Kraft und Gegenkraft | 87% |
| Kräftegleichgewicht | 96% |
| Addition von Kräften/Kräfteparallelogramm/ Komponentenzerlegung | 100% |
| Kraft als Vektor | 91% |
| Hebel | 70% |
| Gewichtskraft und Masse | 96% |
| Gewichtskraft an verschiedenen Orten | 78% |
| Gleichförmige Bewegung und ihre Voraussetzung | 74% |
| Qualitative Beschreibung beschleunigter Bewegungen | 52% |
| Qualitative Beschreibung von Kreisbewegungen | 4% |
| Geschwindigkeit | 83% |
| Zusammenhang von Geschwindigkeit und Beschleunigung | 39% |
| Trägheit | 65% |
| Energie/Arbeit/Leistung | 74% |
| Berücksichtigung von Reibung oder Luftwiderstand | 52% |
| Verwendung und Definition verschiedener Einheiten | 87% |
| Differenzierung zwischen Einheiten und Größen | 83% |
| Lesen und Interpretieren von Diagrammen | 65% |

hinweg wird der Inhalt der Schülerfachwissenstest daher als curriculumvalide für die Jahrgangsstufen 8 und 9 erachtet.

Der Anteil behandelter Unterthemen variierte für die einzelnen Lehrkräfte ($Min$ = 53%, $Max$ = 90%, $M$ = 73%, $SD$ = 10%). Zurückzuführen ist dies auf Unterschiede im Behandlungszeitraum der Unterrichtseinheit Mechanik, der zwischen 12 und 59 Unterrichtsstunden (normiert auf 45-Minuten-Stunden) variierte (vergl. Abschnitt 8.1.2.1 auf Seite 162). Die Länge der Unterrichtseinheit (definiert als die Anzahl der nach Lehrerangaben zwischen Prä- und Post-Test gehaltenen 45-Minuten-Stunden) wird als Prädiktor in alle Modelle zur Erklärung der Klassenunterschiede im Post-Test aufgenommen (vergl. Abschnitt 8.3.1.1 auf Seite 171 zum Kontrollvariablenmodell im Ergebnisteil dieser Arbeit). Unter dieser Voraussetzung

ist der Anteil behandelter Themen kein signifikanter Prädiktor für die Klassenmittelwerte der Post-Testwert der Lernenden am Ende der Unterrichtseinheit Mechanik ($\gamma^{StdYX} = .15 \pm .23$, $\mathrm{KI}_{95\,\%} = [-.29, .60]$, $p_{1\text{-seitig}} = .250$).

Obwohl der Schülerfachwissenstest nicht als curriculumvalide für jede einzelne Klasse angesehen werden kann, ist demnach dennoch ein fairer Vergleich zwischen den Klassen gewährleistet.

**Konstruktvalidität** Zur Beurteilung der Konstruktvalidität kann lediglich auf die Methode der diskriminanten Validierung zurückgegriffen werden. Um das im Schülertest erfasste Konstrukt *Fachwissen in Mechanik* von dem Konstrukt *Intelligenz* abzugrenzen, wurden Korrelationen zwischen den Prä- und Post-Testwerten und den mit dem KFT erhobenen kognitiven Fähigkeiten der Lernenden in der Stichprobe der $N = 610$ an beiden Testzeitpunkten anwesenden Lernenden berechnet (siehe Tabelle 7.10). Der KFT wird in Abschnitt 7.5.5.1 auf Seite 132 beschrieben.

Tabelle 7.10.
*Korrelationen zwischen den Prä- und Post-Testwerten und den kognitiven Fähigkeiten der Lernenden*

| $N = 610$ | $r_{\mathrm{Pearson}}$ | $\mathrm{KI}_{95\,\%}$ | $r_{\mathrm{Spearman}}$ | $\mathrm{KI}_{95\,\%}$ | $\tau_{\mathrm{Kendall}}$ | $\mathrm{KI}_{95\,\%}$ |
|---|---|---|---|---|---|---|
| Prä - KFT | $\mathbf{.32 \pm .04}$ | [.25, .39] | $\mathbf{.28 \pm .04}$ | [.21, .36] | $\mathbf{.197 \pm .027}$ | [.146, .250] |
| Post - KFT | $\mathbf{.31 \pm .04}$ | [.23, .39] | $\mathbf{.30 \pm .04}$ | [.22, .37] | $\mathbf{.206 \pm .027}$ | [.154, .260] |

*Anmerkungen.* Alle Korrelationen sind signifikant mit $p_{1\text{-seitig}} < .001$ und daher fett gedruckt. Da keine negativen Korrelationen zu erwarten sind, wurde einseitig auf Signifikanz getestet. Aufgrund signifikanter Abweichungen von der Normalverteilung bei allen Variablen werden zusätzlich nicht-parametrische Korrelationen berichtet.

Die Korrelationskoeffizienten für die beiden Messzeitpunkte unterscheiden sich kaum. Zu beiden Zeitpunkten besteht höchstens (bezogen auf $r_{\mathrm{Pearson}}$) eine mittlere Korrelation zwischen Schülertest und KFT. Die „wahren“ Korrelationen werden in messfehlerbehafteten Messungen jedoch stets unterschätzt (vergl. Abschnitt 7.4.6 auf Seite 101). Auch die um die Messfehler der Testinstrumente bereinigten Korrelationskoeffizienten ($r_{\mathrm{Pearson,korr.,Prä\text{-}KFT}} = .49$, $r_{\mathrm{Pearson,korr.,Post\text{-}KFT}} = .43$) zeigen allerdings eine deutliche Abgrenzung des mit dem Schülerfachwissenstest erfassten Konstrukts zu der Intelligenz der Lernenden.

**Kriteriumsvalidität** Im Rahmen einer kriterialen Validierung wurden die Korrelationen zwischen den Prä- und Post-Testwerten und den Schulnoten der Lernenden in den Fächern Physik, Mathematik und Deutsch betrachtet (siehe Tabelle 7.11 auf der nächsten Seite). Zu beiden Messzeitpunkten gaben die Lernenden die letzte ihnen bekannte Note in den besagten Fächern an (Note 1 =„sehr gut“ bis Note 6 =„ungenügend“). Je nach Messzeitpunkt handelte es sich um Zeugnis- oder Quartalsnoten.

Im Falle einer validen Erfassung des Wissens, dass den Lernenden im Rahmen der Mechanikeinheit vermittelt wurde, würde man eine höhere Korrelation des

Schülertests zur Physiknote als zur Deutschnote erwarten. Bezüglich der Korrelation zur Mathematiknote wäre eine niedrige Korrelation als zur Physiknote wünschenswert, allerdings, unter Berücksichtigung von Validierungsergebnissen andere Studien (Geller, 2015, S. 98; Schoppmeier, 2013, S. 71), nicht zwingend zu erwarten. Außerdem würde man in Bezug auf die Korrelation zur Physiknote eine höhere Korrelation zum Zeitpunkt des Post-Tests erwarten, da das zu erhebende Wissen hier bereits gelernt werden konnte (Geller, 2015, S. 92).

Tabelle 7.11.
*Korrelationen zwischen den Prä- und Post-Testwerten und den Schulnoten der Lernenden*

| $N = 600$[1] | | Physiknote | Mathematiknote | Deutschnote |
|---|---|---|---|---|
| Prä-Test | $r_{\text{Spearman}}$ | $\mathbf{-.20 \pm .04}$ | $\mathbf{-.20 \pm .04}$ | $\mathbf{-.14 \pm .04}$ |
| | $\text{KI}_{95\,\%}$ | $[-.27, -.12]$ | $[-.27, -.12]$ | $[-.21, -.06]$ |
| Post-Test | $r_{\text{Spearman}}$ | $\mathbf{-.27 \pm .04}$ | $\mathbf{-.31 \pm .04}$ | $\mathbf{-.13 \pm .04}$ |
| | $\text{KI}_{95\,\%}$ | $[-.34, -.19]$ | $[-.38, -.23]$ | $[-.21, -.05]$ |

*Anmerkungen.* Alle Korrelationen sind signifikant mit $p_{\text{1-seitig}} < .001$ und daher fett gedruckt. Da keine negativen Korrelationen zu erwarten sind, wurde einseitig auf Signifikanz getestet. Da Schulnoten nicht als intervallskaliert angenommen werden können, wurden Spearman-Rangkorrelationen berechnet. Negative Korrelationen sind an dieser Stelle aufgrund der Polung der Notenskala im deutschen Schulsystem erwünscht.

[1] In der Stichprobe der $N = 610$ Lernenden, die zu beiden Testzeitpunkten anwesend waren, lagen lediglich für $N = 600$ Lernende vollständige Notenangaben vor.

Die Korrelationen zwischen Schülertest und der Physik- bzw. Mathematiknote zum Zeitpunkt des Prä-Tests sind identisch und auch zum Zeitpunkt des Post-Tests unterscheiden sich die Korrelationen nicht signifikant voneinander ($Z(600) = 1.06$, $p_{\text{1-seitig}} = .290$).[9] Die Korrelation zur Physiknote ist zu beiden Testzeitpunkten höher als zur Deutschnote, wobei der Unterschied in den Korrelationen aber erst für den Post-Test signifikant wird (Prä-Test: $Z(600) = -1.36$, $p_{\text{1-seitig}} = .087$; Post-Test: $Z(600) = -2.83$, $p = .003$). Die Korrelation zur Physiknote ist für den Post-Test zwar höher, unterscheidet sich für Prä- und Post-Test jedoch nicht signifikant voneinander ($Z(600) = 1.425$, $p_{\text{1-seitig}} = .077$). Bereinigt man die Korrelationen um den Messfehler der Schülertestinstrumente, ändern sich die Befunde lediglich dahingehend, dass der Unterschied zwischen der Korrelation zur Physik- bzw. zur Deutschnote auch für den Prä-Test signifikant wird ($Z(600) = -1.934$, $p_{\text{1-seitig}} = .027$).

Erwartungsgemäß korreliert der Schülerfachwissenstest also höher mit der Physiknote als mit der Deutschnote der Lernenden. Wie in anderen Studien auch, gilt dies allerdings nicht für den Vergleich zur Korrelation mit der Mathematiknote der Lernenden. Dieser Umstand ist insofern nicht verwunderlich, da die Physik- und Mathematiknoten ihrerseits relativ hoch korreliert sind (Prä: $r_{\text{Spearman}} = .561 \pm .029$,

[9] Die einseitige Testung ist auch hier angemessen, da die Hypothese beinhaltet, dass die Korrelation zur Physiknote größer als die Korrelation zur Mathematiknote ist.

$\mathrm{KI}_{95\,\%} = [.500, .615]$, $p_{\text{1-seitig}} < .001$; Post: $r_{\text{Spearman}} = .50 \pm .03$, $\mathrm{KI}_{95\,\%} = [.44, .57]$, $p_{\text{1-seitig}} < .001$). Auch der Befund, dass der Unterschied zwischen den Korrelationen zur Physiknote im Prä- und Post-Test nicht signifikant wird, stellt keinen Grund zur Beunruhigung dar. Zwischen den Messzeitpunkten erfolgte nicht zwingend eine Notenvergabe durch die Lehrkräfte. Die Lernenden gaben dementsprechend zum Teil die gleiche Note wie im Prä-Test an, die sich demnach nicht auf den Zeitraum der Unterrichtseinheit bezog. Der Zusammenhang der Physiknote zum Post-Testergebnis der Lernenden wird daher wahrscheinlich eher unterschätzt.

Als weiteres Kriterium zur Beurteilung der Validität des Schülertests wird der Zeitraum für Lerngelegenheiten und damit die Länge der Unterrichtseinheit (definiert als die Anzahl der nach Lehrerangaben zwischen Prä- und Post-Test gehaltenen 45-Minuten-Stunden) betrachtet. Wenn der Schülertest Wissen erfasst, dass im Rahmen der Unterrichtseinheit Mechanik von den Schülerinnen und Schülern gelernt werden konnte, sollte die Länge der Unterrichtseinheit ein bedeutsamer Prädiktor für die Post-Testergebnisse der Lernenden sein. In Abschnitt 8.3.1.1 auf Seite 171 des Ergebnisteils dieser Arbeit wird gezeigt, dass die Länge der Unterrichtseinheit im Rahmen eines Mehrebenenmodells $(65 \pm 18)\%$ der zwischen den Klassen liegenden Varianz im Post-Test aufklärt ($\gamma_{\text{Zeit}}^{\text{StdYX}} = 0.80 \pm 0.12$, $\mathrm{KI}_{95\,\%} = [0.60, 1.01]$, $p_{\text{1-seitig}} < .001$). Auch dieser Befund spricht für eine valide Erfassung des Konstruktes.

Zusammengenommen weisen die Validierungsergebnisse darauf hin, dass der Schülerfachwissenstest das physikalische Fachwissen misst, das Lernenden der Jahrgangsstufe 8 oder 9 im Rahmen einer Unterrichtseinheit zur Mechanik vermittelt wird. Der Test konnte das vermittelte Wissen zwar nicht für jede einzelne Klasse curriculumvalide erfassen, dennoch konnte gezeigt werden, dass der Schülerfachwissenstest einen fairen Vergleich zwischen den Leistungen verschiedener Klassen ermöglicht.

### 7.5.4. Fragebogen zum situationalen Interesse am Unterricht

Im Rahmen des Gesamtprojektes wurde jeweils am Ende der videographierten Unterrichtsstunden die aktuelle Motivation der Lernenden erfasst. Hierfür wurde in Zusammenarbeit mit den Projektpartnern aus der Psychologie der *Fragebogen zur aktuellen Motivation* (FAM) von Rheinberg, Vollmeyer und Burns (2001, S. 66) adaptiert, der sich ursprünglich auf die aktuelle Motivation bei der Bearbeitung von Aufgaben bezieht. Der FAM bildet die Skalen *Herausforderung*, *Misserfolgsbefürchtung*, *Erfolgswahrscheinlichkeit* und *Interesse* ab. Die adaptierte Version des FAM wird im Folgenden als „FAM-Video“ bezeichnet. Wie auch der FAM umfasst der FAM-Video 18 Items. Allerdings konnte jeweils ein Item der Skalen *Erfolgswahrscheinlichkeit* (Item 3) und *Interesse* (Item 11) nicht sinnvoll in den FAM-Video übersetzt werden (vergl. Rheinberg et al., 2001, S. 66). Stattdessen wurden zwei Items zur Erfassung der *Zielklarheit* ergänzt.

Das situationale Interesse der Lernenden am Unterricht wird in dieser Arbeit mit der Interessenskala des FAM-Video erfasst. Die Skala umfasst vier Items, die

auf einer siebenstufigen Likert-Skala (1 =„stimme gar nicht zu“, 7 =„stimme voll zu“) beurteilt werden können:

*FAM-V1*: *„Ich mag solche Unterrichtsstunden wie die heute.“*

*FAM-V3*: *„Im Unterricht heute mochte ich die Rolle des Wissenschaftlers, der Zusammenhänge entdeckt.“*

*FAM-V6*: *„Ich fand diese Unterrichtsstunde sehr interessant.“*

*FAM-V15*: *„Solche Themen wie heute im Unterricht würde ich auch in meiner Freizeit bearbeiten.“*

#### 7.5.4.1. Technische Details zur Auswertung

Das situationale Interesse der Lernenden wurde über den Skalenmittelwert der Interessenskala des FAM-Video gemessen. Die Auswertung der ersten (1M) und zweiten Unterrichtstunde (2M) erfolgte getrennt. Nicht bearbeitete Items wurden durch den Mittelwert aus dem Itemmittelwert in der Stichprobe aller Schülerinnen und Schüler und dem Skalenmittelwert des Lernenden, dessen Datensatz einen fehlenden Wert enthielt, ersetzt. Diese Vorgehensweise hat den Vorteil, dass zum einen berücksichtigt wird, ob einem Item in der Regel eher gar nicht oder voll zugestimmt wurde (vergl. Rost, 2004, S. 327), und zum anderen, ob der Lernenden, der das Item nicht bearbeitet hat, in den restlichen Items der Skala eher Ablehnung oder Zustimmung signalisierte. In der ersten Unterrichtsstunde wurden in 5% der Fälle lediglich drei der vier Items der Interessenskala des FAM-Video bearbeitet. In der zweiten Unterrichtstunde galt dies für 3% der Fälle. Die Abweichung zwischen den Skalenmittelwerten der Lernenden vor und nach der Ersetzung fehlender Werte betrug maximal 0.54 bzw. 0.40 Punkte (entspricht 0.45 bzw. 0.29 Standardabweichungen im situationalen Interesse der Lernenden in der ersten bzw. zweiten Unterrichtsstunde), war aber im Mittel noch weitaus geringer (1M: $N = 31$, $M = -0.05$, $SD = 0.19$; 2M: $N = 31$, $M = -0.04$, $SD = 0.17$).[10]

#### 7.5.4.2. Objektivität, Reliabilität, Validität

Der FAM-Video kann aufgrund der standardisierten Testdurchführung (vergl. Abschnitt 7.2.3.3 auf Seite 87) und des Umstands, dass es sich zum einen um Selbsteinschätzungen auf einer Likertskala handelt und zum anderen lediglich Vergleiche des situationalen Interesses innerhalb der Stichprobe auf Klassenebene

[10] Betrachtet man lediglich den mittleren Fehler auf die Skalenmittelwerte von Lernenden mit fehlenden Werten in einem Item, könnten die Skalenmittelwerte für das situationale Interesse der Lernenden mit zwei Dezimalstellen angegeben werden. Dies würde allerdings eine Präzision suggerieren, die diese Messung wahrscheinlich nicht erfüllt. Die Skalenmittelwerte werden daher stets mit einer Dezimalstelle angegeben (Eine Fehlerberechnung ist an dieser Stelle nicht weiterführend, da keine Populationsmittelwerte geschätzt werden).

erfolgen, bezüglich der Durchführung, der Auswertung und der Interpretation als objektives Testinstrument betrachtet werden.

Die Reliabilität der Interessenskala des FAM-Video ist sowohl für die erste Unterrichtsstunde ($N = 633$, $\alpha = .74$, $\mathrm{KI}_{95\,\%} = [.68, .80]$) als auch für die zweite Unterrichtsstunde ($N = 625$, $\alpha = .80$, $\mathrm{KI}_{95\,\%} = [.74, .85]$) zufriedenstellend bis gut. Die Ersetzung fehlender Werte zeigte keinen Einfluss auf die Reliabilität.

Der FAM ist ein mehrfach validiertes Testinstrument zur Erfassung der aktuellen Motivation von Lernenden bei der Bearbeitung von Aufgaben (vergl. Rheinberg et al., 2001, S. 60-64). In der Regel wird der FAM zur Messung der aktuellen Motivation als unabhängige Variable eingesetzt. Die Validierung der einzelnen FAM-Skalen bezieht sich daher auf deren prognostische Validität in Bezug auf Lernverhalten und Lernleistungen. Über die Validierung mit bekannten Gruppen konnten Rheinberg et al. (2001, S. 61-62) beispielsweise in einer Studie ihre Hypothese bestätigen, dass die Interessenskala des FAM Leistungsvorhersagen beim selbstgesteuerten Verständnislernen, nicht aber beim fragengeführten Faktenlernen mit einem Lernprogramm erlaubt. In einer weiteren Studie konnten die Autoren außerdem zeigen, dass sich die spätere Lernleistung durch das mit dem FAM gemessene Interesse am Aufgabeninhalt einer komplexen Lernaufgabe nur bei langsamen Lernern vorhersagen lässt, deren aktuelle Motivation bei der Aufgabenbearbeitung stärker durch ihr Interesse am Aufgabeninhalt bestimmt ist (Rheinberg et al., 2001, S. 63-64). Die von Rheinberg et al. (2001) zusammengefassten Validitätshinweise deuten auf eine valide Erfassung des situationalen Interesses an Aufgabeninhalten durch den FAM hin. Es wird daher davon ausgegangen, dass auch die Interessenskala des FAM-Video das situationale Interesse der Lernenden am Unterricht valide erfasst.

Das situationale Interesse der Lernenden in der ersten und zweiten Unterrichtsstunde ist nicht konstant, die Maße korrelieren aber (sowohl auf Schülerebene als auch auf Klassenebene) hoch miteinander (vergl. Tabelle 7.12 auf der nächsten Seite). Dieser Befund kann als weiterer Hinweis auf die Validität des FAM-Video interpretiert werden, da es sich hierbei um einen erwartungsgemäßen Befund handelt: Der FAM-Video bezieht sich auf das situationale Interesse in einer konkrete Unterrichtsstunde, das im Gegensatz zum individuellen Fachinteresse kein stabiles Merkmal darstellt (vergl. Abschnitt 5.2 auf Seite 63). Dass das situationale Interesse in den beiden Unterrichtsstunden hoch korreliert, spricht allerdings dafür, dass Lehrkräfte, die das situationale Interesse der Lernenden in *einer* Unterrichtstunde wecken, vermutlich auch in der Lage sind, dies ebenfalls in anderen Unterrichtsstunden zu tun.

### 7.5.5. Erhebung der Kontrollvariablen

Im Rahmen dieser Arbeit soll untersucht werden, ob sich das Fachwissen der Lernenden am Ende der Unterrichtseinheit zur Mechanik durch das Professionswissen der Lehrkräfte und die kognitiv aktivierende Gestaltung des Unterrichts vorhersagen lässt. Das Fachwissen der Schülerinnen und Schüler am Ende einer Unterrichtseinheit hängt davon ab, über welches Wissen bezüglich des Lerngegenstands die Lernenden schon vor Beginn des Unterrichts verfügten. Daher werden

Tabelle 7.12.
*Korrelationen zwischen den Maßen für das situationale Interesse der Lernenden in der 1. und 2. Unterrichtsstunde auf Schülerebene und auf Klassenebene*

| Merkmale | Korrelation auf Schülerebene | Korrelation auf Klassenebene |
|---|---|---|
| $N$ | 600 | 23 |
| $r_{\text{Pearson}}$ | **.619 ± .028** | |
| $\text{KI}_{95\,\%}$ | [.561,.674] | |
| $p_{\text{1-seitig}}$ | < .001 | |
| $r_{\text{Spearman}}$ | **.607 ± .028** | **.63 ± .17** |
| $\text{KI}_{95\,\%}$ | [.546,.660] | [.23,.86] |
| $p_{\text{1-seitig}}$ | < .001 | .001 |
| $\tau_{\text{Kendall}}$ | **.461 ± .023** | **.49 ± .15** |
| $\text{KI}_{95\,\%}$ | [.410,.507] | [.17,.73] |
| $p_{\text{1-seitig}}$ | < .001 | < .001 |

*Anmerkungen.* Signifikante Korrelationen mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Aufgrund signifikanter Abweichungen von der Normalverteilung werden auf Schülerebene zusätzlich und auf Klassenebene lediglich nicht-parametrische Korrelationen berichtet. Auf Klassenebene wurden die in den Klassen gemittelten Werte korreliert.

die Schülerfähigkeiten im Prä-Test als Prädiktor in alle Modelle zur Erklärung der Varianz in den Post-Testergebnissen aufgenommen und das Vorwissen damit kontrolliert. Das Abschneiden der Lernenden im Post-Test wird allerdings auch durch zahlreiche weitere Individualmerkmale beeinflusst. Einen bedeutsamen Einfluss auf Schülerleistung haben die kognitiven Fähigkeiten der Lernenden oder die zuhause gesprochene Sprache (vergl. Fischer et al., 2014b, S. 19; Pöhlmann et al., 2013, S. 324), die daher als Kontrollvariable erhoben wurden. Des Weiteren kann auch die Lesefähigkeit der Lernenden einen Einfluss auf deren Testergebnisse haben (vergl. z. B. Baumert et al., 2010, S. 9). Da der Schülerfachwissenstest ein Multiple-Choice Test ist und die Aufgabenstellungen nur wenig Text enthalten, wurde auf den Einsatz eines Testinstruments zur Erhebung der Lesefähigkeit verzichtet. Nachteile von mehrsprachig aufwachsenden Lernenden, die oft zur Gruppe schwächerer Leserinnen und Leserinnen gehören (Rjosk, McElvany, Anders & Becker, 2011), werden allerdings durch die Kontrolle der zuhause gesprochenen Sprache berücksichtigt. Das Fachwissen der Schülerinnen und Schüler im Post-Test hängt darüber hinaus von dem Zeitraum für Lerngelegenheiten ab. Da die Länge der Unterrichtseinheit in den Klassen stark variierte, wurde die Unterrichtszeit als Kontrollvariable auf Klassenebene erfasst. Um Einschätzen zu können, ob die videographierten Unterrichtstunden authentische Beispiele für das Unterrichtsgeschehen in den untersuchten Klassen darstellen, wurden außerdem Indikatoren für die Repräsentativität des videographierten Unterrichts erfasst. Im Folgenden wird beschrieben, wie die genannten Kontrollvariablen operationalisiert und erhoben wurden.

#### 7.5.5.1. Kognitive Fähigkeiten der Lernenden

Zur Erfassung der kognitiven Fähigkeiten der Lernenden wurde auf ein standardisiertes Testinstrument zurückgegriffen: den *Kognitive Fähigkeitentest* (KFT) von Heller und Perleth (2000). Der KFT besteht aus insgesamt neun Untertests, die den Dimensionen *Verbale Fähigkeiten* (V1-V3), *Quantitative Fähigkeiten* (Q1-Q3) und *Figural-räumliche Fähigkeiten* (N1-N3) zugeordnet werden können. Aus zeitökonomischen Gründen wurde zur Erfassung der kognitiven Fähigkeiten der Lernenden nur eine Skala des KFT eingesetzt. Hierfür wurde in Anlehnung an die IPN-Videostudie die Skala N2 (Figurenanalogien) der Testversion A des KFT ausgewählt (vergl. Seidel, Rimmele & Dalehefte, 2003). Im Folgenden bezieht sich die Bezeichnung „KFT“ daher lediglich auf diese Teilskala.

Der KFT besteht aus 30 Multiple-Choice Aufgaben (Single-Select, 5 Antwortmöglichkeiten), die auf zwei verschiedene Testhefte mit jeweils 25 Aufgaben für die Jahrgangsstufen 8 und 9 verteilt sind. Die Testhefte sind über 20 identische Aufgaben miteinander verankert. In den Aufgaben werden jeweils zwei Figuren gezeigt, die in einem bestimmten Verhältnis zueinander stehen. Aufgabe der Schülerinnen und Schüler ist es, zu einer dritten Figur aus den Antwortmöglichkeiten diejenige Figur herauszufinden, die mit der dritten Figur in gleicher Relation steht, wie die beiden ersten zueinander. Für den KFT war eine feste Bearbeitungszeit von acht Minuten vorgegeben.

**Technische Details zur Auswertung** Für jede korrekt gelöste Aufgabe im KFT wurde ein Punkt vergeben. Da es sich beim KFT um einen Speed-Test handelt (Heller & Perleth, 2000, S. 8) und eine feste Bearbeitungszeit vorgegeben war, wurden nicht bearbeitete Aufgaben als Ausdruck geringerer Fähigkeit betrachtet. Sie wurden daher nicht als fehlende Werte behandelt, sondern stattdessen mit null Punkten bewertet. Aufgrund des durch die verschiedenen Testhefte bedingten Multi-Matrix-Designs, wurde der KFT im Rahmen einer Rasch-Analyse ausgewertet (vergl. Abschnitt 7.4.2 auf Seite 93). Heller und Perleth (2000, S. 19) schreiben „[Bei der Rasch-Analyse des KFT] ergaben sich zwar in den meisten Fällen signifikante Abweichungen vom Rasch-Modell aufgrund einzelner Items, jedoch zeigen die Analysen auch, daß die meisten Items jedes Subtests als Rasch-homogen angesehen werden können“. Für die Schätzung der Personenfähigkeiten im Rahmen eines eindimensionalen dichotomen Rasch-Modells wurden daher keine Aufgaben entfernt, obwohl acht Aufgaben signifikanten Misfit ins Rasch-Modell zeigten. In Abbildung B.4 auf Seite 252 im Anhang findet sich die Wright-Map für die Aufgaben des KFT.

**Objektivität, Reliabilität, Validität** Beim KFT handelt es sich um ein standardisiertes Multiple-Choice-Testinstrument. Daher wird vorausgesetzt, dass der KFT sowohl bezüglich der Durchführung als auch bezüglich der Auswertung und Interpretation als objektives Testinstrument betrachtet werden kann. Die Reliabilitäten des KFT in der hier untersuchten Stichprobe wurden im Rahmen der Rasch-Analyse bestimmt und zeigten gute Werte (vergl. Tabelle 7.13 auf der nächsten Seite).

Tabelle 7.13.
*Reliabilität des Kognitive Fähigkeitentests in der Stichprobe der $N = 23$ Gymnasialklassen*

| | | KFT |
|---|---|---|
| $N_{\text{Personen}}$ | | 640 |
| $N_{\text{Aufgaben}}$ | | 30 |
| Personenreliabilität | Real | .83 |
| | Model | .84 |
| Itemreliabilität | Real | .99 |
| | Model | .99 |

Bei der Normalform des KFT handelt es sich um ein vielfach validiertes Testinstrument (Heller & Perleth, 2000, 27ff). Da der KFT in dieser Arbeit nicht für die individuelle Intelligenzdiagnostik, sondern lediglich zur Erfassung einer Kontrollvariable eingesetzt wird und demnach nur eine begrenzte Testzeit zur Verfügung stand, wurde allerdings nur eine der neun Skalen des KFT eingesetzt. Heller und Perleth (2000, S. 47) konnten zeigen, dass die *Allgemeine Intelligenz* am stärksten durch die nonverbalen Skalen N1-N3 zur Erfassung der figural-räumlichen Fähigkeiten bestimmt wird. Diese Skalen „[...] prüfen das logische Denken mit besonderem Bezug zu anschauungsgebundenen Aspekten räumlichen Vorstellungsvermögens“ (Heller & Perleth, 2000, S. 39). Im Vergleich mit den Skalen N1 und N3 korreliert die Skala N2 in den Jahrgangsstufen 8 und 9 am höchsten mit den sechs anderen Skalen zur Erfassung der verbalen und quantitativen Fähigkeiten ($.35 \leq r_{\text{Pearson}} \leq .59$)(Heller & Perleth, 2000, S. 26). Es wird daher davon ausgegangen, dass die hier eingesetzte Skala N2 als valider Indikator für die kognitiven Fähigkeiten der Lernenden betrachtet werden kann.

#### 7.5.5.2. Zuhause gesprochene Sprache der Lernenden

Der sprachliche Hintergrund der Lernenden wurde über die zuhause gesprochene Sprache operationalisiert (vergl. z. B. Geller, 2015; Rjosk et al., 2011) und über die Frage „Welche Sprache wird bei euch zuhause hauptsächlich gesprochen?“ erfasst. Während die Schülerinnen und Schüler angeben konnten, ob sie zuhause hauptsächlich „deutsch“, „deutsch und andere“ oder „andere“ Sprachen sprechen, wurde die Variable für die Auswertung dichotomisiert und lediglich unterschieden, ob die Lernenden einen einsprachigen („0“) oder mehrsprachigen („1“) Hintergrund haben.

#### 7.5.5.3. Unterrichtszeit

Die Lehrkräfte gaben zum Zeitpunkt der Post-Erhebung an, wie viele Unterrichtsstunden sie in der Unterrichtseinheit zur Mechanik zwischen dem Prä- und dem Post-Test unterrichtet hatten. Die Unterrichtszeit wurde auf Basis dieser Angaben

aus der Anzahl unterrichteter Stunden multipliziert mit der Stundenlänge, die zwischen 45 und 90 Minuten variierte, berechnet. Einige Lehrkräfte schätzten die Anzahl unterrichteter Stunden allerdings lediglich über den Behandlungszeitraum der Unterrichtseinheit ab, ohne dabei Ferienzeiten, Feiertage oder weitere Gründe für Unterrichtsausfall zu berücksichtigen. Die Angaben der Lehrkräfte wurde daher auf Plausibilität überprüft. In drei Fällen (ID 6, ID 20 und ID 23) gaben die Lehrkräfte drei bzw. zwei Stunden mehr an, als theoretisch, nach Abzug von Ferienzeiten und Feiertagen, zwischen Prä- und Post-Erhebung hätten stattfinden können. Die Angaben dieser Lehrkräfte wurden daher nach unten korrigiert. Für eine weitere Lehrkraft (ID 25) fehlte die Angabe zur Anzahl der Stunden und wurde daher durch die Anzahl der Stunden ersetzt, die theoretisch zwischen den Erhebungen hätten stattfinden können.

#### 7.5.5.4. Repräsentativität des videographierten Unterrichts

Um die Repräsentativität der videographierten Unterrichtsstunden für den Unterricht der Lehrkräfte im Allgemeinen einschätzen zu können, wurden zum einen die durch das Filmen des Unterrichts hervorgerufene Nervosität sowohl der Lernenden als auch der Lehrkraft erhoben. Zum anderen wurde erfasst, als wie typisch die eingesetzten Unterrichtsmethoden und als wie ähnlich das Verhalten der Lehrkraft bzw. der Lernenden empfunden wurde. Hierfür wurden jeweils am Ende der videographierten Unterrichtstunden Lehrer- und Schülerfragebögen aus dem QuiP Projekt eingesetzt (Fischer et al., 2014a). Die Nervosität während der ersten zehn Minuten der videographierten Unterrichtsstunde und während der restlichen Unterrichtszeit wurde mit einer fünfstufigen Likertskala (1 =„sehr“, 2 =„ziemlich“, 3 =„etwas“, 4 =„nur wenig“, 5 =„überhaupt nicht nervös“) erhoben. Wie typisch die verwendeten Unterrichtsmethoden und wie ähnlich das Schüler- bzw. Lehrerverhalten im Vergleich zum sonstigen Physikunterricht war, wurde ebenfalls mit einer fünfstufigen Likertskala erhoben (1 =„sehr“, 2 =„größtenteils“, 3 =„einigermaßen“, 4 =„wenig“, 5 =„überhaupt nicht typisch bzw. ähnlich“). Die Lehrkräfte schätzen zusätzlich ein, inwieweit das Verhalten der Klasse in Bezug auf Konzentration, Unruhe, Engagement, Lautstärke und Ablenkung vom üblichen Verhalten abwich (1 =„weniger als sonst“, 2 =„etwa gleich wie sonst“, 3 =„stärker als sonst“).

## 7.6. Beschreibung des videobasierten Ratinginstruments zur Beurteilung der kognitiven Aktivierung im Unterricht

Die kognitiv aktivierende Gestaltung der Unterrichtsstunden wurde in dieser Arbeit auf Basis eines Videoratings beurteilt. Hierfür wurde in Zusammenarbeit mit den Projektpartnern aus der Biologie und der Psychologie[11] das an der Universität Paderborn entwickelte videobasierte Ratinginstrument von Vogelsang (2014)

[11] Mein Dank für die Zusammenarbeit gilt an dieser Stelle Christian Förtsch, Sonja Werner, Tobias Dorfner (Biologie, Arbeitsgruppe Neuhaus, Technische Universität München) und Gerlinde

adaptiert. Das Paderborner Rating zur kognitiven Aktivierung wurde auf Basis des Ratinginstruments zur „Unterstützung bei der Konstruktion von Wissen“ aus der Pythagoras-Studie (Rakoczy & Pauli, 2006) entwickelt[12], das wiederum auf Instrumenten aus der IPN-Videostudie (Widodo & Duit, 2004) und auf Ratinginstrumenten von Clausen (2002), Clausen, Reusser und Klieme (2003) und von Kunter (2005) aufbaut (vergl. Vogelsang, 2014, S. 302). Das Rating erfasst „...[die] Handlungen einer Lehrperson, die Schüler zu aktiven und herausfordernden Lernprozessen anregen“ (Vogelsang, 2014, S. 311).

Im Folgenden wird zunächst die in dieser Studie eingesetzte adaptierte Version des Ratings zur kognitiven Aktivierung beschrieben, Unterschiede zum Paderborner Ratinginstrument und zu dessen Auswertung werden erläutert und das Ratingverfahren sowie das Ratertraining wird beschrieben. Außerdem werden technische Details zur Auswertung der Ratings erläutert. Abschließend wird die Objektivität, Reliabilität und die Validität des Ratings diskutiert.

### 7.6.1. Rating zur kognitiven Aktivierung im Unterricht

Die kognitive Aktivierung im Unterricht wurde im Rahmen eines Overall-Ratings über die Einschätzung von 39 Handlungsindikatoren auf einer dreistufigen Ratingskala (1 =„trifft nicht zu“, 2 =„teils teils“, 3 =„trifft zu“) beurteilt. Die Handlungsindikatoren verteilen sich auf sieben Subskalen, die die in Abschnitt 5.3 auf Seite 65 beschriebenen Merkmalen eines kognitiv aktivierenden Unterrichts beschreiben. In dem für das Ratingverfahren genutzten Ratingmanual (vergl. Anhang A.4 auf Seite 222) wird zunächst die Grundidee jeder Subskala beschrieben. Anschließend werden die Handlungsindikatoren aufgelistet und Beispiele für die möglichen Indikatorausprägungen aufgeführt. Tabelle 7.14 auf der nächsten Seite zeigt eine Übersicht über die zu den einzelnen Subskalen gehörenden Handlungsindikatoren und deren Kurzbeschreibung.

### 7.6.2. Unterschiede zum Paderborner Ratinginstrument

Da mit dem Paderborner Ratingmanual keine zufriedenstellenden Interrater-Übereinstimmungen erzielt werden konnten ($ICC_{\text{2-fakt.,unjust}} = .19 - .65$ auf Subskalenebene, vergl. Vogelsang, 2014, S. 341 und S. xxvii im Anhang), erfolgte das Rating der kognitiven Aktivierung in der Studie von Vogelsang (2014) auf Grundlage einer Konsensbildung zwischen den Ratern und Raterinnen. In dem Versuch eine objektivere Beurteilung der kognitiven Aktivierung auf Basis des Ratingmanuals zu ermöglichen, wurde die im Paderborner Instrument eingesetzte vierstufige Ratingskala in Absprache mit Vogelsang in eine dreistufige Ratingskala überführt (persönliche Kommunikation, Frühjahr 2014). Außerdem wurden im Ratingmanual

Lenske (Psychologie, Arbeitsgruppe Leutner/Wirth, Universität Duisburg-Essen/Ruhruniversität Bochum).

[12] Die Skala „Unterstützung bei der Konstruktion von Wissen“ wird in der Pythagoras-Studie an anderer Stelle auch unter dem Begriff „kognitive Aktivierung“ aufgeführt (vergl. Hugener, 2006, S. 47).

Tabelle 7.14.
*Subskalen und Handlungsindikatoren zur Beurteilung der kognitiven Aktivierung*

| Indikator | Kurzbeschreibung |
|---|---|
| **Skala A** | **Lernstatus bewusst machen** |
| A1 | Bezug zu vorangegangenen Stunden |
| A2[1] | Bezug zu konkreten Zeitpunkten in der Vergangenheit |
| A3 | Verweis auf weiterführende Themen |
| A4 | Expliziter Ausblick auf Inhalt der Stunde |
| A5 | Rückblick auf bereits Gelerntes |
| **Skala B** | **Exploration des Vorwissens und der Vorstellungen** |
| B1[2] | Durchführung von Brainstormings |
| B2 | Frage nach Vorwissen/Vorstellungen ohne Abzielen auf bestimmte Antwort |
| B3 | Anregung, Thema nach eigenem Verständnis zu erläutern |
| B4 | Frage nach Ideen/Vorstellungen ohne Wertung |
| B5[1,2] | Anregung Thema mit bekannten Begriffen zu verbinden |
| **Skala C** | **Exploration der Denkweisen** |
| C1 | Frage wie SuS zu bestimmte Antworten gelangt sind |
| C2 | Forderung von Begründungen für Antworten |
| C3[1,2] | Frage was SuS verstanden haben |
| C4[1,2] | Bei Verständnisschwierigkeiten, Frage nach Denkprozessen |
| C5 | Anregung Sachverhalte in eigenen Worten zu erläutern |
| C6 | Häufig Wie-und Warum-Fragen |
| **Skala D** | **Evolutionärer Umgang mit Schülervorstellungen** |
| D1 | Aufgreifen und Verwenden von Ideen der SuS |
| D2[1] | Unterscheidung Wissenschafts-/Alttagssprache |
| D3[2] | Einführung wissenschaftl. Begriffe ausgehend von SV |
| D4[1,2] | Belastung nicht korrekter Vorstellungen durch Aufzeigen von Widerspruch |
| D5 | Aufforderung, auf Wissenstand aufbauend zu argumentieren/schlussfolgern |
| D6 | SuS in die Irre gehen lassen, bis sie es selbst merken |
| **Skala E** | **Lehrperson als Mediator** |
| E1[1,2] | Beiträge der SuS aufeinander beziehen |
| E2 | Aufforderung, Beiträge selbst aufeinander zu beziehen |
| E3 | Nachfrage bei missverständlichen/unvollständigen Äußerungen |
| E4[1,2] | Unterstützung beim Ausformulieren von Ideen |
| E5 | Forderung von Begründungen |
| E6 | Ball an andere SuS weitergeben, statt Antworten sofort zu bewerten |
| E7 | Zeit zum Finden von Ideen/Antworten |
| E8 | Aktive Beteiligung der SuS durch eigene Beiträge |
| **Skala F** | **Rezeptives Lernverständnis der Lehrperson (negativ gepolt)** |
| F1 | Kleinschrittiges Frageverhalten |
| F2 | Kleinschrittige/Rezeptartige Arbeitsanweisungen |
| F3 | SuS sind lediglich Stichwortgeber |
| F4[1] | Betonung des Auswendiglernen/genauen Wiedergebens |
| **Skala G** | **Herausfordernde Lerngelegenheiten** |
| G1[2] | Aufgaben-/Fragestellungen, die mehr als Ja-/Nein-Antworten verlangen |
| G2 | Schwerpunkt auf Aufgaben-/Fragestellungen, die zum Nachdenken anregen |
| G3 | Aufgaben-/Fragestellungen die Vergleichen & Analysieren erfordern |
| G4[1,2] | Frage nach Hypothesen in Experimentiersituationen |
| G5 | Aufgaben-/Fragestellunge, die nicht nur auswendig gelerntes Wissen abfragen |

[1/2] Indikator wurde nach Analysen in Rating zur 1./2. Unterrichtsstunde entfernt
(vergl. Abschnitt 7.6.5 auf Seite 141)

für jeden Handlungsindikator Beispiele für die jeweiligen Indikatorausprägungen ergänzt. Aufgrund der Zusammenarbeit mit den Projektpartnern aus der Biologie, handelt es sich bei diesen Beispielen um Situationsbeschreibungen aus dem Physik- oder Biologieunterricht.

Handlungsindikatoren, die im Rahmen der Analysen von Vogelsang (2014) als problematisch identifiziert und aus der Skalenbildung ausgeschlossen wurden, wurden nicht in das adaptierte Rating aufgenommen (vergl. Vogelsang, 2014, S. 375 und S. xli im Anhang). Auf Basis der Ergebnisse einer von den Projektpartnern aus der Biologie durchgeführten Pilotierungsstudie wurde außerdem ein Handlungsindikator der Subskala *Herausfordernde Lerngelegenheiten* entfernt.[13] Die Skala *Lehrperson als Mediator* wurde um einen Handlungsindikator („Der Lehrkraft gelingt es, die Schüler durch eigene Beiträge aktiv am Unterricht zu beteiligen“) ergänzt. Außerdem wurde die Formulierung mehrerer Handlungsindikatoren überarbeitet. Für das Ratingverfahren wurden außerdem alle Unterrichtsvideos transkribiert. Die Beurteilung der kognitiven Aktivierung im Unterricht erfolgte daher auf Basis des Videomaterials und der Unterrichtstranskripte. Für die Berechnung der Qualitätsmaße zur kognitiven Aktivierung im Unterricht wurden auf Grundlage der Ergebnisse der Reliabilitätsanalysen außerdem weitere Handlungsindikatoren entfernt (Abschnitt 7.6.5 auf Seite 141). In der Paderborner Studie wurden pro Studienteilnehmer ein bis drei Unterrichtsstunden beurteilt (Vogelsang, 2014, S. xl im Anhang). Für die Auswertungen wurden auf Ebene der Handlungsindikatoren über die vorhandenen Unterrichtstunden gemittelt (Vogelsang, 2014, S. 374). Dieses Vorgehen erlaubt es nicht, die Konstanz der kognitiven Aktivierung über die Unterrichtsstunden hinweg zu analysieren. In der vorliegenden Arbeit werden die zwei videographierten Unterrichtsstunden daher zunächst getrennt ausgewertet.

### 7.6.3. Beschreibung des Ratertrainings

Zunächst wurde eine studentischen Hilfskraft im Umgang mit dem Ratingmanual geschult. Die Hilfskraft besuchte eine eineinhalbtägige Raterschulung, die in Kooperation mit den Projektpartner aus der Psychologie geplant und durchgeführt wurde.[14] Der Theorieteil dieser Veranstaltung beinhaltete die Vermittlung theoretischer Grundlagen zum Verständnis des zu beurteilenden Konstrukts der kognitiven Aktivierung und die Thematisierung typischer Raterfehler. Im Praxisteil der Veranstaltung wurden anhand von Videovignetten die Grundideen der Subskalen und die Beurteilung der Handlungsindikatoren intensiv diskutiert. Den Abschluss der Veranstaltung bildete das gemeinsame Rating einer kompletten Unterrichtsstunde.

Auf die Raterschulung folgte ein dreimonatiges Ratertraining an Unterrichtsvideos von Lehrkräften, die nicht zur untersuchten Stichprobe gehörten. Die Unterrichtsstunden wurden zunächst unabhängig durch die Autorin und die studentische Hilfskraft beurteilt, anschließend erfolgte eine Diskussion und Konsensbildung. Im letzten Monat des Ratertrainings konnte die zweite studentische Hilfskraft

[13] Indikator G2_Herausf5 aus Vogelsang (2014)

[14] Mein Dank gilt an dieser Stelle Gerlinde Lenske.

für das Raterteam angeworben werden. Zu diesem Zeitpunkt zeichnete sich bereits deutlich ab, dass auch mit dem überarbeiteten Ratingmanual zur kognitiven Aktivierung keine zufriedenstellende Interrater-Übereinstimmung erzielt werden konnte (Gründe hierfür werden in Abschnitt 7.6.6 auf Seite 142 zur Objektivität des Ratings diskutiert).

### 7.6.4. Beschreibung des Ratingverfahrens

Die Beurteilung der kognitiven Aktivierung im Unterricht auf Basis des Ratingmanuals erfolgte durch die Autorin und die zwei geschulten studentische Hilfskräfte (Lehramt Physik an Gymnasien und Gesamtschulen, 8. Fachsemester). Die 1. Unterrichtsstunde jeder Lehrkraft wurde von allen drei Ratern und Raterinnen beurteilt, die 2. Unterrichtsstunde lediglich von den studentischen Hilfskräften.

Da auch in dieser Studie keine zufriedenstellende Interrater-Übereinstimmung auf Basis des Ratingmanuals erreicht werden konnte (Abschnitt 7.6.6 auf Seite 142), erfolgte die Beurteilung der kognitiven Aktivierung in Anlehnung an das Vorgehen von Vogelsang (2014, S. 342) für jedes Unterrichtsvideo in einem zweischrittigen Ratingverfahren mit Konsensbildung. Dieses Verfahren wird im Folgenden kurz beschrieben.

**1.Schritt:** Zunächst erfolgte eine unabhängige Beurteilung einer Unterrichtsstunde durch jeden Rater bzw. jede Raterin. Hierfür wurde das Video des Unterrichts betrachtet und das Transkript für Notizen und Hervorhebung relevanter Lehrer- oder Schüleraussagen genutzt. Außerdem wurde auf dem Sitzplan der betrachteten Klasse dokumentiert, wenn die Lernenden sich im Unterricht meldeten oder Beiträge zum Unterrichtsgespräch formulierten. Die Rater und Raterinnen waren angehalten, sich in regelmäßigen Abständen Notizen zu machen und konnten das Unterrichtsvideo jederzeit stoppen oder relevante Unterrichtsstellen erneut betrachten. In der Regel wurde das von der Totalenkamera aufgenommene Video ausgewählt. Es bestand allerdings jederzeit die Möglichkeit relevante Unterrichtsstellen zusätzlich aus Perspektive der Aktionskamera oder der Lehrerkamera zu betrachten. Beim Ausfüllen des Ratings wurden (sofern möglich bzw. sinnvoll) die für die Beurteilung eines Handlungsindikators ausschlaggebende Unterrichtsstellen im Ratingbogen vermerkt.

**2.Schritt:** Im zweiten Schritt des Ratingverfahrens wurden die Ratings jedes einzelnen Handlungsindikators durch die verschiedenen Rater und Raterinnen zunächst verglichen und anschließend so lange diskutiert, bis ein Konsens bestand. Die Konsensbeurteilung, die für alle weiteren Analysen genutzt wurde, wurde in einem Masterrating für jede Unterrichtsstunde festgehalten. Um die Konsensbildung transparent zu machen und zu dokumentieren, erfolgte ein Audiomitschnitt der Diskussionen.

Abhängig vom Rater bzw. der Raterin und von der zu beurteilenden Unterrichtsstunde dauerte das Rating einer Unterrichtsstunde ca. $3-7$ Stunden. Für

die Diskussion und Konsensbildung mussten zusätzlich etwa $2-6$ Stunden pro Unterrichtsvideo aufgewendet werden. Damit den Ratern und Raterinnen zum Zeitpunkt der Besprechung der Unterrichtsstunden diese noch präsent waren, lag zwischen der Durchführung der Ratings und der Diskussion und Konsensbildung in der Regel nicht mehr als $1-4$ Tage. Außerdem erfolgte die Beurteilung einer weiteren Unterrichtsstunde in der Regel erst nach dem Besprechungstermin der zuvor beurteilten Unterrichtstunde. In einigen Fällen wurden zwei Unterrichtsstunden geratet und an einem gemeinsamen Besprechungstermin diskutiert.

### 7.6.5. Technische Details zur Auswertung

Die Auswertung der ersten und zweiten Unterrichtstunde erfolgte getrennt. Im Folgenden beziehen sich die Abkürzungen „1M“ und „2M“ auf die erste bzw. zweite Unterrichtsstunde zur Mechanik. Zunächst wurden die Handlungsindikatoren der negativ gepolten Subskala F (Rezeptives Lernverständnis) umgepolt (F1-F5 →F1n-F5n). Anschließend wurden fehlende Werte ersetzt (siehe nächster Abschnitt). Nach Ausschluss weiterer Indikatoren im Rahmen der Reliabilitätsanalysen wurden auf Subskalenebene Qualitätsmaße für die Unterrichtsstunden über die mittlere Ausprägung in den zu den jeweiligen Skalen gehörenden (und nach den Analysen verbleibenden) Handlungsindikatoren berechnet. Das Qualitätsmaß für die kognitiv aktivierende Gestaltung einer Unterrichtsstunde wurde über die gemittelten Subskalenmaße berechnet.[15] Dieses Vorgehen trägt dem Umstand Rechnung, dass die Subskalen aus unterschiedlich vielen Handlungsindikatoren gebildet wurden. Bei einer Mittelung über alle Handlungsindikatoren würden daher Subskalen mit vielen Handlungsindikatoren ein größeres Gewicht haben, was aus theoretischer Sicht nicht sinnvoll erscheint, da alle Subskalen gleichermaßen zum Konstrukt der kognitiven Aktivierung beitragen sollten – bzw. existieren bisher keine empirischen Befunde zur stärkeren Bedeutung einzelner Merkmale.

Für die Berechnung der Qualitätsmaße kann entweder der Median oder der arithmetische Mittelwert gewählt werden. Für die Verwendung des Medians spricht, dass die Handlungsindikatoren auf einer lediglich dreistufigen Ratingskala beurteilt wurden und demnach nicht sichergestellt werden kann, dass es sich um eine intervallskalierte Ratingskala handelt (vergl. Abschnitt 7.4.1 auf Seite 92 zum Umgang mit Ordinalskalen). Problematisch bei der Verwendung des Medians zur Bestimmung der Subskalenmaße ist allerdings zum einen die teils sehr geringe Anzahl der in den Subskalen verbleibenden Indikatoren und zum anderen die unterschiedliche Anzahl an Indikatoren in den verschiedenen Skalen. Ersteres führt zu einem sehr hohen Informationsverlust, was an folgendem Beispiel erläutert werden soll: Unterrichtsstunde A, B und C haben die Ausprägungen (1,1,1), (1,1,2) und (1,1,3) in den drei Handlungsindikatoren einer Subskala X – nur der Mittelwert würde Unterschiede in der Stundenbewertung abbilden, der Median wäre für alle drei Unterrichtsstunden gleich. Die unterschiedliche Anzahl an Handlungsindikatoren in den verschiedenen Subskalen führt zudem dazu, dass in Subskalen mit gerader

[15] Ob es sich bei diesen Maßen um *Qualitätsmaße* im Sinne von Fenstermacher und Richardson (2005) handelt, wird in Abschnitt 7.6.8 auf Seite 153 zur prädiktiven Validität diskutiert.

Indikatoranzahl eine stärkere Differenzierung zwischen unterschiedlich bewerteten Unterrichtsstunden erfolgt als in Subskalen mit ungerader Indikatoranzahl, da der Median für gerade $N$ auch die Zwischenwerte 1.5 und 2.5 annehmen kann – hierin würde eine gewisse Willkür liegen. Aus den genannten Gründen wird der Mittelwert in dieser Arbeit als geeigneteres Maß angesehen um Unterschiede in der kognitiv aktivierenden Gestaltung der untersuchten Unterrichtsstunden zu beschreiben.[16]

**Umgang mit fehlenden Werten** Eine sinnvolle Beurteilung der Handlungsindikatoren war nicht in allen Fällen möglich. Beispielsweise konnte nicht beurteilt werden, ob eine Lehrkraft ihre Schülerinnen und Schüler bei Verständnisschwierigkeiten nach ihren Denkprozessen fragt (Handlungsindikator C4), wenn in der betrachteten Unterrichtsstunde keine Verständnisschwierigkeiten sichtbar wurden. In solchen Fällen wurden die entsprechenden Handlungsindikatoren mit „nicht beurteilbar“ bewertet. Daraus ergaben sich folgende Schwierigkeiten bei der Auswertung: Zum einen reduzierte sich für einzelne Skalen die für die Berechnung der Reliabilität nutzbare Stichprobegröße erheblich, so dass eine sinnvolle Schätzung der Reliabilität und der Indikatortrennschärfen zum Teil nicht möglich war. Zum anderen würden nicht beurteilte Handlungsindikatoren bei der Bildung der Skalenmittelwerte zu einer unterschiedlichen Gewichtung der restlichen Handlungsindikatoren einer Skala führen. Die Skalenmittelwerte verschiedener Unterrichtsstunden wären damit nicht mehr direkt vergleichbar. Konnte ein Handlungsindikator in mehr als 20% der Unterrichtsstunden (jeweils bezogen auf die erste bzw. zweite Unterrichtsstunde) nicht beurteilt werden, wurde der entsprechende Indikator aus den Analysen ausgeschlossen. In der ersten Unterrichtsstunde traf dies für vier, in der zweiten Unterrichtsstunde für fünf Handlungsindikatoren zu (vergl. Tabelle 7.16 auf Seite 142). Bei Handlungsindikatoren, die nur in Einzelfällen nicht beurteilt werden konnten, wurden die fehlenden Werte durch den Mittelwert aus dem Indikatormittelwert in der Stichprobe aller Unterrichtsstunden und dem Skalenmittelwert der Unterrichtsstunde, deren Datensatz einen fehlenden Wert enthielt, ersetzt. Diese Vorgehensweise hat den Vorteil, dass zum einen berücksichtigt wird, ob der Handlungsindikator in der Regel eher schlecht oder eher gut bewertet wurde (vergl. Rost, 2004, S. 327), und zum anderen, ob die Unterrichtsstunde bezüglich der betroffenen Skala eher gut oder eher schlecht bewertet wurde. In der ersten Stunde musste jeweils ein Missing in den Indikatoren D6 und E8 und in der zweiten Stunde ein Missing in Indikator E8 ersetzt werden. Hieraus resultierende Fehler können wie folgt abgeschätzt werden: Ein Indikator kann mit den Werten 1, 2 oder 3 beurteilt werden. Der maximale Fehler auf einen rekodierten Wert $x$ beträgt $\sigma_x = 3 - x$ für $x < 2$ und $\sigma_{\mathrm{x,max}} = 1 - x$ für $x > 2$. Der maximale Fehler auf den Subskalenmittelwert $\bar{x}$ einer Subskala (SS) mit $N_{\mathrm{Ind.}}$ Handlungsindikatoren

[16] Unterschiede in den Subskalenmittelwerten werden im obigen Beispiel erst in der ersten Dezimalstelle sichtbar, daher erscheint es angebracht diese anzugeben. Die Angabe weiterer Dezimalstellen würde allerdings eine Präzision suggerieren, die diese Messung wahrscheinlich nicht erfüllt. Die Skalenmittelwerte werden daher stets mit einer Dezimalstelle angegeben (Eine Fehlerberechnung ist an dieser Stelle nicht weiterführend, da keine Populationsmittelwerte geschätzt werden).

und dem rekodierten Wert $x$ beträgt dann $\sigma_{\text{SS,max}} = \sigma_{\text{x,max}}/N_{\text{Ind.}}$. Der maximale Fehler auf den Mittelwert $\bar{X}$ der Gesamtskala (GS) in der betreffenden Unterrichtsstunde, der über die $N_{\text{SS}} = 7$ Subskalenmittelwerte gebildet wird, beträgt dann $\sigma_{\text{GS,max}} = \sigma_{\text{SS,max}}/N_{\text{SS}}$.[17]

Tabelle 7.15 enthält eine Übersicht über die rekodierten Werte und deren maximalen Fehler sowie die maximalen Fehler auf die Qualitätsmaße in Einheiten ihrer Standardabweichungen.

Tabelle 7.15.
*Rekodierte Missings in den Handlungsindikatoren des Ratings zur kognitiven Aktivierung und Ergebnisse der Abschätzung des maximalen Fehler auf die Qualitätsmaße ($N_{Ind.,Skala\ D} = 5$, $N_{Ind.,Skala\ E} = 5$)*

| Subskala | Indikator | Fall ID | $x$ | $\sigma_{\text{x,max}}$ | $\sigma_{\text{SS,max}}/SD_{\text{SS}}$ | $\sigma_{\text{GS,max}}/SD_{\text{GS}}$ |
|---|---|---|---|---|---|---|
| Skala D | D6 | ID 29, 1M | 1.8 | 1.2 | 0.6 | 0.1 |
| Skala E | E8 | ID 20, 1M | 1.8 | 1.2 | 0.4 | 0.07 |

*Legende:* $x$ =Rekodierter Wert; $N_{\text{Ind.}}$=Anzahl der in der Skala verbleibenden Indikatoren; $\sigma_{\text{x,max}}$=Maximaler Fehler auf den rekodierten Indikator; $\sigma_{\text{SS,max}}/SD_{\text{SS}}$=Maximaler Fehler auf den Subskalenmittelwert der Unterrichtsstunde mit Missing, dargestellt in Einheiten der Standardabweichung des Subskalenmittelwertes über alle Unterrichtsstunden; $\sigma_{\text{GS,max}}/SD_{\text{GS}}$=Maximaler Fehler auf den Gesamtskalemittelwert der Unterrichtsstunde mit Missing, dargestellt in Einheiten der Standardabweichung des Gesamtskalenmittelwertes über alle Unterrichtsstunden

**Klassische Analyse** Die Analyse des Ratings zur kognitiven Aktivierung wurde für die erste und zweite Unterrichtsstunde getrennt ausgeführt und erfolgte in mehreren Schritten. Zunächst wurden die Handlungsindikatoren, die in mehr als 20% der Unterrichtsstunden nicht beurteilt werden konnten oder die keine Streuung aufwiesen ($SD = 0$), aus der Analyse ausgeschlossen (Schritt 1). Anschließend wurde die Reliabilität der Gesamtskala über alle verbleibenden Handlungsindikatoren bestimmt (1M: $N_{\text{Ind.}} = 35$, $\alpha_{\text{C}} = .91$; 2M: $N_{\text{Ind.}} = 31$, $\alpha_{\text{C}} = .88$)[18], um Indikatoren mit geringer Trennschärfe (Korrigierte Item-Skala Korrelation $< .01$) zu identifizieren (Schritt 2). Sofern mit Blick auf die Validität des Ratings keine Gründe dagegen sprachen, wurden diese Indikatoren aus der Gesamtskala entfernt. Hierbei wurde insbesondere darauf geachtet, dass die Subskalenkonstrukte durch die verbleibenden Handlungsindikatoren noch hinreichend gut repräsentiert wurden. Je nachdem welche Unterrichtsstunde betrachtet wurde (1. oder 2. Stunde Mechanik) und welche Handlungsindikatoren schon vorher aus der Skala ausgeschlossen werden mussten, konnte ein Indikator in der Analyse der 1. Stunde als wichtig für die Subskala und in der Analyse der 2. Stunde als weniger wichtig erachtet werden. Im nächsten Schritt wurden die Subskalen analysiert (Schritt

[17] $\sigma_{\text{SS,max}} = \frac{1}{N_{\text{Ind.}}}\left(\sum_{i=1}^{N_{\text{Ind.}}-1} x_i + \left(x_j + \sigma_{x_j,\text{max}}\right)\right) - \bar{x} = \frac{1}{N_{\text{Ind.}}}\left(\sum_{i=1}^{N_{\text{Ind.}}} x_i + \sigma_{x_j,\text{max}}\right) - \bar{x} = \frac{\sigma_{\text{x,max}}}{N_{\text{Ind.}}}$; Die Herleitung für $\sigma_{\text{GS,max}}$ erfolgt analog.

[18] Da laut einer SPSS-Warnung die Determinante der Kovarianzmatrix null oder annähernd null ist, ist eine Berechnung der Konfidenzintervalle in diesem Fall nicht möglich.

3). In Einzelfällen zeigten in der Analyse der Subskalen weitere Indikatoren eine geringe Trennschärfe. Auch hier wurden diese lediglich dann entfernt, wenn das Subskalenkonstrukt durch die verbleibenden Indikatoren noch hinreichend gut repräsentiert wurde. Wenn die Reliabilität einer Subskala bereits größer als .7 war, wurde auf den Ausschluss weiterer Indikatoren verzichtet. Auf Basis der verbleibenden Handlungsindikatoren wurden die Subskalenmittelwerte gebildet (Schritt 4). Einige Handlungsindikatoren wiesen eine extrem geringe Streuung auf (in Extremfällen erhielten mehr als 20 der 23 Unterrichtsstunden in diesen Handlungsindikatoren die gleiche hohe oder niedrige Beurteilung).[19] Mit Blick auf die Validität wurden diese Indikatoren nicht entfernt. Ein Vergleich der Subskalenmittelwerte mit und ohne Einbezug der Indikatoren zeigte, dass ein Vorteil dieses Vorgehens in einer leicht verbesserten Differenzierung zwischen sehr gut (bzw. sehr schlecht) bewerteten Unterrichtsstunden besteht. Ein Nachteil liegt darin, dass die Differenz der Subskalenmittelwerte zwischen sehr gut und sehr schlecht bewerteten Unterrichtsstunden abnimmt. Das Maß für die Gesamtskala zur kognitiven Aktivierung wurde über den Mittelwert der Subskalenmittelwerte gebildet. Tabelle 7.16 gibt eine Übersicht über die in den einzelnen Analyseschritten entfernten Handlungsindikatoren.

Tabelle 7.16.
*Auffällige Handlungsindikatoren im Rating zur kognitiv aktivierenden Gestaltung der 1. und 2. Unterrichtsstunden*

| Stunde | Indikator | Grund für Auffälligkeit | Verfahrensweise |
|---|---|---|---|
| 1M | C4, D4, E4, G4 | n.b. in > 5 Fällen | entfernt |
| | A2, C3, D2, E1, F4n, B5 | Trennschärfe < .01 | entfernt |
| | A4 | Trennschärfe < .01 | beibehalten |
| 2M | C4, D3, D4, E4, G4 | n.b. in > 5 Fällen | entfernt |
| | B1, E1, G1 | $SD = 0$ | entfernt |
| | C3, B5[1] | Trennschärfe < .01 | entfernt |
| | A4, D1, E6[2] | Trennschärfe < .01 | beibehalten |

*Legende:* n.b.= nicht beurteilbar
[1] Trennschärfe war erst in Subskalenanalyse < .01
[2] Trennschärfe war nur in Analyse der Gesamtskala < .01

## 7.6.6. Objektivität

Um die Durchführungsobjektivität sicherzustellen, wurden Regeln für das Ratingverfahren formuliert (vergl. Anhang A.4 auf Seite 222). Für alle Unterrichtsstunden lagen Videoaufzeichnungen aus verschiedenen Perspektiven sowie Stundentranskripte und Sitzpläne vor, die für die Beurteilung genutzt werden konnten. Die Rater und Raterinnen waren angehalten, die Ratings in ruhiger Arbeitsatmosphäre durchzuführen, beim Beobachten gezielt auf die Handlungsindikatoren des Manuals

[19] Für die Ausprägung „teils teils“ trat dieser Fall nicht auf.

zu achten und die Videos regelmäßig anzuhalten, um sich Notizen zu machen. Die Beurteilung der Handlungsindikatoren erfolgte erst nach der Sichtung der gesamten Unterrichtsstunde. Eine ausführliche Beschreibung des Ratingverfahrens erfolgte bereits in Abschnitt 7.6.4 auf Seite 138.

Um die Auswertungsobjektivität des Ratings zur kognitiven Aktivierung zu beurteilen, wurde mit Hilfe der unjustierten zweifaktoriellen ICC die Interrater-Übereinstimmung bezüglich der Einzelratings auf Subskalen- und Gesamtskalenebene bestimmt (vergl. Abschnitt 7.4.4 auf Seite 98 im Kapitel zu statistische Methoden). Hierfür wurden zunächst (unter Ausschluss der in Tabelle 7.16 auf Seite 142 aufgeführten Handlungsindikatoren) für jeden Rater bzw. jede Raterin auf die zuvor beschriebene Weise die Qualitätsmaße auf Subskalenebene und auf Ebene der Gesamtskala berechnet. Tabelle 7.17 auf der nächsten Seite gibt eine Übersicht über die ICCs auf Subskalenebene für beide Unterrichtsstunden. Die ICCs beziehen sich für die erste Unterrichtsstunde (1M) auf drei und für die zweite Unterrichtsstunde (2M) auf zwei Rater und Raterinnen (vergl. Abschnitt 7.6.4 auf Seite 138 zum Ratingverfahren). Fehlende Werte wurden in den Einzelratings nicht ersetzt. In einigen Fällen wurden daher geringere Fallzahlen zur Bestimmung der ICCs verwendet.

Auf Gesamtskalenebene ist die Übereinstimmung zwischen den Ratern und Raterinnen zwar zufriedenstellend, auf Subskalenebene zeigen sich allerdings wesentlich geringere Übereinstimmungen. Vergleicht man die ICCs mit den von Vogelsang (2014, S. xxvii im Anhang) berichteten Werten auf Subskalenebene, zeigen sich zwar höhere ICCs für die die Subskalen A-C, F und G und für die Gesamtskala, die von Vogelsang berichteten Werte liegen allerdings in den meisten Fällen innerhalb der 95%-Konfidenzintervalle der in der vorliegenden Arbeit berechneten ICCs (Ausnahme bilden die Skalen A (1M), C (2M) und G (1M) – hier konnten signifikante Verbesserungen in der Interrater-Übereinstimmung erzielt werden). Entsprechendes gilt für die ICCs der Skalen D und E, die bei Vogelsang höher waren als in der vorliegenden Arbeit.[20] Auf Ebene der Gesamtskala konnte die Interrater-Übereinstimmung von .48 auf .69 in der ersten Unterrichtsstunde bzw. .64 in der zweiten Unterrichtsstunde erhöht werden. Aber auch hier schließen die Konfidenzintervalle den von Vogelsang berichteten Wert ein. Durch das intensive Ratertraining, die Reduzierung der Ratingskala auf ein dreistufiges Format und der Ausschärfung des Ratingmanuals durch die Formulierung von Beispielen für die verschiedenen Ausprägungen jedes Handlungsindikators konnten insgesamt zwar leichte Verbesserungen in der Interrater-Übereinstimmung erzielt werden, diese werden aber nicht als ausreichend erachtet. Berechnet man die ICCs für die über alle Rater und Raterinnen gemittelten Qualitätsmaße, ergibt sich für die erste Stunde auf Ebene der Gesamtskala $ICC_{\text{2-fakt.,unjust}} = .87$ und damit eine präzise Beschreibung des wahren Wertes (vergl. Wirtz & Caspar, 2002, S. 234). Für die zweite Unterrichtsstunde ergibt sich lediglich $ICC_{\text{2-fakt.,unjust}} = .78$. Auf Subskalenebene gibt es außerdem immer noch Werte $< .6$. Wegen der als nicht ausreichend erachteten Interrater-Übereinstimmung wurde eine Konsensbildung

[20] Anhaltspunkte dafür, warum sich die Interrater-Übereinstimmung in diesen beiden Skalen reduzierte, liegen leider nicht vor.

Tabelle 7.17.
*Interrater-Übereinstimmung ($ICC_{2\text{-fakt.,unjust}}$) für die Subskalenmittelwerte und den Gesamtskalenmittelwert zur kognitiven Aktivierung für die 1. und 2. Unterrichtsstunde zur Mechanik ($N_{Rater,1M} = 3$, $N_{Rater,2M} = 2$)*

| Skala | | 1M | 2M |
|---|---|---|---|
| Skala A: Lernstatus bewusst machen | $N$ | 23 | 23 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .64 | .62 |
| | $\text{KI}_{95\,\%}$ | [.42, .81] | [.30, .82] |
| Skala B: Exploration des Vorwissens | $N$ | 23 | 22 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .57 | .52 |
| | $\text{KI}_{95\,\%}$ | [.34, .77] | [.13, .77] |
| Skala C: Exploration der Denkweisen | $N$ | 23 | 23 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .55 | .71 |
| | $\text{KI}_{95\,\%}$ | [.31, .76] | [.44, .87] |
| Skala D: Evolutionärer Umgang mit Schülervorstellungen | $N$ | 21 | 22 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .36 | .21 |
| | $\text{KI}_{95\,\%}$ | [.11, .63] | [−.21, .57] |
| Skala E: Lehrperson als Mediator | $N$ | 22 | 19 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .48 | .63 |
| | $\text{KI}_{95\,\%}$ | [.22, .72] | [.27, .84] |
| Skala F: Rezeptives Lernverständnis | $N$ | 20 | 20 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .33 | .61 |
| | $\text{KI}_{95\,\%}$ | [.05, .62] | [.23, .81] |
| Skala G: Herausfordernde Lerngelegenheiten | $N$ | 23 | 23 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .46 | .45 |
| | $\text{KI}_{95\,\%}$ | [.21, .69] | [.08, .72] |
| Gesamtskala: Kognitive Aktivierung | $N$ | 17 | 17 |
| | $ICC_{2\text{-fakt.,unjust}}$ | .69 | .64 |
| | $\text{KI}_{95\,\%}$ | [.45, .86] | [.26, .85] |

*Legende: N* =Anzahl gültiger Fälle ohne Missings
*Anmerkung:* Die angegebenen ICCs beziehen sich auf die Skalenmittelwerte der einzelnen Rater und Raterinnen und nicht auf die über alle Rater und Raterinnen gemittelten Skalenmittelwerte.

vorgenommen (Abschnitt 7.6.4 auf Seite 138). Hierbei handelt es sich um ein übliches Vorgehen, wie es auch in zahlreichen anderen Studien durchgeführt wird (vergl. z. B. Krauss, Neubrand et al., 2008, S. 239; Kunter, 2005, S. 214-215; Vogelsang, 2014, S. 328,507).

Die kognitiv aktivierende Gestaltung des Unterrichts verschiedener Lehrkräfte wird nur innerhalb der Stichprobe miteinander verglichen. Es erfolgt keine Bewertung der absoluten Qualität des Unterrichts. Daher kann die Interpretati-

onsobjektivität als weitestgehend gewährleistet betrachtet werden (vergl. Bortz & Döring, 2006, S. 195).

### 7.6.7. Reliabilität

Die Reliabilität des Ratings zur kognitiven Aktivierung wurde über die Berechnung von Cronbachs Alpha geschätzt. Die Reliabilität der Gesamtskala zur kognitiven Aktivierung wurde auf Basis der sieben Subskalenmittelwerte bestimmt. Tabelle 7.18 zeigt die Reliabilitäten für die Subskalen und die Gesamtskala. Die großen Konfidenzintervalle spiegeln die kleine Stichprobe wieder, die für die Reliabilitätsberechnung genutzt werden konnte.

Tabelle 7.18.
*Reliabilität des Ratings zur kognitiv aktivierenden Gestaltung der 1. und 2. Unterrichtsstunde zur Mechanik ($N_{1M} = N_{2M} = 23$)*

| Skala | | 1M | 2M |
|---|---|---|---|
| Skala A: Lernstatus bewusst machen | $N_{\text{Ind.}}$ | 4 | 5 |
| | $\alpha_{\text{C}}$ | .24 | .36 |
| | $\text{KI}_{95\,\%}$ | [−.32, .80] | [−.11, .82] |
| Skala B: Exploration des Vorwissens | $N_{\text{Ind.}}$ | 4 | 3 |
| | $\alpha_{\text{C}}$ | .71 | .64 |
| | $\text{KI}_{95\,\%}$ | [.37, 1.0] | [.18, 1.0] |
| Skala C: Exploration der Denkweisen | $N_{\text{Ind.}}$ | 4 | 4 |
| | $\alpha_{\text{C}}$ | .79 | .72 |
| | $\text{KI}_{95\,\%}$ | [.51, 1.0] | [.40, 1.0] |
| Skala D: Evolutionärer Umgang mit Schülervorstellungen | $N_{\text{Ind.}}$ | 4 | 4 |
| | $\alpha_{\text{C}}$ | .74 | .24 |
| | $\text{KI}_{95\,\%}$ | [.42, 1.0] | [−.29, .77] |
| Skala E: Lehrperson als Mediator | $N_{\text{Ind.}}$ | 6 | 6 |
| | $\alpha_{\text{C}}$ | .77 | .67 |
| | $\text{KI}_{95\,\%}$ | [.53, 1.0] | [.36, 1.0] |
| Skala F: Rezeptives Lernverständnis | $N_{\text{Ind.}}$ | 3 | 4 |
| | $\alpha_{\text{C}}$ | .52 | .63 |
| | $\text{KI}_{95\,\%}$ | [.02, 1.0] | [.24, 1.0] |
| Skala G: Herausfordernde Lerngelegenheiten | $N_{\text{Ind.}}$ | 4 | 3 |
| | $\alpha_{\text{C}}$ | .76 | .83 |
| | $\text{KI}_{95\,\%}$ | [.46, 1.0] | [.48, 1.0] |
| Gesamtskala: Kognitive Aktivierung | $N_{\text{Subskalenmittelwerte}}$ | 7 | 7 |
| | $\alpha_{\text{C}}$ | .91 | .87 |
| | $\text{KI}_{95\,\%}$ | [0.77, 1.0] | [0.70, 1.0] |

*Legende:* $N_{\text{Ind.}}$ =Anzahl in Skala verbleibender Handlungsindikatoren
*Anmerkung.* Von R ausgegebene obere Konfidenzintervallgrenzen > 1 wurden durch durch den Grenzwert 1.0 substituiert.

**Reliabilität der Subskalen** Auf Ebene der Subskalen zeigen sich weitestgehend ausreichende bis gute Reliabilitäten mit Ausnahme der Subskala A und der Subskala D in der zweiten Unterrichtsstunde. Die Reliabilität der Subskala F in der ersten Unterrichtsstunde kann gerade noch als ausreichend bezeichnet werden (Lamberti, 2001, S. 31), zumal die Skala nur drei Handlungsindikatoren enthält. Die geringe Reliabilität der Subskala A (Lernstatus bewusst machen) in beiden Unterrichtsstunden liegt wahrscheinlich darin begründet, dass ein sehr heterogenes Konstrukt erfasst wird – so muss eine Lehrkraft z. B. nicht zwingend einen Rückblick auf bereits Gelerntes geben, wenn sie zuvor einen expliziten Ausblick auf den Stundeninhalt gegeben hat. Für die schlechte Reliabilität der Subskala D (Evolutionärer Umgang mit Schülervorstellungen) in der zweiten Unterrichtsstunde ergibt sich keine augenscheinliche Erklärung. Für Subskalen mit Reliabilitäten mit $\alpha_C < .5$ oder einem Konfidenzintervall, das den Nullpunkt einschließt, werden die Subskalenmittelwerte nicht in Analysen auf Subskalenebene einbezogen.

**Reliabilität der Gesamtskala** Die auf Basis der Subskalenmittelwerte berechneten Reliabilitäten für die Gesamtskala können in beiden Stunden als sehr gut bezeichnet werden. Für die erste Unterrichtsstunde ergeben sich zudem für alle Subskalenmittelwerte Trennschärfen $> .6$ – auch der Mittelwert der nicht reliablen Subskala A zeigt also eine gute Passung in das Gesamtkonstrukt. Für die zweite Stunde ergeben sich geringere Trennschärfen für die Subskalen A (.29), B (.58) und D (.52). Auch diese Subskalen zeigen aber eine ausreichende bis gute Passung in das Gesamtkonstrukt.

## 7.6.8. Validität

Die kognitiv aktivierende Gestaltung des Unterrichts in den zwei videographierten Unterrichtsstunden wird mit dem Ziel beurteilt, Qualitätsmaße zu generieren, die als Indikator dafür interpretiert werden können, wie kognitiv aktivierend eine Lehrkraft grundsätzlich ihren Unterricht gestaltet. Daher wird zunächst diskutiert, inwieweit die videographierten Unterrichtsstunden als repräsentativ für den üblichen Unterricht der Lehrkräfte angenommen werden können. Anschließend wird die Inhaltsvalidität des Ratings zur kognitiven Aktivierung untersucht – hierfür ist zum einen wichtig, dass das Konstrukt der kognitiven Aktivierung durch die Subskalenkonstrukte beschrieben werden kann und zum anderen, dass die Subskalenkonstrukte ihrerseits durch die Handlungsindikatoren hinreichend gut repräsentiert werden. Um Hinweise auf die Konstruktvalidität zu erhalten, werden Korrelationen zu anderen Merkmalen der Unterrichtsqualität untersucht. Um sicherzustellen, dass die Qualitätsmaße für die kognitiv aktivierende Gestaltung des Unterrichts ein Merkmal der Unterrichtsqualität im Sinne von Fenstermacher und Richardson (2005) abbilden, wird abschließend deren prädiktive Validität in Bezug auf Unterrichtserfolg diskutiert.

**Voraussetzungen für eine valide Interpretation der Videodaten** Die kognitiv aktivierende Gestaltung des Unterrichts wird in zwei aufeinanderfolgenden Unter-

richtsstunden beurteilt – interpretiert werden die aus diesen Unterrichtsstunden generierten Qualitätsmaße allerdings als Indikator dafür, wie kognitiv aktivierend eine Lehrkraft grundsätzlich ihren Unterricht gestaltet. Um von einer validen Interpretation ausgehen zu können, sollte zunächst die Repräsentativität der videographierten Unterrichtsstunden sichergestellt werden. Hierfür werden die durch das Filmen des Unterrichts hervorgerufene Nervosität der Unterrichtsakteure und die Einschätzung des Verhaltens der Lehrkraft bzw. der Lernenden im Vergleich zu üblichen Unterrichtsstunden untersucht sowie Angaben zur Unterrichtsplanung ausgewertet (vergl. Abschnitt 7.5.5.4 auf Seite 134). Anschließend wird untersucht, ob die kognitive Aktivierung über die beiden videographierten Unterrichtsstunden konstant ist.

**Nervosität**
In den ersten 10 Minuten des Unterrichts gaben in der ersten Unterrichtsstunde 17% der Lehrkräfte an, ziemlich oder sehr nervös gewesen zu sein, in der zweiten Unterrichtsstunde galt dies nur noch für 4% (also eine Lehrkraft). Für die restliche Zeit der Unterrichtsstunden gab lediglich eine Lehrkraft an, in der ersten Unterrichtsstunde ziemlich nervös gewesen zu sein. Das Verhalten dieser Lehrkraft wurde in dieser Unterrichtsstunde allerdings dennoch von 78% ihrer Schülerinnen und Schüler als größtenteils typisch oder sehr typisch bezeichnet. 83% der Lehrkräfte waren in der restlichen Unterrichtszeit in beiden Stunden nur wenig oder überhaupt nicht nervös. Ein ähnliches Bild ergab sich für die Lernenden. In den ersten 10 Minuten der ersten und zweiten Unterrichtsstunde waren im Mittel 83% bzw. 95% der Lernenden einer Klasse nur wenig oder überhaupt nicht nervös (1M: $SD = 12\%$, $Min = 57\%$, $Max = 96\%$, 2M: $SD = 4\%$, $Min = 85\%$, $Max = 100\%$), in der restlichen Unterrichtszeit galt dies für 95% bzw. 97% der Lernenden einer Klasse (1M: $SD = 5\%$, $Min = 76\%$, $Max = 100\%$, 2M: $SD = 3\%$, $Min = 90\%$, $Max = 100\%$). Es wird daher angenommen, dass die durch die Videographie des Unterrichts hervorgerufene Nervosität der Unterrichtsakteure keine negativen Auswirkungen auf die Repräsentativität der gefilmten Unterrichtsstunden hat.

**Unterrichtsplanung**
Zwei Lehrkräfte gaben an, angesichts der Tatsache gefilmt zu werden, den Unterricht in beiden Unterrichtsstunden bewusst anders geplant und mehr Zeit als gewöhnlich für die Unterrichtsplanung aufgewendet zu haben. Weitere drei Lehrkräfte hatten nur die erste Unterrichtsstunde bewusst anders geplant (zwei von diesen Lehrkräften wendeten mehr Zeit für die Planung des Unterrichts auf als gewöhnlich). Mindestens 73% der Lernenden in den Klassen dieser insgesamt fünf Lehrkräfte schätzen die im Unterricht eingesetzten Methoden und das Verhalten ihrer Lehrkraft dennoch als größtenteils typisch oder sehr typisch ein. Bei 74% der Lehrkräfte hatten die designbedingten Vorgaben für die erste Unterrichtsstunde (Einführung des physikalischen Kraftbegriff, Beinhalten eines Lehrerexperiment und primäres Lehrziel im Kompetenzbereich Fachwissen, vergl. Abschnitt 7.1 auf Seite 81) laut Selbstauskünften keinen Einfluss auf die Planung der Stunde. Für die zweite Unterrichtsstunde, für die es keine Vorgaben gab, galt dies für 95% der Lehrkräfte. Allerdings gaben 57% der Lehrkräfte an, mehr Zeit als gewöhnlich

für die Planung der ersten Stunde aufgewendet zu haben; für die zweite Unterrichtsstunde sank dieser Anteil auf 30%. Den Lernenden schien dies allerdings nicht aufzufallen – der Anteil der Lernenden in den Klassen dieser Lehrkräfte, die die im Unterricht eingesetzten Methoden und das Verhalten ihrer Lehrkraft als größtenteils typisch oder sehr typisch einschätzen, unterschied sich nicht signifikant von dem entsprechenden Anteil in Klassen von Lehrkräften, die nicht mehr Zeit für die Unterrichtsplanung aufgewendet hatten. Die Repräsentativität der Unterrichtsstunden scheint also nicht beeinträchtigt zu sein.

**Unterrichtsmethoden**
Die in der ersten Unterrichtsstunde eingesetzten Unterrichtsmethoden wurden von 57% der Lehrkräfte als größtenteils oder sehr typisch und von weiteren 35% als einigermaßen typisch empfunden. Lediglich zwei Lehrkräfte gaben an, Unterrichtsmethoden eingesetzt zu haben, die nur wenig typisch für ihren üblichen Unterricht seien – bei einer der Lehrkräfte schätzen 83% der Schülerinnen und Schüler in ihrer Klasse diese Methoden dennoch als größtenteils oder sehr typisch an, bei der anderen Lehrkraft galt dies allerdings nur für 48% der Lernenden. Allerdings schätzen 76% der Lernenden die Unterrichtsmethoden zumindest einigermaßen typisch ein. Das Verhalten dieser Lehrkraft wurde außerdem von 88% ihrer Schülerinnen und Schüler als größtenteils oder sehr typisch eingeschätzt. In der zweiten Unterrichtsstunde gaben dieselben Lehrkräfte und eine weitere Lehrkraft an, lediglich einigermaßen typische Unterrichtmethoden eingesetzt zu haben – bei einer Lehrkraft hielten dennoch 80% der Lernenden die Methoden für größtenteils oder sehr typisch. Bei den anderen beiden Lehrkräften galt dies allerdings nur 18% bzw. 38% der Lernenden. Das Lehrerverhalten wurde aber auch hier von 97% bzw. 67% der Lernenden als größtenteils oder sehr typisch eingeschätzt.

Von den Lernenden schätzen im Mittel 73% bzw. 72% der Lernenden einer Klasse die in der ersten bzw. zweiten Unterrichtsstunde eingesetzten Unterrichtsmethoden als größtenteils oder sehr typisch ein (1M: $SD = 20\%$, $Min = 20\%$, $Max = 100\%$, 2M: $SD = 21\%$, $Min = 18\%$, $Max = 100\%$). In jeweils drei ersten bzw. zweiten Unterrichtsstunden wurden die Unterrichtsmethoden von weniger als 50% der Lernenden einer Klasse als *größtenteils* oder *sehr* typisch eingeschätzt. Mindestens 53% der Lernenden einer Klasse schätzten diese Unterrichtsstunden allerdings als zumindest *einigermaßen* typisch ein – mit Ausnahme einer Unterrichtsstunde, die von lediglich 36% der Lernenden einer Klasse als mindestens einigermaßen typisch eingeschätzt wurde. Das Verhalten der Lehrkraft in dieser Unterrichtsstunde wurde allerdings von 75% der Lernenden als größtenteils oder sehr typisch eingeschätzt. Da die kognitiv aktivierende Gestaltung des Unterrichts nicht an bestimmte Unterrichtsmethoden gebunden ist, ergeben sich auch hieraus keine offensichtlichen Einschränkungen für die Repräsentativität der Unterrichtsstunden.

**Verhalten der Lehrkräfte**
Im Mittel schätzen 84% bzw. 82% der Lernenden einer Klasse das Verhalten ihrer Lehrkraft in der ersten bzw. zweiten Unterrichtsstunde als größtenteils oder sehr typisch ein (1M: $SD = 15\%$, $Min = 42\%$, $Max = 100\%$, 2M: $SD = 15\%$, $Min = 48\%$, $Max = 100\%$). Lediglich zwei Unterrichtsstunden (die erste bzw. zweite Unter-

richtsstunde von zwei Lehrkräften) wurden von weniger als 50% der Lernenden als größtenteils oder sehr typisch bezüglich des Lehrerverhaltens eingeschätzt – das Verhalten der Lehrkräfte in diesen Unterrichtsstunden wurde allerdings von 62% bzw. 66% ihrer Schülerinnen und Schüler als zumindest einigermaßen typisch eingeschätzt. Da höchstens eine der zwei pro Lehrkraft videographierten Unterrichtsstunden als weniger typisch durch die Lernenden beurteilt wurde, scheinen die beiden Unterrichtsstunden zusammengenommen den üblichen Unterricht besser zu repräsentieren als sie es einzeln tun würden. Gegebenenfalls durch ein nicht-typisches Verhalten der Lehrkräfte resultierende Verzerrungen in den Qualitätsmaßen zur kognitiven Aktivierung könnten daher durch eine Mittelung der Qualitätsmaße über beide Unterrichtsstunden reduziert werden.

**Verhalten der Lernenden**
In der ersten Unterrichtsstunde schätzten alle Lehrkräfte das Verhalten ihrer Schülerinnen und Schüler als einigermaßen ähnlich im Vergleich mit üblichen Unterrichtsstunden ein – 70% gaben ein größtenteils oder sehr ähnliches Verhalten der Lernenden an. In der zweiten Unterrichtsstunde gaben 70% der Lehrkräfte ein größtenteils oder sehr ähnliches Verhalten, 13% ein einigermaßen ähnlichen und 9% (2 Lehrkräfte) ein wenig ähnliches Verhalten der Lernenden an. Eine der beiden Lehrkräfte gab an, dass ihre Schülerinnen und Schüler weniger konzentriert und unruhiger als üblich waren. Die andere Lehrkraft schätzte die Lernenden weniger konzentriert, weniger engagiert, unruhiger, lauter und abgelenkter als üblicherweise ein. Für die erste Unterrichtsstunde gaben diese beiden Lehrkräfte allerdings ein sehr ähnliches Verhalten der Schülerinnen und Schüler im Vergleich zu üblichen Stunden an. Die Mehrheit der Lehrkräfte gab sowohl für die erste, als auch für die zweite Unterrichtsstunde ein vergleichbares oder positiveres Verhalten der Lernenden bezüglich der Konzentration (100%/67%), des Engagements (87%/78%), der Unruhe (96%/90%), der Lautstärke (96%/95%) und der Abgelenktheit (91%/90%) der Lernenden im Unterricht an. Sofern negativere Verhaltensweisen angegeben wurde, galt dies stets nur für eine der zwei Unterrichtsstunden.

Das Verhalten der Lernenden sollte das Lehrangebot der Lehrkraft in Bezug auf die kognitiv aktivierende Gestaltung nicht maßgeblich beeinflussen, dennoch kann ein Einfluss nicht ausgeschlossen werden. Da insbesondere negativere Verhaltensweisen der Lernenden *entweder* in der ersten *oder* in der zweiten Unterrichtsstunde auftauchten, könnten auch hier beide Unterrichtsstunden zusammengenommen eine bessere Repräsentation des üblichen Unterricht darstellen und eine Mittelung der Qualitätsmaße über beide Unterrichtsstunden erneut dazu beitragen, gegebenenfalls durch ein nicht-typisches Verhalten der Lernenden resultierende Verzerrungen in den Qualitätsmaßen zu reduzieren.

**Konstanz der kognitiven Aktivierung über zwei Unterrichtsstunden**
Auch wenn die beiden videographierten Unterrichtsstunden den üblichen Unterricht einer Lehrkraft hinreichend gut abbilden, besteht bezüglich der kognitiv aktivierenden Gestaltung des Unterrichts folgende Problematik: Auf Basis von Videodaten aus der Pythagoras-Studie konnten Praetorius et al. (2014) zeigen, dass die kognitiv aktivierende Gestaltung des Unterrichts (operationalisiert über die Exploration

der Denkweisen, das rezeptive Lernverständnis der Lehrkraft und herausfordernde Lerngelegenheiten) von Mathematiklehrkräften über mehrere Unterrichtsstunden hinweg erheblich variiert und dass für eine reliable Schätzung der kognitiven Aktivierung mindestens neun Unterrichtsstunden nötig wären.[21] Zugleich werfen die Autoren die Frage auf, ob die von ihnen vorgenommene Operationalisierung der kognitiven Aktivierung das Konstrukt auch über Einführungsstunden hinaus hinreichend gut abbildet:

> Nevertheless, it is obvious that cognitive activation in an introduction lesson is not exactly the same as cognitive activation in a practice lesson. Measuring only one specific aspect of cognitive activation may be sufficient to predict student learning within a single lesson or a short introductory unit, as in the case of the Pythagoras project (Lipowsky et al., 2009); however, if cognitive activation is to be used as a predictor of student learning in a broader sense or even as an indicator of teacher effectiveness that generalizes across classrooms and contents, its operationalization should be revisited. (S. 9)

In der vorliegenden Arbeit wurde das Konstrukt der kognitiven Aktivierung etwas breiter gefasst und zusätzliche Merkmale einer kognitiv aktivierenden Gestaltung des Unterrichts erhoben. Tabelle 7.19 auf der nächsten Seite zeigt die Korrelationen zwischen den Qualitätsmaßen für die erste und zweite Unterrichtsstunde (auf Subskalenebene werden lediglich signifikante Korrelationen berichtet). Da weder die Subskalenmaße noch die Gesamtmaße hoch zwischen beiden Unterrichtstunden korrelieren, stellt sich auch hier die Frage, ob das Rating zur kognitiven Aktivierung lediglich für die Beurteilung von Einführungsstunden geeignet ist, und demnach lediglich die erste Unterrichtsstunde für die Auswertungen berücksichtigt werden sollte, oder ob kognitiv aktivierender Unterricht in beiden Unterrichtsstunden valide über das Rating beurteilt werden kann, das Konstrukt an sich aber nicht stabil über mehrere Unterrichtstunden ist. In den bisherigen Analysen zeigten sich zwar etwas schlechtere Reliabilitätsmaße für die Ratings der zweiten Unterrichtsstunde, deutliche Hinweise darauf, dass die kognitiv aktivierende Gestaltung der zweiten Unterrichtsstunde schlechter durch das Rating erfasst wurde, zeigten sich jedoch nicht.

Zusammenfassend kann Folgendes festgehalten werden: Die Ergebnisse der Lehrer- und Schülerbefragungen weisen darauf hin, dass bis auf wenige Ausnahmen sowohl die Lehrenden als auch die Lernenden in den videographierten Unterrichtsstunden für sie typisches Verhalten zeigten. Es kann also davon ausgegangen werden, dass diese Unterrichtsstunden übliches Unterrichtgeschehen repräsentieren. Beide Unterrichtsstunden zusammengenommen sollten zudem den Unterricht *aller* Lehrkräfte dieser Stichprobe hinreichend gut repräsentieren. Sollten sich in den weiteren Untersuchungen keine deutlichen Bedenken bezüglich der Validität des Ratings für die zweite Unterrichtsstunde ergeben, sollte als Indikator für die kognitiv aktivierende Gestaltung des Unterrichts einer Lehrkraft daher ein über beide

[21] Die Anzahl der notwendigen Unterrichtsstunden wurde über eine sogenannte „Decision Study“ (D-Study) im Rahmen der Generalisierungstheorie bestimmt.

Tabelle 7.19.
*Korrelationen zwischen den Qualitätsmaßen zur kognitiv aktivierenden Gestaltung der 1. und 2. Unterrichtsstunde zur Mechanik für die reliablen Subskalen und die Gesamtskala ($N_{1M} = N_{2M} = 23$)*

| Korrel. Skala | Skala B | Skala C | Skala E | Skala G | Skala F | Gesamtskala |
|---|---|---|---|---|---|---|
| $r_{\text{Pearson}}$ | | | **.48 ± .19** | | | **.38 ± .20** |
| $\text{KI}_{95\,\%}$ | | | [.05,.78] | | | [-.02,.74] |
| $p_{\text{1-seitig}}$ | | | .010 | | | .037 |
| $r_{\text{Spearman}}$ | **.43 ± .20** | .24 ± .22 | **.42 ± .19** | **.52 ± .17** | .26 ± .20 | .34 ± .22 |
| $\text{KI}_{95\,\%}$ | [.02,.77] | [-.20,.64] | [-.02,.74] | [-.14,.78] | [-.16,.61] | [-.08,.75] |
| $p_{\text{1-seitig}}$ | .019 | .133 | .023 | .005 | .117 | .053 |
| $\tau_{\text{Kendall}}$ | **.37 ± .18** | .20 ± .18 | **.32 ± .16** | **.43 ± .15** | .21 ± .17 | **.27 ± .17** |
| $\text{KI}_{95\,\%}$ | [.02,.68] | [-.16,.53] | [-.02,.61] | [-.11,.68] | [-.14,.51] | [-.03,.61] |
| $p_{\text{1-seitig}}$ | .018 | .121 | .024 | .006 | .114 | .036 |

*Anmerkungen.* Signifikante Korrelationen mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Die Qualitätsmaße für die Subskalen B, C, F und G sind in mindestens einer Unterrichtsstunde nicht normalverteilt, daher werden für diese Skalen nur nicht-parametrische Korrelationen berichtet. Auch für normalverteilte Maße werden zusätzlich nicht-parametrische Korrelationen berichtet, da das Intervallskalenniveau der Qualitätsmaße zur kognitiven Aktivierung nicht sichergestellt werden kann (vergl. Abschnitt 7.4.1 auf Seite 92 zum Umgang mit Ordinalskalen).

Unterrichtsstunden gemitteltes Maß verwendet werden. Zwei Unterrichtsstunden sind nach Praetorius et al. (2014) nicht ausreichend, um einen stabilen Schätzer für die kognitive Aktivierung zu erhalten. Sofern die Einzelmaße und das über beide Unterrichtsstunden gemittelte Maß dennoch prädiktiv für den Unterrichtserfolg über einen längeren Zeitraum sind, sollte allerdings auch vor dem Hintergrund der oben zitierten Kritik von Praetorius et al. (2014) von einer validen Interpretation der Qualitätsmaße als Indikator dafür, wie kognitiv aktivierend eine Lehrkraft ihren Unterricht üblicherweise gestaltet, ausgegangen werden können.

**Inhaltsvalidität** Das eingesetzte Rating zur kognitiven Aktivierung stellt eine Adaption mehrerer in anderen Studien eingesetzter Instrumente dar – auch andere Forscher nutzen die hier gewählten Subskalenkonstrukte und Handlungsindikatoren zur Beschreibung eines kognitiv aktivierenden Unterrichts (vergl. z. B. Clausen, 2002; Clausen et al., 2003; Kunter, 2005; Praetorius et al., 2014; Rakoczy & Pauli, 2006; Vogelsang, 2014; Widodo & Duit, 2004). Dies kann im Sinne eines Expertenurteils als Indiz für die Angemessenheit der Beschreibung eines kognitiv aktivierend gestalteten Unterrichts über die durch die Subskalen beschriebenen Merkmale gewertet werden.[22]

Um sicherzustellen, dass die Handlungsindikatoren die Subskalenkonstrukte hinreichend gut repräsentieren, wurde im Rahmen des eigentlichen Ratings zusätzlich zu den Handlungsindikatoren für jede Subskala ein Gesamteindruck geratet. Dabei sollte auf der dreistufigen Ratingskala eingeschätzt werden, wie gut die Grundidee

[22] Ob ein derartiger Unterricht die Lernenden allerdings wirklich „kognitiv aktiviert“, kann nicht beurteilt werden. Zunächst handelt es sich also um normativ für gut befundene Merkmale.

eines Subskalenkonstrukts in der Unterrichtsstunde durch die Lehrkraft umgesetzt wurde. Hierbei durfte explizit von der Bewertung der einzelnen Handlungsindikatoren abgewichen werden – auch eine unterschiedlich starke Gewichtung der Handlungsindikatoren für die Bewertung des Gesamteindruck war möglich, falls einzelne Handlungsindikatoren aus Sicht der Rater und Raterinnen die Umsetzung des Merkmals in einer Unterrichtsstunde stärker bestimmten. Tabelle 7.20 zeigt die Korrelationen zwischen den über die Handlungsindikatoren bestimmten Subskalenmittelwerten und dem für jede Skala gerateten Gesamteindruck für die reliablen Subskalen.[23]

Tabelle 7.20.
*Korrelationen zwischen den Subskalenmittelwerten und den Subskalengesamteindrücken in der 1. und 2. Unterrichtsstunde zur Mechanik ($N_{1M} = N_{2M} = 23$)*

| Korrelierte Subskala | | $r_{\text{Spearman}}$ | $\text{KI}_{95\,\%}$ | $p_{\text{1-seitig}}$ | $\tau_{\text{Kendall}}$ | $\text{KI}_{95\,\%}$ | $p_{\text{1-seitig}}$ |
|---|---|---|---|---|---|---|---|
| Skala B | 1M | **.83 ± .08** | [.62, .93] | < .001 | **.74 ± .07** | [.56, .85] | < .001 |
| | 2M | **.36 ± .10** | [.32, .67] | .045 | **.34 ± .09** | [.30, .62] | .045 |
| Skala C | 1M | **.87 ± .06** | [.69, .94] | < .001 | **.79 ± .06** | [.64, .87] | < .001 |
| | 2M | **.58 ± .13** | [.34, .79] | .002 | **.52 ± .11** | [.31, .71] | .003 |
| Skala D | 1M | **.77 ± .10** | [.54, .89] | < .001 | **.68 ± .09** | [.49, .82] | < .001 |
| Skala E | 1M | **.85 ± .07** | [.69, .93] | < .001 | **.75 ± .07** | [.62, .85] | < .001 |
| | 2M | **.89 ± .05** | [.77, .94] | < .001 | **.81 ± .04** | [.71, .88] | < .001 |
| Skala F | 1M | **.70 ± .12** | [.41, .87] | < .001 | **.64 ± .11** | [.37, .81] | < .001 |
| | 2M | **.83 ± .11** | [.57, .95] | < .001 | **.75 ± .11** | [.51, .89] | < .001 |
| Skala G | 1M | **.84 ± .09** | [.63, .95] | < .001 | **.77 ± .09** | [.58, .90] | < .001 |
| | 2M | **.71 ± .13** | [.45, .98] | < .001 | **.63 ± .11** | [.41, .81] | < .001 |

*Anmerkungen.* Signifikante Korrelationen mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Korrelationen werden nur für reliable Subskalen berichtet. Da die Subskalenmittelwerte für die meisten Subskalen nicht normalverteilt sind, werden nur nicht-parametrische Korrelationen berichtet (Ausnahmen bilden die Skala E und F(2M)).

Bis auf Skala B in der zweiten Unterrichtsstunde korrelieren die Subskalenmittelwerte durchweg hoch mit dem Gesamteindruck bezüglich der Grundidee der jeweiligen Subskala – die Korrelationen für die zweite Unterrichtsstunde sind nur geringfügig kleiner als in der ersten Unterrichtsstunde. Es wird daher angenommen, dass die Handlungsindikatoren eine gute Beschreibung der jeweiligen Subskala darstellen. Die lediglich mittelhohe Korrelation der Skala „Exploration des Vorwissens“ mit dem Gesamteindruck zu dieser Skala in der zweiten Unterrichtsstunde könnte darin begründet liegen, dass die Handlungsindikatoren dieser Subskala eher die Exploration von Vorwissen beschreiben, wie man sie in Einführungsstunden erwarten würde, nicht aber in weiterführenden Stunden.

[23] Für die nicht-reliable Subskala A ergibt sich $r_{\text{Spearman}} = .67 \pm .12$ ($\text{KI}_{95\,\%} = [.40, .83]$, $p_{\text{1-seitig}} < .01$) (1M) bzw. $r_{\text{Spearman}} = .65 \pm .13$ ($\text{KI}_{95\,\%} = [.35, .83]$, $p_{\text{1-seitig}} < .01$) (2M) und für die Subskala D in der 2. Unterrichtsstunde $r_{\text{Spearman}} = .36 \pm .26$ ($\text{KI}_{95\,\%} = [-.27, .74]$, $p_{\text{1-seitig}} < .05$).

Auf Basis der vorangegangen Überlegungen wird von der Inhaltsvalidität des Ratings zur kognitiven Aktivierung ausgegangen. Dies gilt auch für die zweite Unterrichtsstunde, da lediglich eine Subskala wesentlich schlechter durch die Handlungsindikatoren repräsentiert wurde als in der ersten Unterrichtsstunde.[24]

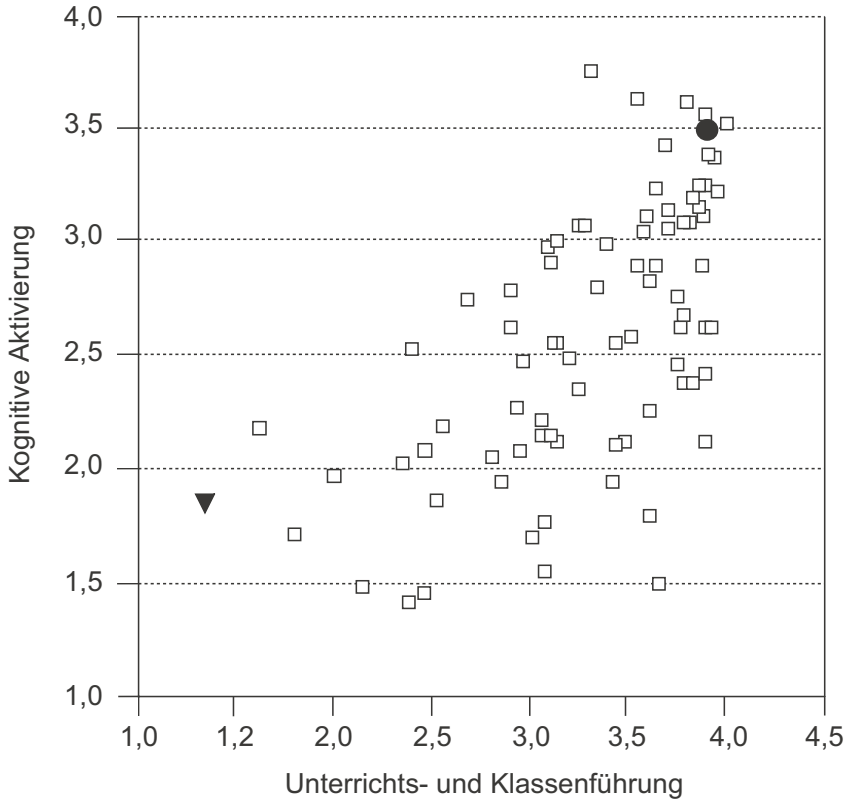
**Konstruktvalidität** Um Hinweise auf die Konstruktvalidität des Ratings zur kognitiven Aktivierung zu finden, wird eine diskriminante Validierung zu den Konstrukten Klassenführung und Vernetztheit der Sachstruktur durchgeführt, die im Rahmen der ProwiN-Videostudie ebenfalls erhoben wurden (vergl. hierzu Lenske et al., 2016; Liepertz, 2016). Da ein Maß für die Vernetztheit der Sachstruktur nur für die erste Unterrichtsstunde bestimmt wurde, werden die entsprechenden Korrelationen nur für diese Unterrichtsstunde berichtet. Tabelle 7.21 auf Seite 156 gibt einen Überblick über die Korrelationen zwischen den Merkmalen. Da keine negativen Zusammenhänge zu erwarten sind, wurden die Korrelationen einseitig auf Signifikanz getestet.

Die kognitive Aktivierung korreliert in mittlerer Höhe mit der Vernetzung der Sachstruktur, aber nicht mit der Klassenführung. Die beobachteten Zusammenhänge sind insofern erwartungskonform, da es sich bei der Vernetzung um ein fachspezifisches Merkmal und bei der Klassenführung um ein allgemeinpädagogisches Merkmal handelt. Die fachspezifischen Merkmale Vernetztheit und kognitive Aktivierung lassen sich deutlich voneinander trennen.

Klassenführung wird als wesentliche Voraussetzung für anspruchsvollen Unterricht angenommen (Helmke, 2009, S. 174) und hat sich im Rahmen der TIMSS-Studie als notwendige (aber nicht hinreichende) Voraussetzung für kognitiv aktivierenden Unterricht herausgestellt (Klieme et al., 2001, S. 53). Abbildungen 7.6a und 7.6b auf der nächsten Seite und auf Seite 155 zeigen den Zusammenhang der beiden Merkmale in der TIMSS-Stichprobe und in der hier untersuchte Stichprobe. Das Ergebnis der TIMSS-Studie, dass es keine Unterrichtsstunden mit niedrigem Qualitätsmaß für die Klassenführung bei gleichzeitig hohem Qualitätsmaß für die kognitive Aktivierung gibt, lässt sich sowohl bezüglich der für die erste und zweite Unterrichtsstunde generierten Qualitätsmaße als auch für die über beide Unterrichtsstunden gemittelten Qualitätsmaße replizieren, was als weiterer Hinweis auf die Validität dieser Maße gewertet werden kann.

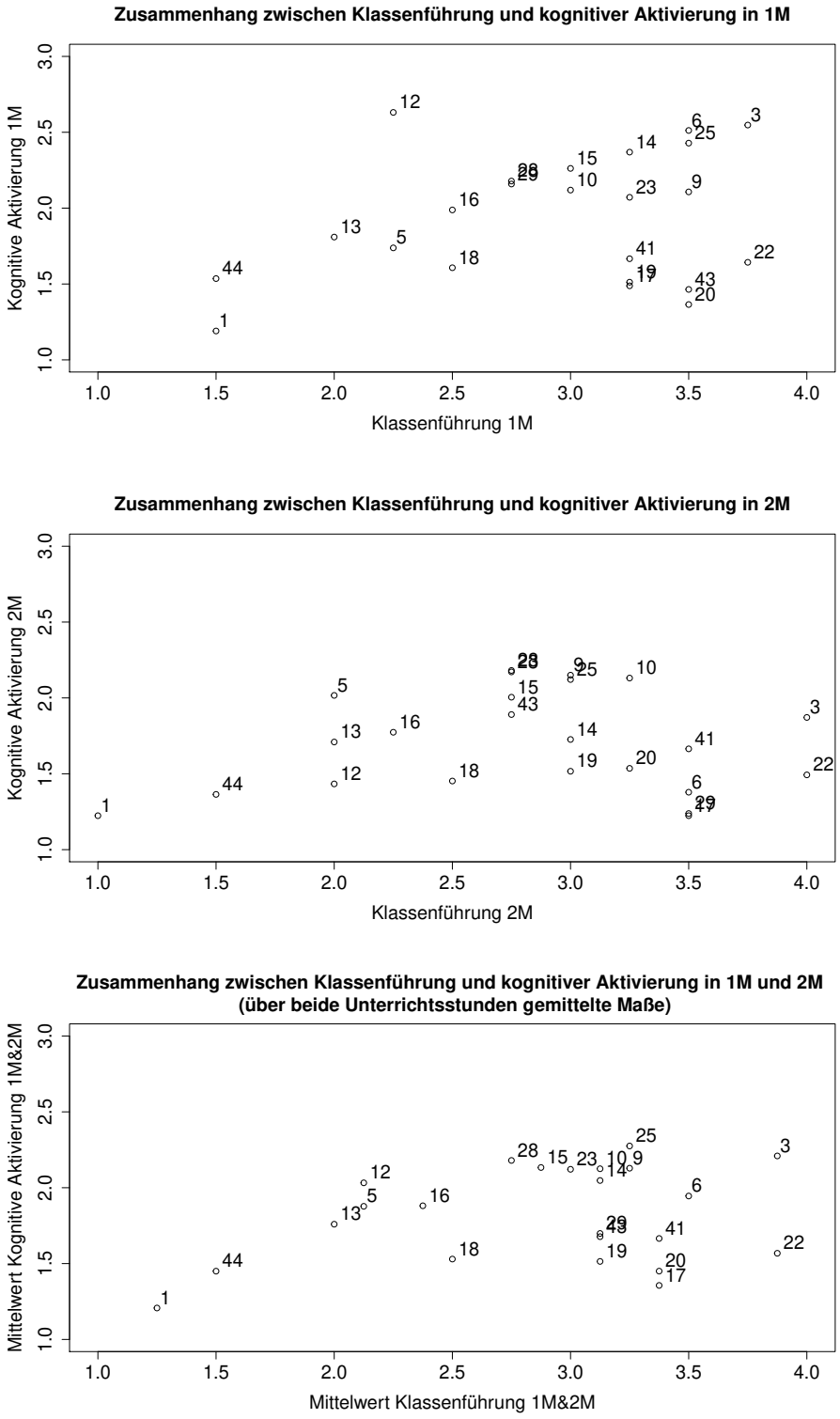
**Prädiktive Validität** Um die prädiktive Validität der Qualitätsmaße zur kognitiv aktivierenden Gestaltung des Unterrichts bezüglich des Unterrichtserfolgs zu untersuchen, wird im Rahmen von Mehrebenenanalysen geprüft, ob die Maße signifikant zur Aufklärung von Klassenunterschieden in den Fachwissensleistungen der Lernenden am Ende der gesamten Unterrichtseinheit zur Mechanik und im situationalen Interesse der Lernenden am Ende der jeweiligen Unterrichtsstunde

[24] Um die angenommen Skalenstruktur zu überprüfen, wäre außerdem die Durchführung exploratorischer oder konfirmatorischer Faktorenanalysen wünschenswert. Dies ist aber wegen der geringen Stichprobengröße nicht möglich, da hierfür mindestens 60 bzw. 200 Fälle notwendig wären (vergl. Bühner, 2006, S. 193,262).




(a)

Abbildung 7.6.
*Scatterplots für den Zusammenhang zwischen Klassenführung und kognitiver Aktivierung im Unterricht: (a) Scatterplot aus der TIMSS-Studie (Abbildung entnommen aus Klieme, Schümer & Knoll, 2001, S. 53). (Fortsetzung auf der nächsten Seite)*




(b)

Abbildung 7.6.
*(Fortsetzung) Scatterplots für den Zusammenhang zwischen Klassenführung und kognitiver Aktivierung im Unterricht: (b) Scatterplots für die Zusammenhänge zwischen den Qualitätsmaßen für kognitive Aktivierung und Klassenführung in der ersten Unterrichtsstunde (oben), in der zweiten Unterrichtsstunde (mittig) und zwischen den über beide Unterrichtsstunden gemittelten Qualitätsmaßen (unten).*

Tabelle 7.21.
*Korrelationen zwischen kognitiver Aktivierung (KA) und Klassenführung (KF) bzw. Vernetztheit (V) der Sachstruktur im Unterricht ($N_{1M} = N_{2M} = 23$)*

| Merkmale | KA - KF | | KA - V |
|---|---|---|---|
| | 1M | 2M | 1M |
| $r_{\text{Pearson}}$ | | $.09 \pm .23$ | $\mathbf{.38 \pm .15}$ |
| $\text{KI}_{95\,\%}$ | | [-.39,.51] | [.03,.62] |
| $p_{\text{1-seitig}}$ | | .335 | .037 |
| $r_{\text{Spearman}}$ | $.15 \pm .26$ | $-.04 \pm .25$ | $\mathbf{.40 \pm .18}$ |
| $\text{KI}_{95\,\%}$ | [-.37,.63] | [-.52,.45] | [-.01,.70] |
| $p_{\text{1-seitig}}$ | .251 | .437 | .029 |
| $\tau_{\text{Kendall}}$ | $.13 \pm .21$ | $-.01 \pm .19$ | $\mathbf{.25 \pm .14}$ |
| $\text{KI}_{95\,\%}$ | [-.27,.55] | [-.38,.35] | [-.02,.51] |
| $p_{\text{1-seitig}}$ | .203 | .468 | .045 |

*Anmerkungen.* Signifikante Korrelationen mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Das Maß für Klassenführung in der ersten Unterrichtsstunde ist nicht normalverteilt, daher wird hier keine parametrische Korrelation berichtet. Auch für normalverteilte Merkmale werden zusätzlich nicht-parametrische Korrelationen berichtet, da das Intervallskalenniveau der Qualitätsmaße zur kognitiven Aktivierung nicht sichergestellt werden kann (vergl. Abschnitt 7.4.1 auf Seite 92 zum Umgang mit Ordinalskalen).

beitragen. Hierdurch soll sichergestellt werden, dass es sich bei diesen Maßen auch um *Qualitätsmaße* im Sinne von Fenstermacher und Richardson (2005) handelt. Die Überprüfung der Zusammenhänge dient der Beantwortung der Forschungsfrage 2.1, weshalb die entsprechenden Mehrebenenmodelle im Ergebnisteil dieser Arbeit in Abschnitt 8.3.1.3 und 8.3.2.2 auf Seite 174 und auf Seite 180 berichtet werden.

**Kognitive Aktivierung und Fachwissensleistung** Nach Kontrolle des Vorwissens, der kognitiven Fähigkeiten, des Geschlechts, der zuhause gesprochenen Sprache und der Unterrichtszeit sind sowohl die Qualitätsmaße für die kognitiv aktivierende Gestaltung der ersten Unterrichtsstunde als auch die Qualitätsmaße für die zweite Unterrichtsstunde prädiktiv für die Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik ($\gamma_{\text{KA1M}}^{\text{StdYX}} = 0.40 \pm 0.22$, $p_{\text{1-seitig}} = .036$; $\gamma_{\text{KA2M}}^{\text{StdYX}} = 0.40 \pm 0.16$, $p_{\text{1-seitig}} = .005$). Zusammen mit der Unterrichtszeit klärt die kognitive Aktivierung $R^2 = (79 \pm 15)\%$ (Modell $2.1\text{a}_{\text{1M}}$) bzw. $R^2 = (80 \pm 20)\%$ (Modell $2.1\text{a}_{\text{2M}}$) der Varianz in den Klassenmittelwerten für die Post-Testwerte auf. Der Fachwissenserwerb der Lernenden wurde über einen weitaus längeren Zeitraum als die zwei videographierten Unterrichtsstunden erhoben. Dass dieser

dennoch signifikant damit zusammenhängt, wie kognitiv aktivierend die zwei Unterrichtsstunden gestaltet wurden, kann als Beleg dafür gewertet werden, dass das Rating zur kognitiven Aktivierung ein Merkmal der Unterrichtsqualität erfasst und dass die kognitiv aktivierende Gestaltung des üblichen Unterrichts einer Lehrkraft hinreichend gut durch die Qualitätsmaße in beiden Unterrichtsstunden beschrieben werden kann. Auch das über beide Stunden gemittelte Qualitätsmaß ist ein signifikanter Prädiktor für die Posttestleistungen der Lernenden ($\gamma^{\mathrm{StdYX}}_{\mathrm{KA1M\&2M}} = 0.46 \pm 0.20$, $p_{\text{1-seitig}} = .010$). Die Varianzaufklärung bezüglich der zwischen den Klassen liegenden Varianz wird durch die Mittelung über beide Unterrichtsstunden erhöht ($R^2 = (85 \pm 17)\%$), im Rahmen der Fehlerabschätzung handelt es sich aber nicht um eine bedeutsame Erhöhung.

**Situationales Interesse** Das Qualitätsmaß zur kognitiv aktivierenden Gestaltung der ersten Unterrichtsstunde ist ein signifikanter Prädiktor für das situationale Interesse am Ende der ersten Unterrichtsstunde ($\gamma^{\mathrm{StdYX}}_{\mathrm{KA1M}} = 0.35 \pm 0.19$, $p_{\text{1-seitig}} = .028$). Weder für die erste Unterrichtsstunde noch für die zweite Unterrichtsstunde tragen die Qualitätsmaße für die kognitive Aktivierung allerdings signifikant zur Aufklärung der Varianz in den Klassenmittelwerten für das situationale Interesse der Lernenden bei (Modell $2.1\mathrm{b}_{\mathrm{1M}}$: $R^2 = (12 \pm 13)\%$, $p_{\text{1-seitig}} = .170$; Modell $2.1\mathrm{b}_{\mathrm{2M}}$: $R^2 = (2 \pm 7)\%$, $p_{\text{1-seitig}} = .366$). Dies ist insofern verwunderlich, weil das situationale Interesse unmittelbar am Ende der jeweiligen Unterrichtstunde erhoben wurde. Die Ergebnisse deuten darauf hin, dass die über das in dieser Studie eingesetzte Rating beurteilte kognitiv aktivierende Gestaltung des Unterrichts nicht mit dem situationalen Interesse der Lernenden zusammenhängt. Da ein Zusammenhang zwischen kognitiv aktivierender Unterrichtsgestaltung und dem situationalen Interesse der Lernenden bisher nicht ausreichend empirisch abgesichert ist (vergl. Abschnitt 3.3.3.2 und 5.3.2 auf Seite 31 und auf Seite 68), stellt dieses Ergebnis allerdings nicht die grundsätzliche Validität des Ratings in Frage – bei der Interpretation der Qualitätsmaße zur kognitiven Aktivierung muss allerdings beachtet werden, dass diese zwar lernförderlichen, aber nicht Interesse generierenden Unterricht beschreiben.

Die Validierungsergebnisse lassen sich wie folgt zusammenfassen: Zunächst scheinen die zwei videographierten Unterrichtsstunden zusammengenommen den üblichen Unterricht der Lehrkräfte hinreichend gut zu repräsentieren. Es wird davon ausgegangen, dass das hier eingesetzte Rating die kognitiv aktivierende Gestaltung des Unterrichts in beiden Unterrichtsstunden inhaltsvalide erfasst, wobei das Rating Merkmale der kognitiv aktivierenden Unterrichtsgestaltung der ersten Unterrichtsstunde etwas besser beschreibt. Zusammenhänge zu anderen Merkmalen der Unterrichtsqualität und zu den in dieser Studie betrachteten Zielkriterien von Unterricht weisen darauf hin, dass die auf Basis des Ratings generierten Qualitätsmaße ein fachspezifischen Merkmal der Unterrichtsqualität erfassen, das Klassenunterschiede in den Fachwissensleistungen (nicht aber im situationalen Interesse) der Lernenden aufklärt. Außerdem konnten im Rahmen der Validierung Ergebnisse der TIMSS Studie zum Zusammenhang zwischen kognitiver

Aktivierung und Klassenführung repliziert werden, was als weiterer Hinweis auf die Validität des Ratings gewertet werden kann – Klassenführung scheint eine notwendige aber nicht hinreichende Bedingung für die Gestaltung eines kognitiv aktivierenden Unterrichts zu sein. An dieser Stelle sei angemerkt, dass lediglich die Qualitätsmaße für die Gesamtskala zur kognitiven Aktivierung validiert wurden – die Subsklalenmaße wurden nicht gesondert validiert.

# 8. Ergebnisse

In diesem Kapitel werden zunächst deskriptive Ergebnisse zu allen in die späteren Analysen einbezogenen Lehrer-, Schüler- und Unterrichtsvariablen vorgestellt. In Abschnitt 8.2 auf Seite 168 werden die Ergebnisse zu den Fachwissenszuwächsen der Lernenden vom Prä- zum Post-Test berichtet. In Abschnitt 8.3 auf Seite 170 werden die Ergebnisse der Mehrebenenanalysen zur Beantwortung der Forschungsfragen 1 und 2.1 zum Zusammenhang zwischen Professionswissen und Unterrichtserfolg und kognitiv aktivierendem Unterricht und Unterrichtserfolg dargestellt. Abschließend wird in Abschnitt 8.4 auf Seite 181 über die Ergebnisse zum Zusammenhang zwischen dem Professionswissen der Lehrkräfte und der kognitiv aktivierenden Gestaltung ihres Unterrichts für die Beantwortung der Forschungsfrage 2.2 berichtet.

## 8.1. Deskriptive Ergebnisse

In den folgenden Abschnitten werden die deskriptiven Ergebnisse zur Beschreibung der Lehrerstichprobe, des Unterrichts und der Schülerstichprobe vorgestellt.

### 8.1.1. Beschreibung der Lehrerstichprobe

Die in der vorliegenden Arbeit untersuchte Lehrerstichprobe umfasst $N = 23$ Lehrkräfte, die Physik am Gymnasium unterrichteten und mit einer 8. oder 9. Klasse an der Studie teilnahmen. Im Folgenden wird der demographische Hintergrund und die Lehrerfahrung und anschließend das Professionswissen der Lehrkräfte beschrieben. In den jeweiligen Abschnitten erfolgt außerdem ein Vergleich mit der Stichprobe der Gymnasiallehrkräfte aus ProwiN I – dieser Vergleich soll Hinweise darauf liefern, ob es sich bei der in dieser Arbeit untersuchten Stichprobe um eine starke Positivauswahl handelt.[1]

#### 8.1.1.1. Demographischer Hintergrund und Lehrerfahrung

In Tabelle 8.1 auf der nächsten Seite sind deskriptive Ergebnisse für den demographischen Hintergrund und die Lehrerfahrung der in dieser Arbeit untersuchten Lehrkräfte (ProwiN II) im Vergleich zur Stichprobe der Gymnasiallehrkräfte aus NRW aus ProwiN I aufgeführt. Die hier untersuchten Lehrkräfte unterrichten im Mittel eine Unterrichtsstunde Physik mehr pro Woche als die Lehrkräfte der ProwiN I-Stichprobe. Bezüglich der Geschlechterverteilung, des Alters und der

[1] Die Stichprobe der Gymnasiallehrkräfte aus ProwiN I stellt allerdings ihrerseits wahrscheinlich eine Positivauswahl dar, da auch die Teilnahme an ProwiN I freiwillig war.

Lehrerfahrung existieren keine nennenswerten Unterschiede zwischen den Stichproben.

Weibliche Physiklehrkräfte sind in der ProwiN II-Stichprobe mit einem Anteil von 35% im Vergleich zur Bundeslandquote, die im Schuljahr 2013/2014 in NRW bei 27% lag, etwas überrepräsentiert (vergl. MSW, 2015).[2] Das Durchschnittsalter der Lehrkräfte lag mit $M = 44$ Jahren knapp unter dem Bundeslanddurchschnitt für Gymnasiallehrkräfte aller Fächer, der in NRW im Schuljahr 2013/2014 bei 45.5 Jahren lag (vergl. MSW, 2015, S. 45). Die Stichprobe umfasst sowohl Lehrkräfte, die am Beginn ihres Berufslebens stehen, als auch Lehrkräfte, die seit mehr als 30 Jahren Physik unterrichten – es wird also ein breites Erfahrungsspektrum abgedeckt. Ihre eigene Schulzeit haben 17% der Lehrkräfte mit einem sehr guten (Abiturnote $< 1.5$) und 70% mit einem guten Abitur ($1.5 <$ Abiturnote $< 2.5$) abgeschlossen (vergl. Tabelle B.1 auf Seite 242 im Anhang). Da keine Vergleichsdaten zur mittleren Abiturdurchschnittsnote von Physiklehrkräften in NRW vorliegen, kann nicht entschieden werden, ob es sich bei den hier untersuchten Lehrkräften bezüglich der Abiturnote, die als Indikator für die allgemeinen kognitiven Fähigkeiten der Lehrkräfte angesehen werden kann (vergl. z. B. Abel & Faust, 2010, S. 51), um eine Positivauswahl handelt.

Tabelle 8.1.
*Demographischer Hintergrund und Lehrerfahrung der Lehrkräfte aus ProwiN II im Vergleich zur Stichprobe der Gymnasiallehrkräfte NRW aus ProwiN I*

| Merkmale | ProwiN II ($N = 23$) | | | | ProwiN I ($N = 79$) | | | |
|---|---|---|---|---|---|---|---|---|
| | ♀ | | ♂ | | ♀ | | ♂ | |
| Geschlecht | 35% | | 65% | | 37% | | 63% | |
| | *M* | *SD* | *Min* | *Max* | *M* | *SD* | *Min* | *Max* |
| Alter [Jahre] | 44.0 | 11.6 | 28.0 | 63.0 | 43.7 | 9.9 | 27.0 | 64.0 |
| Abiturnote[1] | 1.9 | 0.4 | 1.3 | 2.8 | | | | |
| Jahre im Schuldienst[2] | 12.9 | 11.6 | 2.0 | 36.0 | 12.8 | 11.2 | 1.0 | 38.0 |
| Stunden/Woche Physikunterricht[3] | 10.5 | 4.6 | 3.0 | 22.0 | 9.1 | 5.4 | 0.0 | 26.0 |

[1] Angaben zur Abiturnote lagen nur für die ProwiN II-Stichprobe vor.
[2] Entspricht auch der Anzahl an Jahren, in denen Physik unterrichtet wurde.
[3] Anzahl der pro Woche unterrichteten Schulstunden im Unterrichtsfach Physik zum Zeitpunkt der Post-Erhebung.

#### 8.1.1.2. Professionswissen

Als Maß für das fachspezifische Professionswissen der Lehrkräfte wurden im Rasch-Modell Personenfähigkeiten geschätzt. Der Nullpunkt der Fähigkeitsskala wurde

[2] Die Bundeslandquote wurde aus den in MSW (2015, S. 53 bzw. S. 55) angegebenen Zahlen für die Anzahl der Lehrkräfte und die Anzahl der weiblichen Lehrkräfte mit Lehrbefähigungen im Fach Physik für Gymnasien ermittelt.

auf den Mittelwert der Aufgabenschwierigkeiten gelegt, deren Standardabweichung bei $SD = 1.0$ liegt. Daher können die CK- und PCK-Testwerte sowohl negative als auch positive Werte annehmen. Im Rahmen der Rasch-Analyse werden für alle Lehrkräfte untere Grenzwerte für die Standardfehler auf die Personenfähigkeiten ausgegeben.[3] Der über die Gesamtstichprobe der $N = 102$ Gymnasiallehrkräfte aus ProwiN I und II gemittelte Standardfehler auf die Personenfähigkeiten beträgt im CK-Test $M_{Error} = 0.44$ ($SD = 0.07$, $Min = 0.39$, $Max = 0.66$) und im PCK-Test $M_{Error} = 0.44$ ($SD = 0.07$, $Min = 0.42$, $Max = 1.11$). Die CK- und PCK-Testwerte werden daher mit einer Dezimalstelle angegeben (vergl. Abschnitt 7.4.1 auf Seite 90 im Kapitel zu allgemeinen Hinweisen zur Datenanalyse). Der PK-Testwert entspricht dem Anteil gelöster Aufgaben im PK-Test.

Tabelle 8.2.
*Deskriptive Statistik für das Professionswissen der Lehrkräfte aus ProwiN II ($N = 23$) im Vergleich zur Stichprobe der Gymnasiallehrkräfte NRW aus ProwiN I ($N = 79$)*

| Merkmale | ProwiN II | | | | ProwiN I | | | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *Min* | *Max* | *M* | *SD* | *Min* | *Max* |
| CK[1] | 0.4 | 1.0 | −1.5 | 1.9 | 0.0 | 0.8 | −1.8 | 1.9 |
| PCK[1] | 0.0 | 0.5 | −0.7 | 1.0 | −0.1 | 0.8 | −3.6 | 1.2 |
| PK[2] | 72 | 7 | 60 | 83 | 69 | 13 | 23 | 93 |

[1] Angabe in Rasch-Logits
[2] Angabe in % gelöster Aufgaben

Tabelle 8.2 zeigt die deskriptiven Ergebnisse für das Professionswissen der Lehrkräfte in der in dieser Arbeit untersuchten Stichprobe (ProwiN II) und in der Stichprobe der Gymnasiallehrkräfte NRW aus ProwiN I. Die hier untersuchten Lehrkräfte schneiden im Mittel sowohl im fachspezifischen Professionswissen als auch im pädagogischen Wissen etwas besser ab als die in ProwiN I untersuchten Gymnasiallehrkräfte – die Unterschiede sind allerdings klein und nur bezüglich des Fachwissens signifikant (vergl. Tabelle 8.3 auf der nächsten Seite). Da sehr unterschiedlich große Gruppen verglichen werden und zudem die PCK- und PK-Testwerte der Lehrkräfte der ProwiN I-Stichprobe nicht normalverteilt sind, wurden Mann-Whitney-U-Tests gerechnet. Die Effektstärken wurde über $r_{\mathrm{MW}} = z/\sqrt{N}$ berechnet (vergl. Field, 2009, S. 550). Da in erster Linie geprüft werden sollte, ob die hier untersuchten Lehrkräfte *besser* abschneiden, wurde einseitig auf Signifikanz getestet.

Korrelationen zwischen dem CK, PCK und PK der Lehrkräfte in der um die ProwiN I-Lehrkräfte erweiterten Stichprobe wurden bereits in Tabelle 7.5 auf Seite 111 im Abschnitt zur Validierung der Professionswissenstests gezeigt und diskutiert. In der hier untersuchten Stichprobe der $N = 23$ Physiklehrkräfte korrelieren die Professionswissensdimensionen nicht signifikant miteinander (vergl. Tabelle 8.4 auf der nächsten Seite).

[3] Hierbei handelt es sich um die Model Standard Errors aus Winsteps.

Tabelle 8.3.
*Statistiken (U), z-Werte, Effektstärken ($r_{MW}$) und Signifikanzen der Mann-Whitney-U-Tests auf Unterschiede zwischen den CK-, PCK- und PK-Testwerten der ProwiN I- ($N = 23$) und ProwiN II-Lehrkräfte ($N = 79$)*

| Merkmale | CK | PCK | PK |
|---|---|---|---|
| $U$ (23,79) | $\mathbf{(6.9 \pm 1.3) \cdot 10^2}$ | $(68.2 \pm 1.3) \cdot 10^2$ | $(8.4 \pm 1.3) \cdot 10^2$ |
| $z$ | **−1.8** | −0.68 | −0.55 |
| $r_{\text{MW}}$ | **−.18** | −.07 | −.05 |
| $p_{\text{1-seitig}}$ | .039 | .246 | .290 |

*Anmerkungen.* Signifikante Unterschiede mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt.

Tabelle 8.4.
*Korrelationen zwischen den Dimensionen des Professionswissens in der Stichprobe der ProwiN II-Lehrkräfte ($N = 23$)*

| Merkmale | PCK-CK | PCK-PK | CK-PK |
|---|---|---|---|
| $r_{\text{Pearson}}$ | .22 ± .22 | .08 ± .23 | .07 ± .20 |
| $\text{KI}_{95\,\%}$ | [−.22, .62] | [−.34, .54] | [−.33, .45] |
| $p_{\text{1-seitig}}$ | .157 | .366 | .371 |

## 8.1.2. Beschreibung des Unterrichts

Die Fachwissensleistungen der Lernenden wurden vor und nach der Unterrichtseinheit Mechanik erhoben. Innerhalb der Unterrichtseinheit wurden zwei Unterrichtsstunden videographiert: die Einführungsstunde zum Kraftbegriff (1M) sowie die Folgestunde (2M). In diesem Abschnitt werden zunächst die deskriptiven Ergebnisse zum Umfang der Unterrichtseinheit Mechanik und zu den videographierten Unterrichtsstunden und anschließend die deskriptiven Ergebnisse zur kognitiv aktivierenden Gestaltung der Unterrichtsstunden beschrieben.

### 8.1.2.1. Unterrichtszeit in der Unterrichtseinheit Mechanik

Die Unterrichtseinheit Mechanik fand zwischen dem Prä- und dem Post-Test statt. Im Mittel lagen zwischen der Prä- und der Post-Erhebung 175 Tage ($SD = 60$ Tage, $Min = 70$ Tage, $Max = 310$ Tage).[4] Die Lehrkräfte gaben an, wie viele Unterrichtsstunden sie im Rahmen der Mechanikeinheit unterrichtet hatten. Der auf 45-Minuten-Stunden normierte Stundenumfang der Mechanikeinheit variierte zwischen 12 und 59 Unterrichtsstunden ($M = 34$, $SD = 10$) – hierbei handelt es sich um die tatsächlich stattgefundenen Unterrichtsstunden. Über das Verhältnis von

[4] Umgerechnet in Wochen entspricht dies: $M = 25.0$ Wochen, $SD = 8.5$ Wochen, $Min = 10.0$ Wochen, $Max = 44.3$ Wochen.

tatsächlich stattgefundenen zu theoretisch möglichen Unterrichtsstunden (also der Anzahl an Unterrichtsstunden, die nach Abzug von Ferienzeiten und Feiertagen zwischen Prä- und Post-Erhebung hätte stattfinden können) wurde der Anteil ausgefallener Unterrichtstunden geschätzt: Dieser variierte zwischen 0% und 59% ($M = 12\%$, $SD = 16\%$) (vergl. Tabelle B.2 auf Seite 243 im Anhang).

#### 8.1.2.2. Kognitive Aktivierung im videographierten Unterricht

Die videographierten Unterrichtsstunden fanden zu unterschiedlichen Zeitpunkten innerhalb der Unterrichtseinheit Mechanik statt. Der Zeitraum zwischen dem Prä-Test und der Aufzeichnung der Unterrichtsstunde zur Einführung des Kraftbegriffes variierte zwischen 1 und 91 Tagen ($M = 26$ Tage, $SD = 22$ Tage).[5] Bei 30% der Lehrkräfte wurden 45-Minuten-Stunden aufgezeichnet, bei jeweils 13% der Lehrkräfte 60- bzw. 67.5-Minuten-Stunden, bei 4% der Lehrkräfte 70-Minuten-Stunden und bei 39% der Lehrkräfte 90-Minuten-Stunden (vergl. Tabelle B.1 auf Seite 242 im Anhang).

Tabelle 8.5 auf der nächsten Seite zeigt die deskriptiven Ergebnisse für die Qualitätsmaße zur kognitiv aktivierenden Gestaltung der ersten und zweiten Unterrichtsstunde und für die über beide Unterrichtsstunden gemittelten Qualitätsmaße. Korrelationen zwischen den Qualitätsmaßen in der ersten und zweiten Unterrichtsstunde wurden bereits in Tabelle 7.19 auf Seite 151 im Abschnitt zur Validierung des videobasierten Ratinginstruments gezeigt und diskutiert.

### 8.1.3. Beschreibung der Schülerstichprobe

Insgesamt nahmen $N = 661$ Schülerinnen und Schüler an mindestens einem der vier Erhebungstermine (Prä-Test, Video 1M, Video 2M, Post-Test) teil. In diesem Abschnitt wird zunächst der demographische Hintergrund der Lernenden beschrieben. Anschließend erfolgt eine Beschreibung der Leistungen der Schülerinnen und Schüler im Prä- und Post-Test zum Fachwissen in Mechanik und im Kognitive Fähigkeitentest (KFT). Außerdem wird über deskriptive Ergebnisse zum situationalen Interesse der Lernenden berichtet. Die Beschreibung der Schülerstichprobe erfolgt sowohl auf Schülerebene als auch auf Klassenebene.

#### 8.1.3.1. Demographischer Hintergrund

Von den $N = 660$ Lernenden, die am Prä- oder Post-Test teilnahmen (und für die daher Angaben zum demographischen Hintergrund vorliegen), waren 57% weiblich und 43% männlich. Im Mittel waren die Lernenden zum Zeitpunkt der Erhebungen $M = 13.8$ Jahre alt ($SD = 0.7$ Jahre, $Min = 11.5$ Jahre, $Max = 17.0$ Jahre). 79% der Schülerinnen und Schüler gaben an, zuhause ausschließlich deutsch zu sprechen, 18% gaben an, deutsch und andere Sprachen zu sprechen und 3% gaben an, ausschließlich andere Sprachen zu sprechen.[6]

[5] Umgerechnet in Wochen entspricht dies: $M = 3.7$ Wochen, $SD = 3.2$ Wochen, $Min = 0$ Wochen, $Max = 13$ Wochen.

[6] Die entsprechenden Angaben für die $N = 610$ Lernenden, die am Prä- und Post-Test teilgenommen hatten, unterscheiden sich kaum von den hier berichteten Werten. Der Anteil der

Tabelle 8.5.
*Deskriptive Statistik für die Qualitätsmaße zur kognitiv aktivierenden Gestaltung der 1./2. videographierten Unterrichtsstunde (1M/2M) und für die über beide Unterrichtsstunden gemittelten Qualitätsmaße (1M&2M) (N = 23)*

| Skala | | 1M | 2M | 1M&2M |
|---|---|---|---|---|
| Skala A:[1] Lernstatus bewusst machen | *M* | 1.7 | 1.4 | 1.5 |
| | *SD* | 0.4 | 0.3 | 0.3 |
| | *Min* | 1.0 | 1.0 | 1.0 |
| | *Max* | 2.8 | 2.2 | 2.1 |
| Skala B: Exploration des Vorwissens | *M* | 1.8 | 1.3 | 1.6 |
| | *SD* | 0.6 | 0.4 | 0.4 |
| | *Min* | 1.0 | 1.0 | 1.0 |
| | *Max* | 3.0 | 2.3 | 2.3 |
| Skala C: Exploration der Denkweisen | *M* | 1.7 | 1.5 | 1.6 |
| | *SD* | 0.6 | 0.5 | 0.5 |
| | *Min* | 1.0 | 1.0 | 1.0 |
| | *Max* | 2.8 | 2.5 | 2.5 |
| Skala D:[2] Evolutionärer Umgang mit Schülervorstellungen | *M* | 1.6 | 1.4 | 1.5 |
| | *SD* | 0.5 | 0.4 | 0.4 |
| | *Min* | 1.0 | 1.0 | 1.0 |
| | *Max* | 2.8 | 2.5 | 2.5 |
| Skala E: Lehrperson als Mediator | *M* | 2.0 | 2.0 | 2.0 |
| | *SD* | 0.5 | 0.4 | 0.4 |
| | *Min* | 1.0 | 1.2 | 1.1 |
| | *Max* | 2.8 | 2.7 | 2.7 |
| Skala F: Kein rezeptives Lernverständnis | *M* | 2.5 | 2.5 | 2.5 |
| | *SD* | 0.4 | 0.4 | 0.3 |
| | *Min* | 1.3 | 1.5 | 1.7 |
| | *Max* | 3.0 | 3.0 | 3.0 |
| Skala G: Herausfordernde Lerngelegenheiten | *M* | 2.2 | 1.8 | 2.0 |
| | *SD* | 0.5 | 0.5 | 0.4 |
| | *Min* | 1.5 | 1.0 | 1.4 |
| | *Max* | 3.0 | 2.7 | 2.8 |
| Gesamtskala: Kognitive Aktivierung | *M* | 1.9 | 1.7 | 1.8 |
| | *SD* | 0.4 | 0.3 | 0.3 |
| | *Min* | 1.2 | 1.2 | 1.2 |
| | *Max* | 2.6 | 2.2 | 2.3 |

*Anmerkungen.* Die kognitiv aktivierende Unterrichtsgestaltung wurde auf einer dreistufigen Ratingskala eingeschätzt (1 =„trifft nicht zu“, 2 =„teils teils“, 3 =„trifft zu“).

[1] Subskala war in beiden Unterrichtsstunden nicht reliabel ($\alpha_{C,1M} = .24$, $\alpha_{C,2M} = .36$)

[2] Subskala war in der 2. Unterrichtsstunde nicht reliabel ($\alpha_{C,2M} = .24$)

Auf Klassenebene variierte der Anteil der Mädchen zwischen 16% und 100% ($M = 57\%$, $SD = 19\%$; zwei Lehrkräfte (ID 16, ID 23) unterrichteten an einem Mädchengymnasium). Das mittlere Alter der Lernenden in den Klassen variierte zwischen 13.3 und 14.9 Jahren ($M = 13.8$ Jahre, $SD = 0.5$ Jahre). Der Anteil der Lernenden, die angaben, zuhause nicht oder nicht nur deutsch zu sprechen, variierte zwischen den Klassen zwischen 0% und 47% ($M = 21\%$, $SD = 12\%$) (vergl. Tabelle B.2 auf Seite 243 im Anhang).

#### 8.1.3.2. Fachwissensleistungen und kognitive Fähigkeiten

Die in diesem Abschnitt aufgeführten Ergebnisse beziehen sich auf die Leistungen der $N = 610$ Lernenden, die sowohl am Prä- als auch am Post-Test teilgenommen hatten. Als Maß für die Fachwissensleistungen der Schülerinnen und Schüler im Prä- und Post-Test und für deren Leistungen im Kognitive Fähigkeitentest (KFT) wurden im Rasch-Modell Personenfähigkeiten geschätzt. Der Nullpunkt der Fähigkeitsskala wurde auf den Mittelwert der Aufgabenschwierigkeiten gelegt, deren Standardabweichung bei $SD = 1.0$ liegt. Daher können die Prä-, Post- und KFT-Testwerte sowohl negative als auch positive Werte annehmen. Im Rahmen der Rasch-Analyse werden für jeden Lernenden untere Grenzwerte für die Standardfehler auf die Personenfähigkeiten ausgegeben.[7] Der über die Gesamtstichprobe der $N = 610$ am Prä- und Post-Test anwesenden Schülerinnen und Schüler gemittelte Standardfehler auf die Personenfähigkeiten beträgt für den Prä-Test $M_{Error} = 0.56$ ($SD = 0.04$, $Min = 0.51$, $Max = 0.85$), für den Post-Test $M_{Error} = 0.57$ ($SD = 0.11$, $Min = 0.51$, $Max = 1.87$) und für den KFT $M_{Error} = 0.60$ ($SD = 0.28$, $Min = 0.45$, $Max = 1.85$). Die Prä-, Post- und KFT-Testwerte werden daher mit einer Dezimalstelle angegeben (vergl. Abschnitt 7.4.1 auf Seite 90 im Kapitel zu allgemeinen Hinweisen zur Datenanalyse).

Tabelle 8.6 auf der nächsten Seite zeigt die deskriptiven Ergebnisse für die Leistungen der Lernenden im Prä- und Post-Test und deren kognitive Fähigkeiten auf Schülerebene. Außerdem werden deskriptive Ergebnisse für die innerhalb der Klassen gemittelten Werte berichtet.

#### 8.1.3.3. Situationales Interesse

Tabelle 8.7 auf Seite 167 zeigt die deskriptiven Ergebnisse zum situationalen Interesse der Lernenden am Unterricht in der ersten und zweiten Unterrichtsstunde sowie die über beide Unterrichtsstunden gemittelten Maße. Korrelationen zwischen den Maßen für das situationale Interesse der Lernenden in der ersten und zweiten Unterrichtsstunde wurden bereits in Tabelle 7.12 auf Seite 131 im Abschnitt zur Validierung des Fragebogens zum situationalen Interesse gezeigt und diskutiert.

Mädchen beträgt in dieser Stichprobe 56% und die maximale Altersangabe 16 Jahre, alle anderen Werten sind identisch.

[7] Hierbei handelt es sich um die Model Standard Errors aus Winsteps.

Tabelle 8.6.
*Deskriptive Statistik für die Fachwissensleistungen und die kognitiven Fähigkeiten der Lernenden auf Schülerebene und für die auf Klassenebene gemittelten Werte*

| | *M* | *SD* | *Min* | *Max* |
|---|---|---|---|---|
| Schülerebene ($N = 610$) | | | | |
| Prä-Test | 0.3 | 0.8 | −2.2 | 3.3 |
| Post-Test | 0.7 | 0.9 | −1.8 | 5.3 |
| KFT | 1.0 | 1.7 | −3.9 | 5.1 |
| Klassenebene ($N = 23$) | | | | |
| Prä-Test | 0.3 | 0.3 | −0.2 | 0.9 |
| Post-Test | 0.7 | 0.4 | −0.1 | 1.2 |
| KFT | 1.0 | 0.8 | −0.3 | 2.5 |
| Gültige Fälle innerhalb der Klassen | 26.5 | 3.6 | 19 | 32 |
| Missings innerhalb der Klassen[1] | 2.2 | 1.8 | 0 | 7 |

*Anmerkung.* Alle Testwerte sind in Rasch-Logits angegeben.
[1] Gibt an, wie viele der Schülerinnen und Schüler einer Klasse im Mittel an mindestens einem der beiden Testzeitpunkte fehlten.

Tabelle 8.7.
*Deskriptive Statistik für das situationale Interesse der Lernenden am Unterricht in der 1. und 2. videographierten Unterrichtsstunde (1M/2M) und für die über beide Unterrichtsstunden gemittelten Maße (1M&2M)*

| | | *M* | *SD* | *Min* | *Max* |
|---|---|---|---|---|---|
| 1M | Schülerebene ($N_{1M} = 633$)<br>Situationales Interesse | 4.3 | 1.2 | 1.0 | 7.0 |
| | Klassenebene ($N = 23$)<br>Situationales Interesse | 4.3 | 0.5 | 3.4 | 5.0 |
| | Gültige Fälle innerhalb der Klassen | 27.5 | 3.3 | 20 | 32 |
| | Missings innerhalb der Klassen | 1.2 | 1.1 | 0 | 4 |
| 2M | Schülerebene ($N_{2M} = 625$)<br>Situationales Interesse | 4.2 | 1.4 | 1.0 | 7.0 |
| | Klassenebene ($N = 23$)<br>Situationales Interesse | 4.2 | 0.7 | 2.8 | 5.1 |
| | Gültige Fälle innerhalb der Klassen | 27.2 | 3.3 | 20 | 33 |
| | Missings innerhalb der Klassen | 1.6 | 1.0 | 0 | 4 |
| 1M&2M | Schülerebene ($N_{1M\&2M} = 600$)<br>Situationales Interesse | 4.2 | 1.1 | 1.0 | 7.0 |
| | Klassenebene ($N = 23$)<br>Situationales Interesse | 4.2 | 0.6 | 3.1 | 5.1 |
| | Gültige Fälle innerhalb der Klassen | 26.1 | 3.3 | 20 | 31 |
| | Missings innerhalb der Klassen | 2.7 | 1.6 | 0 | 7 |

*Anmerkungen.* Das situationale Interesse wurde von den Lernenden auf einer siebenstufigen Likertskala eingeschätzt (1 =„stimme gar nicht zu“, 7 =„stimme voll zu“). Die Anzahl an Missings innerhalb der Klassen gibt an, wie die $N_{1M} = 661-633 = 28$, $N_{2M} = 661-625 = 36$ bzw. $N_{1M\&2M} = 661-600 = 61$ in der 1./2. Unterrichtsstunde bzw. in einer der beiden Unterrichtstunden fehlenden Schülerinnen und Schüler auf die Klassen verteilt sind.

## 8.2. Ergebnisse zum Fachwissenszuwachs der Lernenden

Der Fachwissenszuwachs der Lernenden wurde aus der Differenz zwischen den Post-und Prä-Testwerten der Lernenden im Schülerfachwissenstest berechnet. Von den $N = 610$ Schülerinnen und Schülern, die an beiden Testzeitpunkten anwesend waren, schnitten 66% im Post-Test besser ab als im Prä-Test (Fachwissenszuwachs $> 0$) und 34% genauso gut oder schlechter (Fachwissenzuwachs $\leq 0$). Der mittlere Zuwachs im Fachwissen der Lernenden über die Unterrichtseinheit Mechanik beträgt $M = 0.4$ ($SD = 0.9$, $Min = -3.1$, $Max = 4.4$) und entspricht damit einer halben Standardabweichung im Prä-Test (vergl. Tabelle 8.6 auf Seite 166). Im Mittel schnitten die Lernenden im Post-Test signifikant besser ab als im Prä-Test ($t(609) = 10.501$, $p_{\text{1-seitig}} < .001$) – es handelt sich um einen signifikanten Effekt mit mittlerer Effektstärke ($d = 0.43 \pm 0.04$, $\text{KI}_{95\,\%} = [0.35, 0.50]$). Cohens $d$ wurde über den Mittelwert und die Standardabweichung der Differenzwerte bestimmt. Aufgrund signifikanter Abweichungen der Verteilung der Differenzwerte von der Normalverteilung wurde zusätzlich ein Wilkoxon-Vorzeichen-Rang-Test gerechnet, der ebenfalls einen signifikanten Effekt mit mittlerer Effektstärke anzeigt ($T(609) = (1331.26 \pm 0.43) \cdot 10^2$, $z = 9.74$, $r_{\text{W}} = .39$, $p_{\text{asympt.,1-seitig}} < .001$). Die Effektstärke wurde über $r_{\text{W}} = z/\sqrt{N}$ berechnet (vergl. Field, 2009, S. 558).

Zwischen den Fachwissenszuwächsen in den einzelnen Klassen zeigen sich deutliche Unterschiede: in 9 der 23 Klassen sind die Fachwissenszuwächse nicht signifikant ($p_{\text{1-seitig}} > .05$). Die mittleren Fachwissenszuwächse in den Klassen variieren zwischen $M_{\text{ID 44}} = 0.0$ ($SD = 0.9$, $Min = -1.2$, $Max = 1.9$) und $M_{\text{ID 9}} = 0.8$ ($SD = 0.8$, $Min = -1.1$, $Max = 2.4$) und die Effektstärken zwischen $d_{\text{ID 44}} = -0.04 \pm 0.21$ ($\text{KI}_{95\,\%} = [-0.47, 0.35]$) und $d_{\text{ID 9}} = 1.06 \pm 0.26$ ($\text{KI}_{95\,\%} = [0.56, 1.52]$). In den Klassen mit den IDs 1, 5, 10, 14, 18 und 25 waren die Differenzwerte zwischen Prä- und Post-Testwerten nicht normalverteilt. Für die Signifikanzbestimmung wurden in diesen Klassen Wilkoxon-Vorzeichen-Rang-Tests gerechnet. In allen anderen Fällen wurde die Signifikanz der Differenzwerte über t-Tests bestimmt. Effektstärken wurden für alle Klassen über Cohens $d$ bestimmt, um diese vergleichen zu können. Abbildung 8.1 auf der nächsten Seite zeigt Unterschiede in den Fachwissenszuwächsen in den untersuchten Klassen.

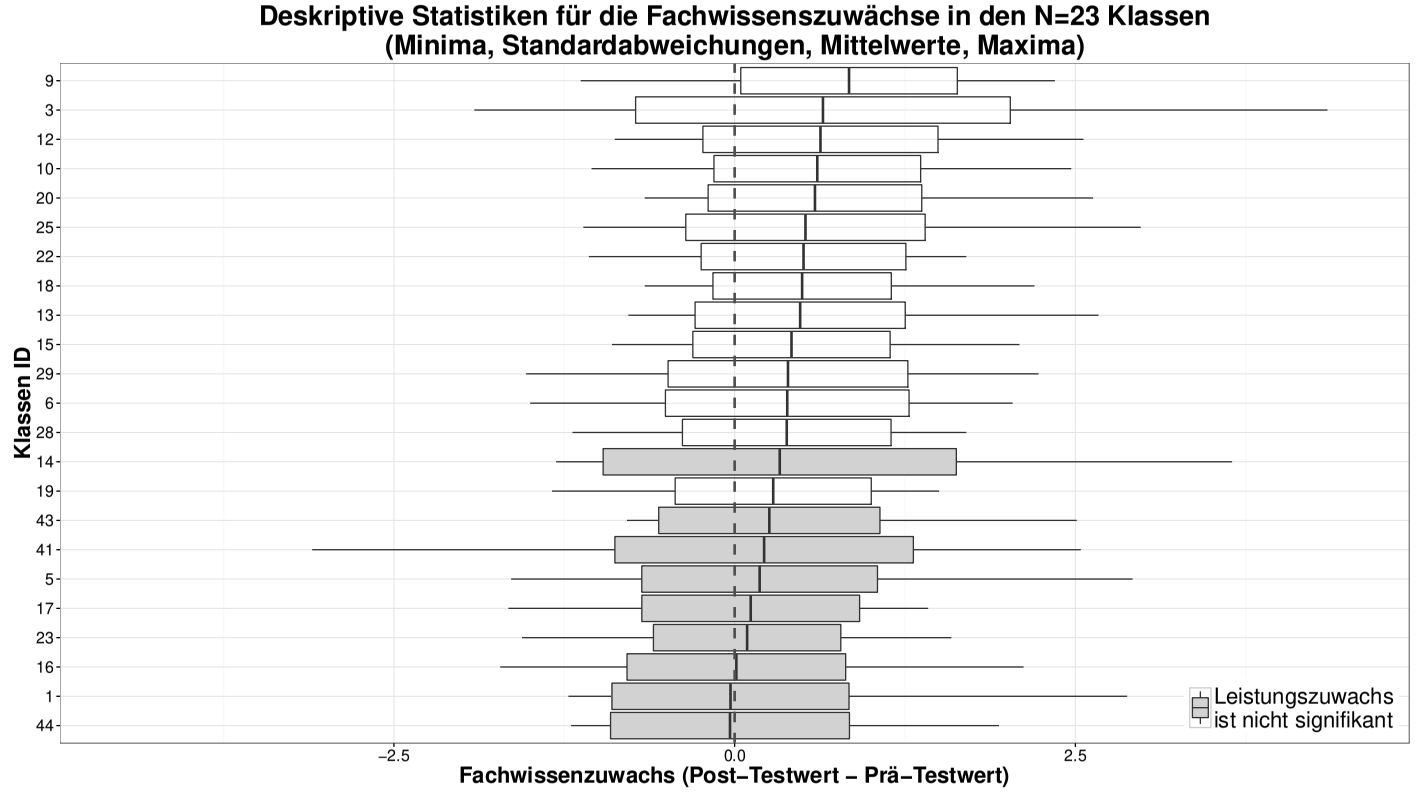



Abbildung 8.1.

*Fachwissenszuwächse in den $N = 23$ Klassen. Dargestellt ist der mittlere Fachwissenszuwachs der Lernenden einer Klasse (vertikale durchgezogene Linien), dessen Standardabweichung (Boxen) sowie die Minima und Maxima (linkes bzw. rechtes Ende der horizontalen Linien). Graue Boxen kennzeichnen Klassen in denen die Leistungszuwächse nicht signifikant größer null sind ($p_{1\text{-}seitig} > .05$). Die gestrichelte vertikale Hilfslinie zeigt einen Fachwissenszuwachs von null an.*

## 8.3. Ergebnisse der Mehrebenenanalysen

In diesem Abschnitt werden die Ergebnisse der Mehrebenenanalysen zur Beantwortung der Forschungsfragen 1 und 2.1 berichtet. Es soll geklärt werden, ob das Professionswissen der Lehrkräfte und die kognitiv aktivierende Gestaltung der videographierten Unterrichtsstunden Varianz im Fachwissen der Lernenden am Ende der Unterrichtseinheit Mechanik und im situationalen Interesse der Lernenden am Ende der videographierten Unterrichtsstunden aufklären. Für die Modelle zum Fachwissen der Lernenden wurde der Datensatz der $N = 610$ Lernenden genutzt, die am Prä- und Post-Test anwesend waren. Für die Modelle zum situationalen Interesse der Lernenden wurde der Datensatz der jeweils in der betreffenden Unterrichtsstunde anwesenden bzw. in beiden Unterrichtstunden anwesenden Lernenden genutzt ($N_{1M} = 633$, $N_{2M} = 625$, $N_{1M\&2M} = 600$). Jeweils eine Beispielsyntax aus dem für die Mehrebenenanalysen verwendeten Programm Mplus (L. K. Muthén & Muthén, 2007) für die Modelle zum Fachwissen und zum situationalen Interesse der Lernenden findet sich in Abbildung B.5 auf Seite 253 im Anhang. Die Steigungskoeffizienten für die in die Modelle aufgenommenen Prädiktoren wurden einseitig auf Signifikanz getestet, da bezüglich aller Zusammenhänge Hypothesen über deren Richtungen vorliegen. Ebenso wurden Residualvarianzen und die Varianzaufklärung $R^2$ einseitig auf Signifikanz getestet, da hier von Interesse ist, ob die entsprechenden Werte größer als null sind. An dieser Stelle sei angemerkt, dass es sich bei *allen* im Folgenden berichteten Modellen um saturierte Modelle handelt. Die in diesem Kapitel verwendete Notation für Mehrebenenmodelle ist angelehnt an die Notation von Geiser (2011, S. 203):[8]

| | |
|---|---|
| $Y_{ij}$: | Wert eines Lernenden $i$ aus Klasse $j$ auf der abhängigen Variablen auf Level 1 |
| $X_{ij}$: | Wert eines Lernenden $i$ aus Klasse $j$ auf einer Level-1-Prädiktorvariable $X$ |
| $\beta_{0j}$: | Random Intercept (Achsenabschnitt) in der Level-1-Regression für Klasse $j$ |
| $\beta_X$: | Konstanter Slope (Steigungskoeffizient) für eine Level-1-Prädiktorvariable $X$ in den Level-1-Regressionen |
| $r_{ij}$: | Residuum eines Lernenden $i$ aus Klasse $j$ in der Level-1-Regression |
| $\sigma^2_{r_{ij}}$: | Residualvarianz auf Level 1 |
| $W_j$: | Wert einer Klasse oder Wert der Lehrkraft einer Klasse $j$ auf einer Level-2-Prädiktorvariable $W$ |
| $\gamma_{00}$: | Mittelwert (Grand Mean) bzw. Intercept in der Level-2-Regression auf den Random Intercept auf Level 1 |
| $\gamma_W$: | Konstanter Slope (Steigungskoeffizient) für eine Level-2-Prädiktorvariable $W$ in der Level-2-Regression auf den Random Intercept auf Level 1 |
| $u_{0j}$: | Residuum einer Klasse $j$ in der Level-2-Regression für den Random Intercept auf Level 1 |
| $\sigma^2_{u_{0j}}$: | Residualvarianz auf Level 2 im Random Intercept auf Level 1 |

[8] Um Redundanzen zu vermeiden, tauchen die hier eingeführten Symbole – mit Ausnahme der Steigungskoeffizienten – nicht separat im Abkürzungsverzeichnis dieser Arbeit auf.

### 8.3.1. Prädiktoren für die Fachwissensleistungen im Posttest

Für eine anschauliche Interpretation wurden die Post-Testwerte der Lernenden für die Mehrebenenanalysen z-standardisiert. Zunächst wurde ein sogenanntes *Random-Intercept-Only*-Modell (Nullmodell) geschätzt, um die Varianzanteile auf Schüler- und Klassenebene und die Intraklassenkorrelation (ICC) und damit den Anteil der zwischen den Klassen liegenden Varianz an der Gesamtvarianz zu bestimmen. Dieses Modell enthält noch keine Prädiktoren. Auf Level-1 (Schülerebene) wird der Post-Testwert $Y_{ij}$ eines Lernenden $i$ in der Klasse $j$ durch den mittleren Post-Testwert $\beta_{0j}$ seiner Klasse und die Abweichung $r_{ij}$ seines Post-Testwerts vom Klassenmittelwert beschrieben:

$$\text{Level-1: } Y_{ij} = \beta_{0j} + r_{ij}. \tag{8.1}$$

Die Varianz auf Schülerebene wird über die Varianz der $r_{ij}$ beschrieben (Residualvarianz auf Schülerebene) und beträgt $\sigma^2_{r_{ij}} = 0.90 \pm 0.09$ ($\text{KI}_{95\,\%} = [0.73, 1.13]$, $p_{\text{1-seitig}} < .001$). Auf Level-2 (Klassenebene) wird der mittlere Post-Testwert ($\beta_{0j}$) einer Klasse wiederum durch den Gesamtmittelwert $\gamma_{00}$ über die Lernenden aller Schulklassen und die Abweichung $u_{0j}$ des Klassenmittelwerts vom Gesamtmittelwert beschrieben:

$$\text{Level-2: } \beta_{0j} = \gamma_{00} + u_{0j}. \tag{8.2}$$

Die Varianz auf Klassenebene wird über die Varianz der $u_{0j}$ beschrieben (Residualvarianz auf Klassenebene) und beträgt $\sigma^2_{u_{0j}} = 0.10 \pm 0.03$ ($\text{KI}_{95\,\%} = [0.04, 0.18]$, $p_{\text{1-seitig}} < .001$). Da der Post-Testwert z-standardisiert wurde, entspricht die Varianz auf Klassenebene der $ICC_{\text{1-fakt.,unjust}}$. Demnach können $(10 \pm 3)\%$ der Gesamtvarianz in den Post-Testwerten durch Prädiktoren auf Klassenebene aufgeklärt werden.

#### 8.3.1.1. Kontrollvariablenmodell (KV-Modell)

Im nächsten Schritt wurde ein *Random-Intercept-and-Means-as-Outcomes*-Modell geschätzt, in dem der durch die Kontrollvariablen (KV) erklärte Anteil der Varianz in den Post-Testwerten bestimmt wurde. Auf Schülerebene wurden der Prä-Testwert, der KFT-Testwert, das Geschlecht (0=♀,1=♂) und die zuhause gesprochene Sprache (0=nur deutsch, 1=deutsch und andere oder andere) und auf Klassenebene die Unterrichtszeit (gemessen als Anzahl der 45-Minuten-Stunden) als Prädiktoren in das Modell aufgenommen.[9] Vor den Analysen wurden die Prä-Testwerte und KFT-Testwerte auf Schülerebene und die Unterrichtszeit auf Klassenebene z-standardisiert. Die Regressionsgleichung auf Schülerebene enthält neben den zufällig zwischen den Klassen variierenden Klassenmittelwerten $\beta_{0j}$

[9] Tabellen B.6 und B.7 auf Seite 246 und auf Seite 247 im Anhang geben einen Überblick über die Korrelationen zwischen den Prädiktoren.

der Post-Testwerte für jeden Prädiktor $X$ einen konstanten Steigungskoeffizienten (Slope) $\beta_X$ und lautet wie folgt:

$$\text{Level-1: } Y_{ij} = \beta_{0j} + \beta_{\text{Prä}} \cdot X_{\text{Prä},ij} + \beta_{\text{KFT}} \cdot X_{\text{KFT},ij} + \beta_{\text{Geschl.}} \cdot X_{\text{Geschl.},ij} + \beta_{\text{Sprache}} \cdot X_{\text{Sprache},ij} + r_{ij}. \quad (8.3)$$

Die Level-2-Regressionsgleichung für die Klassenmittelwerte $\beta_{0j}$ enthält neben dem mittleren Intercept über alle Klassen $\gamma_{00}$ einen konstanten Slope $\gamma_{\text{Zeit}}$ für die Unterrichtszeit $W_{\text{Zeit}}$:

$$\text{Level-2: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + u_{0j}. \quad (8.4)$$

Die Ergebnisse für die Steigungskoeffizienten, den Intercept, die Residualvarianzen und die durch die Prädiktoren erklärte Varianz ($R^2$) auf Schüler- und Klassenebene finden sich in der Spalte „KV“ in Tabelle 8.8 auf Seite 176. Für eine anschauliche Interpretation der geschätzten Werte wird der Intercept aus der unstandardisierten Modelllösung berichtet, für kontinuierliche Prädiktoren werden vollstandardisierte Steigungskoeffizienten berichtet (StdYX) und für dichotome Prädiktoren halbstandardisierte Steigungskoeffizienten (StdY). Die vollständige Regressionsgleichung lässt sich daher nicht aus den hier aufgeführten Regressionskoeffizienten zusammensetzen. Vollständig standardisierte Steigungskoeffizienten werden in Mplus auf Schülerebene über $\beta^{\text{StdYX}} = \beta \cdot SD(X_{ij})/SD(Y_{ij})$, halbstandardisierte Koeffizienten über $\beta^{\text{StdY}} = \beta/SD(Y_{ij})$ und die vollstandardisierten Koeffizienten auf Klassenebene über $\gamma^{\text{StdYX}} = \gamma \cdot SD(W_j)/SD(\beta_{0j})$ berechnet (vergl. L. K. Muthén & Muthén, 2007, S. 577).

**Die Ergebnisse können wie folgt interpretiert werden:** Durch die z-Standardisierung der Prädiktorvariablen entspricht der mittlere Intercept über alle Klassen $\gamma_{00} = -0.07 \pm 0.06$ dem Schätzwert für den z-standardisierten Post-Testwert einer Schülerin ($X_{\text{Geschl.},ij} = 0$), die zuhause nur deutsch spricht ($X_{\text{Sprache},ij} = 0$), die sowohl im Prä-Test als auch im KFT durchschnittlich abgeschnitten hat ($X_{\text{Prä},ij} = X_{\text{KFT},ij} = 0$) und die zu einer Klasse gehört, die das Thema Mechanik über einen Zeitraum von 34 45-Minuten-Stunden (Mittelwert der Stundenanzahl in der Unterrichtseinheit Mechanik, entspricht $W_{\text{Zeit},j} = 0$) behandelt hat. Die vollstandardisierten Steigungskoeffizienten $\beta^{\text{StdYX}}_{\text{Prä}} = 0.36 \pm .04$ und $\beta^{\text{StdYX}}_{\text{KFT}} = 0.23 \pm 0.05$ geben an, um wie viele Standardabweichungen sich der geschätzte Post-Testwerts eines Lernenden erhöht, wenn er im Prä-Test oder im KFT um eine Standardabweichung besser als der Durchschnitt abgeschnitten hat. Die halbstandardisierten Steigungskoeffizienten $\beta^{\text{StdY}}_{\text{Geschl.}} = 0.34 \pm .07$ und $\beta^{\text{StdY}}_{\text{Sprache}} = -0.35 \pm 0.09$ geben an, um wie viele Standardabweichungen in den geschätzten Post-Testwerten Jungen besser abschneiden als Mädchen und Lernende, die zuhause nicht oder nicht nur deutsch sprechen, schlechter abschneiden als Lernende, die zuhause nur deutsch sprechen. Der vollstandardisierte Steigungskoeffizient für die Unterrichtszeit $\gamma^{\text{StdYX}}_{\text{Zeit}} = 0.80 \pm 0.12$ gibt an, um wie viele Standardabweichungen im geschätzten Klassenmittelwert für die Post-Testwerte

Lernende besser abschneiden, deren Klassen zehn 45-Minuten-Stunden (Standardabweichung der Stundenanzahl in der Unterrichtseinheit Mechanik) mehr Unterricht in Mechanik erhalten haben als der Durchschnitt.

Auf Schülerebene werden $R^2 = (34 \pm 3)\%$ der Varianz in den Post-Testwerten durch die Kontrollvariablen aufgeklärt. Der Anteil der zwischen den Klassen liegenden Varianz sinkt auf $ICC_{\text{1-fakt.,unjust}} = .04$. Die Unterrichtszeit erklärt $R^2 = (65 \pm 18)\%$ der Varianz zwischen den Klassen. Die Residualvarianzen sind dennoch sowohl auf Schülerebene als auch auf Klassenebene signifikant von Null verschieden.

#### 8.3.1.2. Professionswissensmodelle (Modelle 1a-c)

In diesem Abschnitt sollen die Hypothesen H1a-c für die Beantwortung der Forschungsfrage 1 überprüft werden:

> H1a-c:
> *Unterschiede in den Fachwissensleistungen der Lernenden werden (nach Kontrolle des Vorwissens, der kognitiven Fähigkeiten, des Geschlechts, der zuhause gesprochenen Sprache und der Unterrichtszeit) durch Unterschiede im a) CK, b) PCK und c) PK der Lehrkräfte erklärt. Höhere Testergebnisse der Lehrkräfte im CK, PCK bzw. PK hängen mit höheren Fachwissensleistungen der Lernenden zusammen.*

Für die Überprüfung der Hypothesen wurden drei *Random-Intercept-and-Means-as-Outcomes*-Modelle geschätzt, die im Vergleich zum KV-Modell jeweils einen zusätzlichen Prädiktor $W$ auf Klassenebene enthielten: den z-standardisierten CK-Testwert (Modell 1a), PCK-Testwert (Modell 1b) oder PK-Testwert (Modell 1c) der Lehrkräfte.[10] Die Regressionsgleichungen auf Klassenebene lauten also:

Level-2:

$$\text{1a: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{CK}} \cdot W_{\text{CK},j} + u_{0j}, \tag{8.5}$$

$$\text{1b: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{PCK}} \cdot W_{\text{PCK},j} + u_{0j}, \tag{8.6}$$

$$\text{1c: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{PK}} \cdot W_{\text{PK},j} + u_{0j}. \tag{8.7}$$

Die Ergebnisse für die Steigungskoeffizienten, den Intercept, die Residualvarianzen und die durch die Prädiktoren erklärte Varianz ($R^2$) auf Schüler- und Klassenebene finden sich in der Spalte „Professionswissensmodelle“ in Tabelle 8.8 auf Seite 176.

Lediglich im Modell 1c wird im Vergleich zum KV-Modell zusätzliche Varianz auf Klassenebene aufgeklärt. Die Unterrichtszeit und das PK der Lehrkräfte erklären $R^2 = (76 \pm 17)\%$ der Varianz in den Klassenmittelwerten der Post-Testwerte. Der vollstandardisierte Steigungskoeffizient $\gamma_{\text{PK}}^{\text{StdYX}} = 0.34 \pm 0.18$ gibt an, um wie viele Standardabweichungen im geschätzten Klassenmittelwert der Post-Testwerte

[10] Tabelle B.8 auf Seite 247 im Anhang gibt einen Überblick über die Korrelationen zwischen der Unterrichtszeit und den zusätzlichen Prädiktoren auf Klassenebene.

Lernende besser abschneiden, die von einer Lehrkraft unterrichtet werden, deren PK um eine Standardabweichung vom Durchschnitt abweicht. Der Effekt ist in etwa halb so groß wie der Effekt durch die Unterrichtszeit ($\gamma_{\text{Zeit}}^{\text{StdYX}} = 0.75 \pm 0.13$). Der Koeffizient $\gamma_{\text{Zeit}}^{\text{StdYX}}$ ist im Modell 1c etwas kleiner als im KV-Modell. Die Unterschiede sind im Rahmen der Fehlerabschätzung aber nicht von Bedeutung. Die Korrelation zwischen der Unterrichtszeit und dem PK-Testwert ist nicht signifikant ($N = 23$, $r_{\text{Pearson}} = .18 \pm .21$, $\text{KI}_{95\,\%} = [-.25, .53]$, $p = .425$).

Der Vorteil, den Lernende haben, deren Lehrkraft im PK-Test eine Standardabweichung besser abgeschnitten hat als der Durchschnitt, ist also vergleichbar mit einem Lernvorsprung von ca. fünf 45-Minuten-Stunden Mechanikunterricht (entspricht einer halben Standardabweichung in der Unterrichtszeit).[11]

| | |
|---|---|
| H1a (CK ↔ Schülerfachwissen) | abgelehnt |
| H1b (PCK ↔ Schülerfachwissen) | abgelehnt |
| H1c (PK ↔ Schülerfachwissen) | angenommen |

#### 8.3.1.3. Modelle zur kognitiven Aktivierung (Modelle 2.1a$_{\text{1M/2M/1M\&2M}}$)

In diesem Abschnitt soll die Hypothese H2.1a für die Beantwortung der Forschungsfrage 2.1 überprüft werden:

> H2.1a:
> *Unterschiede in den Fachwissensleistungen der Lernenden werden (nach Kontrolle des Vorwissens, der kognitiven Fähigkeiten, des Geschlechts, der zuhause gesprochenen Sprache und der Unterrichtszeit) durch Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts erklärt. Höhere Ausprägungen in der kognitiven Aktivierung hängen mit höheren Fachwissensleistungen der Lernenden zusammen.*

Für die Überprüfung der Hypothese wurden drei *Random-Intercept-and-Means-as-Outcomes*-Modelle geschätzt, die im Vergleich zum KV-Modell jeweils einen zusätzlichen Prädiktor $\boldsymbol{W}$ auf Klassenebene enthielten: das z-standardisierte Qualitätsmaß für die kognitiv aktivierende Gestaltung der ersten Unterrichtsstunde (1M), der zweiten Unterrichtsstunde (2M) oder das über beide Unterrichtsstunden

[11] Dieser Vergleich dient lediglich dazu, die Größenordnung des PK-Effekts grob einschätzen zu können und sollte nicht überinterpretiert werden. Alle in die Mehrebenenanalysen einbezogenen Messwerte sind fehlerbehaftet. Es erfolgt weder eine Fortpflanzung der Messunsicherheiten, noch können systematische Fehler bei der Messung aller Variablen abgeschätzt werden.

gemittelte Qualitätsmaß (1M&2M).[12] Die Regressionsgleichungen auf Klassenebene lauten also:

Level-2:

$$2.1a_{1M}\text{: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{KA1M}} \cdot W_{\text{KA1M},j} + u_{0j}, \tag{8.8}$$

$$2.1a_{2M}\text{: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{KA2M}} \cdot W_{\text{KA2M},j} + u_{0j}, \tag{8.9}$$

$$2.1a_{1M\&2M}\text{: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{KA1M\&2M}} \cdot W_{\text{KA1M\&2M},j} + u_{0j}. \tag{8.10}$$

Die Ergebnisse für die Steigungskoeffizienten, den Intercept, die Residualvarianzen und die durch die Prädiktoren erklärte Varianz ($R^2$) auf Schüler- und Klassenebene finden sich in der Spalte „Modelle zur kognitiven Aktivierung“ in Tabelle 8.8 auf der nächsten Seite. In allen drei Modellen wird im Vergleich zum KV-Modell zusätzliche Varianz auf Klassenebene aufgeklärt. Die Qualitätsmaße für die kognitiv aktivierende Gestaltung der ersten bzw. zweiten Unterrichtsstunde sind signifikante Prädiktoren für die Post-Testwerte der Lernenden – zusammen mit der Unterrichtszeit klären sie $R^2 = (79 \pm 15)\%$ (Modell $2.1a_{1M}$) bzw. $R^2 = (80 \pm 20)\%$ (Modell $2.1a_{2M}$) der Varianz in den Klassenmittelwerten der Post-Testwerte auf. Die größte Varianzaufklärung kann durch Modell $2.1a_{1M\&2M}$ realisiert werden ($R^2 = (85 \pm 17)\%$), weshalb dieses Modell hier näher beschrieben werden soll.

Der vollstandardisierte Steigungskoeffizient $\gamma^{\text{StdYX}}_{\text{KA1M\&2M}} = 0.46 \pm 0.20$ gibt an, um wie viele Standardabweichungen im geschätzten Klassenmittelwert der Post-Testwerte Lernende besser abschneiden, deren Unterricht im Qualitätsmaß zur kognitiv aktivierenden Gestaltung um eine Standardabweichung besser bewertet wurde als der Durchschnitt. Der Effekt ist etwas mehr als halb so groß wie der Effekt durch die Unterrichtszeit ($\gamma^{\text{StdYX}}_{\text{Zeit}} = 0.74 \pm 0.12$). Der Koeffizient $\gamma^{\text{StdYX}}_{\text{Zeit}}$ ist im Modell $2.1a_{1M\&2M}$ etwas kleiner als im KV-Modell. Die Unterschiede sind im Rahmen der Fehlerabschätzung aber nicht von Bedeutung. Die Korrelation zwischen der Unterrichtszeit und dem über beide Unterrichtsstunden gemittelten Qualitätsmaß für die kognitive Aktivierung ist nicht signifikant ($N = 23$, $r_{\text{Pearson}} = .13 \pm .16$, $\text{KI}_{95\,\%} = [-.16, .44]$, $p = .568$).

Der Vorteil, den Lernende haben, deren Unterricht im Qualitätsmaß zur kognitiven Aktivierung um eine Standardabweichung besser bewertet wurde als der Durchschnitt, ist vergleichbar mit einem Lernvorsprung von ca. sechs 45-Minuten-Stunden Mechanikunterricht (entspricht 0.6 Standardabweichungen in der Unterrichtszeit).[13]

| H2.1a (KA ↔ Schülerfachwissen) angenommen |
|---|

[12] Tabelle B.8 auf Seite 247 im Anhang gibt einen Überblick über die Korrelationen zwischen der Unterrichtszeit und den zusätzlichen Prädiktoren auf Klassenebene.

[13] Auch dieser Vergleich dient lediglich dazu, die Größenordnung des Effekts der kognitiven Aktivierung grob einschätzen zu können und sollte nicht überinterpretiert werden – alle in die Mehrebenenanalysen einbezogenen Messwerte sind fehlerbehaftet. Es erfolgt weder eine Fortpflanzung der Messunsicherheiten, noch können systematische Fehler bei der Messung aller Variablen abgeschätzt werden.

Tabelle 8.8.
*Ergebnisse der Mehrebenenregressionen auf die Post-Testwerte der Lernenden im Fachwissen. Im Vergleich zum Kontrollvariablenmodell (KV) enthalten die Modelle 1a-c als zusätzlichen Prädiktor $\boldsymbol{W}$ auf Klassenebene das CK, PCK bzw. PK der Lehrkräfte und die Modelle 2.1a$_{1M/2M/1M\&2M}$ das Qualitätsmaß für die kognitiv aktivierende Gestaltung der 1./2. Unterrichtsstunde bzw. das über beide Unterrichtsstunden gemittelte Qualitätsmaß*

| | | | | Professionswissensmodelle | | | Modelle zur kognitiven Aktivierung | | |
|---|---|---|---|---|---|---|---|---|---|
| | Modellbezeichnung | | KV | 1a | 1b | 1c | 2.1a$_{1M}$ | 2.1a$_{2M}$ | 2.1a$_{1M\&2M}$ |
| | Erweiterung von KV auf Level-2 | | | ($\boldsymbol{W} = \boldsymbol{W}_{CK}$) | ($\boldsymbol{W} = \boldsymbol{W}_{PCK}$) | ($\boldsymbol{W} = \boldsymbol{W}_{PK}$) | ($\boldsymbol{W} = \boldsymbol{W}_{KA1M}$) | ($\boldsymbol{W} = \boldsymbol{W}_{KA2M}$) | ($\boldsymbol{W} = \boldsymbol{W}_{KA1M\&2M}$) |
| Schülerebene ($N = 610$) | Prä-Test | $\beta^{StdYX}_{Prä}$ | **0.36 ± 0.04** | **0.36 ± 0.04** | **0.36 ± 0.04** | **0.35 ± 0.04** | **0.36 ± 0.04** | **0.36 ± 0.04** | **0.36 ± 0.04** |
| | | $KI_{95\,\%}$ | [0.28, 0.44] | [0.28, 0.44] | [0.28, 0.44] | [0.28, 0.43] | [0.28, 0.44] | [0.29, 0.44] | [0.29, 0.44] |
| | KFT | $\beta^{StdYX}_{KFT}$ | **0.23 ± 0.05** | **0.23 ± 0.05** | **0.23 ± 0.05** | **0.23 ± 0.04** | **0.23 ± 0.05** | **0.24 ± 0.05** | **0.24 ± 0.04** |
| | | $KI_{95\,\%}$ | [0.15, 0.32] | [0.15, 0.32] | [0.15, 0.31] | [0.16, 0.31] | [0.15, 0.31] | [0.16, 0.32] | [0.16, 0.32] |
| | Geschlecht | $\beta^{StdY}_{Geschl.}$ | **0.34 ± 0.07** | **0.34 ± 0.07** | **0.34 ± 0.07** | **0.33 ± 0.07** | **0.33 ± 0.07** | **0.34 ± 0.07** | **0.33 ± 0.07** |
| | (0 =♀) | $KI_{95\,\%}$ | [0.20, 0.47] | [0.20, 0.47] | [0.20, 0.47] | [0.19, 0.46] | [0.20, 0.47] | [0.20, 0.47] | [0.20, 0.47] |
| | Sprache | $\beta^{StdY}_{Sprache}$ | **−0.35 ± 0.09** | **−0.35 ± 0.08** | **−0.35 ± 0.09** | **−0.36 ± 0.09** | **−0.35 ± 0.08** | **−0.34 ± 0.09** | **−0.34 ± 0.08** |
| | (0 =deutsch) | $KI_{95\,\%}$ | [−0.51, −0.19] | [−0.51, −0.19] | [−0.51, −0.19] | [−0.52, −0.20] | [−0.51, −0.19] | [−0.50, −0.18] | [−0.50, −0.18] |
| | Residual- | $\sigma^2_{r_{ij}}$ | **0.64 ± 0.07** | **0.64 ± 0.07** | **0.64 ± 0.07** | **0.64 ± 0.07** | **0.64 ± 0.07** | **0.64 ± 0.07** | **0.64 ± 0.07** |
| | varianz | $KI_{95\,\%}$ | [0.52, 0.76] | [0.52, 0.76] | [0.52, 0.76] | [0.52, 0.76] | [0.52, 0.76] | [0.52, 0.76] | [0.52, 0.76] |
| | Varianz- | $R^2$ | **0.34 ± 0.03** | **0.34 ± 0.03** | **0.33 ± 0.03** | **0.330 ± 0.029** | **0.335 ± 0.029** | **0.341 ± 0.029** | **0.339 ± 0.029** |
| | aufklärung | | | | | | | | |
| Klassenebene ($N = 23$) | Intercept | $\gamma_{00}$ | −0.07 ± 0.06 | −0.07 ± 0.06 | −0.07 ± 0.06 | −0.06 ± 0.06 | −0.07 ± 0.06 | −0.07 ± 0.05 | −0.07 ± 0.06 |
| | | $KI_{95\,\%}$ | [−0.17, 0.04] | [−0.18, 0.04] | [−0.18, 0.04] | [−0.16, 0.04] | [−0.17, 0.04] | [−0.17, 0.03] | [−0.17, 0.04] |
| | Unterrichts- | $\gamma^{StdYX}_{Zeit}$ | **0.80 ± 0.11** | **0.81 ± 0.11** | **0.81 ± 0.11** | **0.75 ± 0.13** | **0.69 ± 0.14** | **0.85 ± 0.12** | **0.74 ± 0.12** |
| | zeit | $KI_{95\,\%}$ | [0.60, 1.01] | [0.61, 1.00] | [0.60, 1.02] | [0.51, 0.98] | [0.43, 0.95] | [0.63, 1.07] | [0.52, 0.97] |
| | | $p_{1\text{-seitig}}$ | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 |
| | $\boldsymbol{W}$ | $\gamma^{StdYX}_{W}$ | | 0.07 ± 0.19 | −0.15 ± 0.16 | **0.34 ± 0.18** | **0.40 ± 0.22** | **0.40 ± 0.16** | **0.46 ± 0.20** |
| | | $KI_{95\,\%}$ | | [−0.30, 0.44] | [−0.45, 0.15] | [0.00, 0.68] | [−0.03, 0.83] | [0.10, 0.69] | [0.07, 0.84] |
| | | $p_{1\text{-seitig}}$ | | .363 | .158 | .024 | .036 | .005 | .010 |
| | Residual- | $\sigma^2_{u_{0j}}$ | **0.016 ± 0.007** | **0.015 ± 0.007** | **0.014 ± 0.008** | 0.011 ± 0.007 | 0.009 ± 0.008 | 0.009 ± 0.008 | 0.007 ± 0.008 |
| | varianz | $KI_{95\,\%}$ | [0.001, 0.030] | [0.001, 0.030] | [−0.001, 0.030] | [−0.002, 0.024] | [−0.005, 0.024] | [−0.007, 0.025] | [−0.009, 0.022] |
| | | $p_{1\text{-seitig}}$ | .018 | .020 | .032 | .052 | .109 | .128 | .193 |
| | Varianz- | $R^2$ | **0.65 ± 0.18** | **0.65 ± 0.16** | **0.67 ± 0.18** | **0.76 ± 0.17** | **0.79 ± 0.15** | **0.80 ± 0.20** | **0.85 ± 0.17** |
| | aufklärung | $p_{1\text{-seitig}}$ | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 |

*Legende:* StdYX:= indiziert vollstandardisierte Steigungskoeffizienten; StdY:= indiziert halbstandardisierte Steigungskoeffizienten
*Anmerkungen.* Signifikante Werte mit $p_{1\text{-seitig}} < .05$ sind fett gedruckt. Alle Steigungskoeffizienten, Residualvarianzen und $R^2$ auf Schülerebene sind signifikant mit $p_{1\text{-seitig}} < .001$.

### 8.3.2. Prädiktoren für das situationale Interesse der Lernenden

Für eine anschauliche Interpretation wurden auch die Maße für das situationale Interesse der Lernenden in der ersten (1M) und zweiten Unterrichtsstunde (2M) und das über beide Unterrichtsstunden gemittelte Maß (1M&2M) für die Mehrebenenanalysen z-standardisiert. Zunächst wurden *Random-Intercept-Only*-Modelle (Nullmodelle) geschätzt, um die Residualvarianzen auf Schüler- und Klassenebene und die ICCs und damit den Anteil der zwischen den Klassen liegenden Varianz an der Gesamtvarianz zu bestimmen. Die Modelle enthalten noch keine Prädiktoren. Auf Level-1 (Schülerebene) wird das situationale Interesse $Y_{ij}$ eines Lernenden $i$ in der Klasse $j$ durch das mittlere situationale Interesse $\beta_{0j}$ seiner Klasse und die Abweichung $r_{ij}$ seines situationalen Interesses vom Klassenmittelwert beschrieben:

Level-1:

$$\text{1M: } Y_{ij,1\text{M}} = \beta_{0j,1\text{M}} + r_{ij,1\text{M}}, \tag{8.11}$$

$$\text{2M: } Y_{ij,2\text{M}} = \beta_{0j,2\text{M}} + r_{ij,2\text{M}}, \tag{8.12}$$

$$\text{1M\&2M: } Y_{ij,1\text{M\&2M}} = \beta_{0j,1\text{M\&2M}} + r_{ij,1\text{M\&2M}}. \tag{8.13}$$

Die Varianz auf Schülerebene wird über die Varianz der $r_{ij}$ beschrieben (Residualvarianz auf Schülerebene). Auf Level-2 (Klassenebene) wird das mittlere situationale Interesse ($\beta_{0j}$) in einer Klasse wiederum durch den Gesamtmittelwert $\gamma_{00}$ über die Lernenden aller Schulklassen und die Abweichung $u_{0j}$ des Klassenmittelwerts vom Gesamtmittelwert beschrieben:

Level-2:

$$\text{1M: } \beta_{0j,1\text{M}} = \gamma_{00,1\text{M}} + u_{0j,1\text{M}}, \tag{8.14}$$

$$\text{2M: } \beta_{0j,2\text{M}} = \gamma_{00,2\text{M}} + u_{0j,2\text{M}}, \tag{8.15}$$

$$\text{1M\&2M: } \beta_{0j,1\text{M\&2M}} = \gamma_{00,1\text{M\&2M}} + u_{0j,1\text{M\&2M}}. \tag{8.16}$$

Die Varianz auf Klassenebene wird über die Varianz der $u_{0j}$ beschrieben (Residualvarianz auf Klassenebene). Die Residualvarianzen sind in Tabelle 8.9 auf der nächsten Seite aufgeführt. Da das situationale Interesse z-standardisiert wurde, entspricht die Varianz auf Klassenebene der $ICC_{\text{1-fakt.,unjust}}$. Demnach können $(17 \pm 4)$% bzw. $(19 \pm 5)$% der Gesamtvarianz im situationalen Interesse der Lernenden in der ersten und zweiten Unterrichtsstunde und $(20 \pm 5)$% der Gesamtvarianz im über beide Unterrichtsstunden gemittelten situationalen Interesse der Lernenden durch Prädiktoren auf Klassenebene aufgeklärt werden. In die Modelle zum situationalen Interesse der Lernenden wurden keine Kontrollvariablen als Prädiktoren einbezogen.

#### 8.3.2.1. Professionswissensmodelle (Modelle 1d-f)

In diesem Abschnitt sollen die Hypothesen H1d-f für die Beantwortung der Forschungsfrage 1 überprüft werden:

Tabelle 8.9.
*Residualvarianzen in den Nullmodellen für das situationale Interesse der Lernenden in der 1. und 2. Unterrichtsstunde Mechanik und für das über beide Unterrichtsstunden gemittelte situationale Interesse*

| | | | 1M | 2M | 1M&2M |
|---|---|---|---|---|---|
| Schüler | Stichprobe | $N$ | 633 | 625 | 600 |
| | Residualvarianz | $\sigma^2_{r_{ij}}$ | $\mathbf{0.83 \pm 0.07}$ | $\mathbf{0.80 \pm 0.06}$ | $\mathbf{0.80 \pm 0.07}$ |
| | | $\mathrm{KI}_{95\,\%}$ | $[0.70, 0.96]$ | $[0.69, 0.92]$ | $[0.66, 0.94]$ |
| Klassen | Stichprobe | $N$ | 23 | 23 | 23 |
| | Residualvarianz | $\sigma^2_{u_{0j}}$ | $\mathbf{0.17 \pm 0.04}$ | $\mathbf{0.19 \pm 0.05}$ | $\mathbf{0.20 \pm 0.05}$ |
| | | $\mathrm{KI}_{95\,\%}$ | $[0.10, 0.24]$ | $[0.10, 0.29]$ | $[0.10, 0.29]$ |

*Legende:* StdYX:= indiziert vollstandardisierte Steigungskoeffizienten;
*Anmerkungen.* Signifikante Werte mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Alle Residualvarianzen sind signifikant größer als null mit $p_{\text{1-seitig}} < .001$.

H1d-f:
*Unterschiede im situationalen Interesse der Lernenden werden durch Unterschiede im d) CK, e) PCK und f) PK der Lehrkräfte erklärt. Höhere Testergebnisse im CK, PCK bzw. PK hängen mit höheren Ausprägungen des situationalen Interesses der Lernenden im Unterricht zusammen.*

Da die beiden videographierten Unterrichtsstunden zusammengenommen den Unterricht besser repräsentieren als die Einzelstunden (vergl. Abschnitt 7.6.8 auf Seite 146), wurden für die Überprüfung der Hypothesen *Random-Intercept-and-Means-as-Outcomes*-Modelle für das über beide Unterrichtsstunden gemittelte situationale Interesse geschätzt, die als Prädiktor $W$ auf Klassenebene den z-standardisierten CK-Testwert (Modell 1d), PCK-Testwert (Modell 1e) oder PK-Testwert (Modell 1f) der Lehrkräfte enthielten. Die Regressionsgleichung auf Schülerebene wird auch in diesen Modellen durch Gleichung (8.13) auf Seite 177 beschrieben und enthält lediglich den zufällig zwischen den Klassen variierenden Klassenmittelwert $\beta_{0j,\text{1M\&2M}}$ für das situationale Interesse der Lernenden. Die Level-2-Regressionsgleichung für die Klassenmittelwerte $\beta_{0j}$ lauten wie folgt:

Level-2:

$$\text{1d: } \beta_{0j,\text{1M\&2M}} = \gamma_{00,\text{1M\&2M}} + \gamma_{\text{CK}} \cdot W_{\text{CK},j} + u_{0j,\text{1M\&2M}}, \tag{8.17}$$

$$\text{1e: } \beta_{0j,\text{1M\&2M}} = \gamma_{00,\text{1M\&2M}} + \gamma_{\text{PCK}} \cdot W_{\text{PCK},j} + u_{0j,\text{1M\&2M}}, \tag{8.18}$$

$$\text{1f: } \beta_{0j,\text{1M\&2M}} = \gamma_{00,\text{1M\&2M}} + \gamma_{\text{PK}} \cdot W_{\text{PK},j} + u_{0j,\text{1M\&2M}}. \tag{8.19}$$

Die Ergebnisse für die Residualvarianz auf Schülerebene und die vollstandardisierten Steigungskoeffizienten, den Intercept, die Residualvarianzen und die durch den jeweiligen Prädiktor erklärte Varianz ($R^2$) auf Klassenebene finden sich in der Spalte „Professionswissensmodelle“ in Tabelle 8.10 auf der nächsten Seite. Auch hier werden für eine anschauliche Interpretation der geschätzten Werte der

Tabelle 8.10.
*Ergebnisse der Mehrebenenregressionen auf das situationale Interesse der Lernenden in der 1. und 2. Unterrichtsstunde Mechanik. Die Modelle 1d-f (Professionswissensmodelle) enthalten auf Klassenebene als Prädiktor $W$ das CK, PCK bzw. PK der Lehrkräfte, die Modelle 2.1b$_{1M}$ und 2.1b$_{2M}$ (Modelle zur kognitiven Aktivierung) das Qualitätsmaß für die kognitiv aktivierende Gestaltung der 1. bzw. 2. Unterrichtsstunde*

| Modellbezeichnung | | | Professionswissensmodelle 1d ($W = W_{CK}$) | 1e ($W = W_{PCK}$) | 1f ($W = W_{PK}$) | Modelle zur kognitiven Aktivierung 2.1b$_{1M}$ ($W = W_{KA1M}$) | 2.1b$_{2M}$ ($W = W_{KA2M}$) |
|---|---|---|---|---|---|---|---|
| Schülerebene | Stichprobe | $N$ | 600 | 600 | 600 | 633 | 625 |
| | Residual-varianz | $\sigma^2_{r_{ij}}$ | **0.80 ± 0.07** | **0.80 ± 0.07** | **0.80 ± 0.07** | **0.83 ± 0.07** | **0.80 ± 0.06** |
| | | KI$_{95\,\%}$ | [0.66, 0.94] | [0.66, 0.94] | [0.66, 0.94] | [0.70, 0.96] | [0.69, 0.92] |
| | | $p_{1\text{-seitig}}$ | < .001 | < .001 | < .001 | < .001 | < .001 |
| Klassenebene ($N = 23$) | Intercept | $\gamma_{00}$ | 0.00 ± 0.10 | 0.00 ± 0.10 | 0.00 ± 0.10 | 0.01 ± 0.09 | −0.01 ± 0.10 |
| | | KI$_{95\,\%}$ | [−0.20, 0.19] | [−0.20, 0.19] | [−0.19, 0.19] | [−0.16, 0.18] | [−0.20, 0.18] |
| | $W$ | $\gamma^{StdYX}_{W}$ | −0.13 ± 0.21 | 0.06 ± 0.21 | 0.20 ± 0.21 | **0.35 ± 0.19** | 0.15 ± 0.23 |
| | | KI$_{95\,\%}$ | [−0.53, 0.27] | [−0.34, 0.46] | [−0.21, 0.61] | [−0.01, 0.71] | [−0.28, 0.59] |
| | | $p_{1\text{-seitig}}$ | .261 | .383 | .166 | .028 | .246 |
| | Residual-varianz | $\sigma^2_{u_{0j}}$ | **0.19 ± 0.05** | **0.20 ± 0.05** | **0.19 ± 0.05** | **0.15 ± 0.04** | **0.19 ± 0.05** |
| | | KI$_{95\,\%}$ | [0.10, 0.29] | [0.10, 0.29] | [0.10, 0.27] | [0.08, 0.22] | [0.11, 0.27] |
| | | $p_{1\text{-seitig}}$ | < .001 | < .001 | < .001 | < .001 | < .001 |
| | Varianz-aufklärung | $R^2$ | 0.02 ± 0.06 | 0.004 ± 0.025 | 0.04 ± 0.09 | 0.12 ± 0.13 | 0.02 ± 0.07 |
| | | $p_{1\text{-seitig}}$ | .375 | .441 | .314 | .170 | .366 |

*Anmerkung.* Signifikante Werte mit $p_{1\text{-seitig}} < .05$ sind fett gedruckt.

Intercept aus der unstandardisierten Modelllösung und die vollstandardisierten Steigungskoeffizienten (StdYX) für die kontinuierlichen Prädiktoren berichtet. Die vollständige Regressionsgleichung lässt sich daher nicht aus den hier aufgeführten Regressionskoeffizienten zusammensetzen. Vollständig standardisierte Steigungskoeffizienten auf Klassenebene werden in Mplus über $\gamma_{\text{StdYX}} = \gamma \cdot SD(W_j)/SD(\beta_{0j})$ (vergl. L. K. Muthén & Muthén, 2007, S. 577) berechnet.

Der Intercept gibt den geschätzten Klassenmittelwert des z-standardisierten situationalen Interesses der Lernenden für eine Klasse an, die von einer Lehrkraft unterrichtet wurde, die über ein durchschnittliches CK, PCK oder PK verfügt. Die vollstandardisierten Steigungskoeffizienten geben an, um wie viele Standardabweichungen im geschätzten Klassenmittelwert des situationalen Interesses Lernende besser abschneiden, die von einer Lehrkraft unterrichtet werden, deren CK, PCK oder PK um eine Standardabweichung vom Durchschnitt abweicht.

In keinem der Modelle wird ein signifikanter Anteil der Varianz auf Klassenebene aufgeklärt.

| | |
|---|---|
| H1d (CK ↔ Situationales Interesse der Lernenden) | abgelehnt |
| H1e (PCK ↔ Situationales Interesse der Lernenden) | abgelehnt |
| H1f (PK ↔ Situationales Interesse der Lernenden) | abgelehnt |

#### 8.3.2.2. Modelle zur kognitiven Aktivierung (Modelle 2.1b$_{\text{1M/2M}}$)

In diesem Abschnitt soll die Hypothese H2.1b für die Beantwortung der Forschungsfrage 2.1 überprüft werden:

H2.1b
*Unterschiede im situationalen Interesse der Lernenden werden durch Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts erklärt. Höhere Ausprägungen in der kognitiven Aktivierung hängen mit höheren Ausprägungen des situationalen Interesses der Lernenden im Unterricht zusammen.*

Da das situationale Interesse der Lernenden am Ende der videographierten Unterrichtsstunden erhoben wurde und sich konkret auf den Unterricht in der jeweiligen Stunde bezieht, wurde die Hypothese 2.1b getrennt für beide Unterrichtsstunden überprüft. Hierfür wurde jeweils ein *Random-Intercept-and-Means-as-Outcomes*-Modell für das situationale Interesse der Lernenden in der ersten (1M) bzw. zweiten (2M) Unterrichtsstunde geschätzt, das als Prädiktor $\boldsymbol{W}$ auf Klassenebene das z-standardisierte Qualitätsmaß für die kognitiv aktivierende Gestaltung der ersten bzw. zweiten Unterrichtsstunde enthielt. Die Regressions-

gleichungen auf Schülerebene werden durch Gleichungen (8.11) und (8.12) auf Seite 177 beschrieben. Die Regressionsgleichungen auf Klassenebene lauten:

Level-2:

$$2.1b_{1M}\text{: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{KA1M}} \cdot W_{\text{KA1M},j} + u_{0j}, \tag{8.20}$$

$$2.1b_{2M}\text{: } \beta_{0j} = \gamma_{00} + \gamma_{\text{Zeit}} \cdot W_{\text{Zeit},j} + \gamma_{\text{KA2M}} \cdot W_{\text{KA2M},j} + u_{0j}. \tag{8.21}$$

Die Ergebnisse für die Residualvarianz auf Schülerebene und die vollstandardisierten Steigungskoeffizienten, den Intercept, die Residualvarianzen und die durch den jeweiligen Prädiktor erklärte Varianz ($R^2$) auf Klassenebene finden sich in der Spalte „Modelle zur kognitiven Aktivierung“ in Tabelle 8.10 auf Seite 179.

Weder in der ersten Unterrichtsstunde noch in der zweiten Unterrichtsstunde erklärt das Maß für die kognitiv aktivierende Gestaltung des Unterrichts einen signifikanten Anteil der Varianz im situationalen Interesse der Lernenden auf Klassenebene. Zwar wird der vollstandardisierte Steigungskoeffizient $\gamma_{\text{KA1M}} = 0.35 \pm 0.19$ signifikant ($p_{\text{1-seitig}} = .028$), die Varianzaufklärung ist allerdings auch in diesem Fall nicht signifikant von 0 verschieden ($p_{\text{1-seitig}} = .169$).

| H2.1b (KA ↔ Situationales Interesse der Lernenden) abgelehnt |
|---|

## 8.4. Ergebnisse zum Zusammenhang zwischen Professionswissen und kognitiv aktivierend gestaltetem Unterricht

In diesem Abschnitt sollen die Hypothesen H2.2a-e für die Beantwortung der Forschungsfrage 2.2 überprüft werden:

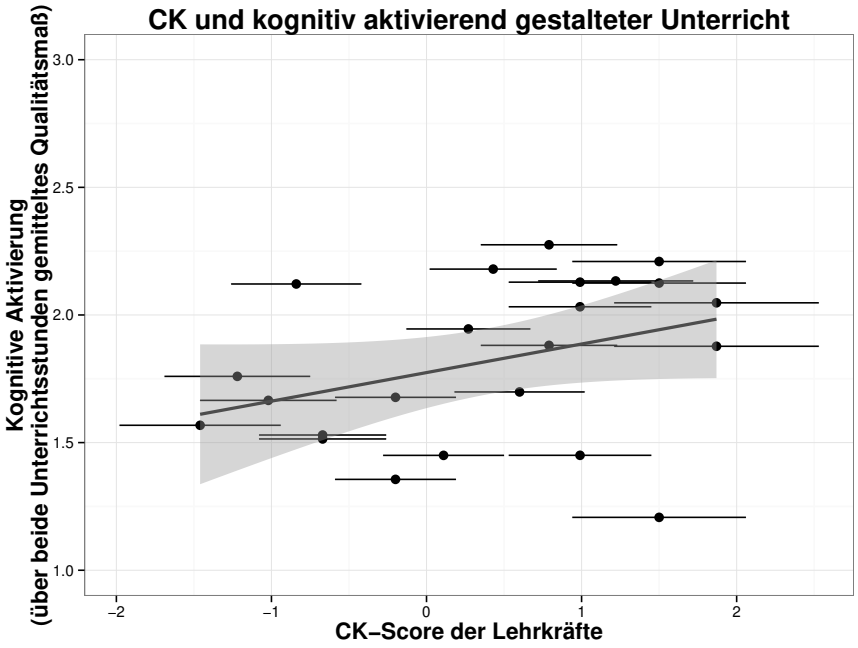
| | |
|---|---|
| H2.2a-b: | *Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts werden durch Unterschiede im a) CK und b) PCK der Lehrkräfte erklärt. Höhere Ausprägungen im CK bzw. PCK hängen mit höheren Ausprägungen in der kognitiven Aktivierung zusammen.* |
| H2.2c: | *Kognitive Aktivierung hängt stärker mit PCK als mit CK zusammen.* |
| H2.2d: | *Falls ein Zusammenhang zwischen PK und kognitiver Aktivierung existiert, ist dieser schwächer als die Zusammenhänge zwischen CK bzw. PCK und kognitiver Aktivierung.* |

Da die beiden videographierten Unterrichtsstunden zusammengenommen den Unterricht besser repräsentieren als die Einzelstunden (vergl. Abschnitt 7.6.8 auf Seite 146) und das über beide Unterrichtsstunden gemittelte Qualitätsmaß für die kognitive Aktivierung die größte Varianzaufklärung im Mehrebenenmodell zur Erklärung der Post-Testwerte der Lernenden im Fachwissen liefert (vergl.

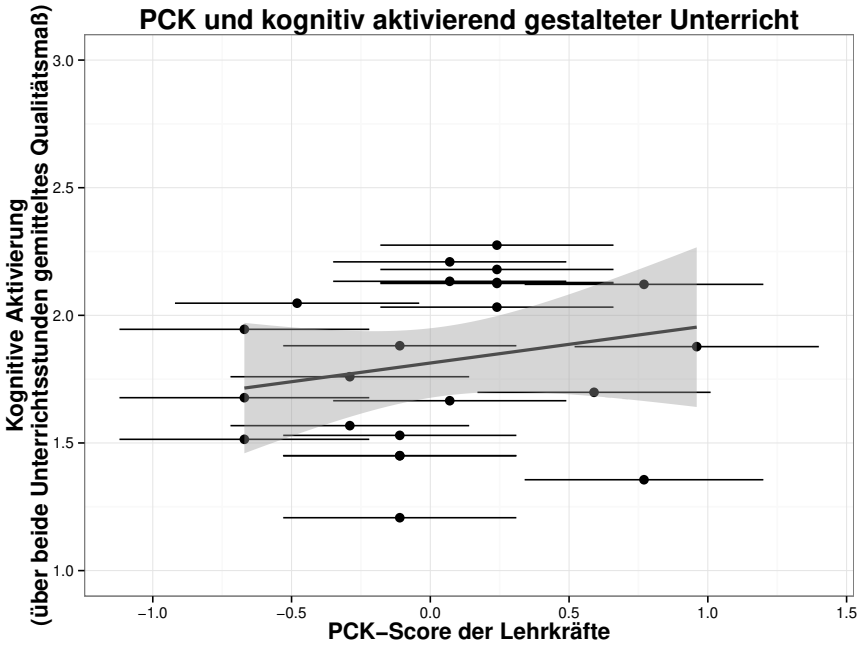
Abschnitt 8.3.1.3 auf Seite 174), wurden für die Überprüfung der Hypothesen Korrelationen zwischen den CK-, PCK- und PK-Testwerten und dem über beide Unterrichtsstunden gemittelten Qualitätsmaß für die kognitive Aktivierung im Unterricht berechnet (vergl. Abschnitt 8.4 auf Seite 185). Abbildungen 8.2a bis 8.2c auf der nächsten Seite und auf Seite 184 zeigen die entsprechenden Scatterplots.

Das Fachwissen und das pädagogische Wissen der Lehrkräfte korrelieren in mittlerer Höhe mit der kognitiv aktivierenden Gestaltung ihres Unterrichts – die Korrelationen sind signifikant mit $p_{\text{1-seitig}} < .05$ und unterscheiden sich nicht. Anzumerken ist, dass für die Korrelation zwischen PK und kognitiver Aktivierung eine einseitige Testung auf Signifikanz gerechtfertigt werden muss – auf Basis der Ausführungen in Abschnitt 5.3.5 auf Seite 73 im Kapitel zur Ableitung des eigenen Forschungsansatzes konnte zwar vermutet werden, dass die beiden Merkmale positiv zusammenhängen, es war allerdings nicht genug Evidenz für die Formulierung einer eindeutigen Hypothese vorhanden. Da allerdings keinesfalls ein negativer Zusammenhang zu erwarten wäre, wurde auch in diesem Fall einseitige auf Signifikanz getestet. Zwischen dem fachdidaktischen Wissen der Lehrkräfte und der kognitiv aktivierenden Unterrichtsgestaltung gibt es keinen signifikanten Zusammenhang.

| | |
|---|---|
| H2.2a ($r_{\text{CK-KA}} > 0$) | angenommen |
| H2.2b ($r_{\text{PCK-KA}} > 0$) | abgelehnt |
| H2.2c ($r_{\text{PCK-KA}} > r_{\text{CK-KA}}$) | abgelehnt |
| H2.2d ($r_{\text{PCK-KA,CK-KA}} > r_{\text{PK-KA}}$) | abgelehnt |

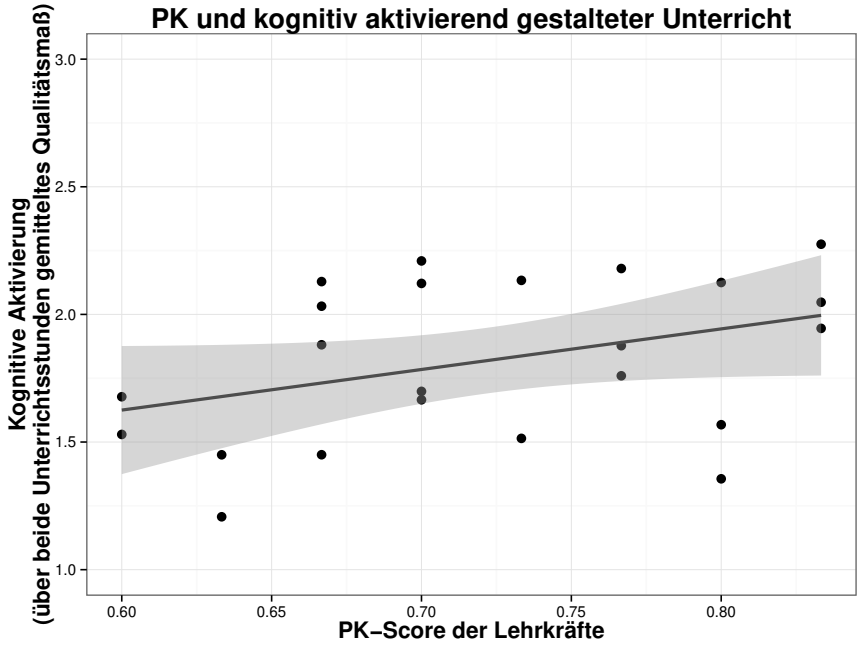



(a)




(b)

Abbildung 8.2.
*Scatterplots mit Regressionslinien und deren* 95%*-Konfidenzregionen (grau schraffierter Bereich) für den Zusammenhang zwischen dem (a) CK und (b) PCK der Lehrkräfte und dem über beide Unterrichtsstunden gemittelten Qualitätsmaß für die kognitiv aktivierende Gestaltung des Unterrichts. Die Fehlerbalken kennzeichnen die unteren Grenzwerte für die Standardfehler auf die Personenfähigkeiten im Rasch-Modell. Da das Rating zur kognitiven Aktivierung klassisch ausgewertet wurde, können für dieses Maß keine Fehlerbalken angegeben werden.*
*(Fortsetzung auf der nächsten Seite)*




(c)

Abbildung 8.2.
*(Fortsetzung) Scatterplot mit Regressionslinie und deren* 95%*-Konfidenzregion (grau schraffierter Bereich) für den Zusammenhang zwischen dem (c) PK der Lehrkräfte und dem über beide Unterrichtsstunden gemittelten Qualitätsmaß für die kognitiv aktivierende Gestaltung des Unterrichts. Da der PK-Test und das Rating zur kognitiven Aktivierung klassisch ausgewertet wurden, können für diese Maße keine Fehlerbalken angegeben werden.*

Tabelle 8.11.
*Korrelationen zwischen dem Professionswissen der Lehrkräfte (CK, PCK, PK) und dem über die 1. und 2. Unterrichtsstunde gemittelten Qualitätsmaß für die kognitive Aktivierung (KA) ($N = 23$)*

| Merkmale | CK-KA$_{1M\&2M}$ | PCK-KA$_{1M\&2M}$ | PK-KA$_{1M\&2M}$ |
|---|---|---|---|
| $r_{\text{Pearson}}$ | **.36 ± .19** | .21 ± .19 | **.38 ± .19** |
| $\text{KI}_{95\,\%}$ | [.02,.72] | [-.14,.57] | [-.04,.68] |
| $p_{\text{1-seitig}}$ | .044 | .165 | .038 |
| $r_{\text{Spearman}}$ | **.37 ± .19** | .30 ± .19 | **.36 ± .19** |
| $\text{KI}_{95\,\%}$ | [-.02,.68] | [-.11,.63] | [-.04,.68] |
| $p_{\text{1-seitig}}$ | .041 | .084 | .047 |
| $\tau_{\text{Kendall}}$ | .25 ± .14 | .16 ± .14 | **.26 ± .14** |
| $\text{KI}_{95\,\%}$ | [.00,.51] | [-.11,.42] | [-.03,.52] |
| $p_{\text{1-seitig}}$ | .050 | .150 | .045 |

*Anmerkungen.* Signifikante Korrelationen mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Es werden zusätzlich nicht-parametrische Korrelationen berichtet, da das Intervallskalenniveau der Qualitätsmaße zur kognitiven Aktivierung nicht sichergestellt werden kann (vergl. Abschnitt 7.4.1 auf Seite 92 zum Umgang mit Ordinalskalen).

# 9. Diskussion und Ausblick

Die vorliegende Arbeit stellt einen der wenigen Versuche dar, den Zusammenhang zwischen dem Professionswissen von Physiklehrkräften und gutem und erfolgreichem Unterrichten zu untersuchen. Ziel der Untersuchung ist es, herauszufinden, ob das mit den im Rahmen des Projektes „Professionswissen in den Naturwissenschaften“ (ProwiN) entwickelten Testinstrumenten erfasste Fachwissen, fachdidaktische Wissen und pädagogische Wissen von Physiklehrkräften prädiktiv für deren Unterrichtsqualität und Unterrichtserfolg ist. Vor dem Hintergrund eines wachsenden Forschungsinteresses am Professionswissen von Lehrkräften und damit verbundener Bemühungen dieses Wissen quantitativ mit Hilfe schriftlicher Testinstrumente zu erfassen, kommt der Frage nach der Relevanz des erhobenen Wissens eine besondere Bedeutung zu. Die Untersuchung dieser Fragestellung impliziert zudem die Untersuchung der Grundannahme, dass das Professionswissen von Lehrkräften eine wichtige Voraussetzung für erfolgreiches Unterrichten darstellt. Wie bei der Formulierung der Forschungsfragen und Hypothesen bereits erläutert wurde, können nicht gefundene Zusammenhänge in Bezug auf diese Grundannahme allerdings nicht eindeutig interpretiert werden. So könnte zwar ein Zusammenhang zwischen dem Wissen und Handeln der Lehrkraft im Unterricht bestehen, sich dieser aber nicht bis auf die Zielkriterien auswirken, oder aber ein Zusammenhang zwischen dem in einem Testinstrument abfragbaren und damit explizierbaren Wissen einer Lehrkraft und ihrem Handeln als solcher schon nicht nachweisbar sein. Nicht gefundene Zusammenhänge können also lediglich anzeigen, dass das für Unterrichtserfolg möglicherweise relevante Professionswissen von den Testinstrumenten nicht erfasst wurde.

In diesem Kapitel, das den Abschluss der vorliegenden Arbeit bildet, werden zunächst die zentralen Ergebnisse der vorliegenden Studie zusammengefasst. In Abschnitt 9.2 auf Seite 189 werden die Voraussetzungen für eine valide Interpretation der Ergebnisse diskutiert. Ziel dieses Unterkapitels ist, einen Eindruck von der Belastbarkeit und Aussagekraft der Ergebnisse der vorliegenden Arbeit zu vermitteln, noch *bevor* diese inhaltlich diskutiert werden. Sowohl an der internen Validität als auch an der externen Validität der Studie kann Kritik geäußert werden. Diese soll vorweg genommen werden, um zu gewährleisten, dass bei der inhaltlichen Diskussion der Ergebnisse in Abschnitt 9.3 auf Seite 206 die Grenzen der vorliegenden Studie vom Leser stets im Blick behalten werden. Abschließend wird in Abschnitt 9.4 auf Seite 211 ein Fazit gezogen, in dem der Beitrag der vorliegenden Arbeit für den wissenschaftlichen Diskurs formuliert wird und Empfehlungen für weitere Untersuchungen gegeben werden. Auf Empfehlungen für die Lehrerausbildung oder die pädagogische Praxis wird hingegen verzichtet, da es

mit Blick auf die Belastbarkeit der Ergebnisse nicht angemessen erscheint, diese auszusprechen.

## 9.1. Kurzzusammenfassung der Ergebnisse

Im Rahmen von Korrelationsanalysen wurde überprüft, ob das mit den ProwiN-Tests gemessene Professionswissen von Physiklehrkräften mit der kognitiv aktivierenden Gestaltung ihres Unterrichts zusammenhängt. Im Rahmen von Mehrebenenanalysen wurde überprüft, ob das Professionswissen der Lehrkräfte und die kognitiv aktivierende Gestaltung des Unterrichts (unter Kontrolle des Vorwissens, der kognitiven Fähigkeiten der Lernenden, des Geschlechts, der zuhause gesprochenen Sprache sowie der Unterrichtszeit) Varianz in den Klassenmittelwerten der Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik oder im situationalen Interesse der Lernenden im Unterricht aufklären.

- Die Maße für die kognitiv aktivierende Gestaltung der beiden videographierten Unterrichtsstunden tragen signifikant zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik, nicht aber zur Aufklärung der Varianz im situationalen Interesse der Lernenden am Unterricht der jeweiligen videographierten Unterrichtsstunde bei.

- Das Fachwissen der Lehrkräfte hängt signifikant mit der kognitiv aktivierenden Unterrichtsgestaltung zusammen, liefert aber keinen signifikanten Beitrag zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik oder im situationalen Interesse der Lernenden am Unterricht.

- Das fachdidaktische Wissen der Lehrkräfte hängt weder signifikant mit der kognitiv aktivierenden Unterrichtsgestaltung zusammen, noch liefert es einen signifikanten Beitrag zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik oder im situationalen Interesse der Lernenden am Unterricht.

- Das pädagogische Wissen der Lehrkräfte hängt signifikant mit der kognitiv aktivierenden Unterrichtsgestaltung zusammen und liefert einen signifikanten Beitrag zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik, nicht aber im situationalen Interesse der Lernenden am Unterricht.

Auf den ersten Blick könnten diese Ergebnisse Hinweise auf die prädiktive Validität des Tests zum pädagogischen Wissen liefern, während sie die prädiktive Validität der fachspezifischen Professionswissenstests in Bezug auf Unterrichtserfolg in Frage stellen könnten. Unterschiede in den Klassenmittelwerten im situationalen Interesse der Lernenden werden durch keine der drei Professionswissensdimensionen aufgeklärt.

Unterrichtserfolg ist nach dem Angebots-Nutzungsmodell von Helmke (2009, S. 73) allerdings durch zahlreiche weitere Variablen beeinflusst und hängt nicht zuletzt auch davon ab, ob das von der Lehrkraft bereitgestellte Lehrangebot von den Lernenden auch wirklich genutzt wird. Letzteres kann eine Lehrkraft zwar unterstützen, nicht aber garantieren. Nach dem konstruktivistischen Lernverständnis erfordert die Nutzung des Lehrangebots eine aktive Auseinandersetzung der Lernenden mit dem Lerngegenstand. Diese kann durch die kognitiv aktivierende Gestaltung des Unterrichts durch die Lehrkraft unterstützt werden.

Die Ergebnisse könnten Hinweise darauf liefern, dass mit der kognitiv aktivierenden Gestaltung des Unterrichts ein Merkmal der Unterrichtsqualität im Sinne von Fenstermacher und Richardson (2005) erfasst wird, sofern Unterrichtserfolg lediglich über die Fachwissensleistung der Lernenden modelliert wird. Den Hypothesen folgend, sollte insbesondere das mit den ProwiN-Tests gemessene fachdidaktische Wissen, aber auch das Fachwissen der Lehrkräfte positiv mit der kognitiv aktivierenden Gestaltung ihres Unterrichts zusammenhängen, während für das pädagogische Wissen, sofern überhaupt vorhanden, kleinere Effekte erwartet werden. Die Ergebnisse der vorliegenden Arbeit, die signifikante positive Zusammenhänge lediglich für das Fachwissen und das pädagogische Wissen mit gleicher Effektstärke anzeigen, nicht aber für das fachdidaktische Wissen, überraschen daher und könnten die prädiktive Validität des PCK-Tests auch in Bezug auf Unterrichtsqualität in Frage stellen.

## 9.2. Voraussetzungen für eine valide Interpretation der Ergebnisse

Um eine valide Interpretation der Ergebnisse der vorliegenden Studie zu gewährleisten, wird vor der inhaltlichen Diskussion der Ergebnisse zunächst deren Belastbarkeit und Aussagekraft diskutiert. Bei der vorliegenden Studie handelt es sich um eine quasiexperimentelle Feldstudie, in der eine nicht zufällig gezogene kleine Stichprobe von $N = 23$ Lehrkräfte und ihren Klassen in ihrem natürlichen Umfeld untersucht wurde. Sowohl an der internen Validität als auch an der externen Validität der Studie kann daher berechtigte Kritik geäußert werden. Es erscheint daher nicht angemessen, die Ergebnisse und mögliche Konsequenzen für das Forschungsfeld zu interpretieren und lediglich im Anschluss auf Einschränkungen der Studie hinzuweisen. Vor der inhaltlichen Diskussion erfolgt daher der Versuch die Auswirkungen der methodischen Probleme dieser Studie auf die einzelnen Ergebnisse abzuschätzen, um im Anschluss herausarbeiten zu können, welche Ergebnisse dennoch einen Beitrag für den wissenschaftlichen Diskurs leisten können.

Die folgenden Abschnitte dienen dazu, sich Antworten auf zwei grundsätzliche Fragestellungen zu nähern:

1. Inwieweit können valide Aussagen über die Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg in der in der hier untersuchten Stichprobe gemacht werden?

2. Wie wahrscheinlich ist es, dass ähnliche Ergebnisse, die zu gleichen inhaltlichen Interpretationen führen würden, auch in anderen Stichproben beobachtet werden könnten?

### 9.2.1. Diskussion der internen Validität der Untersuchung

Nach Bortz und Döring (2006, S. 53) liegt interne Validität vor, „wenn Veränderungen in den abhängigen Variablen eindeutig auf den Einfluss der unabhängigen Variablen zurückzuführen sind bzw. wenn es neben der Untersuchungshypothese keine besseren Alternativerklärungen gibt".

Das Design der vorliegenden Studie ist prinzipiell günstig für die externe Validität der Untersuchung, jedoch ungünstig für die interne Validität (vergl. Bortz & Döring, 2006, S. 58), da lediglich korrelative, nicht aber kausal bedingte Zusammenhänge untersucht werden können und der Einfluss von Störvariablen nicht gänzlich eliminiert werden kann. In quasiexperimentellen Felduntersuchungen kann nicht zweifelsfrei ausgeschlossen werden, dass es sich bei beobachteten Zusammenhängen um Scheinzusammenhänge handelt (Bortz & Döring, 2006, S. 526).

Um die interne Validität der Untersuchung zu erhöhen, wurde versucht die Versuchsbedingungen so konstant wie möglich zu halten. So wurden die Lehrkräfte stets zum gleichen Zeitpunkt – zu Beginn der Unterrichtseinheit Mechanik – bezüglich ihres Professionswissens getestet. Über den Einsatz von Testleitermanualen wurde außerdem versucht die Testungsbedingungen bei allen Testungen möglichst konstant zu halten. Der Zeitraum zwischen den Testungen variierte abhängig von den Begebenheiten an der jeweiligen Schule – durch die Kontrolle der tatsächlichen Unterrichtszeit in den Mehrebenenmodellen für die Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik wurde allerdings versucht, den Einfluss dieser Störvariablen zu eliminieren. Auch in Bezug auf den videographierten Unterricht wurde versucht vergleichbare Versuchsbedingungen herzustellen: Das Thema der ersten Videostunde wurde vorgegeben und die Lehrkräfte wurden gebeten, ein Lehrerexperiment in ihren Unterricht zu integrieren sowie als primäres Lernziel einen Fachwissenserwerb aufseiten der Lernenden anzustreben. Für die zweite Unterrichtsstunde wurden allerdings keine Vorgaben gemacht (vergl. Abschnitte 7.1 und 7.2 auf Seite 81 und auf Seite 82 zum Untersuchungsdesign und der Durchführung der Studie).

Die Länge der videographierten Unterrichtsstunden variierte abhängig von den Begebenheiten an der jeweiligen Schule. Stender, Geller, Neumann und Fischer (2013) konnten zeigen, dass die Stundenlänge – die sogenannte Unterrichtstaktung – Einfluss auf Aspekte der lernprozessorientierten Sequenzierung von Unterricht haben kann. In Bezug auf die interne Validität der vorliegenden Studie wird die unterschiedliche Unterrichtstaktung in den untersuchten Klassen allerdings nicht als problematisch erachtet, da ein bedeutsamer Einfluss der Unterrichtstaktung auf die kognitiv aktivierende Gestaltung des Unterrichts aus theoretischer Sicht nicht anzunehmen ist. Empirisch untersucht wurde dieser Zusammenhang bisher allerdings nicht.

Ferner müssen sich Videostudien mit der Kritik auseinandersetzen, dass Lehrkräfte ihren Unterricht, in dem Wissen gefilmt zu werden, bewusst anders planen

könnten und daher lediglich „Best-Practice“-Stunden analysiert werden könnten. Die Repräsentativität des videographierten Unterrichts für den herkömmlichen Unterricht der Lehrkräfte wurde bereits ausführlich in Abschnitt 7.6.8 auf Seite 146 zur Validierung des Ratings zur kognitiven Aktivierung diskutiert – mit dem Ergebnis, dass der videographierte Unterricht von den Unterrichtsakteuren in den meisten Fällen als typisch eingeschätzt wurde. Selbst für den Fall, dass es sich dennoch um Best-Practice-Beispiele für den Unterricht der Lehrkräfte handelt, wäre dies für die Interpretation der Ergebnisse zum Zusammenhang zwischen Professionswissen und Unterrichtsqualität nicht ausschlaggebend. In diesem Fall müsste lediglich davon ausgegangen werden, dass die Lehrkräfte ihre Handlungsressourcen, zu denen laut Hypothesen ihr Professionswissen gehört, für die kognitiv aktivierende Gestaltung der videographierten Unterrichtsstunden voll ausgeschöpft hätten.

Die Datenerhebung in der vorliegenden Studie erfolgte sequenziell: Zunächst wurde das Professionswissen der Lehrkräfte erhoben, danach wurde das Videomaterial für die Einschätzung der kognitiv aktivierenden Gestaltung des Unterrichts aufgezeichnet und zum Schluss der Unterrichtserfolg gemessen. Beobachtete Zusammenhänge können daher zumindest als Hinweis auf Kausalitäten gewertet werden, da die unabhängigen Variablen vor den abhängigen Variablen erhoben wurden (vergl. Bortz & Döring, 2006, S. 523). So kann beispielsweise das Professionswissen der Lehrkräfte die kognitiv aktivierende Gestaltung des Unterrichts beeinflussen, umgekehrt können Erfahrungen, die die Lehrkraft in diesen Unterrichtsstunden gesammelt hat, das erhobene Wissen allerdings nicht nachträglich verändern.

Um dem Umstand Rechnung zu tragen, dass *natürliche* Gruppen untersucht wurden, die Lernenden also nicht zufällig auf die Klassen verteilt waren, wurden auf Schülerebene Variablen kontrolliert, die auf die Schülerleistung wirken, aber ihrerseits nicht im Rahmen der Unterrichtseinheit durch das professionelle Wissen der Lehrkraft oder die Unterrichtsqualität beeinflusst werden können. Kontrolliert wurden das Vorwissen, die kognitiven Fähigkeiten, das Geschlecht und die von den Lernenden zuhause gesprochene Sprache (vergl. Abschnitt 5.2 auf Seite 63).

Die untersuchte Stichprobe stellt allerdings auch auf Klassenebene keine Zufallsstichprobe dar. Die Lehrkräfte konnten sich freiwillig zu einer Teilnahme an der Studie bereit erklären, was die Frage aufwirft, ob es sich bei den hier untersuchten Lehrkräften um eine Positivauswahl handelt – schließlich ist anzunehmen, dass Lehrkräfte, die sich als weniger kompetent wahrnehmen, sich weniger häufig bereit erklären ihr Wissen testen zu lassen und darüber hinaus ihren Unterricht filmen zu lassen. Dieser Aspekt wird in Abschnitt 9.2.2 auf Seite 193 ausführlicher diskutiert, da sich hieraus eher Konsequenzen für die externe Validität der Untersuchung ergeben. In Bezug auf die interne Validität muss berücksichtigt werden, dass Variablen die ebenfalls einen Einfluss auf die Qualität und den Erfolg des Unterrichts der teilnehmenden Lehrkräfte haben könnten, in der untersuchten Stichprobe nicht zufällig verteilt sein könnten. Hierzu gehören beispielsweise die im Modell zur professionellen Handlungskompetenz von Baumert und Kunter (2006) aufgeführten motivational-selbstregulativen Merkmale wie Überzeugungen und Werthaltungen, motivationale Orientierungen und selbstregulative Fähigkeiten oder aber die Selbstwirksamkeitserwartung der Lehrkräfte. Einige dieser Variablen wurden im Gesamtprojekt zwar erhoben, im Rahmen der vorliegenden Arbeit aber

nicht ausgewertet. Es kann daher nicht ausgeschlossen werden, dass Zusammenhänge zwischen den in dieser Studie untersuchten unabhängigen und abhängigen Variablen auf die Existenz von konfundierenden Variablen zurückzuführen sind (also Variablen die die gemeinsame Varianz in den unabhängigen und abhängigen Variablen verursachen). Auch andere Merkmale der Unterrichtsqualität, wie z. B. Klassenführung, könnten als konfundierende Variablen wirken.

#### 9.2.1.1. Diskussion der Messfehler

Ein weiteres Problem für die interne Validität und damit auch für die Interpretation der Ergebnisse liegt in den zum Teil erheblichen Messungenauigkeiten der Testinstrumente. Mit Blick auf die im Methodenteil dieser Arbeit berichteten Reliabilitäten gilt dies insbesondere für den PCK-Test (Pers. Rel. = .59) sowie die Schülerfachwissenstest (Prä-Test/Post-Test Pers. Rel. = .51/.61) und etwas weniger ausgeprägt für den PK-Test ($\alpha_C = .67$). Mögliche Gründe für die niedrigen Reliabilitäten wurden bereits in Abschnitt 7.5.1.6 und 7.5.3.4 auf Seite 108 und auf Seite 122 diskutiert. Bei der Interpretation der Ergebnisse muss berücksichtigt werden, welche Auswirkungen die niedrigen Reliabilitäten der Messinstrumente auf die Ergebnisse haben könnten.

Messfehler führen in der Regel zu einer Unterschätzung von Zusammenhängen (vergl. z. B. OECD, 2012, S. 105; Rost, 2004, S. 389). Aufgrund dessen kann beispielsweise nicht ausgeschlossen werden, dass in der vorliegenden Stichprobe lediglich wegen der großen Messfehler auf die PCK-Testwerte wesentlich geringere Zusammenhänge zwischen PCK und kognitiver Aktivierung als zwischen CK oder PK und kognitiver Aktivierung beobachtet wurden. Vergleicht man um Messfehler bereinigte Korrelationen der Professionswissensdimensionen mit der kognitiven Aktivierung, ändert dies allerdings nichts an dem Umstand, dass zwischen fachdidaktischem Wissen und kognitiver Aktivierung die geringsten Zusammenhänge existieren ($r_{\text{Pearson,korr.,CK-KA}} = .42$, $r_{\text{Pearson,korr.,PCK-KA}} = .27$, $r_{\text{Pearson,korr.,PK-KA}} = .46$).[1] Die Interpretation der Ergebnisse bliebe in diesem Fall unbeeinflusst von den Messfehlern.

Unklar ist allerdings, wie sich die Messfehler auf die Ergebnisse der Mehrebenenanalysen auswirken, in denen sowohl die abhängigen als auch die unabhängigen Variablen fehlerbehaftet sind. Woodhouse, Yang, Goldstein und Rasbash (1996, S. 211) konnten zeigen, dass die Intraklassenkorrelation und damit der Anteil der zwischen den Klassen liegenden Varianz an der Gesamtvarianz in der abhängigen Variable schon bei Reliabilitäten von .85 massiv unterschätzt wird. Messungenauigkeiten in den Prädiktorvariablen können hingegen sowohl zu einer Überschätzung als auch zu einer Unterschätzung von Zusammenhängen führen (vergl. z. B. Kromrey et al., 2006; Woodhouse et al., 1996).

Angeführt werden kann an dieser Stelle lediglich, dass die Messinstrumente trotz geringer Reliabilitäten erwartungskonforme Ergebnisse im Zuge der Überlegungen für die Testvalidierung lieferten (vergl. Abschnitt 7.5.1.7, 7.5.2.6 und 7.5.3.5 auf

[1] Zur Bereinigung wurde an dieser Stelle lediglich durch die Quadratwurzel der Reliabilitäten der Professionswissenstests geteilt (vergl. Abschnitt 7.4.6 auf Seite 101 im Kapitel zu den statistischen Methoden).

Seite 109, auf Seite 115 und auf Seite 123). Die mit den Schülertests gemessenen Leistungszuwächse sind außerdem vergleichbar mit den in anderen Studien gemessenen Leistungszuwächsen (vergl. Abschnitt 9.2.3.1 auf Seite 199). Auch scheinen sowohl die Validierungsergebnisse als auch die Ergebnisse zum Zusammenhang zwischen Professionswissen und den Fachwissensleistungen der Lernenden robust gegen Veränderungen an den Messinstrumenten zu sein. So konnten Validierungsergebnisse von Kirschner (2013) repliziert werden, obwohl Änderungen an den in der ersten Phase entwickelten Testinstrumenten vorgenommen werden mussten, die zu niedrigeren Reliabilitäten der ProwiN-Tests für das fachspezifische Professionswissen führten (vergl. Abschnitt 7.5.1.4 auf Seite 105). Auch liegen inzwischen Daten aus dem zweiten im Rahmen von ProwiN Physik durchgeführten Dissertationsprojekt vor, in dem die hier untersuchte Stichprobe um 12 Lehrkräfte auf insgesamt $N = 35$ Physiklehrkräfte mit ihren Klassen erweitert wurde (vergl. Liepertz, 2016). Obwohl die fachspezifischen Professionswissenstest und die Schülertests von Liepertz (2016) etwas anders ausgewertet werden (für die Rasch-Analysen werden andere Programme und damit auch andere Schätzalgorithmen genutzt, im Schülertest werden andere Items aus den Analysen ausgeschlossen), was sich auch auf die Reliabilitäten der Testinstrumente auswirkt (diese sind zum Teil noch geringer als in der vorliegenden Arbeit), ergeben sich keine Änderungen für die Interpretation der Ergebnisse: Auch in der erweiterten Stichprobe trägt das Fachwissen der Lehrkräfte nicht signifikant zur Varianzaufklärung in den Fachwissensleistungen der Lernenden bei. Das fachdidaktische Wissen leistet zwar einen signifikanten Beitrag zur Varianzaufklärung – der Zusammenhang zwischen fachdidaktischem Wissen und Unterrichtserfolg ist allerdings negativ (Liepertz, Cauet, Borowski & Fischer, 2015). Letzteres deutet sich bereits in der in der vorliegenden Arbeit untersuchten Stichprobe an: Der Regressionskoeffizient $\gamma_{\text{PCK}}^{\text{StdYX}} = -0.15$ ist ebenfalls negativ. Lenske et al. (2016) berichten für diese Stichprobe ebenfalls einen signifikanten Beitrag des pädagogischen Wissens für die Varianzaufklärung in den Post-Testleistungen der Lernenden, obwohl 11 der 30 in der vorliegenden Arbeit zur Berechnung der PK-Testwerte genutzten Items aus den Analysen ausgeschlossen werden, was ebenfalls zu einer leichten Veränderung in der Reliabilität führt.

## 9.2.2. Diskussion der externen Validität der Untersuchung

Nach Bortz und Döring (2006, S. 53) ist eine Untersuchung extern valide, „wenn das in einer Stichprobenuntersuchung gefundene Ergebnis auf andere Personen, Situationen oder Zeitpunkte generalisiert werden kann“.

Das Hauptproblem der vorliegenden Studie besteht in der kleinen, für die Untersuchung der Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg zur Verfügung stehenden Stichprobe von lediglich $N = 23$ Lehrkräften mit ihren Klassen und spiegelt sich in den hieraus resultierenden großen Standardfehlern und Konfidenzintervallen für alle auf Klassenebene berechneten Zusammenhangsmaße oder Regressionskoeffizienten wider. Die exakte Größe dieser Zusammenhänge in anderen Studien mit anderen Stichproben replizieren zu können, erscheint daher unwahrscheinlich. Der ursprüngliche Untersuchungsplan sah eine wesentlich größere Stichprobe von 40 Lehrkräften für die Untersuchung

der Forschungsfragen vor. Dieses Vorhaben scheiterte allerdings daran, dass sich innerhalb des zweijährigen Erhebungszeitraumes trotz aller Bemühungen lediglich die hier untersuchten 23 Lehrkräfte zu einer Teilnahme bereit erklärten.

Abgesehen von der beschränkten Größe der Stichprobe, soll zunächst kurz deren Zusammensetzung diskutiert werden. Wie bereits erwähnt, ist anzunehmen, dass es sich bei den in dieser Studie untersuchten Lehrkräften um eine Positivauswahl handelt. Tatsächlich stellt die untersuchte Stichprobe im Vergleich zu der in ProwiN I untersuchten Gymnasiallehrerstichprobe aus Nordrhein-Westfalen (NRW) eine leichte Positivauswahl bezüglich ihres mit den ProwiN-Tests gemessenen Fachwissens, nicht aber bezüglich ihres fachdidaktischen und pädagogischen Wissens dar (vergl. Abschnitt 8.1.1.2 auf Seite 160). Die ProwiN I-Stichprobe stellt allerdings ihrerseits keine Zufallsstichprobe dar und es erscheint nicht unwahrscheinlich, dass es sich auch hier bereits um eine Positivauswahl handelt. Aus den bereits genannten Gründen ist zudem anzunehmen, dass die Stichprobe auch bezüglich der Qualität des Unterrichts eine Positivauswahl darstellt. Ein Vergleich mit einer Referenzgruppe ist diesbezüglich leider nicht möglich. Angeführt werden kann nur, dass die Qualitätsmaße für die kognitiv aktivierende Unterrichtsgestaltung zumindest keine Deckeneffekte anzeigen (vergl. Tabelle 8.5 auf Seite 164).

Problematisch ist, dass in Extremgruppen im Vergleich zur Grundgesamtheit Zusammenhänge zwischen Merkmalen unterschiedlich stark ausgeprägt sein können, was zu einer Unterschätzung der Zusammenhänge in der Grundgesamtheit führen kann (Bortz & Döring, 2006, S. 509). Vergleicht man zum Beispiel die Korrelationen zwischen den Professionswissensdimensionen in der ProwiN II-Stichprobe der $N = 23$ Lehrkräfte (vergl. Tabelle 8.2 auf Seite 161) mit den entsprechenden Korrelationen in der um die Gymnasiallehrkräfte aus NRW der ProwiN I-Stichprobe erweiterten Stichprobe der $N = 102$ Physiklehrkräfte, die für die Reliabilitäts- und Validierungsanalysen genutzt wurde (vergl. Tabelle 7.5 auf Seite 111), zeigen sich deutliche Unterschiede: Die Korrelationskoeffizienten sind in der ProwiN II-Stichprobe wesentlich kleiner und keiner der Zusammenhänge wird signifikant. Dies könnte zum einen daran liegen, dass die Lehrkräfte bezüglich des Fachwissens im Vergleich mit der ProwiN I-Stichprobe eine Positivauswahl darstellen. Zum anderen sind die Streuungen der PCK- und PK-Testwerte etwas geringer als in der ProwiN I-Stichprobe, was ebenfalls zu einer Unterschätzung der Zusammenhänge führen könnte.

Mit Blick auf diese Einschränkungen stellt sich die Frage, inwieweit die Kriterien zur Annahme oder Ablehnung der in dieser Arbeit untersuchten Hypothesen dem Zufall unterliegen könnten, und welche Konsequenzen sich hieraus auf die Interpretierbarkeit der Ergebnisse ergeben. Grundsätzlich kann es bei der Hypothesentestung zu zwei Fehlschlüssen kommen: Die Nullhypothese kann abgelehnt werden, obwohl Sie wahr ist (Fehler 1. Art), oder sie kann angenommen werden, obwohl sie falsch ist (Fehler 2. Art).

#### 9.2.2.1. Fehler 1. Art: Diskussion der signifikanten Zusammenhänge

Der Fehler 1. Art wird über die Festlegung des Alphaniveaus gesteuert, das angibt, ab welchem $p$-Wert eine Hypothese angenommen wird. Die Wahrscheinlichkeit

dafür, dass es sich bei den in der vorliegenden Stichprobe beobachteten signifikanten Zusammenhängen um Zufallsprodukte handelt, liegt demnach theoretisch bei unter 5%. Wenn man mehrere Hypothesen in der gleichen Stichprobe testet, kommt es allerdings zu einer Kumulierung des Alphafehlers – die Wahrscheinlichkeit eines der signifikanten Ergebnisse nur durch Zufall zu erhalten, nimmt zu (vergl. z. B. Field, 2009, S. 348). Dies gilt insbesondere, wenn die Hypothesen nicht sauber aus der Theorie abgeleitet werden (vergl. Bortz, 2005, S. 130).

Dass der Fehler 1. Art ein ernstzunehmendes Risiko darstellt und Artefakte in kleinen Stichproben ein Problem darstellen und zu Fehlinterpretationen führen können, zeigt folgende Kuriosität in der in dieser Arbeit untersuchten Stichprobe: Der Zusammenhang zwischen der kognitiv aktivierenden Gestaltung des Unterrichts und dem situationalen Interesse der Lernenden im Unterricht wurde in der vorliegenden Studie getrennt für beide Unterrichtsstunden überprüft. Beide Unterrichtsstunden zusammengenommen repräsentieren den Unterricht allerdings besser, als die einzelnen Unterrichtsstunden es tun (vergl. Abschnitt 7.6.8 auf Seite 146). Daher ließe sich argumentieren, dass der Zusammenhang zwischen den über beide Unterrichtsstunden gemittelten Maßen für die kognitiv aktivierende Gestaltung des Unterrichts und das situationale Interesse der Lernenden überprüft werden sollte. In der Tat stellt das über beide Unterrichtsstunden gemittelte Qualitätsmaß für die kognitive Aktivierung in einem entsprechenden Mehrebenenmodell einen signifikanten Prädiktor für das über beide Unterrichtsstunden gemittelte situationale Interesse der Lernenden dar ($\gamma^{\mathrm{StdYX}}_{\mathrm{KA1M\&2M}} = 0.45 \pm 0.17$, $\mathrm{KI}_{95\,\%} = [0.13, 0.76]$, $p_{\text{1-seitig}} = .003$), der Beitrag zur Varianzaufklärung in den Klassenmittelwerten wird nur knapp nicht signifikant ($R^2 = (20 \pm 15)\%$, $p_{\text{1-seitig}} = .083$). Bei genauerem Hinschauen stellt sich allerdings heraus, dass der Zusammenhang auf einen nicht sinnvoll zu interpretierenden Zusammenhang zwischen der kognitiv aktivierenden Gestaltung der ersten Unterrichtsstunde und dem am Ende der zweiten Unterrichtsstunde erhobenen situationalen Interesse der Lernenden zurückgeht – die Varianzaufklärung in einem entsprechend spezifizierten Modell liegt bei $R^2 = (26 \pm 14)\%$ ($p_{\text{1-seitig}} = .025$). Offensichtlich handelt es sich hierbei um ein Artefakt.[2]

Eine Möglichkeit die Alphafehlerkumulierung zu berücksichtigen besteht in einer Korrektur des Alphaniveaus (beispielsweise nach Bonferoni, oder weniger konservativ nach Holm) hin zu einem strengeren Signifikanzkriterium (vergl. z. B. Bortz & Lienert, 2008, S. 39). Mit einer Absenkung des Alphaniveaus sinkt zwar die Wahrscheinlichkeit für einen Fehler 1. Art, die Wahrscheinlichkeit für einen Fehler 2. Art steigt allerdings an (vergl. z. B. Field, 2009, S. 56). Mit Blick auf die ohnehin geringe Teststärke (vergl. Abschnitt zum Fehler 2. Art) und weil jede der in dieser Arbeit getesteten Hypothesen spezifisch aus der Theorie abgeleitet wurde (vergl. hierzu Bortz, 2005, S. 130), wurde in dieser Arbeit auf eine Korrektur des Alphaniveaus verzichtet.

[2] Im Folgenden wird sich herausstellen, dass die Klassenunterschiede im situationalen Interesse nicht bedeutsam sind. Der berichtete Zusammenhang wurde an dieser Stelle lediglich angeführt, um die Notwendigkeit der folgenden Diskussion zu unterstreichen.

Der Zusammenhang zwischen der kognitiv aktivierenden Gestaltung des Unterrichts und dem Unterrichtserfolg lässt sich unabhängig davon nachweisen, ob hierfür das Qualitätsmaß für die erste Unterrichtsstunde, die zweite Unterrichtsstunde oder das über beide Unterrichtsstunden gemittelte Qualitätsmaß verwendet wird (wobei letzteres natürlich nicht unabhängig von den ersten beiden ist). Die Wahrscheinlichkeit dafür, dass zwei oder gar alle drei Zusammenhänge Zufallsprodukte sind, sinkt also wieder. Gleiches gilt für die Ergebnisse zum pädagogischen Wissen. Auch hier zeigen sich konsistente Ergebnisse: PK hängt sowohl mit Unterrichtsqualität als auch mit Unterrichtserfolg zusammen.

Ob ein Zusammenhang signifikant wird, hängt zudem von dessen Effektstärke ab. Diese wiederum wird in kleinen Stichproben wesentlich stärker durch Ausreißer beeinflusst als es in größeren Stichproben der Fall ist, was zu einer Überschätzung tatsächlich in der Grundgesamtheit vorhandener Zusammenhänge führen kann. Auf extreme Ausreißer würden signifikante Abweichungen von der Normalverteilung hinweisen, was in der ProwiN II-Stichprobe weder für die Verteilung der Testwerte im Professionswissen noch für die Qualitätsmaße zur kognitiven Aktivierung der Fall ist (vergl. Tabelle B.3 auf Seite 244 im Anhang). Um einschätzen zu können, ob dennoch lediglich Ausreißer für die beobachteten Zusammenhänge verantwortlich sein könnten, lohnt ein Blick auf die nicht-parametrischen Zusammenhangsmaße – so ist der Rangkorrelationskoeffizient $\tau_{\text{Kendall}}$ besonders robust gegen Ausreißer (vergl. Bortz & Lienert, 2008, S. 301). Für den Zusammenhang zwischen Fachwissen und pädagogischem Wissen mit der kognitiven Aktivierung wurden die Werte für $\tau_{\text{Kendall}}$ in Abschnitt 8.4 auf Seite 185 berichtet. Es sei angemerkt, dass die Werte für $\tau_{\text{Kendall}}$ nicht nur im vorliegenden Fall, sondern auch generell, wesentlich kleiner sind als für Spearman-Rangkorrelationen oder Pearson-Korrelationen und nicht direkt mit diesen vergleichbar sind (Field, 2009, S. 193). Für das pädagogische Wissen zeigt auch die Rangkorrelation nach Kendall einen signifikanten Zusammenhang zur kognitiv aktivierenden Unterrichtsgestaltung an. Im Falle des Fachwissens verfehlt der Rangkorrelationskoeffizient mit $p_{\text{1-seitig}} = .050$ so knapp das Signifikanzniveau von $p < .05$, dass auch in diesem Fall nicht davon auszugehen ist, dass der beobachtete Zusammenhang zwischen Fachwissen und kognitiv aktivierendem Unterricht in der in dieser Arbeit untersuchten Stichprobe lediglich durch Ausreißer bedingt ist.

Um grob abzuschätzen, ob Ausreißer einen Einfluss auf die Ergebnisse der Mehrebenenanalysen zum Zusammenhang zwischen pädagogischen Wissen und kognitiver Aktivierung mit Unterrichtserfolg haben könnten, wurden zusätzlich zu den in der vorliegenden Arbeit berichteten Ergebnissen unter Vernachlässigung der Mehrebenenstruktur residuale Lernzuwächse zunächst aus einer Regression auf Schülerebene berechnet, in der die durch die Kontrollvariablen (Prä-Testwert, KFT-Testwert, Geschlecht, zuhause gesprochene Sprache) erklärte Varianz aus den Post-Testwerten der Lernenden herausgerechnet wurde. Diese Residuen wurden auf Klassenebene über den Mittelwert aggregiert und anschließend in einer Regression auf Klassenebene um die durch die Unterrichtszeit erklärte Varianz bereinigt. Die Pearson-Korrelationen zwischen diesen effektiven Leistungszuwächsen und den Qualitätsmaßen für die kognitive Aktivierung oder dem PK-Testwert der Lehrkräfte liegen in sehr ähnlicher Größenordnung zu den vollstandardisierten Re-

gressionskoeffizienten aus den Mehrebenenanalysen. Die gegen Ausreißer robusten Rangkorrelationen nach Kendall zeigen ebenfalls signifikante Zusammenhänge an (vergl. Tabelle B.9 auf Seite 248 im Anhang). Festgehalten werden kann also, dass es keine Hinweise dafür gibt, dass die Ergebnisse der Mehrebenenanalysen durch Ausreißer verzerrt wurden.

Bezüglich des Zusammenhangs zwischen pädagogischem Wissen und Unterrichtserfolg muss berücksichtigt werden, dass im Rahmen der Mehrebenenanalysen aufgrund der kleinen Stichprobengröße die Standardfehler auf die Regressionskoeffizienten unterschätzt werden könnten, was die Alphafehlerwahrscheinlichkeit erhöhen könnte (vergl. Abschnitt 7.4.5 auf Seite 99). Verwiesen werden kann an dieser Stelle zwar auf die Ergebnisse von Lenske et al. (2016), die auch in der erweiterten ProwiN II-Stichprobe einen signifikanten Beitrag des pädagogischen Wissens zur Varianzaufklärung in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik beobachten, allerdings ist selbst in der erweiterten Stichprobe noch mit einer Unterschätzung der Standardfehler zu rechnen. Korrekte Schätzungen werden erst ab Stichprobengrößen von mindestens 50 Klassen erwartet (Maas & Hox, 2004, S. 135).

#### 9.2.2.2. Fehler 2. Art: Diskussion der nicht signifikanten Zusammenhänge

Die Betrachtung des Fehlers 1. Art hat in vielen Disziplinen eine lange Forschungstradition – so ist die Signifikanz von Ergebnissen oftmals Voraussetzung dafür, dass diese publiziert werden können (obgleich sich diesbezüglich ein Wandel vollzieht, vergl. z. B. Novella, 2015; Shrout, 1997). Weitaus seltener wird der Fehler 2. Art, die sogenannte Betafehlerwahrscheinlichkeit, diskutiert, obwohl auch diese Ursache von Fehlinterpretationen sein kann (vergl. z. B. Bortz & Döring, 2006, S. 637; Stelzl, 2006, S. 14). Während der Betafehler die Wahrscheinlichkeit angibt, mit der eine Hypothese, obwohl sie wahr ist, verworfen wird, gibt $1 - \beta$, die sogenannte Teststärke an, mit welcher Wahrscheinlichkeit eine richtige Hypothese auch als solche erkannt wird (Bortz & Lienert, 2008, S. 50). Bei gegebenem Alphaniveau und gegebener Effektstärke, hängt die Teststärke unmittelbar von der Stichprobengröße ab (Bortz & Döring, 2006, S. 603).

Für die Interpretation der Ergebnisse der vorliegenden Arbeit ist von Interesse, ob die nicht signifikanten Zusammenhänge zwischen dem fachspezifischen Professionswissen und Unterrichtserfolg und zwischen PCK und kognitiver Aktivierung die Validität der Messinstrumente (bzw. der Modellierung) in Frage stellen oder lediglich das Resultat einer nicht ausreichenden Teststärke darstellen. Als überraschendstes Ergebnis kann sicherlich der nicht signifikante Zusammenhang zwischen PCK und kognitiver Aktivierung bewertet werden, zumal Zusammenhänge zum Fachwissen und pädagogischen Wissen in der vorliegenden Stichprobe durchaus beobachtet werden konnten. Aus theoretischer Sicht und mit Blick auf das im PCK-Test abgefragte Wissen sollte der PCK-Testwert der Lehrkräfte im Vergleich mit den anderen Professionswissensdimensionen am stärksten mit kognitiver Aktivierung zusammenhängen. Gegen die Validität des PCK-Tests würde also sprechen, wenn dieser auch in der Grundgesamtheit geringer oder lediglich genauso stark mit der kognitiv aktivierenden Unterrichtsgestaltung zusammenhängt wie CK oder PK.

Die nicht beobachteten Zusammenhänge zwischen PCK und kognitiver Aktivierung können allerdings nur dann als Hinweis darauf interpretiert werden, dass der PCK-Test kein handlungsrelevantes Wissen erfasst, wenn die Wahrscheinlichkeit dafür, einen Zusammenhang in der Größenordnung des Zusammenhangs zwischen PK und kognitiver Aktivierung ($r_{\text{Pearson}} = .38$, vergl. Abschnitt 8.4 auf Seite 185) nachzuweisen, mindestens über 50% liegt (vergl. hierzu auch Bortz & Döring, 2006, S. 637). Die Teststärke für diesen Fall wurde mit Hilfe des Programms GPower 3.1.7 (Faul, Erdfelder, Lang & Buchner, 2007) berechnet. Sie liegt bei $1 - \beta = 58\%$ (mit $N = 23$, $\alpha_{\text{1-seitig}} = .05$, $r_{\text{Pearson}} = .38$) und damit zwar deutlich unterhalb der gemeinhin als angemessen erachteten Teststärke von 80% (vergl. z. B. Bortz & Lienert, 2008, S. 51), aber immerhin über 50% (was reinem Zufall entsprechen würde).

Teststärken im Rahmen von Mehrebenenmodellierungen zu bestimmen, stellt ein recht kompliziertes Unterfangen dar (vergl. z. B. Nezlek, 2008, S. 855). Für eine grobe Abschätzung der Teststärke soll deshalb auch hier auf die im letzten Abschnitt eingeführten effektiven Leistungszuwächse zurückgegriffen werden (deren Berechnung allerdings die Mehrebenenstruktur der Daten vernachlässigt). Setzt man als Vergleichswert wieder die Korrelation zwischen PK und dem effektiven Leistungszuwachs an, die bei $r_{\text{Pearson}} = .34$ liegt (vergl. Tabelle B.9 auf Seite 248 im Anhang), beträgt die Teststärke 49%. Dass in der vorliegenden Stichprobe keine Zusammenhänge zwischen dem fachspezifischen Professionswissen und Unterrichtserfolg gemessen werden konnten, könnte daher Zufall sein. Auch an dieser Stelle kann allerdings auf die Ergebnisse von Liepertz et al. (2015) verwiesen werden, die auch bei der erweiterten Stichprobe der $N = 35$ Physiklehrkräfte mit ihren Klassen keine positiven Zusammenhänge zwischen dem fachspezifischen Professionswissen der Lehrkräfte und den Fachwissensleistungen der Lernenden zeigen. Die Teststärke liegt in dieser Stichprobe bei 65% (als Vergleichswert wurde erneut $r_{\text{Pearson}} = .34$ verwendet). In der erweiterten Stichprobe zeigt sich außerdem ein negativer Zusammenhang zwischen fachdidaktischem Wissen und Unterrichtserfolg, der sich in der vorliegenden Stichprobe lediglich andeutet.[3]

Unter Berücksichtigung der Ergebnisse von Liepertz scheinen die nicht gefundenen Zusammenhänge zwischen PCK und kognitiver Aktivierung sowie zwischen PCK und Unterrichtserfolg – trotz geringer Teststärke – die Frage nach der prädiktiven Validität des PCK-Tests für gutes und erfolgreiches Unterrichten nicht ganz ungerechtfertigt aufzuwerfen. Bei der Interpretation des nicht signifikanten Zusammenhangs zwischen CK und Unterrichtserfolg muss allerdings beachtet werden, dass selbst in der erweiterten Stichprobe von Liepertz die Wahrscheinlichkeit dafür, einen möglicherweise vorhandenen Zusammenhang nicht nachweisen zu können und die Nullhypothese fälschlicherweise abzulehnen, bei mindestens 35% liegt. Darüber hinaus könnte der Zusammenhang zwischen CK und Unterrichtserfolg aufgrund der Positivauswahl der Lehrkräfte unterschätzt werden, was die Teststärke weiter

[3] Angemerkt sei an dieser Stelle, dass für die Berechnungen bei Liepertz et al. (2015) ein anderes Programm und damit andere Schätzverfahren für die Durchführung der Mehrebenenanalysen benutzt wurden.

herabsetzen würde.

### 9.2.3. Diskussion der Bedeutsamkeit der Varianz im Unterrichtserfolg und in der Unterrichtsqualität

Eine weitere Voraussetzung für eine valide Interpretation der Ergebnisse zum Zusammenhang zwischen Professionswissen und gutem und erfolgreichem Unterrichten ist die Einschätzung ihrer Aussagekraft. Die inhaltliche Diskussion der Ergebnisse wird sich an der Leitfrage orientieren, ob die ProwiN-Professionswissenstests für Physiklehrkräfte Wissen messen, das prädiktiv für die Unterrichtsqualität und den Unterrichtserfolg der hier untersuchten Physiklehrkräfte ist. Die Interpretation der Ergebnisse basiert also darauf, ob durch das Professionswissen der Lehrkräfte Unterschiede zwischen den mittleren Fachwissensleistungen und dem mittleren situationalen Interesse der Lernenden verschiedener Klassen sowie in der kognitiv aktivierenden Gestaltung des Unterrichts aufgeklärt werden können. Um die Ergebnisse in Bezug auf die Leitfrage angemessen diskutieren zu können, muss allerdings sichergestellt werden, dass die in den abhängigen Variablen beobachtete Varianz zwischen den Klassen nicht nur statistisch signifikant sondern auch bedeutsam ist – schließlich hätte eine Aufklärung *unbedeutsamer* Unterschiede zwischen den Klassen durch das Professionswissen der Lehrkräfte nur wenig Relevanz. Im Folgenden werden daher zunächst die deskriptiven Ergebnisse zu den abhängigen Variablen diskutiert, um zu klären, welche Ergebnisse mit Blick auf die eigentliche Leitfrage diskutierenswert erscheinen. An dieser Stelle sei darauf hingewiesen, dass die Interpretation dieser Ergebnisse natürlich ebenfalls den zuvor diskutierten Einschränkungen unterliegt.

#### 9.2.3.1. Schülerfachwissen

Der mittlere Zuwachs im Fachwissen der Lernenden über die Unterrichtseinheit Mechanik entspricht einer halben Standardabweichung im Prä-Test. Der höchste von einer Klasse erreichte mittlere Fachwissenszuwachs entspricht genau einer Standardabweichung im Prä-Test (vergl. Abschnitt 8.2 auf Seite 168). Eine Mittelwertdifferenz von einer Standardabweichung im Prä-Test bedeutet, dass ein Lernender, dessen Personenfähigkeit im Prä-Test dem Mittelwert der Gesamtstichprobe entspricht, im Post-Test lediglich eine Fähigkeit erreicht, die die besten 16% aller Lernenden bereits im Prä-Test erreichten.[4] Vor diesem Hintergrund erscheinen die in der vorliegenden Arbeit gemessenen Fachwissenszuwächse sehr niedrig.

Es stellt sich die Frage, ob die durch die niedrige Reliabilität der Schülertests hervorgerufenen Messungenauigkeiten möglicherweise vorhandene größere Leistungsunterschiede zwischen Prä- und Post-Erhebung überdecken. Diese Frage lässt sich nicht mit Bestimmtheit verneinen. Allerdings kann an dieser Stelle darauf

[4] Basis für diese Überlegung bildet die Eigenschaft der Standardabweichung normalverteilter Merkmale, dass im Intervall $M \pm SD$ um den Mittelwert einer Verteilung gerade 68% der Beobachtungen liegen (vergl. z. B. Bortz, 2005, S. 42).

verwiesen werden, dass in anderen Studien sehr ähnliche Ergebnisse zu Schülerleistungszuwächsen in den naturwissenschaftlichen Unterrichtsfächern und speziell im Unterrichtsfach Physik gefunden wurden:

In der Messwiederholungsstudie von PISA 2003/2004 entsprach der mittlere Zuwachs in der naturwissenschaftlichen Kompetenz von Schülerinnen und Schülern vom Ende der Jahrgangsstufe 9 zum Ende der Jahrgangsstufe 10 lediglich einer viertel Standardabweichung in den Kompetenzen zum ersten Messzeitpunkt ($d = .24$, vergl. Walter, Senkbeil, Rost, Carstensen & Prenzel, 2006, S. 112). Diese Zahlen beziehen sich allerdings auf eine an allen Schulformen erhobene Schülerstichprobe und nicht nur – wie in der vorliegenden Studie – auf Gymnasialschülerinnen und -schüler, was die sogar noch geringere als in der vorliegenden Studie beobachtete Effektstärke erklären könnte. Der Zeitraum zwischen den Messzeitpunkten in PISA 2003/2004 war größer als in der vorliegenden Studie, in der im Mittel lediglich ein halbes Jahr zwischen Prä- und Post-Erhebung lagen (vergl. Abschnitt 8.1.2.1 auf Seite 162). Im Vergleich mit den PISA-Tests scheint der Schülerfachwissenstest Leistungszuwächse daher sogar etwas besser aufzulösen.

Niedrigere Leistungszuwächse zeigte auch die Studie „Bildungsprozesse und psychosoziale Entwicklung im Jugendalter und jungen Erwachsenenalter“ (BIJU): Zwischen den Jahrgangsstufen 7 und 10 entsprach der Leistungszuwachs in den Physikleistungen von Lernenden am Gymnasium 1.7 Standardabweichungen der Leistungen zum ersten Erhebungszeitpunkt (Köller & Baumert, 2008, S. 742) – unter der Annahme einer linearen Leistungsentwicklung würde dies einem Leistungszuwachs von 0.6 Standardabweichungen pro Schuljahr (bzw. 0.3 Standardabweichungen pro Halbjahr) entsprechen.[5]

Für das Bundesland Nordrhein-Westfalen (NRW), in dem die vorliegende Studie durchgeführt wurde, konnten in der QuiP-Studie für Lernende der Jahrgangsstufe 9/10 an allen Schulformen sogar gar keine signifikanten Leistungszuwächse über den Zeitraum eines halben Schuljahres im physikalischen Fachwissen zum Thema „Elektrizitätslehre“ gemessen werden – in Finnland und der Schweiz wurden unter Verwendung des gleichen Testinstruments Leistungszuwächse mit einer Effektstärke von $d = .61$ bzw. $d = .32$ gemessen (Geller, Neumann, Boone & Fischer, 2014, S. 3058-3059). Ergebnisse des IQB-Bundesländervergleichs 2012 zeigten zudem, dass Schülerinnen und Schüler der Jahrgangsstufe 9 an Gymnasien in NRW die zweitniedrigsten Kompetenzstände im physikalischen Fachwissen erreichten und diese signifikant niedriger als der Bundesdurchschnitt waren (Pant et al., 2013, S. 151).

Unter Berücksichtigung dieser Ergebnisse sind für Gymnasialschülerinnen und -schüler (insbesondere im Bundesland NRW) höhere Fachwissenszuwächse als die in der vorliegenden Studie gemessenen gar nicht zu erwarten. Der hier eingesetzte Schülerfachwissenstest scheint das im Rahmen der betrachteten Unterrichtseinheit zu erwerbende Wissen sogar etwas besser abzubilden als andere Testinstrumente.

Problematisch für die Untersuchung des Einflusses von Merkmalen auf Klassenebene, wie dem Professionswissen der Lehrkräfte oder der kognitiv aktivierenden

[5] Nach Köller und Baumert (2008, S. 741) zeigten lineare Wachstumsmodelle in der untersuchten Stichprobe eine zufriedenstellende Passung.

Gestaltung des Unterrichts, ist der verhältnismäßig kleine Anteil der zwischen den Klassen liegenden Varianz an der Gesamtvarianz in den Post-Testwerten der Lernenden. Dieser lag in der vorliegenden Studie bei 10% und verringerte sich nach Berücksichtigung der Kontrollvariablen auf Schülerebene auf 4% (vergl. Abschnitt 8.3.1 auf Seite 171). Im Vergleich zu anderen Studien sind auch diese Werte allerdings wenig überraschend. In der QuiP-Studie betrug der Anteil der zwischen den Klassen liegenden Varianz nach Kontrolle des Prä-Tests, der kognitiven Fähigkeiten und der zuhause gesprochenen Sprache (sowie einiger nicht bedeutsamer Testhefteffekte) zwar noch 13%, zu beachten ist allerdings, dass Ländereffekte und Schulformeffekte hier noch nicht herausgerechnet wurden (Geller, 2015, S. 102). Entsprechende Vergleichszahlen für die naturwissenschaftliche Kompetenz in PISA 2003/2004 oder das physikalische Fachwissen der Lernenden im IQB-Bundesländervergleich liegen leider nicht vor. In der COACTIV-Studie betrug der Anteil der zwischen den Klassen liegenden Varianz an der Gesamtvarianz der mathematischen Kompetenz von Lernenden am Ende der Jahrgangsstufe 10 beachtliche 46% (Baumert et al., 2010, S. 159) – nach Kontrolle der mathematischen Kompetenz am Ende der Jahrgangsstufe 9, der kognitiven Fähigkeiten, der Lesefähigkeit, des sozioökonomischen Status, des Bildungsgrads der Eltern und der Schulform verblieb allerdings auch hier nur noch ein Anteil von 5% der Gesamtvarianz in der mathematischen Kompetenz der Lernenden, der überhaupt durch Prädiktoren auf Klassenlevel erklärt werden kann (Baumert et al., 2010, S. 162).

Die Aussage, dass 10% der Gesamtvarianz in den Post-Testleistungen der Lernenden auf Unterschiede zwischen den Klassen zurück geht, besitzt wenig Aussagekraft, solange die absolute Höhe der Gesamtvarianz nicht inhaltlich interpretiert wird. Um einschätzen zu können, wie bedeutend die Unterschiede in den Fachwissensleistungen der Lernenden auf Klassenebene sind, können die deskriptiven Ergebnisse zu den Post-Testwerten sowie die Ergebnisse zu den Fachwissenszuwächsen der Lernenden in Beziehung gesetzt werden. Die Standardabweichung der Post-Testwerte auf Klassenebene ($SD = 0.4$) ist gleich dem durchschnittlichen Fachwissenszuwachs über die gesamte Unterrichtseinheit Mechanik (vergl. Tabelle 8.6 und Abschnitt 8.2 auf Seite 166 und auf Seite 168). Zwischen den besten und schlechtesten 16% der untersuchten Klassen bestehen also Unterschiede in den mittleren Fachwissensleistungen der Lernenden, die doppelt so groß wie die in dieser Studie gemessenen Leistungszuwächse über die gesamte Unterrichtseinheit sind. Die Klassenunterschiede in den Post-Testleistungen der Lernenden scheinen also durchaus bedeutsam zu sein.

Dieser Vergleich dient allerdings lediglich dazu, die Bedeutsamkeit der Klassenunterschiede in den Fachwissensleistungen der Lernenden grob einschätzen zu können und sollte nicht überinterpretiert werden. Es erfolgt weder eine Fortpflanzung der Messunsicherheiten, noch können systematische Fehler bei der Messung aller Variablen abgeschätzt werden. Auch wenn dies prinzipiell möglich wäre, erfolgt daher an dieser Stelle bewusst keine Umrechnung der Klassenunterschiede in Leistungsvorsprünge gemessen in Schuljahren oder Unterrichtsstunden, wie sie in PISA, COACTIV oder anderen Large-Scale-Studien vorgenommen wird. In Abschnitt 9.2.3.3 auf Seite 204 wird eine solche Umrechnung zwar durchgeführt,

dort wird aber durch die Gegenüberstellung zweier unterschiedlicher Umrechnungsverfahren deutlich gemacht, dass die konkreten Zahlen lediglich dahingehend interpretiert werden sollten, ob sie auf eine Bedeutsamkeit der Unterschiede hinweisen, und nicht in ihrer absoluten Höhe bewertet werden sollten.

#### 9.2.3.2. Situationales Interesse der Lernenden

Der Anteil der Gesamtvarianz im situationalen Interesse, der durch die Klassenzugehörigkeit der Lernenden erklärt werden kann, liegt je nach betrachteter Unterrichtsstunde bei $17 - 20\%$ (vergl. Abschnitt 8.3.2 auf Seite 177). Im PLUS-Projekt ergaben sich ein etwas höherer Anteil von $26\%$ für die zwischen den Klassen liegende Varianz im themenspezifischen situationalen Interesse von Lernenden der Jahrgangsstufe $6$ an Hauptschulen und Gymnasien (Fricke, 2015, S. 169). Im QuiP-Projekt lagen lediglich $3\%$ der Gesamtvarianz im individuellen Fachinteresse Lernender der Jahrgangsstufe 9/10 aller Schulformen zwischen den Klassen (Keller, Neumann & Fischer, 2014, S. 136).

Mit Blick auf die deskriptiven Ergebnisse zum situationalen Interesse der Lernenden stellt sich erneut die Frage nach der Bedeutsamkeit der Klassenunterschiede. Da kaum Unterschiede in den Maßen für die erste Unterrichtsstunde, die zweite Unterrichtsstunde und in den über beide Unterrichtsstunden gemittelten Maßen bestehen, wird exemplarisch die deskriptive Statistik für die über beide Unterrichtsstunden gemittelten Maße diskutiert (vergl. Tabelle 8.7 auf Seite 167).

Der Mittelwert von $M = 4.2$ für das situationale Interesse entspricht der mittleren Kategorie der siebenstufigen Zustimmungsskala, auf der die Lernenden ihr situationales Interesse einschätzen konnten. Werte $\leq 3$ signalisieren eher Ablehnung und können als eher niedriges situationales Interesse interpretiert werden, Werte $\geq 5$ signalisieren eher Zustimmung und damit ein eher höheres situationales Interesse. Auf Klassenebene streuen die Interessensmaße zwischen den Werten $Min = 3.1$ und $Max = 5.1$, was lediglich einer sehr schwachen Tendenz zur Ablehnung oder Zustimmung entspricht. Die Standardabweichung auf Klassenebene $SD = 0.6$ zeigt an, dass in $68\%$ der Klassen die Lernenden im Mittel weder Interesse noch Desinteresse signalisierten. Die Klassenunterschiede im situationalen Interesse der Lernenden scheinen daher nicht wirklich bedeutsam zu sein.

Über die Gründe dafür, dass in dieser Studie keine praktisch bedeutsamen Unterschiede zwischen dem mittleren situationalen Interesse der Lernenden in den Klassen gemessen wurden, kann nur gemutmaßt werden. Es könnte sein, dass das situationale Interesse der Lernenden in der Jahrgangsstufe $8$ und $9$ tatsächlich kaum von Merkmalen auf Klassenebene beeinflusst wird. Sucht man diesbezüglich erneut den Vergleich mit anderen Studien, ergibt sich die Schwierigkeit, dass oftmals lediglich die zwischen den Klassen liegende Varianz berichtet wird, nicht aber die Streuung von Interessensmaßen auf Klassenebene. Bei Betrachtung der deskriptiven Ergebnisse der PLUS-Studie für das themenspezifische situationale Interesse von Lernenden der Jahrgangsstufe $6$ am Gymnasium am Ende einer Unterrichtseinheit zeigt sich ebenfalls keine große Streuung der Maße auf Schülerebene (eingeschätzt wurde das Interesse auf einer vierstufigen Likertskala: $M = 2.52$, $SD = .80$, vergl. Fricke, 2015, S. 137), die Unterschiede in den Klassenmittelwerten könnten ähnlich

gering wie in der vorliegenden Arbeit sein. Auch die in der Studie von Daniels (2008, S. 221) zur Interessenentwicklung im Jugendalter gemessene Streuung des physikbezogenen themenspezifischen Interesses von Lernenden der Jahrgangsstufen 7 und 10 ist ähnlich gering. Interessant ist in diesem Kontext auch der Befund von Kunter, Baumert und Köller (2007), die zwar einen Zusammenhang zwischen den durch Lernende der Jahrgangsstufe 7 und 8 wahrgenommenen Merkmalen der Unterrichtsqualität (untersucht wurden Regelklarheit und Monitoring als Aspekte der Klassenführung) und mathematikbezogenen Interesse auf Individualebene nachweisen konnten, nicht aber auf Klassenebene: Die auf Klassenebene aggregierten Maße klärten keine Varianz auf.

Es könnte auch sein, dass die geringen Unterschiede im situationalen Interesse der Lernenden darauf zurückzuführen sind, dass keine Zufallsstichprobe untersucht wurde. Man könnte vermuten, dass Lehrkräfte, die ihre Klassen als eher uninteressiert und unmotiviert einschätzen, weniger gewillt waren mit diesen Klassen an der Studie teilzunehmen. Allerdings würde man in diesem Fall dennoch erwarten, Klassen mit höheren Ausprägungen im situationalen Interesse zu beobachten, was in der vorliegenden Stichprobe nicht der Fall war. Des Weiteren können auch messtheoretische Gründe für die geringen Klassenunterschiede verantwortlich sein. Zu nennen wäre hier die „Tendenz zur Mitte“ oder aber der Umstand, dass nicht klar ist, welche Bezugsnorm die Lernenden bei der Einschätzung ihres situationalen Interesses wählten. So könnten die Lernenden ihr Interesse am videographierten Unterricht (obwohl dies nicht gefordert wurde) womöglich im Vergleich zu ihrem situationalen Interesse in anderen Unterrichtsstunden eingeschätzt haben, mit dem Resultat, dass sie den videographierten Unterricht im Mittel weder interessanter noch uninteressanter einschätzen als den restlichen Unterricht.

Hinterfragt werden könnte allerdings auch die valide Erfassung des situationalen Interesses. Das situationale Interesse wurde über die vier Items der Interessenskala des Fragebogens zur aktuellen Motivation im Unterricht (FAM-Video) erfasst, der sich in der ursprünglichen Fassung von Rheinberg et al. (2001) auf die aktuelle Motivation von Lernenden bei der Bearbeitung von Aufgaben bezieht (vergl. Abschnitt 7.5.4 auf Seite 128). Nun könnte es sein, dass die geringe Varianz zwischen den Klassen aus der Beschränkung auf die Subskala zum situationalen Interesse resultiert. Auch wenn sie in der vorliegenden Arbeit nicht betrachtet wurden, liegen für die untersuchte Stichprobe auch die Daten aus dem vollständigen Fragebogen vor. Diese zeigen allerdings eine sogar noch geringere Streuung und Spannweite auf Klassenebene. Da die Originalfassung des FAM in der Regel zur Messung der aktuellen Motivation als unabhängiger Variable eingesetzt wird, wurde die Interessenskala bisher insbesondere mit Blick auf ihre prognostische Validität in Bezug auf Lernverhalten und Lernleistungen untersucht. Nimmt man das über beide Unterrichtsstunden gemittelte Maß für das situationale Interesse der Lernenden als zusätzlichen Prädiktor auf Individualebene in das Kontrollvariablenmodell zur Erklärung der Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik auf, trägt dieses nicht signifikant zur Varianzaufklärung auf Schülerebene bei. Die Übersetzung des auf Aufgaben bezogenen FAM in den FAM-Video, der sich auf Unterricht und damit auf ein Klassenmerkmal bezieht, könnte daher zu Einbußen in der Validität geführt haben.

#### 9.2.3.3. Kognitiv aktivierende Unterrichtsgestaltung

Über die Analyse der Zusammenhängen zwischen den Testwerten der Lehrkräfte in den Professionswissenstests und den Qualitätsmaßen zur kognitiv aktivierenden Gestaltung des Unterrichts sollte die prädiktive Validität der ProwiN-Professionswissenstests in Bezug auf Unterrichtsqualität überprüft werden. Hierfür musste zunächst sichergestellt werden, dass mit der kognitiven Aktivierung ein Merkmal erfasst wird, das prädiktiv für Unterrichtserfolg ist. Als wichtigstes Ergebnis zur kognitiv aktivierenden Gestaltung des Unterrichts der untersuchten Lehrkräfte kann daher festgehalten werden, dass die Qualitätsmaße für die erste und zweite Unterrichtsstunde und das über beide Unterrichtsstunden gemittelte Qualitätsmaß unter Kontrolle des Vorwissens, der kognitiven Fähigkeiten der Lernenden, des Geschlechts, der zuhause gesprochenen Sprache sowie der Unterrichtszeit ein signifikanter Prädiktor für das Fachwissen der Lernenden am Ende der Unterrichtseinheit Mechanik sind (vergl. Tabelle 8.8 auf Seite 176). Aus den methodischen Überlegungen in den vorangegangenen Abschnitten ergeben sich keine Hinweise darauf, dass diese Zusammenhänge überschätzt werden. Wie sich die Messfehler der Kontrollvariablen auf die Zusammenhänge auswirken ist allerdings unklar. Dass die Zusammenhänge unabhängig von der Wahl des Qualitätsmaßes zur Varianzaufklärung in den Fachwissensleistungen der Lernenden beitragen, kann allerdings als Hinweis auf die Belastbarkeit der Ergebnisse gewertet werden.

Die Qualitätsmaße für die erste und zweite Unterrichtsstunde leisten allerdings keinen signifikanten Beitrag zur Aufklärung der Klassenunterschiede im situationalen Interesse der Lernenden am Unterricht in der jeweiligen Unterrichtsstunde. In der ersten Unterrichtsstunde ist der Regressionskoeffizient für das Qualitätsmaß zur kognitiven Aktivierung zwar signifikant größer als null, dies gilt jedoch nicht für den entsprechenden Koeffizienten in der zweiten Unterrichtsstunde (vergl. Tabelle 8.10 auf Seite 179). Da im letzten Abschnitt festgestellt wurde, dass die mit der Interessenskala des FAM-Video gemessenen Unterschiede auf Klassenebene keine praktische Relevanz haben, bildet die Ablehnung der Hypothese zum Zusammenhang zwischen kognitiver Aktivierung und situationalem Interesse kein starkes Argument gegen die prädiktive Validität der Qualitätsmaße zur kognitiv aktivierenden Unterrichtsgestaltung in Bezug auf Unterrichtserfolg. Die kognitive Aktivierung, so wie sie in dieser Studie gemessen wurde, kann demnach als Qualitätsmerkmal eines Unterrichts angesehen werden, der in positivem Zusammenhang mit Schülerleistungen im Fachwissen zu stehen scheint. Die Frage, ob sich dieser Unterricht ebenfalls günstig auf das situationale Interesse der Lernenden im Unterricht auswirkt und damit die Grundlage für ein gesteigertes Fachinteresse der Schülerinnen und Schülern am Unterrichtsfach Physik schafft, muss allerdings weiterhin als offen angesehen werden.

In den vorangegangen Überlegungen wurde die kognitive Aktivierung als unabhängige Variable betrachtet. Im Bezug auf die eigentliche Fragestellung stellt sie aber eine abhängige Variable dar. Daher muss auch an dieser Stelle die Bedeutsamkeit der Klassenunterschiede in der kognitiven Aktivierung diskutiert werden.

Bei Betrachtung der deskriptiven Ergebnisse für das über beide Unterrichtsstunden gemittelte Qualitätsmaß (verg. Tabelle 8.5 auf Seite 164) fällt zum einen auf, dass durch die Spannweite der Maße ($Min = 1.2$, $Max = 2.3$) die dreistufige Ratingskala (1 =„trifft nicht zu“, 2 =„teils teils“, 3 =„trifft zu“) nach oben nicht voll ausgenutzt wird. Auch die Standardabweichung der Maße wirkt mit $SD = 0.3$ auf den ersten Blick eher gering, da sie nur 15% der maximal möglichen Spannbreite der Skala abdeckt. Sie ist allerdings in vergleichbarer Größenordnung wie die in der Studie von Vogelsang beobachtete Standardabweichung, die bezogen auf die dort verwendete vierstufige Ratingskala ebenfalls nur 13% der maximal möglichen Spannbreite abdeckte (vergl. Vogelsang, 2014, S. 412).[6]

Eine Standardabweichung in der kognitiv aktivierenden Unterrichtsgestaltung geht in der in dieser Arbeit untersuchten Stichprobe mit einem Unterschied von einer halben Standardabweichung in den Klassenmittelwerten der Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik einher ($\gamma^{\mathrm{StdYX}}_{\mathrm{KA1M\&2M}} = 0.46 \pm 0.20$, vergl. Tabelle 8.8 auf Seite 176). Um die praktische Bedeutung dieses Zusammenhangs zu veranschaulichen, können zwei unterschiedliche Überlegungen angeführt werden. Erstens kann ein Bezug zu den Überlegungen in Abschnitt 9.2.3.1 auf Seite 199 gezogen werden: Eine halbe Standardabweichung in den Post-Testleistungen der Lernenden auf Klassenebene entspricht in etwa der Hälfte des Leistungszuwachs über die gesamte Unterrichtseinheit Mechanik, die im Mittel 34 Unterrichtsstunden umfasste. Zweitens kann der Regressionskoeffizient für den Zusammenhang zwischen kognitiver Aktivierung und Fachwissensleistung mit dem Regressionskoeffizienten der Unterrichtszeit verglichen werden, der fast doppelt so groß ist ($\gamma^{\mathrm{StdYX}}_{\mathrm{Zeit}} = 0.74 \pm 0.12$). Demnach würde eine Standardabweichung im Qualitätsmaß zur kognitiven Aktivierung mit einem Unterschied in den Klassenmittelwerten einhergehen, der in etwa so groß ist, wie ein durch sechs Unterrichtsstunden bedingter Unterschied in den Klassenmittelwerten (vergl.Abschnitt 8.3.1.3 auf Seite 174).

Nun führen diese beiden Betrachtungsmöglichkeiten zu sehr unterschiedlichen Zahlen für die in Anzahl an Unterrichtsstunden umgerechneten Leistungsdifferenzen, die mit einer Standardabweichung im Qualitätsmaß zur kognitiv aktivierende Gestaltung des Unterrichts einhergehen (17 vs. 6 Unterrichtsstunden). Dies ist sicherlich nicht nur dem Umstand geschuldet, dass beim Vergleich mit den Fachwissenszuwächsen keine Berücksichtigung der Kontrollvariablen erfolgt. Es wurde bereits darauf hingewiesen, dass derartige Vergleiche lediglich einer groben Einschätzung der Bedeutsamkeit der Zusammenhänge dienen und die Werte mit Blick auf die methodischen Unsicherheiten nicht überinterpretiert werden sollten. Beide Betrachtungsweisen legen allerdings zumindest nahe, dass die gemessenen Unterschiede in der kognitiv aktivierenden Gestaltung des Unterrichts auch praktische Relevanz haben.

In den vorangehenden Abschnitten wurde der vielfältige Einfluss methodischer Probleme auf die zentralen Ergebnisse der vorliegenden Arbeit diskutiert. Diese

[6] Die Abdeckung der Skala wurde aus der von Vogelsang (2014, S. 412) berichteten Standardabweichung in der Stichprobe der Lehramtsanwärter berechnet.

sollten bei der im nächsten Unterkapitel folgenden inhaltlichen Diskussion der Ergebnisse im Blick behalten werden. Obwohl in Bezug auf einige mögliche Kritikpunkte Gegenargumente formuliert werden konnten, lassen sich auf Grundlage der Ergebnisse keine abschließenden und allgemeingültigen Aussagen über die Relevanz des mit den ProwiN-Professionswissenstests gemessenen Wissens für gutes und erfolgreiches Unterrichten treffen. Aus den Überlegungen im letzten Abschnitt folgt, dass durch Merkmale auf Klassenebene lediglich bedeutsam erscheinende Unterschiede in den Fachwissensleistungen der Lernenden und in der kognitiv aktivierenden Gestaltung des Unterrichts aufgeklärt werden können, während die durch Klassenmerkmale zu erklärenden Unterschiede im situationalen Interesse der Lernenden nicht bedeutsam sind.

## 9.3. Diskussion der zentralen Ergebnisse zur prädiktiven Validität der ProwiN-Professionswissenstests

Die folgende Interpretation der Ergebnisse orientiert sich an der Frage, ob die ProwiN-Professionswissenstests für Physiklehrkräfte Wissen messen, das prädiktiv für die Unterrichtsqualität und den Unterrichtserfolg der hier untersuchten Physiklehrkräfte ist. Unter Berücksichtigung der methodischen Überlegungen werden die Ergebnisse für die drei Professionswissensdimensionen getrennt diskutiert. Dabei werden nur die Ergebnisse diskutiert, die im letzten Abschnitt als praktisch relevant eingestuft wurden. Bezüglich des Zusammenhangs zwischen Professionswissen und Unterrichtserfolg werden daher lediglich die Ergebnisse der Mehrebenenmodelle mit der Fachwissensleistung der Lernenden als abhängiger Variable diskutiert. Der Unterrichtserfolg wird nicht mehr multikriterial modelliert.

### 9.3.1. Fachwissen der Lehrkräfte

> Das Fachwissen der Lehrkräfte hängt signifikant mit der kognitiv aktivierenden Unterrichtsgestaltung zusammen, liefert aber keinen signifikanten Beitrag zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik.

**Rekapitulation der methodischen Überlegungen** Im Rahmen der methodischen Überlegungen ergaben sich keine Hinweise darauf, dass der Zusammenhang zwischen CK und kognitiver Aktivierung in der vorliegenden Stichprobe überschätzt wird. Nicht ganz ausgeschlossen werden kann allerdings, dass es sich um einen zufällig in der untersuchten Stichprobe bestehenden Zusammenhang handelt.

Der nicht beobachtete Zusammenhang zwischen Fachwissen und Unterrichtserfolg muss vorsichtig bewertet werden. Zum einen handelt es sich bei den hier untersuchten Lehrkräften bezüglich des Fachwissens eindeutig um eine Positivauswahl, was zu einer Unterschätzung von Zusammenhängen führen könnte. Zum

anderen könnten Zusammenhänge aufgrund der niedrigen Teststärke in der vorliegenden Stichprobe „übersehen“ werden. Dies gilt selbst für die von Liepertz et al. (2015) untersuchte erweiterte Stichprobe, in der ebenfalls kein Zusammenhang zwischen dem Fachwissen der Lehrkräfte und den Fachwissensleistungen der Lernenden beobachtet wurde. Unklar ist zudem, wie sich die Messungenauigkeiten in den Kontrollvariablen auf diesen Zusammenhang auswirken. Des Weiteren könnten andere, durch das Fachwissen der Lehrkräfte unbeeinflusste Merkmale der Unterrichtsqualität, wie z. B. Klassenführung, einen weitaus größeren Einfluss auf Unterrichtserfolg haben und einen möglicherweise vorhandenen Effekt des Fachwissens überdecken. Dies zu prüfen, war in der vorliegenden Arbeit nicht möglich.

**Mögliche inhaltliche Interpretationen** Die Ergebnisse der SII-Studie zum Fachwissen von Mathematikgrundschullehrkräften legen einen nicht-linearen Effekt des Fachwissens auf Unterrichtserfolg nahe – oberhalb eines gewissen Mindestmaßes an Fachwissen ließen sich keine Zusammenhänge zwischen dem Fachwissen der Lehrkräfte und Schülerleistungen beobachten (Hill et al., 2005, S. 396). Auch die Ergebnisse der PLUS-Studie (Ohle et al., 2011) könnten hiermit erklärt werden (vergl. Abschnitt 4.3 auf Seite 41). In der Studie von Sadler et al. (2013) zeigten sich Zusammenhänge zwischen dem Fachwissen von Middle-School-Lehrkräften und dem Fachwissen ihrer Schülerinnen und Schülern der Jahrgangsstufe 7 und 8 nur auf Aufgabenebene, nicht aber auf Testebene (Lehrkräfte und Lernende bearbeiteten die gleichen Aufgaben) – also lediglich bezüglich sehr konkreter Fachinhalte. Der letzte Befund könnte ein Hinweis darauf sein, dass ein Mindestmaß an Fachwissen, das zum erfolgreichen Unterrichten notwendig sein könnte, auf dem Niveau des im Unterricht vermittelten Wissens liegt.

Unter dieser Annahme würde für die in der vorliegenden Arbeit untersuchte Schülerstichprobe in der Jahrgangsstufe 8 und 9, insbesondere das Schulwissen der Lehrkräfte auf Sekundarstufen-I-Niveau eine Rolle für die Fachwissensleistungen der Lernenden spielen. In diesem Kontext wäre eine mögliche Interpretation für das Ergebnis der vorliegenden Arbeit, dass der ProwiN-Fachwissenstest lediglich oberhalb des für erfolgreiches Unterrichten notwendigen Mindestmaß an Fachwissen differenziert. Im ProwiN-Professionswissenstest wurde zwar kein universitäres Wissen abgefragt, neben Schulwissen, das im Unterricht in der Sekundarstufe I vermittelt wird, wurde allerdings auch Schulwissen auf Oberstufenniveau sowie vertieftes Schulwissen abgefragt. Es könnte daher sein, dass der Fachwissenstest im Bereich des Schulwissens auf Sekundarstufen-I-Niveau nicht ausreichend differenziert. Möglich wäre allerdings auch, dass aufgrund der Positivauswahl alle Lehrkräfte der Stichprobe über das Mindestmaß an Fachwissen verfügen, oberhalb dessen kein Zusammenhang zur Schülerleistung mehr zu erwarten wäre. Anzumerken ist außerdem, dass im Fachwissenstest zwar primär Wissen aus dem Inhaltsbereich Mechanik erfasst wird, sich die Aufgaben allerdings nicht unmittelbar auf die von den Lehrkräften in der Unterrichtseinheit Mechanik behandelten Themen beziehen. Letzteres könnte mit Blick auf die Ergebnisse von Sadler et al. (2013) ebenfalls ein möglicher Grund dafür sein, dass in der vorliegenden Stichprobe keine Hinweise

auf die Relevanz des mit dem ProwiN-Fachwissenstest erhobenen Wissens für Unterrichtserfolg gefunden wurden.

Das mit dem ProwiN-Fachwissenstest gemessene Wissen – also auch über das im Unterricht vermittelte Fachwissen hinausgehendes Wissen – scheint allerdings Einfluss auf die Unterrichtsqualität zu haben. Die Lehrkräfte der Stichprobe, die über ein höheres Fachwissen verfügen, scheinen eher in der Lage zu sein, ihren Unterricht kognitiv aktivierend zu gestalten: Sie schaffen herausforderndere Lerngelegenheiten, stellen eher Verbindungen zu bereits Gelerntem und neu zu Lernendem her und zeigen den Lernenden auch Ungereimtheiten in ihren Vorstellungen auf. Darüber hinaus scheinen sie sich kompetent genug zu fühlen, ihren Unterricht weniger rezeptiv zu organisieren und die Denkweisen der Lernenden zu ergründen – und ihren Unterricht damit weniger vorhersehbar zu gestalten. Hier zeigt sich ein wesentlicher Unterschied zu der Studie von Vogelsang (2014): Zwischen dem mit dem Paderborner Testinstrument erfassten Fachwissen, das auch universitäres Wissen einschloss, und kognitiver Aktivierung konnte kein Zusammenhang nachgewiesen werden.

Die nicht beobachteten Zusammenhänge zum Unterrichtserfolg scheinen nicht dadurch bedingt zu sein, dass das in der vorliegenden Arbeit gemessene Fachwissen grundsätzlich keine Handlungsressource für die Lehrkräfte bilden kann. Möglich wäre, dass das Fachwissen der Lehrkräfte, das über das im Unterricht zu vermittelnde Wissen hinaus geht, zwar noch indirekt über die kognitiv aktivierende Gestaltung des Unterrichts auf Unterrichtserfolg wirkt, dieser Einfluss allerdings nicht groß genug ist, um sich bis auf die Zielkriterien von Unterricht auszuwirken.

### 9.3.2. Fachdidaktisches Wissen der Lehrkräfte

> Das fachdidaktische Wissen der Lehrkräfte hängt weder signifikant mit der kognitiv aktivierenden Unterrichtsgestaltung zusammen noch liefert es einen signifikanten Beitrag zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik.

**Rekapitulation der methodischen Überlegungen** Die durch die niedrigere Reliabilität des PCK-Tests bedingten Messfehler sowie die recht geringe Streuung der PCK-Testwerte könnten zu einer Unterschätzung des Zusammenhangs zwischen fachdidaktischem Wissen und kognitiver Aktivierung führen. Es konnte allerdings gezeigt werden, dass selbst bei Berücksichtigung der Messfehler das mit dem PCK-Test erfasste Wissen geringer mit der kognitiven Aktivierung korreliert als das Fachwissen und pädagogische Wissen der Lehrkräfte. Nicht eingeschätzt werden kann, wie sich die Messfehler im PCK-Test und in den Kontrollvariablen auf das Ergebnis für den Zusammenhang zwischen fachdidaktischem Wissen und Unterrichtserfolg auswirken. Auch muss erneut auf die Problematik der niedrigen Teststärke bei der Interpretation nicht signifikanter Zusammenhänge hingewiesen werden. In dem Wissen, dass die sich andeutenden negativen Effekte des fachdidaktischen Wissens auf die Fachwissensleistungen der Lernenden am Ende der

Unterrichtseinheit Mechanik in der hier untersuchten Stichprobe in der erweiterten, von Liepertz et al. (2015) untersuchten Stichprobe signifikant werden, erscheint es allerdings recht unwahrscheinlich, dass in einer genaueren Messung und bei höherer Teststärke ein positiver Zusammenhang gemessen werden könnte.

Es wäre möglich, dass der Effekt des fachdidaktischen Wissens auf Unterrichtserfolg erst unter Kontrolle der Klassenführung beobachtet werden kann, wie es in der PLUS-Studie der Fall war (vergl. Lange, 2010). Dies konnte in der vorliegenden Arbeit nicht untersucht werden. Klassenführung stellt auch auch eine wichtige Voraussetzung für die Sicherung anspruchsvollen und kognitiv aktivierenden Unterrichts dar (vergl. Helmke, 2009, S. 174; Klieme et al., 2001, S. 53). Ein Zusammenhang zwischen CK und kognitiver Aktivierung konnte allerdings *ohne* Kontrolle der Klassenführung beobachtet werden. Es erscheint daher unwahrscheinlich, dass der Zusammenhang zwischen PCK und kognitiver Aktivierung lediglich unter Kontrolle der Klassenführung nachweisbar ist – schließlich sollte das PCK der Lehrkräfte sogar stärker mit kognitiver Aktivierung zusammenhängen als ihr CK (vergl. Abschnitt 5.3.4 und Abschnitt 6.2 auf Seite 71 und auf Seite 77).

Zusammengenommen könnten die Ergebnisse unter Berücksichtigung der methodischen Einschränkungen Hinweise darauf liefern, dass mit dem PCK-Test kein Wissen erhoben wird, das als relevant für gutes oder erfolgreiches Unterrichten erachtet werden kann.

**Mögliche inhaltliche Interpretationen** Noch besteht Uneinigkeit darüber, wie fachdidaktisches Wissen zu modellieren ist und welche und wie viele Wissensfacetten als relevant für erfolgreiches Unterrichten erachtet werden (vergl. Abschnitt 2.3.2 auf Seite 16). Die im ProwiN-Projekt für die Modellierung von PCK genutzten Facetten *Wissen über Schülervorstellungen* und *Wissen über Instruktionsstrategien und Repräsentationen* (letzteres wurde in ProwiN über das Wissen über Experimente sowie Wissen über Konzepte operationalisiert) stellen aber zumindest einen Konsens dar – sie finden sich in nahezu allen Modellierungen wieder (vergl. Tabelle 2.1 auf Seite 18). Dies gilt insbesondere auch im deutschsprachigen Raum. Auch wenn Einigkeit über die Wichtigkeit dieser Facetten besteht, handelt es sich dennoch lediglich um normativ gesetzte Facetten, deren Relevanz für erfolgreiches Unterrichten nicht empirisch abgesichert ist.

Die Ergebnisse der vorliegenden Arbeit könnten die Frage aufwerfen, ob die Modellierung des schriftlich abprüfbaren fachdidaktischen Wissens überdacht werden muss. Diese Frage würde sich gleich auf zwei Ebenen stellen: So könnten die nicht gefundenen Zusammenhänge ein Hinweis darauf sein, dass für die Beschreibung des fachdidaktischen Wissens nicht die richtigen Facetten ausgewählt wurden, oder aber die hier vorgenommene Operationalisierung der Facetten das Problem darstellt. Andere Studien, wie die PLUS-Studie (Lange, 2010), die QuiP-Studie (Ergönenç et al., 2014) oder die Studie von Vogelsang (2014) liefern heterogene und bisher nicht eindeutige Ergebnisse zum Zusammenhang zwischen fachdidaktischem Wissen von Physiklehrkräften und Unterrichtsqualität oder Unterrichtserfolg. Da die Modellierung des fachdidaktischen Wissens in diesen Studien unterschiedliche Facetten beinhaltet, diese aber zum Teil überlappen, ist es, wie bereits erwähnt

wurde, nicht möglich Rückschlüsse auf die Relevanz einzelner Facetten zu ziehen. Aufschluss hierüber können die Ergebnisse der ProwiN-Videostudie in den anderen naturwissenschaftlichen Fächern liefern, da im ProwiN-Projekt ein gemeinsames Modell für die Entwicklung der Testinstrumente verwendet wurde und demnach die selben Facetten in den PCK-Test berücksichtigt wurden. Sollten die in der Chemie oder der Biologie eingesetzten Tests zum fachdidaktischen Wissen prädiktiv valide für gutes und erfolgreiches Unterrichten sein, könnte das ein Hinweis auf die Relevanz der Facetten und darauf sein, dass im Testinstrument für Physiklehrkräfte innerhalb der Facetten nicht das relevante Wissen adressiert wurde. Umgekehrt würde, sofern sich auch in den anderen Fächern das gemessene Wissen nicht als relevant für gutes und erfolgreiches Unterrichten herausstellen sollte, die Relevanz der Facetten aus dem ProwiN-Modell in Frage gestellt.

Ein weiterer Aspekt, der an dieser Stelle diskutiert werden sollte, ist die in der Studie von Sadler et al. (2013) zur Sprache gebrachte „grain size“:

> The reason that many prior studies of the influence of teacher knowledge on student learning may not have found significant effects may lie, at least partially, in their painting with too broad a brush. The grain size of analysis of teachers' knowledge may be important. Our own initial analysis of total test scores (not shown) captured neither the nuances of a teacher's strengths and weaknesses nor the effects that these nuances have on student learning. (S. 1041)

Die Autoren konnten zeigen, dass Lernende von Lehrkräften, die typische Fehlvorstellungen in Antwortmöglichkeiten eines Multiple-Choice-Tests erkannten, diese falschen Antwortmöglichkeiten weniger häufig ankreuzten. Ähnlich wie im Falle des Fachwissens könnte es sein, dass Effekte auf den Unterrichtserfolg nur dann groß genug sind um beobachtet zu werden, wenn das erfasste Wissen in wesentlich engerem Bezug zu den konkret unterrichteten Inhalten steht, PCK also noch weitaus themenspezifischer betrachtet werden muss, als dies in der Regel mit den für Large-Scale-Studien entwickelten Testinstrumenten geschieht. Ähnliche Überlegungen finden sich auch in dem kürzlich veröffentlichten Professionswissensmodell von Gess-Newsome (2015, S. 31), das unter Beteiligung zahlreicher internationaler Professionswissensforscher entwickelt wurde: PCK wird hier explizit als themenspezifisches Wissen modelliert.

### 9.3.3. Pädagogisches Wissen der Lehrkräfte

| |
|---|
| Das pädagogische Wissen der Lehrkräfte hängt signifikant mit der kognitiv aktivierenden Unterrichtsgestaltung zusammen und liefert einen signifikanten Beitrag zur Aufklärung der Varianz in den Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik. |

**Rekapitulation der methodischen Überlegungen** Im Rahmen der methodischen Überlegungen ergaben sich keine Hinweise darauf, dass die Zusammenhänge

zwischen dem pädagogischen Wissen und kognitiver Aktivierung in der vorliegenden Stichprobe überschätzt werden. Auch ist es unwahrscheinlich, dass der beobachtete Zusammenhang lediglich daraus resultiert, dass das pädagogische Wissen einen Einfluss auf die Klassenführung hat, die ihrerseits als Voraussetzung für kognitive Aktivierung angesehen wird. Würde die Klassenführung eine konfundierende Variable für den Zusammenhang zwischen PK und kognitiver Aktivierung darstellen, müsste sie mit der kognitiven Aktivierung korrelieren, was nicht der Fall ist (vergl. Tabelle 7.21 auf Seite 156). Unklar ist allerdings, wie sich Messungenauigkeiten in den Kontrollvariablen auf den Zusammenhang zwischen pädagogischem Wissen und den Fachwissensleistungen der Lernenden auswirken.

**Mögliche inhaltliche Interpretationen** Obwohl lediglich deklaratives pädagogisches Wissen erhoben wurde, dem im Vergleich zum konditional-prozeduralen pädagogischen Wissen eine geringere Bedeutung für die Handlungsrelevanz zugesprochen wird (Lenske et al., 2016), tragen die PK-Testwerte der Lehrkräfte in der untersuchten Stichprobe zur Varianzaufklärung der Fachwissensleistungen der Lernenden am Ende der Unterrichtseinheit Mechanik bei und stehen in Zusammenhang mit der kognitiv aktivierenden Gestaltung des Unterrichts. Der Einfluss des pädagogischen Wissens auf Unterrichtsqualität im Physikunterricht wurde bisher lediglich in der Studie von Vogelsang (2014) untersucht, in der ebenfalls positive Zusammenhänge beobachtet werden konnten (die Aussagekraft dieser Ergebnisse ist allerdings begrenzt, vergl. Abschnitt 4.3 auf Seite 46). Zum Zusammenhang zwischen pädagogischem Wissen und Unterrichtserfolg im Physikunterricht existieren bisher keine Vergleichsstudien.

Da das pädagogische Wissen der Lehrkräfte zur Varianzaufklärung in den Fachwissensleistungen der Lernenden beiträgt, scheint es grundsätzlich möglich zu sein, dass Zusammenhänge zwischen explizierbarem Professionswissen und Unterrichtserfolg bestehen. Dass in der vorliegenden Studie Zusammenhänge zwischen dem pädagogischen und damit fachunabhängigen Wissen der Lehrkräfte, nicht aber zwischen dem fachspezifischen Professionswissen der Lehrkräfte und Unterrichtserfolg beobachtet werden können, könnte ein weiterer Hinweis darauf sein, dass, sobald der Einfluss fachspezifischen Wissens untersucht wird, ein noch themenspezifischerer Fokus gewählt werden muss. Es könnte allerdings auch sein, dass der Einfluss des fachspezifischen Professionswissens erst dann wirksam wird, wenn Lehrkräfte gleichzeitig über genug pädagogisches Wissen verfügen, um die für Unterrichtserfolg notwendigen Rahmenbedingungen im Unterricht zu schaffen.

## 9.4. Fazit und Ausblick

Im Kapitel zur Ableitung des eigenen Forschungsansatzes wurde argumentiert, dass trotz zahlreicher offener Fragen bezüglich der Modellierung von Professionswissen und trotz fehlendem Konsens darüber, ob überhaupt Zusammenhänge zwischen explizierbarem Wissen und Handeln bestehen, die Untersuchung der Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg eine der wenigen Möglichkeiten darstellt, herauszufinden, welches Wissen relevant für gutes

und erfolgreiches Unterrichten ist. Die vorliegende Arbeit hat allerdings deutlich gemacht, welche Probleme die Untersuchung dieser Zusammenhänge in sich birgt: Eindeutige Aussagen über die Relevanz des mit den ProwiN-Testinstrumenten gemessenen Wissens können auf Basis der Ergebnisse der hier vorgestellten Studie nicht getroffen werden.

**Fachwissen** Insbesondere die Ergebnisse zum Fachwissen sind nicht eindeutig. Einerseits deutet sich aufgrund des Zusammenhangs zur kognitiv aktivierenden Gestaltung des Unterrichts die Handlungsrelevanz des erfassten Wissens an. Andererseits scheint der Fachwissenstest aber nicht ausreichend in dem für erfolgreiches Unterrichten relevanten Wissensbereich zu differenzieren und Unterschiede im Fachwissen der in der vorliegenden Arbeit untersuchten Gymnasiallehrkräfte, die möglicherweise mit größerer Effektstärke auf Unterrichtserfolg wirken könnten (und daher auch in kleineren Stichproben beobachtbar sein müssten) nicht zu erfassen. Es können allerdings keine Rückschlüsse darauf gezogen werden, ob der ProwiN-Fachwissenstest prinzipiell nicht in der Lage ist, derartige Unterschiede aufzulösen oder ob dies lediglich für die in der vorliegenden Arbeit untersuchte Stichprobe der Fall ist, die bezüglich ihres Fachwissens eine Positivauswahl darstellt. Belastbare Aussagen zur prädiktiven Validität des Fachwissenstests können daher nicht getroffen werden.

**Fachdidaktisches Wissen** In Bezug auf das fachdidaktische Wissen der Lehrkräfte wurden keine Zusammenhänge zur Unterrichtsqualität und zum Unterrichtserfolg gefunden. Diese Ergebnisse werfen die folgenden Fragen auf:

- Ist das im ProwiN-Test abgefragte Wissen nicht relevant für gutes und erfolgreiches Unterrichten?
- Erfasst der Test das fachdidaktische Wissen von Gymnasiallehrkräften nicht reliabel genug, um Unterschiede im Wissen der Lehrkräfte aufzulösen, die womöglich mit Unterschieden in der Qualität ihres Unterrichts oder im Lernerfolg ihrer Schülerinnen und Schüler einhergehen?
- Oder aber: Hat schriftlich abprüfbares fachdidaktisches Wissen tatsächlich keinen Einfluss auf Unterrichtsqualität und Unterrichtserfolg?

Unabhängig von den Antworten auf diese Fragen, scheint es nicht möglich zu sein, auf Basis der PCK-Testwerte Rückschlüsse darauf zu ziehen, ob eine Lehrkraft über Wissen verfügt, das als Voraussetzung für guten und erfolgreichen Unterricht angesehen werden kann. Dass einzig die niedrige Teststärke und die geringe Streuung der PCK-Testwerte in dieser Studie für die nicht beobachteten Zusammenhänge verantwortlich ist, kann zwar nicht ausgeschlossen werden, erscheint aber mit Blick auf die von Liepertz et al. (2015) berichteten Ergebnisse nicht sehr wahrscheinlich.

Der PCK-Test wurde in der ersten Projektphase des ProwiN-Projektes über Expertenbefragungen, Abgleich mit Fachcurricula, den Vergleich bekannter Gruppen mit zu erwartenden Fähigkeitsunterschieden und durch Zusammenhangsanalysen zwischen den anderen Dimensionen des Professionswissens validiert (vergl.

Abschnitt 5.1.2 auf Seite 58). Die Ergebnisse der vorliegenden Arbeit könnten daher als Hinweise darauf gewertet werden, dass diese „herkömmliche“ Validierung, auf die sich viele Studien bei der Validierung von Professionswissenstests beschränken, nicht auszureichen scheint. Dies gilt zumindest dann, wenn ein solches Testinstrument mit dem Ziel eingesetzt wird, Maßnahmen zur Vermittlung handlungsrelevanten Wissens zu evaluieren, die Lehrkräfte dazu befähigen sollen guten und erfolgreichen Unterricht zu gestalten.

**Pädagogisches Wissen** Die Ergebnisse zum pädagogischen Wissen könnten darauf hinweisen, dass dieses Wissen eine wesentliche Rolle für gutes und erfolgreiches Unterrichten spielt. Das ProwiN-Testinstrument scheint das handlungsrelevante Wissen von Lehrkräften valide abzubilden. Anders als im Fall des fachdidaktischen Wissens, wo die Auswahl der als wichtig erachteten Wissensfacetten eher normativ erfolgt, kann bei der Modellierung des pädagogischen Wissens auf die umfangreichen empirischen Befunde aus der Prozess-Produkt-Forschung zurückgegriffen werden. Bezüglich der Identifizierung handlungsrelevanten pädagogischen Wissens scheint die Professionswissensforschung daher weiter zu sein als im Falle des fachspezifischen Professionswissens.

### 9.4.1. Empfehlungen für künftige Untersuchungen

Im Zuge der Diskussion der Ergebnisse wurden immer wieder Bezüge zu den Ergebnissen anderer Studien gezogen. Viele dieser Studien haben allerdings ähnliche methodische Probleme wie die vorliegende Untersuchung – sei es bezüglich der Reliabilität der eingesetzten Professionswissens- oder Schülertests, bezüglich der Stichprobengrößen und Stichprobenziehung oder bezüglich der designbedingten Einschränkungen, die oftmals keine Untersuchung kausaler Zusammenhänge ermöglichen. Die Aussagekraft der Ergebnisse ist in diesen Studien in vielen Fällen in ähnlicher Weise eingeschränkt wie in der vorliegenden Arbeit. Darüber hinaus wird das Professionswissen der Lehrkräfte in allen Studien unterschiedlich modelliert und erfasst (vergl. Abschnitt 4.3 auf Seite 46).

Um einen kumulativen Erkenntnisgewinn bezüglich der Bedeutsamkeit des Professionswissens von Physiklehrkräften (bzw. des in den einzelnen Testinstrumenten abgefragten Wissens) für gutes und erfolgreiches Unterrichten zu ermöglichen, sollte versucht werden, die Ergebnisse dieser Studien systematisch zu replizieren. Dies mag zwar auf den ersten Blick nicht interessant erscheinen, wäre aber unerlässlich, wenn belastbare Aussagen getroffen werden sollen.

Würde man diesen Weg beschreiten, müsste dabei auch ein besonderes Augenmerk auf die „Hilfsinstrumente“ gelegt werden, also die Messinstrumente zur Erhebung der abhängigen Variablen wie beispielsweise Schülertests oder Instrumente zur Erhebung der Unterrichtsqualität – schließlich ist eine Argumentationskette immer nur so stark wie ihr schwächstes Glied.

In Bezug auf Schülerfachwissenstests sind insbesondere die niedrigen Reliabilitäten bei Prä-Testerhebungen ein Problem (vergl. z. B. Geller, 2015, S. 96; Ohle, 2010, S. 86; Sadler et al., 2013, S. 1031). Da Schülertests meist nicht im Fokus einer Arbeit stehen, wird wesentlich weniger Zeit und Muße in deren Entwicklungsarbeit gesteckt. Schülerfachwissen stellt zudem kein eng definiertes psychologisches Konstrukt dar, was eine reliable Messung erschwert. So zeigen

selbst die in großen Schulleistungsstudien wie PISA 2003 eingesetzten Instrumente keine zufriedenstellenden Reliabilitäten, wenn nicht zahlreiche Kontrollvariablen in Hintergrundmodellen berücksichtigt werden (vergl. z. B. Walter et al., 2006, S. 98). Multi-Matrix-Designs bei der Testheftbearbeitung, wie sie in der vorliegenden Arbeit genutzt wurden, verstärken diese Problematik noch (Linacre, 2011, S. 618).

Ein weiteres kritisches Element stellen die videobasierten Instrumente zur Erfassung der Unterrichtsqualität dar. In der Regel werden große Anstrengungen unternommen, die Objektivität dieser Messinstrumente sicherzustellen. Wie auch im Falle der vorliegenden Studie ist dies allerdings nicht immer eine Garantie für eine objektive Messung. Weitaus seltener wird die Validität solcher Messinstrumente diskutiert. In der vorliegenden Arbeit wurde versucht, das Videorating zur kognitiven Aktivierung zumindest in Ansätzen zu validieren. Derartige Bemühungen sollten in Replikationsstudien ebenfalls weiter verfolgt werden.

Um zu belastbaren Aussagen zu gelangen, müssten darüber hinaus Wege gefunden werden, wie eine Untersuchung von größeren Zufallsstichproben realisiert werden könnte. Der Einfluss von Störvariablen kann in Felduntersuchungen nicht gänzlich ausgeschlossen werden – die Untersuchung größerer Stichproben ermöglicht es allerdings, diese zumindest mit Hilfe statistischer Verfahren zu kontrollieren. In größeren Stichproben als der hier untersuchten Stichprobe könnten so beispielsweise auch Einflüsse des fachspezifischen Professionswissens unter Kontrolle des pädagogischen Wissens oder unter Kontrolle allgemeinpädagogischer Merkmale der Unterrichtsqualität untersucht werden.

Zu leisten sind diese Punkte sicherlich nicht im Rahmen einzelner Doktorarbeiten. Vielmehr müssten derartige Vorhaben von größeren Forschungsverbünden durchgeführt werden, die Mittel und Wege haben, große Untersuchungen zu realisieren, zu deren Teilnahme Lehrkräfte verpflichtet werden könnten, und die darüber hinaus dazu bereit sein müssten, sich mit einem nur langsam voranschreitenden – aber kumulativen - Erkenntnisgewinn zu begnügen.

Mit Blick auf die etwas vielversprechenderen Ergebnisse zur Relevanz des pädagogischen Wissens, stellt sich die Frage, ob in Bezug auf die Identifikation fachspezifischen Wissens, das relevant für gutes und erfolgreiches Unterrichten sein könnte, einen Schritt zurück gegangen werden sollte. Anstatt die Relevanz des aus normativen Gesichtspunkten für wichtig erachteten Wissen zu untersuchen, könnte versucht werden, dieses Wissen im Rahmen eines modifizierten Expertenansatzes zu identifizieren. Eine Möglichkeit dies zu tun, wäre auf Basis von Ergebnissen zum Unterrichtserfolg eine Einteilung von Lehrkräften in erfolgreich und weniger erfolgreich Unterrichtende vorzunehmen. Der Unterricht dieser Lehrkräfte könnte in Bezug auf unterschiedliche Qualitätsmerkmale verglichen werden, und es könnte versucht werden, kritische Unterrichtssituationen zu identifizieren, in denen erfolgreiche und weniger erfolgreiche Lehrkräfte unterschiedliche Reaktionen zeigen. Aus diesen Beobachtungen ließen sich möglicherweise Rückschlüsse darauf ziehen, welches Wissen als Handlungsressource in den entsprechenden Situationen gedient haben könnte. Die sich hieraus ergebenen Erkenntnisse könnten einerseits Ansatzpunkte für die Entwicklung neuer Testinstrumente liefern, andererseits könnte im Rahmen von Interventionsstudien untersucht werden, ob über die Vermittlung des so identifizierten Wissens ein Beitrag zur Ausbildung guter und erfolgreicher Lehrkräfte geleistet werden kann.

# Appendizes

# A. Manuale und Testhefte

## Inhalt

A.1. Testleitermanuale . . . . . 218
A.1.1. Prä-Erhebung . . . . . 218
A.1.2. Post-Erhebung . . . . . 219
A.1.3. Fachspezifisches Professionswissen . . . . . 220
A.2. Schülerfachwissenstest . . . . . 221
A.3. Professionswissenstests . . . . . 221
A.4. Ratingmanual zur kognitiven Aktivierung . . . . . 222

# A.1. Testleitermanuale

## A.1.1. Prä-Erhebung

**Testleitermanual ProwiN-Prä-Test:**

Lesen Sie alles Kursive wortwörtlich vor, um zu gewährleisten, dass alle Testungen in gleicher Weise ablaufen. Handlungsanweisungen sind fettgedruckt.

**Vor Beginn der Testung:**

**Warten bis alle Schüler Platz genommen haben und zur Ruhe gekommen sind.**

**Sinngemäß wiedergeben:**
In der Forschung ist es immer ganz wichtig, dass alles korrekt abläuft und Testungen immer gleich durchgeführt werden. Deswegen muss ich euch alle Erklärungen, die gleich folgen, vorlesen. Wundert euch also nicht darüber.

*Mein Name ist _____ und das ist _____, wir kommen von der Universität Duisburg-Essen.*
*Wir wollen herausfinden, was genau im Lehramtsstudium an den Universitäten unterrichtet werden muss, damit der Physikunterricht für euch besser gestaltet werden kann. Um diesem Ziel näher zu kommen, müssen wir einige Untersuchungen durchführen, und dafür brauchen wir eure Hilfe.*
*Heute werdet ihr zwei Tests ausfüllen. Zuerst einen ganz kurzen Test zu euren kognitiven Fähigkeiten. Danach bekommt ihr den eigentlichen Test zur Mechanik. Dazu erzähl ich euch dann gleich etwas.*
*Bei beiden Tests müsst ihr nur ankreuzen. Stifte bekommt ihr von uns. Die müssen wir nachher aber wieder einsammeln.*
*Es ist ganz wichtig, dass ihr die Aufgaben allein bearbeitet. Wenn ihr voneinander abschreibt, verfälscht ihr die Daten und wir haben ein großes Problem, weil wir falsche Ergebnisse bekommen. Niemand erfährt eure Ergebnisse, weder eure Lehrer noch eure Eltern, also bitte bitte alleine arbeiten.*
*Für den ersten Test bekommt ihr jetzt zwei Testhefte. Ein Heft mit den Aufgaben und ein Heft, in das ihr eure Antworten eintragen könnt. Das Aufgabenheft müssen wir wiederverwenden, deswegen schreibt bitte nichts darein. Lasst die Hefte bitte erstmal zu!*

**Testhefte austeilen.**

*Ihr könnt jetzt den Antwortbogen aufschlagen. Beim Ausfüllen ist es wichtig, dass ihr die Kästchen sauber und mittig ankreuzt. Die Hefte werden später eingescannt und können nicht ausgewertet werden, wenn ihr nicht ordentlich ankreuzt. Wenn ihr euch vertan habt und eine Antwort verändern wollt, malt das falsch angekreuzte Kästchen vollständig aus, kreuzt die richtige Antwort an. Die richtige Antwort könnt ihr zusätzlich auch noch einkringeln, dann ist es ganz eindeutig. Füllt jetzt bitte die Angaben oben auf der Seite aus. Bei der Testform könnt ihr Testform A ankreuzen.*

**Kurz warten.**

*So, schlagt bitte die erste Seite vom Aufgabenheft auf. Wir gehen jetzt das Beispiel gemeinsam durch.*

**Erste Seite vom KFT Aufgabenheft vorlesen und Beispiel durchgehen.**

*Gibt es noch Fragen?*

**Kurz warten. Fragen gegebenenfalls beantworten.**

*Ok, dann könnt ihr jetzt anfangen!*

**Zeit stoppen (8min). Anfangszeit KFT ins Protokoll eintragen.**

*So, die Zeit ist um. Die Testhefte bitte zumachen. Zieht die Klebezettel mit euren Namen bitte ab. Wir sammeln die Tests jetzt ein und teilen den Schülertest aus. Die neuen Testhefte bitte noch nicht aufschlagen.*

**Testhefte einsammeln und Schülertesthefte austeilen.**

**(2. Person kann während der folgenden Erklärung PK Bogen an Lehrkraft aushändigen)**

*Der Test, den ihr jetzt bearbeitet, behandelt das Thema Mechanik. Ihr hattet das Thema noch nicht, deswegen kann es gut sein, dass ihr einige Fragen noch nicht beantworten könnt. Das ist aber nicht schlimm. Wir wollen einfach nur wissen, was ihr jetzt schon zu dem Thema wisst. Der Test besteht aus Multiple Choice Aufgaben, bei denen immer nur eine Antwort richtig ist. Auch hier gilt wieder: sauber und mittig ankreuzen. Wenn ihr euch vertan habt und eine Antwort verändern wollt, malt das falsch angekreuzte Kästchen vollständig aus und kreuzt die richtige Antwort an und macht einen Kringel darum.. Ein Beispiel dafür seht ihr auch gleich im Testheft. Versucht bitte alle Aufgaben zu beantworten. Auch wenn ihr eine Antwort nicht wisst oder euch nicht sicher seid, setzt bitte bei jeder Aufgabe ein Kreuz. Ihr habt 30 min Zeit.*

**Kurz warten.**

*Die Zeit läuft jetzt!*

**Zeit stoppen (30min).**

***Anfangszeit ins Testprotokoll eintragen.***

***Zeit notieren, nach der die erste Abgabe erfolgt. Schüler, die fertig sind, bekommen Beschäftigungsblatt (Sudoka+Mandala). Stifte werden zusammen mit Testheften eingesammelt.***

***Zeit notieren, nach der der PK Bogen abgegeben wird.***

**Nach Ablauf der Zeit:**

*Die 30 min sind jetzt um. Klappt das Testheft jetzt bitte zu.*

***Am Ende für die Mitarbeit bedanken!!!***

## A.1.2. Post-Erhebung

**Testleitermanual ProwiN-Post-Test:**

Lesen Sie alles Kursive wortwörtlich vor, um zu gewährleisten, dass alle Testungen in gleicher Weise ablaufen. Handlungsanweisungen sind fettgedruckt.

**Vor Beginn der Testung:**

**Warten bis alle Schüler Platz genommen haben und zur Ruhe gekommen sind.**

**Sinngemäß wiedergeben:**
Ihr kennt das bereits, aber nochmal zur Erinnerung: In der Forschung ist es immer ganz wichtig, dass alles korrekt abläuft und Testungen immer gleich durchgeführt werden. Deswegen muss ich euch alle Erklärungen, die gleich folgen, vorlesen. Wundert euch also nicht darüber.

*Mein Name ist_____ und das ist _____.*
*Wie ihr ja bereits wisst, wollen wir herausfinden, was genau im Lehramtsstudium an den Universitäten unterrichtet werden muss, damit der Physikunterricht für euch besser gestaltet werden kann. Um diesem Ziel näher zu kommen, brauchen wir heute noch ein letztes Mal eure Hilfe.*
*Ihr werdet heute einen Test und eine kurze Befragung ausfüllen. Zuerst bekommt ihr den eigentlichen Nachtest zur Mechanik. Den Ablauf kennt ihr ja schon vom Vortest, ich erzähl euch gleich aber auch nochmal etwas dazu. Danach wollen wir etwas über euer Interesse an Physik erfahren.*
*Bei beiden Testheften müsst ihr nur ankreuzen. Stifte bekommt ihr wieder von uns. Die müssen wir nachher aber wieder einsammeln.*
*Der Test, den ihr jetzt gleich bearbeitet, behandelt das Thema Mechanik, also das Thema, das ihr in den letzten Wochen durchgenommen habt. Die Klebezettel mit euren Namen könnt ihr abziehen. Es ist ganz wichtig, dass ihr die Aufgaben allein bearbeitet. Wenn ihr voneinander abschreibt, verfälscht ihr die Daten und wir haben ein großes Problem, weil wir falsche Ergebnisse bekommen. Niemand erfährt eure Ergebnisse, weder eure Lehrer noch eure Eltern, also bitte alleine arbeiten.*
*Wir teilen jetzt die Testhefte für den Schülertest aus. Lasst die Testhefte bitte noch geschlossen, weil wir noch kurz etwas erklären wollen und dann gemeinsam anfangen*

**Schülertests austeilen.**

*Der Test besteht aus Multiple Choice Aufgaben, bei denen immer nur eine Antwort richtig ist. Auch hier gilt wieder: sauber und mittig ankreuzen. Wenn ihr euch vertan habt und eine Antwort verändern wollt, malt das falsch angekreuzte Kästchen vollständig aus, kreuzt die richtige Antwort an und macht einen Kringel darum. Ein Beispiel dafür seht ihr auch gleich im Testheft. Versucht bitte alle Aufgaben zu beantworten. Auch wenn ihr eine Antwort nicht wisst oder euch nicht sicher seid, setzt bitte bei jeder Aufgabe ein Kreuz. Ihr habt 30 min Zeit.*

**Kurz warten.**

*Die Zeit läuft jetzt!*

**Zeit stoppen (30min).**

**Anfangszeit ins Testprotokoll eintragen.**
**Zeit notieren, nach der die erste Abgabe erfolgt. Schüler, die fertig sind, bekommen Beschäftigungsblatt (Sudoka+Mandala). Stifte werden zusammen mit Testheften eingesammelt.**

**Der Lehrer kann jetzt mit der Beantwortung des Lehrerfragebogens anfangen. Hierfür bekommt auch er einen schwarzen Fineliner. Die Regeln für das ankreuzen gelten auch für den Lehrer, der kurz darauf hingewiesen werden sollte. (Anfangszeit für die Bearbeitung des Lehrerfragebogens unter Bemerkungen ins Protokoll eintragen.)**

**Nach Ablauf der Zeit:**

*Die 30 min sind jetzt um. Klappt das Testheft jetzt bitte zu.*

**Fragebögen einsammeln. Endzeit ins Protokoll eintragen.**

*Wir teilen jetzt die Fragebögen zum Fachinteresse und eurer Meinung vom Physikunterricht aus. Die Fragebögen bitte noch nicht aufschlagen. Die Klebezettel mit euren Namen könnt ihr wieder abziehen.*

**Testhefte austeilen.**

*Bei dieser Befragung geht es um eure ganz persönliche Meinung, arbeitet also bitte allein. Beim Ausfüllen ist es wichtig, dass ihr die Kästchen sauber und mittig ankreuzt. Die Hefte werden später eingescannt und können nicht ausgewertet werden, wenn ihr nicht ordentlich ankreuzt. Wenn ihr euch vertan habt und eine Antwort verändern wollt, malt das falsch angekreuzte Kästchen vollständig aus und kreuzt die richtige Antwort an. Die richtige Antwort könnt ihr zusätzlich auch noch einkringeln, dann ist es ganz eindeutig.*
*Für den Fragebogen habt ihr ca. 20 min Zeit.*

*Gibt es noch Fragen?*

**Kurz warten. Fragen gegebenenfalls beantworten.**

*Ok, dann könnt ihr jetzt anfangen!*

**Anfangszeit Fachinteresse ins Protokoll eintragen. Testbögen nach Schüler ID sortieren.**

**Zeit notieren, nach der der Lehrerfragebogen abgegeben wird.**

**Abwarten bis alle Schüler fertig sind mit ausfüllen.**

***Am Ende für die Mitarbeit bedanken!!!***

# A.1.3. Fachspezifisches Professionswissen

ProwiN

## Testleitermanual Befragung zum Professionswissen

Der unten stehende fettgedruckte Text soll inhaltlich wiedergegeben werden. Er muss nicht wortwörtlich wiedergegeben werden.
Kommentare und Anweisungen sind *kursiv gedruckt* und dürfen nicht mit vorgelesen werden.

**Das Testheft enthält zwei Aufgabenblöcke. Einen Aufgabenblock zum Fachdidaktischen Wissen und einen zum Fachwissen. Für die Bearbeitung des ersten Aufgabenblocks sind 45min vorgesehen, für den zweiten Aufgabenblock 40min. Bitte blättern Sie erst zum nächsten Teil weiter, wenn ich Sie dazu auffordere. Falls Sie früher fertig sind, können Sie Ihre Antworten in dem gerade bearbeiteten Teil noch einmal durchsehen, bis die Zeit um ist.**

**Der erste Aufgabenblock beginnt mit einem kurzen Speedtest, das heißt, Sie werden gebeten, zwei Aufgaben jeweils innerhalb einer Minute zu bearbeiten.**

**Bitte schlagen Sie jetzt die erste Seite mit der kurzen Einleitung auf und lesen Sie diese durch, danach fangen wir mit dem ersten Aufgabenteil an.**

*Warten Sie etwa 1,5 Minuten und sagen Sie dann:*

**Haben Sie noch Fragen?**

*Klären Sie ggf. die Fragen.*

### *Fachdidaktisches Wissen*

*Achtung: Speedtest!!!*

**Wir beginnen jetzt mit dem Speedtest. Bitte schlagen Sie zur ersten Aufgabe um und beginnen jetzt.**

*Beginnen Sie jetzt, 1 (!) Minute Bearbeitungszeit für die erste Speedtestaufgabe zu nehmen. Notieren Sie die Anfangszeit im Erhebungssitzungsprotokoll.*

*Nach genau einer Minute sagen Sie:*

**STOPP, bitte auf die nächste Seite zur zweiten Aufgabe blättern. Bitte starten Sie jetzt mit der Bearbeitung von Aufgabe 2.**

*Beginnen Sie jetzt wieder, genau 1 (!) Minute Bearbeitungszeit für die zweite Speedtestaufgabe zu nehmen.*

*Nach genau einer Minute sagen Sie:*

**STOPP, bitte auf die nächste Seite zur vierten Aufgabe blättern. Sie haben jetzt noch insgesamt 43 Minuten Zeit, um die restlichen Aufgaben des ersten Aufgabenblocks zu bearbeiten.**

---

*Beginnen Sie jetzt, 43 Minuten Bearbeitungszeit für den dritten Teil zu nehmen.*

*Nach 33 Minuten sagen Sie:*

**Sie haben jetzt noch 10 Minuten Zeit, bevor wir den dritten Aufgabenblock beenden.**

*Nachdem die letzten 10 Minuten um sind, notieren Sie die aktuelle Zeit im Erhebungssitzungsprotokoll und sagen dann:*

**Die Zeit ist um. Auch wenn Sie mit dem dritten Teil noch nicht fertig sein sollten, möchte ich Sie bitten, jetzt die Seite mit dem Deckblatt für Kapitel IV aufzuschlagen.**

### *Fachwissen*

**Für diesen letzten Teil haben Sie 40 Minuten Zeit. Bitte fangen Sie jetzt an, die Aufgaben zu bearbeiten.**

*Beginnen Sie jetzt, 40 Minuten Bearbeitungszeit für den vierten und letzten Teil zu nehmen.*

*Nach 30 Minuten sagen Sie:*

**Sie haben jetzt noch 10 Minuten Zeit, bevor wir den letzten Aufgabenblock beenden.**

*Nachdem die letzten 10 Minuten um sind, notieren Sie die aktuelle Zeit im Erhebungssitzungsprotokoll und sagen dann:*

**Die Zeit ist um. Auch wenn Sie mit dem letzten Teil noch nicht fertig sein sollten, möchte ich Sie bitten, jetzt mit der Bearbeitung aufzuhören und das Heft zu schließen.**

*Sammeln Sie das Testhefte ein.*

**Vielen Dank für Ihre Mitarbeit!**

Anmerkung: Manual wurde auf Grundlage des ProwiN Testleiterskriptes vom 10.02.11 erstellt.

## A.2. Schülerfachwissenstest

An dieser Stelle befanden sich in der bei der Fakultät für Physik der Universität Duisburg-Essen eingereichten Fassung der Dissertation die Testhefte des Schülerfachwissenstests. Eine Kurzbeschreibung aller Aufgaben, Angaben zu deren Herkunft und eine Übersicht über deren Kennzahlen finden sich bei Cauet (2015). Die Testhefte können bei der Autorin angefragt werden (Aktuelle Kontaktdaten unter http://eva.cauet.de).

## A.3. Professionswissenstests

Die im Rahmen der ersten Projektphase des ProwiN-Projekts entwickelten Professionswissenstests wurden bisher nicht veröffentlicht. Nähere Informationen über die Testinstrumente finden sich bei Kirschner (2013) und Lenske et al. (2015). Die Testhefte können bei der Autorin angefragt werden (Aktuelle Kontaktdaten unter http://eva.cauet.de).

# A.4. Ratingmanual zur kognitiven Aktivierung

## Kategoriensystem „Kognitive Aktivierung“

(in Anlehnung an Vogelsang, 2014)

Das von Christoph Vogelsang entwickelte Kategoriensystem „Dimension: Aktivierung/Konstruktion von Wissen“ (veröffentlicht in Vogelsang, C. (2014), *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften: Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*. Studien zum Physik- und Chemielernen. Berlin: Logos.) wurde im Rahmen des Projekts ProwiN an die projektspezifischen Forschungsinteressen adaptiert.

**Video:** _____

**Rater:** _____

**Datum:** _____

**Ziele**

Dieses Manual dient der Beurteilung der Unterrichtsqualitätsdimensionen **Kognitive Aktivierung** und **Strukturierung** in Bezug auf das im Video jeweils ersichtliche *Unterrichtsangebot*. Die Aufgabe der RaterInnen ist es, für die beiden ***Qualitätsdimensionen*** eine Beurteilung vorzunehmen. Der Schwerpunkt liegt auf der Dimension Kognitive Aktivierung. Deshalb wird diese anhand von mehreren Subskalen detaillierter erfasst.

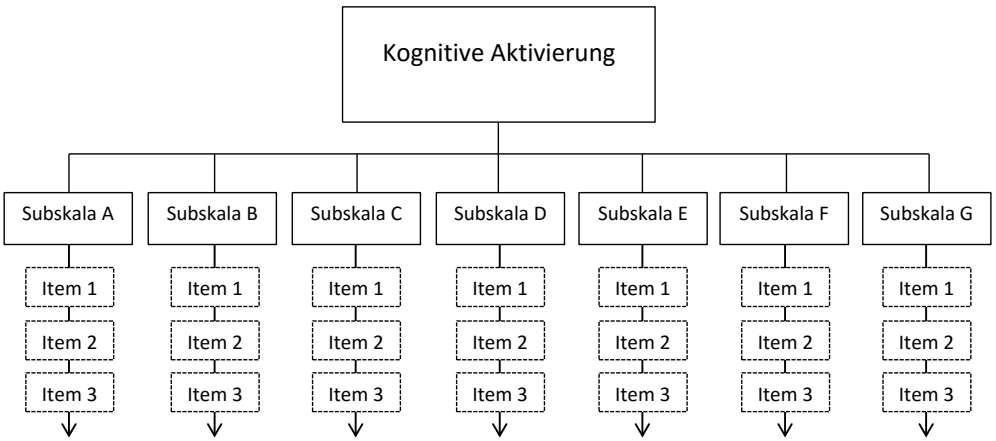
**Aufbau**

Jede Dimension und jedes seiner Merkmale (jede Subskala) wird zunächst kurz beschrieben, wobei die entsprechende ***Grundidee*** dargestellt wird. Diese Grundidee beschreibt jeweils einen idealtypischen Unterricht beziehungsweise einen Aspekt idealtypischen Unterrichts, wie er vor dem Hintergrund von Forschungen zur Unterrichtsqualität angenommen wird. Es ist also ein **empirisch begründetes, normatives Idealbild**, das sich durchaus auch stark von subjektiv erlebten, eigenen Schul- und Unterrichtserfahrungen unterscheiden kann.

Zur Konkretisierung wird jede Grundidee anschließend differenziert in mehrere, möglichst handlungsnah formulierte ***Items* (*Indikatoren)* (**Operationalisierung). Zu jedem Item werden jeweils Hinweise genannt, wann das Item zutrifft bzw. nicht oder nur zum Teil zutrifft. Konkretisiert werden diese Abstufungen anhand von Videovignetten im ***Ratertraining***. Des Weiteren dient ein ***Referenzvideo*** mit einem Masterrating als Richtschnur.

Jeder Skala (oder Subskala) sind zur möglichst objektiven Beurteilung mehrere Items zugeordnet. Die einzelnen Items dienen damit als Bezugspunkt für die Beurteilung und bilden eine Messskala zur Erfassung der Unterrichtsqualität. In ihrer Gesamtheit bilden die Skalen ein Abbild des idealtypischen Lehrerhandelns entsprechend der Grundidee. Die folgende Abbildung 1 zeigt einen grafischen Überblick über die **generelle Struktur** der Beurteilungskategorien.

*Abbildung 1: „Grundstruktur des Beurteilungsbogens" am Beispiel Kognitive Aktivierung*



**Vorgehen bei der Beurteilung**

Da sich die Items auf einen idealtypischen Unterricht beziehen, besteht die Beurteilung darin, einzuschätzen, inwiefern der tatsächlich beobachtete Unterricht beziehungsweise das Handeln der Lehrperson mit diesem Idealbild übereinstimmt. Für jede Subskala liegt deshalb für jedes Item eine dreistufige ***Antwortskala*** vor, wobei die Skala von „1 = *geringe Ausprägung/trifft nicht zu*“ über „2 = mittlere Ausprägung/teils teils“ bis „3 = *starke Ausprägung/trifft zu*“ reicht. Eine geringe Ausprägung bedeutet, dass ein Indikator gar nicht oder nur sehr wenig im Unterricht beobachtet werden kann. Eine hohe Ausprägung bedeutet, dass ein Indikator sehr deutlich und sehr stark ausgeprägt im Unterricht beobachtet werden kann und dem Idealbild vom Unterricht sehr nahe kommt. Der Fokus liegt in den meisten Fällen auf der Beobachtung des Handelns der Lehrperson, da eine Beurteilung für das *Unterrichtsangebot* vorgenommen werden soll. Einige Indikatoren beziehen sich dennoch eher „indirekt“ auf ein Merkmal und legen den Beobachtungsfokus stärker auch auf das Schülerverhalten. Bei einigen Merkmalen werden im Manual diesbezüglich spezifische Beurteilungshinweise gegeben.

Sollte ein Item aufgrund der situativen Bedingungen seines Auftretens (beispielsweise wird nach Verhalten bei Verständnisschwierigkeiten gefragt, es treten aber keine im Video auf) nicht beurteilbar sein, so soll dies mit dem Kürzel *„n.b.“* auf dem Ratingbogen gekennzeichnet werden.

Zusätzlich zur Einschätzung der einzelnen Items wird bei jeder (Sub-)skala eine Einschätzung des **Gesamteindrucks** hinsichtlich des Merkmals erfragt, wobei eine analoge Skala verwendet wird. Wichtig hierbei ist, dass sich der Gesamteindruck auf die Grundidee des Merkmals (der Subskala) und ihre Umsetzung bezieht. Für die Beurteilung können durchaus einzelne Indikatoren stärker ins Gewicht fallen als andere, wenn sie die Umsetzung des Merkmals im Unterricht stärker bestimmen.

Bei der Beurteilung können subjektive Unsicherheiten auftreten, da bei der Beurteilung häufig verschiedene **Beobachtungsaspekte** zu beurteilen und abzuwägen sind. Diese sind im Folgenden beschrieben:

(1) *Häufigkeit* des gezeigten Handels (Wie oft zeigt die Lehrperson diese Handlungsweise?)

(2) *Intensität* des gezeigten Handelns (Wie stark ausgeprägt zeigt die Lehrperson diese Handlungsweise?)

(3) *Verteilung* des gezeigten Handelns in der Klasse (Gegenüber wie vielen Schülerinnen und Schülern zeigt die Lehrperson diese Handlungsweise?)

(4) *Adäquatheit* des gezeigten Verhaltens (Entspricht das Verhalten den Anforderungen der Situation?)

Die Grundregel für das Beurteilen lautet daher, dass sich der **Eindruck** für jedes Item und der Gesamteindruck bezüglich des jeweiligen Merkmals aus allen dieser vier Beobachtungsaspekte zusammensetzen sollen. Die einzelnen Beobachtungsaspekte können unterschiedlich gewichtet werden, weshalb bei einigen Merkmalen zusätzlich Beurteilungshinweise angegeben werden. Die zusätzliche Beurteilung des Gesamteindrucks dient dabei als eine Art Sicherung für den Fall, dass beispielsweise zwar die einzelnen Items nicht sehr stark ausgeprägt beurteilt werden können, der Beurteilende aber dennoch das Merkmal in seiner Grundidee sehr ausgeprägt beurteilen würde, auch wenn dem keine der angeführten Indikatoren zu Grunde liegen, sondern andere Handlungsweisen der Lehrkraft.

Für das Gedächtnis ist es eine enorme Belastung, sämtliche Situationen, die für die Beurteilung einzelner Items wichtig sein können, zu speichern. Deshalb darf das Video jeder Zeit gestoppt werden, um sich Notizen zu machen. Ein Stopp alle 15-20min. ist Pflicht. Die Notizen sind im **Beobachtungsprotokoll** festzuhalten. Dieses dient als Basis bzw. als Gedächtnisstütze bei der anschließenden Beurteilung des gesamten Videos. Es ist jeder Zeit erlaubt, einzelne Situationen erneut zu beobachten. Dies ist insbesondere sinnvoll, wenn man als Rater feststellt, dass man in einer Situation einen recht selektiven **Beobachtungsfokus** hatte (z.B. man fokussierte sehr stark auf störenden Schüler, wobei eventuell weitere Reize unbeobachtet blieben) oder etwas akustisch nicht auf Anhieb zu verstehen war.

**Beurteilungszeitraum und -material**

Jede Beurteilung (*Rating*) bezieht sich immer auf eine **gesamte Schulstunde als Analyseeinheit**. Technisch liegt in den meisten Fällen zusätzlich zu einem Video, welches auf die jeweiligen Aktionen des Geschehens fokussiert (*Aktionskamera*: fokussiert auf die Lehrkraft oder die Schülerinnen und Schüler, die gerade aktiv etwas zum Geschehen beitragen), noch ein zweites Video vor, das die gesamte Klasse in einer Totale zeigt (*Totale*). Beide Videos beziehen sich auf dieselbe Analyseeinheit. In der Regel reicht zur Beurteilung der Dimensionen *Kognitive Aktivierung* und *Strukturierung* die Aktionskamera. Bei Unsicherheiten kann jederzeit die Totale hinzugezogen werden.

**Grundregeln für RaterInnen**

Sie finden nachfolgend ein paar Hinweise, die Ihnen den Ablauf des Kodierens mit Hilfe dieses Beurteilungsbogens beschreiben und die Sie bei der Beurteilung unterstützen sollen.

(1) Sorgen Sie für **passende Rahmenbedingungen** während des Kodierens. Kodieren Sie an einem Ort, der wenig Ablenkung und ausreichend **Ruhe** bietet. Dies ist zum einen hilfreich, um die Konzentration über die volle Länge eines Unterrichtsvideos aufrecht zu erhalten. Zum anderen ist es notwendig, dass Sie in einer lärmarmen Umgebung arbeiten, damit Sie auch alle Äußerungen auf den Unterrichtsvideos verstehen können. Kodieren Sie daher auch immer mit **Kopfhörer**.

(2) Schauen Sie konzentriert das zu beurteilende Video. Stoppen Sie das Video, wann immer sie das Bedürfnis dazu haben, legen sie jedoch mindestens 3 Notizphasen ein (jeweils nach spätestens 15-20min). Bitte notieren Sie sich nichts während des Videoschauens, da Sie dann den Blick vom Video abwenden. Versuchen Sie beim Beobachten gezielt auf die Indikatoren des Manuals zu achten (d.h. selektiver Fokus auf die zu beurteilenden Aspekte).

(3) Nehmen Sie anschließend eine Beurteilung der einzelnen Merkmale und Indikatoren vor. Führen Sie sich hierbei zunächst immer die **Grundidee** der zu beurteilenden Facette vor Augen. Beurteilen Sie danach die einzelnen Items. Bitte bleiben Sie bei der Einschätzung möglichst **„dicht" an den Indikatoren**. Versuchen Sie möglichst objektiv zu bleiben. Geben Sie danach eine Einschätzung Ihres **Gesamteindrucks** ab. Denken Sie hierbei nochmals an die Grundidee (eventuell unterscheiden sich an dieser Stelle Ihre Beurteilungen der Indikatoren und des Gesamteindrucks). Denken Sie auch daran, dass es primär auf das Handeln der Lehrkraft ankommt und eine Beurteilung der Qualität des Lehrerhandelns gesucht wird. Da jedes Item auf einer dreistufigen Skala beurteilt werden soll, überlegen sie zunächst, ob eine generelle *Tendenz* zu erkennen war. Sollte dies nicht der Fall sein, trifft die mittlere Kategorie zu.

(4) Natürlich sind immer Interpretationen und subjektive Einschätzungen nötig. Sollte Ihnen die Beurteilung einiger Items sehr schwer fallen und Sie sich absolut **nicht sicher** sein, markieren Sie diese Items zusätzlich mit dem Kürzel *„n.s."*. Markieren sie Items, die sie nicht beurteilen konnten, weil die durch das Item angesprochene Situation in der gesamten Unterrichtsstunde nicht aufgetreten ist mit dem Kürzel *„n.b."*. **Fragen Sie bei Unklarheiten** direkt bei den Projektverantwortlichen nach.

(5) Beurteilen Sie alle Skalen **nacheinander bis zum Ende**. Versuchen Sie möglichst den gesamten Bogen **in einem Durchgang** zu beurteilen. Sollten Sie merken, dass Ihre Konzentration stark nachlässt, machen Sie eine kurze Pause (ca. 5 Minuten) und beurteilen Sie dann weiter.

(6) Machen Sie eine **längere Pause** (mind. 10 Minuten), bevor Sie nach einem Video **ein weiteres Video** beurteilen. Versuchen Sie dann das Video unabhängig zur vorherigen Stunde zu beobachten und zu beurteilen. Sie vermeiden damit Kontrasteffekte und vermeiden es, die Bewertung der „neuen" Lehrperson mit Ihren Eindrücken der „alten" zu überlagern.

(7) Da es im Laufe der Auswertungen Ihrer Beurteilungen zu Unstimmigkeiten kommen kann, seien Sie **für Nachfragen bereit**. Insbesondere wenn Sie und eine weitere Person zu einigen stark abweichenden Beurteilungen einzelner Items kommen, wird versucht im Gespräch eine möglichst "gute" Übereinstimmung zu erzielen.

(8) Wenn noch irgendetwas unklar sein sollte oder Sie weitere Unterstützung gebrauchen könnten, melden Sie sich bitte jederzeit bei den Projektverantwortlichen.

**A) Lernstatus im gesamten Thema bewusst machen**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit die Lehrperson sich bemüht, den Schülerinnen und Schülern deutlich zu machen, auf welchem Wissen neu zu erwerbende Begriffe und Konzepte aufbauen. Es werden also Zusammenhänge zwischen früher Gelerntem und neu zu Lernendem aufgezeigt und so die Verknüpfung des Wissens gefördert. Grundlegend für alle Indikatoren ist, dass die Schülerinnen und Schüler angeregt werden, die neuen Inhalte in ihr bereits bestehendes Wissenssystem einzuordnen. Das Bewusstmachen des Lernstatus wirkt nach konstruktivistischem Lernverständnis förderlich für die Konstruktion und Vernetzung von Wissen.

*Quellen:* Kunter (2005), Rakoczy & Pauli (2006), adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| A1 | Die Lehrperson bezieht sich auf inhaltliche Ideen, Probleme, Konzepte oder Begriffe aus vorangegangenen Stunden und verweist explizit darauf wie diese mit den „aktuellen" Inhalten verbunden werden.<br>*Beispiele/Indikatoren:*<br>(+): „In der letzten Stunde habt ihr gelernt, welche Keimungsbedingungen ein Samen zum Wachstum benötigt. Heute erarbeiten wir, welche funktionellen Strukturen ein Samen zum Wachstum besitzt, um unter den entsprechenden Bedingungen zu keimen."<br>(-): „Heute zeichnen wir einen Samen und beschriften ihn."<br>(°):„Bislang habt ihr gelernt, welche Keimungsbedingungen ein Samen zum Wachstum benötigt. Wir beschäftigen uns heute weiterhin mit dem Samen." | ☐ | ☐ | ☐ |
| A2 | Die Lehrperson bezieht sich auf konkrete Zeitpunkte, an denen in der Vergangenheit ein Begriff im Unterricht auftauchte, und verbindet diesen Inhalt mit dem „aktuellen" Inhalt.<br>*Beispiele/Indikatoren:*<br>(+): „In der Grundschule habt ihr euch mit Pflanzen in verschiedenen Lebensräumen beschäftigt. Dieses Thema werden wie heute wieder aufgreifen und uns mit dem Wachstum von Pflanzen unter verschiedenen Bedingungen beschäftigen."<br>(-): „In der Grundschule habt ihr euch mit Pflanzen beschäftigt, heute machen wir damit weiter." (Bezug ist zu allgemein, Verbindung wird nicht aufgezeigt.)<br>(°):„Ihr habt in der Grundschule bereits etwas über Pflanzen in verschiedenen Lebensräumen gelernt. Dieses Thema werden wir heute erneut aufgreifen und vertiefen.<br>(Bezug ist zwar da, aber die Verbindung zum aktuellen Thema ist zu oberflächlich.) | ☐ | ☐ | ☐ |
| A3 | Die Lehrperson verweist auf Inhalte und Themen, die sich aus dem aktuellen Unterricht ergeben und die in zukünftigen Stunden besprochen werden.<br>*Beispiele/Indikatoren:*<br>(+): „Heute haben wir gelernt, welche Strukturen ein Samen hat. Die einzelnen Strukturen können verschiedene Funktionen haben. Damit werden wir uns in der nächsten Stunde beschäftigen."<br>(-): „Heute haben wir gelernt, welche Strukturen ein Samen hat. Damit machen wir nächste Woche weiter."<br>(°):„Heute haben wir gelernt welche Strukturen ein Samen hat. In der nächsten Stunde beschäftigen wir uns mit den Funktionen der einzelnen Strukturen." (auf den Zusammenhang wird nicht explizit verwiesen) | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| A4 | Die Lehrperson gibt einen expliziten Ausblick darauf, welche Inhalte in der Stunde thematisiert werden.[1)]<br>*Beispiele/Indikatoren:*<br>(+): „Heute werden wir uns mit dem Thema ‚Photosynthese' beschäftigen. Wir betrachten dabei die Versuche von Priestley und erarbeiten die Gleichung der Photosynthese."<br>(-):„Heute werden wir uns erstmals mit dem Thema „Photosynthese" beschäftigen. Schlagt dazu bitte euer Biologiebuch auf Seite 93 auf." (Der neue Begriff steht bezuglos im Raum, keine Anschlussfähigkeit erzeugt, d.h. es wird einfach ein Begriff genannt)<br>(-): „Wir haben das Thema Photosynthese und machen auch heute damit weiter." (Inhalt der Stunde wird nicht expliziert)<br>(°): „Heute werden wir in das Thema ‚Photosynthese' einsteigen. Unter Photosynthese versteht man ... (altersadäquate Begriffsklärung folgt). Zur Einführung in die Thematik betrachten wir uns zunächst einmal diese Pflanze." (Thema wird zwar erklärt und es wird auch klar, dass die Stunde der Einstieg in die Thematik sein wird, aber alles weitere fehlt).<br>(°): „Heute werden wir über Kraft sprechen und was das überhaupt ist." (Es wird klar, dass die Stunde der Einstieg in die Thematik sein wird und der Klärung des neuen Begriffs/des neuen Konzeptes dienen soll, aber alles weitere fehlt.) | ☐ | ☐ | ☐ |
| A5 | Die Lehrperson gibt im Verlauf oder am Ende der Stunde einen Rückblick auf bereits Gelerntes (bezogen auf die Inhalte der „aktuellen" Stunde). Zentrale Erkenntnisse werden hervorgehoben. [1)] [2)]<br>*Beispiele/Indikatoren:*<br>(+): „In der heutigen Stunde habt ihr gelernt, welche Versuche Priestley durchgeführt hat, welche Schlussfolgerungen sich daraus ergeben und wie aus diesen Erkenntnissen notwendigerweise die Gleichung der Photosynthese resultiert. Diese Gleichung ist eine wichtige Grundlage für die folgenden Stunden."<br>(-): „Heute haben wir uns mit den Versuchen von Priestley beschäftigt. Morgen...."<br>(°): „Heute haben wir die Versuche von Priestley behandelt und haben gesehen, welche Schlussfolgerungen Priestley aus diesen Versuchen gezogen hat." (Zentrale Erkenntnis wird umschrieben und nicht explizit betont, d.h. der Rückblick ist nicht prägnant oder unvollständig)<br>(°) Der Rückblick ist kein Ganzes, sondern vielmehr aus Fragmenten zusammen gesetzt oder der Rückblick enthält zuvor nicht besprochene Informationen. | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Lernstatus bewusst machen*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| A6 | Gutes Bewusstmachen des Lernstatus innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**Anmerkungen:**

- [1)] diese Merkmale lassen sich auch dem Merkmal „Strukturierung des Unterrichts" zuordnen. Der Fokus hierbei ist allerdings eher das „Aktivieren" der Lernenden als die Beurteilung einer Struktur.
- [2)] werden zum Ende des Unterrichts Merksätze formuliert, die nicht als Zusammenfassung oder Rückblick auf Gelerntes in den Unterricht eingebettet werden, ist mit *trifft nicht zu* zu bewerten.

**B) Exploration des Vorwissens und der Vorstellungen**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit die Lehrperson das Vorwissen der Lernenden im Unterricht aktiviert und mit einbezieht. Zum Vorwissen zählen – neben dem Wissen aus vorhergehendem Unterricht – auch außerunterrichtliche Vorstellungen der Schülerinnen und Schüler zu naturwissenschaftlichen Begriffen und Konzepten, sowie Erfahrungen im Zusammenhang mit dem Unterrichtsgegenstand. Diese Exploration kann beispielsweise dadurch geschehen, dass die Lehrperson die Lernenden direkt nach ihren Ideen und Vorstellungen fragt, ohne gleich eine Beurteilung der geäußerten Vorstellungen vorzunehmen. Grundlegend für alle Indikatoren ist, dass die Lehrperson versucht zu erfahren, was die Schülerinnen und Schüler „in ihren Köpfen" haben. Die Handlungen beziehen sich daher hauptsächlich auf deklaratives und prozedurales Wissen und weniger auf die kognitiven Denkprozesse (siehe nächstes Merkmal). Im naturwissenschaftlichen Unterricht bilden die Alltagsvorstellungen der Lernenden einen wichtigen Einflussfaktor für das Verständnis der fachlichen Inhalte. Da Lernen als Konstruktion von Bedeutungen auf Basis schon bekannter Ideen und Konzepte verstanden wird, wirkt ein Handeln der Lehrperson, dass das bekannte Wissen bewusst macht, fördernd auf die Konstruktions- und Verknüpfungsprozess.

*Quellen:* Kunter (2005), Rakoczy & Pauli (2006), adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| B1 | Die Lehrperson führt im Unterricht „Brainstormings" zu Begriffswissen oder zu Ideen der Schülerinnen und Schüler durch.<br>*Beispiele/Indikatoren:*<br>(+): „Was fällt euch zum Begriff Kraft ein?", „Was stellt ihr euch unter Kraft vor?", auch kurze Brainstormings)<br>(-): Ausbleiben des positiven Indikators<br>(°): Es werden nur wenige Vorstellungen und Ideen erfasst bzw. es kommen nur wenige Äußerungen vor. | ☐ | ☐ | ☐ |
| B2 | Die Lehrperson fragt nach Vorwissen (auch Begriffswissen) und Vorstellungen der Schülerinnen und Schüler, ohne auf eine bestimmte Antwort abzuzielen.<br>*Beispiele/Indikatoren:*<br>(+): „Was wisst ihr schon über physikalische Größen?"<br>(-): Ausbleiben des positiven Indikators<br>(°): Es werden nur wenige Äußerungen zugelassen oder das Vorwissen wird an weiteren, offensichtlich sinnvollen, Stellen nicht erfragt. | ☐ | ☐ | ☐ |
| B3 | Die Schülerinnen und Schüler werden angeregt, das Unterrichtsthema (bzw. Aspekte des Themas) nach ihrem Verständnis zu erläutern.<br>*Beispiele/Indikatoren:*<br>(+): „Versucht bitte den Ablauf der Photosynthese in euren eigenen Worte zu erklären."<br>(-): „Wiederhole die Definition" oder „Lese den Merksatz vor" (d.h. ein Umschreiben/ Erklären in eigenen Worten bleibt aus).<br>(°): Schülerbeschreibung ist unvollständig und bleibt unkommentiert.<br>(°): Aufforderung erfolgt, aber die Beschreibung ist stark rezeptiv (d.h. keine adäquate Umsetzung der Aufforderung erfolgt). | ☐ | ☐ | ☐ |
| B4 | Die Lehrperson fragt nach Ideen und Vorstellungen der Schülerinnen und Schüler, ohne eine Wertung der Äußerungen vorzunehmen.<br>*Beispiele/Indikatoren:*<br>(+): Die Lehrkraft lässt die geäußerten Ideen/Vorstellungen unkommentiert und gibt den Ball an die SuS zurück: „Was denkt ihr zu dieser Äußerung?" oder „Ja, möglich, weitere Ideen?" (d.h. kein frühzeitiges Kategorisieren in *richtig* oder *falsch*).<br>(-): „Das ist nicht ganz richtig." oder „Ja, genau, das wollte ich hören." | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| | (°): Geringe Anzahl unkommentierter oder nicht bewerteter Schüleräußerungen (Anzahl muss am Stundentyp relativiert werden). | | | |
| B5 | Die Lehrperson regt die Schülerinnen und Schüler dazu an das „aktuelle“ Thema mit ihnen schon bekannten Begriffen in Verbindung zu setzen.<br>*Beispiele/Indikatoren:*<br>(+): „Kraft ist eine physikalische Größe. Denkt einmal an andere physikalische Größen. Welche Aspekte kann man auf Kraft übertragen?“<br>(+): „An was erinnert euch das? So ähnlich haben wir das schon einmal gesehen.“<br>(-): „Letzte Stunde haben wir uns mit Arbeit beschäftigt. Heute beginnen wir mit Kraft.“<br>(°): Die Schülerreaktion ist sehr, sehr verhalten bzw. lediglich ein Schüler/eine Schülerin stellt eine Verbindung her (d.h. der Impuls ist ersichtlich, aber die Wirkung nicht wirklich) und die Lehrkraft interveniert nicht (d.h. keine erneute Anregung). | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Exploration des Vorwissens*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| B6 | Gute Exploration des Vorwissens innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**C) Exploration der Denkweisen der Schülerinnen und Schüler**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit die Lehrperson versucht, die Gedankengänge der Lernenden bezüglich des Lerngegenstands zu erfahren und mit in den Unterricht einzubeziehen, damit sie einen fachlichen Begriff so einführen kann, dass es dem Verständnis der Schülerinnen und Schüler entspricht. Dabei geht es nur um eine Exploration und nicht um eine Beurteilung der Denkweisen. Hierbei geht es also weniger um die Diagnose von Vorwissen, sondern um eine Diagnose des Lern- und Denkprozesses bezüglich des fachlichen Inhalts an sich. Die Exploration von Denkweisen kann unterstützend bei der Konstruktion neuer Bedeutungen wirken, da sie den Konstruktionsprozess zum einen bewusst macht und zum anderen der Lehrperson ermöglicht, auf die ablaufenden Prozesse zu reagieren. Da Lernen als Konstruktionsprozess verstanden wird, können Handlungen der Lehrperson, die den Prozess bewusst machen, förderlich auf den Konstruktions- und Verknüpfungsprozess wirken.

*Quellen:* Kunter (2005), Rakoczy & Pauli (2006), adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| C1 | Die Lehrperson versucht die Denkweisen von Schülerinnen und Schülern zu verstehen, indem sie fragt, wie sie zu bestimmten Antworten gelangt sind.<br>*Beispiele/Indikatoren:*<br>(+): „Ja kannst du mal erklären, wie du darauf kommst, dass Wasser eine Keimungsbedingung von Samen ist?"<br>(-): Aufforderungen/Nachfragen diesbezüglich bleiben aus und die SuS sind es nicht gewohnt, „automatisch" eine Explikation ihrer Denkweise zu liefern.<br>(-): Denkweisen werden nur bei falschen Antworten erfragt (Abgrenzung zu C4).<br>(°): In Ansätzen erkennbar/nicht häufig genug. | ☐ | ☐ | ☐ |
| C2 | Die Lehrperson fordert von den Schülerinnen und Schülern Begründungen für ihre Antworten (z.B. im Klassengespräch).<br>*Beispiele/Indikatoren:*<br>(+): „Kannst du deine Antwort auch begründen?" „Warum ist das so?"<br>(-): Aufforderungen/Nachfragen diesbezüglich bleiben aus und die SuS sind es nicht gewohnt, „automatisch" ihre Antworten unmittelbar zu begründen.<br>(°): Ist erkennbar, aber nicht häufig genug | ☐ | ☐ | ☐ |
| C3 | Die Lehrperson erkundigt sich bei den Schülerinnen und Schülern, was sie verstanden haben (bzw. was sie nicht verstanden haben).<br>*Beispiele/Indikatoren:*<br>(+): „Hat jemand noch offenen Fragen?", „Was ist momentan noch unklar für euch?"<br>(-): So, das ist soweit klar, ne?! („rhetorische Frage"), Lehrperson wirkt nicht wirklich an Antwort interessiert.<br>(°): Kommt in Relation zur Stunde zwar vor, aber nicht oft genug. | ☐ | ☐ | ☐ |
| C4 | Die Lehrperson fragt bei Verständnisschwierigkeiten, nach den Denkprozessen der Schülerinnen und Schüler.<br>*Beispiele/Indikatoren:*<br>(+): „Was denkst du denn passiert, wenn ich einem Samen bei der Keimung die Wassermenge verringere?", „Erkläre bitte deinen Gedankengang hierzu.", „Wieso glaubst du, dass das so sein muss?"<br>(-): Nachfragen bei Verständnisschwierigkeiten bleiben aus. Fehlern/Fehlvorstellungen wird nicht auf den Grund gegangen.<br>(°): Nachfragen bei Verständnisschwierigkeiten erfolgen, aber nicht regelmäßig/systematisch.<br>Sind keine Verständnisschwierigkeiten ersichtlich, ist das Item nicht beurteilbar (=> n.b.). | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| C5 | Die Lehrperson regt die Schülerinnen und Schüler an, Sachverhalte mit eigenen Worten zu erläutern. (WICHTIG: Keine Wiederholungen von bereits gelernten Argumentationen oder Routinen bzw. reine Beschreibung von Beobachtungen, es geht um Erklärungen aus der Sicht und mit Worten der Lernenden)<br>*Beispiele/Indikatoren:*<br>(+): „Könnt ihr mir das in euren eigenen Worten erklären ?“, „Diese Erklärung beinhaltet Fachbegriffe. Könnte ihr mir ohne die Verwendung der Fachbegriffe erklären, was darunter zu verstehen ist?“<br>(-): Aufforderungen/Nachfragen diesbezüglich bleiben aus.<br>(°): Aufforderungen/Nachfragen erfolgen in Relation zur Stunde nicht häufig genug. | ☐ | ☐ | ☐ |
| C6 | Die Lehrperson stellt im Klassengespräch häufig Wie- und Warum-Fragen. (WICHTIG: nicht im Sinne von Kontroll- oder Disziplinierungsmaßnahmen, es geht um das Anregen des Denkens der Schülerinnen und Schüler).<br>Wichtig: Hier geht es um Fragen, die Erklärungen, Begründungen oder Prozesse betreffen bzw. elaboriertere Fragestellungen.<br>*Beispiele/Indikatoren:*<br>(+): „Wie könnten wir unsere Vermutung hierzu überprüfen?“, „Warum können Mutationen sowohl das Überleben sichern als auch das Überleben erschweren?“<br>(-): „Wie sieht die Tabelle aus, also was steht in den Spalten?“<br>(°): In Relation zur Stunden kommt dieser Fragentypus nicht oft genug vor. | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Exploration der Denkweisen der Schülerinnen und Schüler*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| C7 | Gute Exploration der Denkweisen innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**D) Evolutionärer Umgang mit Schülervorstellungen**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit die Lehrperson die Vorstellungen und das Vorwissen der Schülerinnen und Schüler nutzt, um die fachlichen Inhalte beziehungsweise die fachliche Sichtweise auf die Inhalte zu erarbeiten. Dies kann beispielsweise dadurch geschehen, indem die Lehrperson an bestehende Vorstellungen anknüpft, aber auch durch das Erzeugen eines kognitiven Konfliktes mit dem Ziel einer Veränderung schon bestehender mentaler Konzepte oder der Unterstützung der Konstruktion neuen Wissens. Hierzu sind verschiedene Interaktionen zwischen Lehrperson und Lernenden möglich, in die die Vorstellungen der Schülerinnen und Schüler mit einbezogen werden. Nach der konstruktivistischen Sichtweise von Lernen und den Ergebnissen der Schülervorstellungsforschung bildet ein „evolutionärer" Umgang mit Schülervorstellungen im Unterricht einen Indikator für die Unterstützung der Konstruktion von Wissen.

*Quellen:* Kunter (2005), Rakoczy & Pauli (2006), Clausen (2002) adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| D1 | Die Lehrperson greift Vorstellungen und Ideen der Schülerinnen und Schüler auf und verwendet sie im weiteren Unterricht.<br>*Beispiele/Indikatoren:*<br>(+): „Ihr habt gesagt, dass ein Samen zur Keimung Licht benötigt. Das überprüfen wir mal."<br>(-): „Wir haben letzte Stunde gelesen, dass ein Samen Licht benötigt. Heute schauen wir, warum dies so ist."<br>(°): In Relation zur Stunde, werden nur marginal Ideen/ Vorstellungen der SuS erfasst und/oder aufgegriffen.<br>(Nicht Wiedergabe des Gelernten oder Routinen durch die Lernenden. Kein Sammeln von Lösungen oder Antworten.) | ☐ | ☐ | ☐ |
| D2 | Die Lehrperson macht den Unterschied zwischen fachlicher „Wissenschaftssprache" und Alltagssprache deutlich.<br>*Beispiele/Indikatoren:*<br>(+): „Wenn Biologen von Licht sprechen, meinen sie Energie und nicht Helligkeit.<br>(-): Unterschiede diesbezüglich werden weder implizit noch explizit thematisiert.<br>(°): In Relation zu den verwendeten oder eingeführten Fachbegriffen, wird nicht genug darauf eingegangen oder die Abgrenzung erfolgt nur einseitig („In der Physik versteht man unter Licht eine energetische Quelle" => Bedeutung in der Alltagssprache wird nicht expliziert). | ☐ | ☐ | ☐ |
| D3 | Die Lehrperson führt ausgehend von den Vorstellungen der Lernenden Schritt für Schritt die wissenschaftlichen Begriffe ein (meist durch Fragetechniken).<br>Wichtig: Der Annäherungsprozess muss ersichtlich sein.<br>*Beispiele/Indikatoren:*<br>(+): Die Lehrkraft greift Vorstellungen der Lernenden auf: „Was versteht ihr unter...?", „Was könnte dieser Begriff euer Meinung nach bedeuten?", „In welchem Zusammenhang habt ihr den Begriff schon einmal gehört?", „Ja, das trifft die wissenschaftliche Definition noch nicht ganz. Noch etwas fehlt..."<br>(-): Wir beschäftigen uns heute mit Kraft. In der Physik versteht man unter Kraft Folgendes:... (Definition durch die Lehrkraft erfolgt)."<br>(°):Die Annäherung ist bruchhaft. Schülervorstellungen werden zwar erfasst, aber direkt im Anschluss wird von der Lehrkraft die Definition ohne konkreten Bezug zu den Schülervorstellungen vorgenommen. | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| D4 | Die Lehrperson versucht fachlich nicht korrekte Vorstellungen und Ideen der Schülerinnen und Schüler zu „belasten“, indem sie beispielsweise ein Experiment durchführt, dass diesen Vorstellungen widerspricht.<br>*Beispiele/Indikatoren:*<br>(+): „Pflanzen mit roten Blättern betreiben keine Photosynthese!? Ist das wirklich so? Woher nimmt die Pflanze dann die Energie? Warum sind denn die Blätter überhaupt rot?“<br>(-): „Es gibt auch Pflanzen mit roten Blättern. Diese absorbieren andere Wellenlängen des Lichts. Ist klar, oder?!“<br>(°): Es wird zwar eine typische Fehlvorstellung belastet, jedoch ohne den SuS diese Fehlvorstellung zunächst bewusst zu machen. Es werden also keine Vermutungen von den SuS eingeholt.<br>(n.b.) Eine Belastung ist an keiner Stelle erforderlich oder sinnvoll. | ☐ | ☐ | ☐ |
| D5 | Die Lehrperson fordert die Schülerinnen und Schüler dazu auf, auf ihren Vorstellungen oder auf ihrem aktuellen Wissensstand aufbauend zu argumentieren und Schlussfolgerungen zu ziehen.<br>*Beispiele/Indikatoren:*<br>(+): „Was schlussfolgert ihr hieraus?“, „Worauf gründet deine Vermutung?“, „Warum glaubst du, dass dies so ist?“<br>(-): Aufforderungen diesbezüglich bleiben aus.<br>(°): In Relation zur Stunde kommen Aufforderungen derart nicht oft vor. | ☐ | ☐ | ☐ |
| D6 | Die Lehrperson lässt die Schülerinnen und Schüler mit ihren Vorstellungen auch mal in die Irre gehen, bis sie es selbst merken.<br>*Beispiele/Indikatoren:*<br>(+): „Ja, das könnte eine Erklärung sein...“, „Nehmen wir an, du hast Recht, was folgt aus deiner Behauptung?“<br>(-): „Nein, das ist nicht korrekt.“<br>(°): Die Lehrkraft lässt die SuS zwar auch einmal in die Irre gehen, aber beschleunigt den Prozess durch eigenes Eingreifen sehr stark.<br>Die Formulierung *auch mal* macht deutlich, dass das geforderte Lehrerverhalten nicht häufig erfolgen muss bzw. eine Unterrichtsphase, in der dies stattfindet, ausreicht, um dem Item zuzustimmen. | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Evolutionärer Umgang mit Schülervorstellungen*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| D7 | Guter evolutionärer Umgang mit Schülervorstellungen innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**Anmerkungen:**

- Hilfreich hierfür ist zu überlegen: Wo sind die in der Stunde erarbeiteten Begriffe hergekommen? Wurden tatsächlich Ideen und Konzepte der Schülerinnen und Schüler verwendet? Waren ihre Vorstellungen der Ausgangspunkt für eine Entwicklung der fachlichen Begriffe?

**E) Lehrperson als Mediator**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit die Lehrperson in ihren Interaktionen mit den Schülerinnen und Schülern Bedingungen für eine soziale Ko-Konstruktion von neuem Wissen ermöglicht. Sowohl im Klassengespräch als auch bei Schülerarbeitsphasen kann diese beispielsweise durch das Einfordern von Begründungen und Stellungnahmen gefördert werden. Die Lehrperson nimmt dabei eine die Äußerungen der Lernenden moderierende Haltung ein. Eine Förderung des sozialen Aushandelns von Bedeutungen bildet nach dem konstruktivistischen Lernverständnis einen Indikator für die Konstruktion von Wissen.

*Quellen:* Clausen, Reusser & Klieme (2003), Rakoczy & Pauli (2006), adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| E1 | Die Lehrperson bezieht Beiträge verschiedener Schülerinnen und Schüler aufeinander.<br>*Beispiele/Indikatoren:*<br>(+): „Mareike hat behauptet…, Torsten hingegen, dass…"<br>(-): Indikator bleibt aus.<br>(°): Indikator kommt vor, aber in Relation zur Stunde zu selten. | ☐ | ☐ | ☐ |
| E2 | Die Lehrperson fordert die Schülerinnen und Schüler auf, ihre Beiträge selbst aufeinander zu beziehen.<br>*Beispiele/Indikatoren:*<br>(+): „Wenn du daran denkst, was Katrin gesagt hat, was würdest du darauf antworten?" , „Versucht in eurer Argumentation die Argumente eurer Mitschüler einzubeziehen, zu stärken oder zu entkräften."<br>(+): Die SuS beziehen sich „automatisch", d.h. unaufgefordert aufeinander.<br>(-): Aufforderung bleibt aus und die SuS beziehen sich nicht automatisch aufeinander.<br>(°): In Relation zur Stunde kommt ein Bezugnehmen bzw. die Aufforderung hierzu nicht oft genug vor. | ☐ | ☐ | ☐ |
| E3 | Missverständliche, unvollständige oder unklare Äußerungen werden nicht ignoriert oder lediglich kommentiert, sondern es wird nachgefragt.<br>*Beispiele/Indikatoren:*<br>(+): „Wie genau meinst du das, wenn du sagst: ‚Eine Voraussetzung für die Photosynthese ist Licht.' Was genau ist Licht?", „Ja, das ist noch nicht ganz verständlich, versuche deinen Aussage zu konkretisieren.", „Meinst du damit, dass...?"<br>(-): „Das reicht mir nicht. Wer kann es besser?", „Marcel, was ist deine Antwort?"<br>(°): Nachfragen erfolgen, aber nicht regelmäßig oder systematisch. | ☐ | ☐ | ☐ |
| E4 | Die Lehrperson unterstützt die Schülerinnen und Schüler bei der Ausformulierung bzw. verbalen Ausführungen von Ideen.<br>*Beispiele/Indikatoren:*<br>(+): „Beziehe noch ... mit ein.", „Kannst du den letzten Punkt konkreter ausführen?", „Könnte man hier noch hinzufügen, dass...?", „Meinst du damit, dass...?"<br>(-): Unterstützungsmaßnahmen bleiben aus.<br>(°): Nachfragen/unterstützende Impulse erfolgen, aber in Relation zur Stunde nicht oft genug. Die Lehrperson unterstützt die SuS lediglich indem sie deren Äußerungen konkretisiert.<br>(n.b.) Es werden keine Ideen von den SuS geäußert oder Unterstützungsmaßnahmen sind an keiner Stelle erforderlich.<br><br>Es geht nicht darum zu erfassen, inwiefern die Lehrperson die SuS beim Finden von Lösungen und Antworten unterstützt. Es geht darum, wie die Lehrperson SuS unterstützt, die Probleme haben ihre Gedanken zu versprachlichen. | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| E5 | Die Lehrperson fordert Begründungen für Behauptungen und Vorschläge ein bzw. es werden auch ohne Aufforderung Argumente genannt, um Vorschläge und Behauptungen zu begründen.<br>*Beispiele/Indikatoren:*<br>(+): „Bitte begründe deine Aussage!“, „Hast du hierfür eine Begründung?“<br>(-): Begründungen/Argumente werden weder eingefordert noch werden sie automatisch genannt.<br>(°): In Relation zur Stunde werden Begründungen/Argumente nicht häufig genug eingefordert oder geäußert. | ☐ | ☐ | ☐ |
| E6 | Die Lehrperson liefert nicht sofort bei einer Schülerantwort eine Bewertung, sondern gibt den Ball an andere Schülerinnen und Schüler oder die ganze Klasse weiter.<br>*Beispiele/Indikatoren:*<br>(+): „Aha, was denkt ihr darüber?“,<br>(-): „Nein, das stimmt nicht.“, „Ja, völlig korrekt, das ist die Lösung.“<br>(°): In Relation zur Stunde, gibt die Lehrkraft zwar Äußerungen ohne Wertung zur Diskussion an die Klasse weiter, jedoch in Relation zur Stunde nicht oft genug. | ☐ | ☐ | ☐ |
| E7 | Die Lehrperson gibt den Lernenden Zeit, Ideen und Antworten zu finden<br>*Beispiele/Indikatoren:*<br>(+): Wartezeit bei Schülerantworten.<br>(-): Lehrkraft nimmt sofort den ersten Schüler dran, der sich meldet.<br>(°): Überwiegend lässt die Lehrkraft genügend Zeit, jedoch hin und wieder nicht.<br>Es geht nicht nur um Ideen und Antworten während des Unterrichtsgesprächs, d.h. auch schriftliche Arbeitsaufträge sollten einbezogen werden, sofern sie kognitiv aktivierende Aufgaben beinhalten (d.h. Tafelabschrieb zählt nicht dazu).<br>Es geht nicht darum, ob eine Lehrkraft überhaupt Raum für Schülerideen und Schülerantworten einräumt, sondern ob sie genügend Zeit lässt, wenn Sie Raum dafür schafft. | ☐ | ☐ | ☐ |
| E8 | Der Lehrkraft gelingt es, die Schüler durch eigene Beiträge aktiv am Unterricht zu beteiligen (Gruppenfokus).<br>*Beispiele/Indikatoren:*<br>(+): deutlich mehr als die Hälfte<br>(-): deutlich weniger als die Hälfte<br>(°): etwa die Hälfte<br>(n.b.): SuS haben zwar keine Möglichkeit sich vor der Klasse zu äußern, sind aber nicht nur passive Zuhörer, sondern anderweitig aktiv, z.B. in Kleingruppen. | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Lehrperson als Mediator*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| E9 | Hohe Mediationsfunktion innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**Anmerkungen:**

- Fokus bei der Beurteilung sollte dabei immer darauf liegen, inwiefern die Lehrperson Kooperation und das Lernen der gesamten Klasse „miteinander“ ermöglicht.

**F) Rezeptives Lernverständnis der Lehrperson**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit im Handeln der Lehrperson ein rezeptives Verständnis von Lernen, im Gegensatz zu einem konstruktivistischen Lernen, erkennen lässt. Hierzu gehören beispielsweise das Festhalten an genauen Vorstellungen, wie Aufgaben zu bearbeiten oder an sehr engen Vorgaben, wie Experimente durchzuführen sind, sowie ein enges Frageverhalten. Nach dem konstruktivistischen Lernverständnis bildet ein rezeptives, starres Vorgehen einen negativen Indikator für die Unterstützung der Konstruktion von Wissen.

*Quellen:* Lipowsky & Rakoczy (2006), adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| F1 | Die Lehrperson zeigt ein kleinschrittiges Frageverhalten (stellt also häufig Fragen, die nur eine Antwort zulassen oder die mit einem Wort oder Begriff beantwortet werden können).<br>*Beispiele/Indikatoren:*<br>(+): „Wie heißt diese Struktur im Samen?"- „Wie nennt man den grünen Farbstoff in Pflanzen?", Verwendung von Ja-/Nein-Fragen<br>(-): „Erläutere die Funktion der gesuchten Struktur."<br>(°): Es kommen auch Wie- und Warum-Fragen vor, aber der oben beschriebene Fragetypus dominiert, oder die gestellten Wie-und Warum-Fragen erfordern lediglich den Rückbezug auf eine zuvor gelernte Definition. | ☐ | ☐ | ☐ |
| F2 | Die Lehrkraft gibt kleinschrittige, rezeptartige Arbeitsanweisungen.<br>*Beispiele/Indikatoren:*<br>(+): Die SuS arbeiten die Anweisungen ab, ohne dass eine anspruchsvolle Eigenleistung zu erkennen ist.<br>(-): Die Arbeitsanweisungen fordern die SuS zum Mitdenken heraus. Die Arbeitsanweisungen fordern Ideen der Schüler ein.<br>(°): Es gibt zwar auch Arbeitsanweisungen, die die SuS zum Mitdenken auffordern, die SuS haben aber kaum Möglichkeiten eigene Ideen einzubringen. | ☐ | ☐ | ☐ |
| F3 | Die Schülerinnen und Schüler nehmen die Rollen von Stichwortgebern ein.<br>(*Beispiel:*<br>(+): „Und hier haben wir?" – „Einen Samen." – „Genau, und das ist…" Die SuS müssen lediglich die Sätze der Lehrperson vervollständigen.<br>(-): „Erläutere die Zusammenhänge der Samenstrukturen."<br>(°): Die SuS fungieren zum Teil als Stichwortgeber. Es kommen aber auch elaboriertere Fragen vor. | ☐ | ☐ | ☐ |
| F4 | Die Lehrperson betont das genaue Auswendiglernen und Wiedergeben von fachlichen Inhalten.<br>(*Beispiele:*<br>(+): sehr viele Merksätze, starre Begriffserklärungen, die Lernenden wiederholen Inhalte auf genau eine Weise, etc.<br>(-): Verallgemeinerungen, Zusammenhänge, Konzepte<br>(°): Das Erarbeiten von Zusammenhängen und Verallgemeinerungen ist zwar zum Teil vorhanden, aber dennoch liegt ein Schwerpunkt auf dem Auswendiglernen. Keine eindeutige Tendenz ableitbar.<br>Wichtig: Ein Merksatz ist per se kein Hinweis auf ein rezeptives Lernverständnis. Es geht in dieser Dimension eher um eine Überbetonung des Auswendiglernens eines genau festgelegten Satzes im Gegensatz zum „Verstehen" eines Begriffs. | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Rezeptives Lernverständnis der Lehrperson*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| F5 | Hohes rezeptives Lernverständnis innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**Anmerkungen:**

- Rezept wird hier als Metapher für eine genau festgelegte Schrittfolge von Lernen verstanden.
- Dieses Merkmal ist ein bewusst negatives Merkmal zur Kontrastierung.

**G) Herausfordernde Lerngelegenheiten**

*Grundidee:* In diesem Merkmal wird erfasst, inwieweit im Unterrichtshandeln der Lehrperson herausfordernde Lerngelegenheiten beobachtbar sind. Solche Lerngelegenheiten zeichnen sich dadurch aus, dass sie die Lernenden zum Nachdenken und Überlegen bringen und so kognitive Konflikte erzeugen, in denen die Lernenden erkennen, dass ihr bisheriges Wissen und ihre bisherigen Vorstellungen nicht ausreichend sind und neue Konzepte beziehungsweise das neu zu lernende Wissen plausibler und nützlicher sind. Dies geschieht beispielsweise durch offene und komplexe Aufgabenstellungen, Fragen oder durch offenere Experimentierumgebungen. Der Fokus der Beurteilung liegt dabei auf den Instruktionen der Lehrperson, nicht darauf, ob die Schülerinnen und Schüler die Aufgabenstellungen lösen. Herausfordernde Lerngelegenheiten motivieren und fördern den aktiven Konstruktionsprozess beim Lernen und bilden daher ein Merkmal für die „Aktivierung/Konstruktion von Wissen".

*Quellen:* Rakoczy & Pauli (2006), Widodo (2001) adaptiert

**Bitte geben Sie an, für wie stark ausgeprägt Sie die folgenden Unterrichtshandlungen halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| G1 | Die Lehrperson stellt Aufgaben- oder Fragestellungen, die mehr als nur Ja- oder Nein-Antworten bedürfen.<br>*Beispiele/Indikatoren:*<br>(+): „Warum benötigt eine Pflanze Licht?"<br>(-): „Kann eine Pflanze ohne Licht leben?"<br>(°): Beide Fragetypen kommen vor und es ist keine Tendenz ersichtlich. | ☐ | ☐ | ☐ |
| G2 | Die Lehrperson legt einen Schwerpunkt auf Aufgaben- und Fragestellungen, die zum Nachdenken anregen.<br>*Beispiele/Indikatoren:*<br>(+): Antworten der SuS sind nicht spontan verfügbar. SuS müssen nachdenken.<br>(-): Die SuS können auf Basis ihres bisherigen Wissens oder ohne nennenswerten Denkaufwand (z.B. auf Grundlage des Hefteintrags) die Fragen unmittelbar beantworten. Oder die SuS sind so mit der Aufgaben- bzw. Fragestellung überfordert, dass sie gar nicht erst anfangen darüber nachzudenken.<br>(°): Es kommen oben geschilderte Fragestellungen vor, aber es ist keine Tendenz ersichtlich.<br>(°): Es kommen oben geschilderte Fragestellungen vor, aber diese sind teilweise überfordernd oder den SuS wird häufig schon geholfen, bevor sie wirklich nachdenken konnten. | ☐ | ☐ | ☐ |
| G3 | Es werden Aufgaben- oder Fragestellungen verwendet, die kognitiv anspruchsvolle Aktivitäten des Vergleichens und Analysierens erfordern.<br>*Beispiele/Indikatoren:*<br>(+): „Vergleicht mal einen Kaktus mit einem Baum. Ein Kaktus hat ja keine Blätter. Betreibt er dennoch Photosynthese?"<br>(-): „Nennt mir Organismen, die Photosynthese betreiben."<br>(°): Es erfolgen Vergleiche und Analyseprozesse, aber diese sind entweder nicht wirklich anspruchsvoll oder sie überfordern die SuS.<br>(°): Die geforderten Aktivitäten kommen in Relation zur Stunde nicht oft genug vor. | ☐ | ☐ | ☐ |
| G4 | Die Lehrperson erfragt in Experimentiersituationen nach Hypothesen der Schülerinnen und Schüler.<br>*Beispiele/Indikatoren:*<br>(+): „Was sind eure Vermutungen diesbezüglich?", „Versucht eure Vermutungen als wissenschaftliche Hypothesen zu formulieren.", „Welches Ergebnis erwartet ihr?"<br>(-): Wir machen jetzt ein Experiment. Bitte geht wie in der Beschreibung vor und führt die Arbeitsschritte der Reihe nach durch.<br>(°): Die Lehrkraft fragt nach Hypothesen, gibt sich aber mit einem Schülerbeitrag | ☐ | ☐ | ☐ |

| | | | | |
|---|---|---|---|---|
| | zufrieden und fragt nicht erneut nach.<br>(n.b.) Es kommt keine Experimentiersituation in der Sequenz vor oder die vorkommenden Experimentiersituationen sind so gestaltet, dass eine Hypothesenbildung nicht sinnvoll erscheint. (Bsp.: Die Lehrperson veranschaulicht den SuS anhand von kleinen Versuchen die Wirkungen von Kraft und die SuS sollen anschließend beschreiben wir Kraft aussieht.) | | | |
| G5 | Die Lehrperson stellt Aufgaben- oder Fragestellungen, die nicht nur auswendig gelerntes Wissen abfragen (oder auf reine Beobachtung abzielen).<br>*Beispiele/Indikatoren:*<br>(+): „Wenn wir betrachten, was wir bisher gelernt haben, zu welchen Bereichen wissen wir noch wenig?“, „Welche Schlussfolgerungen lässt dieses Ergebnis zu?“<br>(-): „Was ist das Symbol für Arbeit?“, „Nennt mir ...“, „Wie lautet der Merksatz zu...“<br>(°): Es kommen zwar die geforderten Aufgaben- und Fragestellungen vor, jedoch in Relation zur Stunde nicht häufig genug. | ☐ | ☐ | ☐ |

***Gesamteindruck:*** *Herausfordernde Lerngelegenheiten*

**Bitte geben Sie an, für wie stark ausgeprägt Sie das gesamte Merkmal halten.**

| | | *trifft nicht zu* | *teils teils* | *trifft zu* |
|---|---|---|---|---|
| G6 | Hohe Herausforderung innerhalb der gesamten Analyseeinheit (Stunde). | ☐ | ☐ | ☐ |

**Anmerkungen:**

- Hilfreich ist es, sich bei der Beurteilung die Frage zu stellen, wie komplex beziehungsweise „herausfordernd“ die den Schülerinnen und Schülern gestellten Aufgaben- und Fragestellungen ausfallen. Ziel ist immer die Anregung zum Nachdenken und Verständnis.
- Eine zu komplexe Aufgaben- oder Fragestellung kann zur Überforderung der Lernenden führen und ist nicht mehr herausfordernd oder aktivierend.

# B. Ergänzende Tabellen und Abbildungen

## B.1. Tabellen


B.1. Stichprobenübersicht 1: Merkmale der Lehrkräfte, Testwerte in den Professionswissenstests, Merkmale der videographierten Unterrichtsstunden und Qualitätsmaße zur kognitiven Aktivierung . . . 242
B.2. Stichprobenübersicht 2: Klassenmerkmale, auf Klassenebene aggregierte Testwerte und Maße zum situationalen Interesse der Lernenden sowie Merkmale der Unterrichtseinheit . . . . . 243
B.3. Prüfung der Normalverteilung aller betrachteter Variablen . . . . 244
B.4. Interrater-Übereinstimmung für die PCK- und CK-Aufgaben . . . 245
B.5. Korrelationen zwischen den Dimensionen des Professionswissens in der um die ProwiN I-Gymnasiallehrkräfte aus NRW erweiterten Stichprobe der ProwiN II-Lehrkräfte . . . . . 246
B.6. Korrelationen zwischen den Level-1-Prädiktoren in den Mehrebenenmodellen für die Post-Testwerte der Lernenden im Fachwissen 246
B.7. Nicht-parametrische Effektstärken für die Zusammenhänge der metrischen Level-1-Prädiktoren mit den dichotomen Level-1-Prädiktoren in den Mehrebenenmodellen für die Post-Testwerte der Lernenden im Fachwissen . . . . . 247
B.8. Korrelationen zwischen den Level-2-Prädiktoren in den Mehrebenenmodellen für die Post-Testwerte der Lernenden im Fachwissen 247
B.9. Pearson-Korrelationen und gegen Außreißer robuste Rangkorrelationen nach Kendall zwischen den Dimensionen des Professionswissens und den effektiven Leistungszuwächsen . . . . . 248

Table B.1.

*Stichprobenübersicht 1: Merkmale der Lehrkräfte, Testwerte in den Professionswissenstests, Merkmale der videographierten Unterrichtsstunden und Qualitätsmaße zur kognitiv aktivierenden Gestaltung der 1. und 2. videographierten Unterrichtsstunde (1M/2M) und über beide Unterrichtsstunden gemittelte Qualitätsmaße (1M&2M)*

| | Merkmale der Lehrkräfte | | | | | | | Videomerkmale | | | Qualitätsmaße Kognitive Aktivierung[3] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Alter [Jahre] | Geschlecht [♀/♂] | Abitur-note | Jahre im Schuldienst | Physikst. pro Woche | CK [R-L][1] | PCK [R-L][1] | PK [%][2] | Stundenlänge [min] | Tage zw. Prä-1M | Tage zw. 1M-2M | 1M | 2M | 1M&2M |
| 1 | 62 | ♀ | 1.9 | 36 | 12 | 1.5 | −0.1 | 63 | 90.0 | 36 | 7 | 1.2 | 1.2 | 1.2 |
| 3 | 42 | ♂ | 1.4 | 14 | 7 | 1.5 | 0.1 | 70 | 45.0 | 7 | 164[4] | 2.5 | 1.9 | 2.2 |
| 5[5] | 33 | ♂ | 1.5 | 2 | 10 | 1.9 | 1.0 | 77 | 45.0 | 28 | 53[4] | 1.7 | 2.0 | 1.9 |
| 6 | 41 | ♀ | 1.9 | 10 | 17 | 0.3 | −0.7 | 83 | 70.0 | 21 | 7 | 2.5 | 1.4 | 1.9 |
| 9 | 42 | ♂ | 2.7 | 11 | 7 | 1.0 | 0.2 | 67 | 45.0 | 28 | 3 | 2.1 | 2.2 | 2.1 |
| 10 | 31 | ♂ | 1.7 | 2 | 7 | 1.5 | 0.2 | 80 | 45.0 | 28 | 3 | 2.1 | 2.1 | 2.1 |
| 12 | 46 | ♂ | 2.1 | 15 | 15 | 1.0 | 0.2 | 67 | 60.0 | 22 | 6 | 2.6 | 1.4 | 2.0 |
| 13 | 41 | ♀ | 2.8 | 2 | 4 | −1.2 | −0.3 | 77 | 60.0 | 55 | 1 | 1.8 | 1.7 | 1.8 |
| 14 | 63 | ♂ | 2.2 | 36 | 9 | 1.9 | −0.5 | 83 | 90.0 | 7 | 7 | 2.4 | 1.7 | 2.0 |
| 15 | 38 | ♂ | 1.3 | 6 | 16 | 1.2 | 0.1 | 73 | 90.0 | 23 | 5 | 2.3 | 2.0 | 2.1 |
| 16 | 57 | ♂ | 1.6 | 31 | 8 | 0.8 | −0.1 | 67 | 90.0 | 91 | 7 | 2.0 | 1.8 | 1.9 |
| 17 | 60 | ♀ | 1.7 | 12 | 3 | −0.2 | 0.8 | 80 | 67.5 | 70 | 7 | 1.5 | 1.2 | 1.4 |
| 18 | 30 | ♀ | 2.1 | 3 | 14 | −0.7 | −0.1 | 60 | 90.0 | 28 | 7 | 1.6 | 1.5 | 1.5 |
| 19 | 49 | ♂ | 2.3 | 15 | 10 | −0.7 | −0.7 | 73 | 60.0 | 1 | 6 | 1.5 | 1.5 | 1.5 |
| 20 | 60 | ♂ | 1.6 | 31 | 9 | 1.0 | −0.1 | 67 | 90.0 | 7 | 7 | 1.4 | 1.5 | 1.5 |
| 22 | 28 | ♀ | 1.9 | 3 | 10 | −1.5 | −0.3 | 80 | 90.0 | 18 | 28[6] | 1.6 | 1.5 | 1.6 |
| 23 | 35 | ♂ | 2.2 | 6 | 5 | −0.8 | 0.8 | 70 | 90.0 | 7 | 7 | 2.1 | 2.2 | 2.1 |
| 25 | 29 | ♀ | 1.3 | 3 | 11 | 0.8 | 0.2 | 83 | 45.0 | 56 | 7 | 2.4 | 2.1 | 2.3 |
| 28 | 36 | ♂ | 2.2 | 6 | 22 | 0.4 | 0.2 | 77 | 90.0 | 17 | 11[6] | 2.2 | 2.2 | 2.2 |
| 29 | 43 | ♂ | 1.4 | 10 | 14 | 0.6 | 0.6 | 70 | 67.5 | 12 | 2 | 2.2 | 1.2 | 1.7 |
| 41 | 35 | ♀ | 1.8 | 8 | 7 | −1.0 | 0.1 | 70 | 45.0 | 14 | 5 | 1.7 | 1.7 | 1.7 |
| 43 | 60 | ♂ | 1.8 | 31 | 15 | −0.2 | −0.7 | 60 | 67.5 | 7 | 7 | 1.5 | 1.9 | 1.7 |
| 44[7] | 50 | ♂ | 2.5 | 4 | 10 | 0.1 | −0.1 | 63 | 45.0 | 19 | 2 | 1.5 | 1.4 | 1.5 |

[1] Angabe in Rasch-Logits
[2] Angabe in % gelöster Aufgaben
[3] Die kognitiv aktivierende Unterrichtsgestaltung wurde auf einer dreistufigen Ratingskala eingeschätzt (1 =„trifft nicht zu“, 2 =„teils teils“, 3 =„trifft zu“).
[4] In diesen Klassen war eine Aufzeichnung aufeinanderfolgender Unterrichtsstunden nicht möglich. Hier wurde zu einem späteren Zeitpunkt eine Einführungsstunde in ein anderes, von den Lehrkräften frei gewähltes, Unterthema der Mechanik aufgezeichnet.
[5] Diese Lehrkraft stand kurz vor Abschluss ihres Vorbereitungsdienstes (Referendariats) im Rahmen der Ordnung zur berufsbegleitenden Ausbildung von Seiteneinsteigerinnen und Seiteneinsteigern und der Staatsprüfung (OBAS).
[6] In diesen Klassen lag aufgrund von Stundenausfall und aufgrund der Herbstferien ein größerer Zeitraum zwischen den Videoerhebungen, dennoch handelte es sich um aufeinanderfolgende Unterrichtsstunden.
[7] Seiteneinsteiger ohne Lehrbefähigung im Fach Physik.

Table B.2.

*Stichprobenübersicht 2: Klassenmerkmale, auf Klassenebene aggregierte Testwerte der Schülerinnen und Schüler (SuS) und Maße zum situationalen Interesse der Lernenden in der 1. und 2. videographierten Unterrichtsstunde (1M/2M) und über beide Unterrichtsstunden gemittelte Maße (1M&2M) sowie Merkmale der Unterrichtseinheit*

| | Klassenmerkmale[1] | | | | Aggr. Testwerte SuS[2] | | | Aggr. Interessensmaße SuS[3] | | | Merkmale Unterrichtseinheit | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Jahrgangs-stufe | Anz. SuS | Ant. ♀ [%] | Ant. mehr-sprachig [%] | Prä [R-L][4] | Post [R-L][4] | Differenz Post-Prä [R-L][4] | KFT [R-L][4] | 1M | 2M | 1M&2M | Tage zw. Prä-Post | Stundenausfall [%] | Unterrichtszeit [Anz. 45-min-St.] |
| 1 | 8 | 31 | 58 | 19 | 0.3 | 0.3 | 0.0 | 1.6 | 3.4 | 3.4 | 3.4 | 85 | 7 | 26 |
| 3 | 8 | 28 | 52 | 7 | 0.3 | 0.9 | 0.6 | 0.7 | 4.4 | 4.5 | 4.5 | 198 | 2 | 43 |
| 5 | 8 | 29 | 52 | 34 | 0.1 | 0.3 | 0.2 | 0.8 | 3.8 | 3.6 | 3.7 | 107 | 8 | 23 |
| 6 | 9 | 30 | 48 | 37 | 0.6 | 1.0 | 0.4 | 2.2 | 4.8 | 4.5 | 4.7 | 126 | 0 | 23 |
| 9 | 8 | 34 | 69 | 9 | −0.2 | 0.7 | 0.8 | 0.5 | 4.7 | 5.1 | 5.0 | 197 | 19 | 38 |
| 10 | 8 | 31 | 14 | 10 | 0.4 | 1.0 | 0.6 | 0.2 | 4.2 | 4.5 | 4.3 | 197 | 16 | 38 |
| 12 | 8 | 33 | 41 | 18 | 0.2 | 0.8 | 0.6 | 0.0 | 4.6 | 4.9 | 4.7 | 183 | 5 | 53 |
| 13 | 8 | 32 | 41 | 31 | 0.3 | 0.8 | 0.5 | 0.9 | 3.6 | 4.0 | 3.8 | 202 | 4 | 59 |
| 14 | 9 | 28 | 32 | 0 | 0.9 | 1.2 | 0.3 | 2.5 | 4.9 | 4.8 | 4.9 | 168 | 14 | 36 |
| 15 | 9 | 30 | 43 | 7 | 0.5 | 0.9 | 0.4 | 1.6 | 4.0 | 3.4 | 3.7 | 149 | 0 | 31 |
| 16 | 8 | 25 | 100 | 28 | −0.1 | −0.1 | 0.0 | 0.5 | 3.5 | 3.1 | 3.3 | 310 | 49[5] | 25 |
| 17 | 8 | 30 | 55 | 47 | 0.1 | 0.2 | 0.1 | 0.4 | 3.7 | 4.6 | 4.2 | 203 | 0 | 35 |
| 18 | 8 | 29 | 64 | 28 | −0.2 | 0.3 | 0.5 | 0.4 | 4.8 | 4.7 | 4.8 | 203 | 26 | 34 |
| 19 | 9 | 31 | 60 | 19 | 0.7 | 1.0 | 0.3 | 2.5 | 4.1 | 3.6 | 3.8 | 248 | 58[5] | 31 |
| 20 | 8 | 28 | 85 | 11 | 0.3 | 0.9 | 0.6 | 0.6 | 4.7 | 3.7 | 4.1 | 168 | 0 | 41 |
| 22 | 8 | 31 | 48 | 16 | 0.6 | 1.1 | 0.5 | 0.5 | 4.6 | 4.5 | 4.6 | 228 | 20 | 40 |
| 23 | 9 | 29 | 100 | 17 | 0.3 | 0.4 | 0.1 | 1.4 | 5.0 | 5.1 | 5.1 | 98 | 0 | 21 |
| 25 | 8 | 31 | 57 | 26 | 0.1 | 0.7 | 0.5 | 0.0 | 4.6 | 4.7 | 4.7 | 269 | 3 | 30 |
| 28 | 8 | 23 | 59 | 26 | 0.2 | 0.6 | 0.4 | −0.3 | 4.1 | 3.7 | 3.9 | 166 | 5 | 40 |
| 29 | 8 | 24 | 68 | 33 | 0.1 | 0.5 | 0.4 | 1.2 | 4.3 | 4.7 | 4.5 | 110 | 0 | 38 |
| 41 | 9 | 20 | 42 | 0 | 0.7 | 0.9 | 0.2 | 1.9 | 5.0 | 3.9 | 4.4 | 138 | 10 | 26 |
| 43 | 9 | 25 | 64 | 32 | 0.0 | 0.2 | 0.3 | 1.4 | 4.9 | 4.1 | 4.5 | 70 | 11 | 12 |
| 44 | 9 | 28 | 48 | 21 | 0.5 | 0.4 | 0.0 | 1.7 | 3.4 | 2.8 | 3.1 | 194 | 27 | 30 |

[1] Angaben beziehen sich auf die $N = 660$ SuS, die bei Prä- *oder* Post-Test anwesend waren und für die daher Daten zum demographischen Hintergrund vorlagen.

[2] Aggregierte Daten der $N = 610$ bei Prä- *und* Post-Test anwesenden SuS.

[3] Aggregierte Daten der $N_{1M} = 633$, $N_{2M} = 625$ bzw. $N_{1M\&2M} = 600$ in der 1./2. Unterrichtsstunde bzw. in beiden Unterrichtstunden anwesenden SuS. Das situationale Interesse wurde von den Lernenden auf einer siebenstufigen Likertskala eingeschätzt (1 =„stimme gar nicht zu“, 7 =„stimme voll zu“).

[4] Angabe in Rasch-Logits

[5] Diese Lehrkräfte waren über einen längeren Zeitraum krank. Der Physikunterricht wurde in ihrer Abwesenheit nicht vertreten.

Table B.3.
*Freiheitsgerade (df), Statistiken (W) und Signifikanzen (p) der Shapiro-Wilk-Tests auf Normalverteilung sowie Werte und z-Werte der Schiefe und Kurtosis für alle betrachteten Variablen*

| Auf Normalverteilung | Shapiro-Wilk | | | Schiefe | | Kurtosis | | Sign. Abw. |
|---|---|---|---|---|---|---|---|---|
| geprüfte Variablen | *df* | *W* | *p* | Wert | *z* | Wert | *z* | von Normal |
| **Testwerte LK** | | | | | | | | |
| ProwiN II | | | | | | | | |
| CK | 23 | .946 | .242 | -0.3±0.5 | -0.64 | -1.1±1.0 | -1.15 | nein |
| PCK | 23 | .949 | .278 | 0.3±0.5 | 0.54 | -0.3±1.0 | -0.31 | nein |
| PK | 23 | .940 | .181 | 0.0±0.5 | 0.05 | -1.1±1.0 | -1.20 | nein |
| ProwiN I | | | | | | | | |
| CK | 79 | .977 | .167 | 0.33±0.28 | 1.22 | -0.3±0.6 | -0.52 | nein |
| PCK | 79 | .913 | .000 | -1.38±0.28 | -5.08 | 4.5±0.6 | 8.36 | ja[1] |
| PK | 79 | .924 | .000 | -1.14±0.28 | -4.20 | 2.0±0.6 | 3.76 | ja[1] |
| ProwiN I+II | | | | | | | | |
| CK | 102 | .975 | .053 | 0.23±0.24 | 0.95 | -0.7±0.5 | -1.37 | nein |
| PCK | 102 | .915 | .000 | -1.41±0.24 | -5.90 | 5.2±0.5 | 10.88 | ja[1] |
| PK | 102 | .919 | .000 | -1.22±0.24 | -5.10 | 2.7±0.5 | 5.60 | ja[1] |
| **Video-Maße** | | | | | | | | |
| Kognitive Aktivierung$_{1M}$ | 23 | .955 | .366 | 0.0±0.5 | 0.10 | -1.2±1.0 | -1.26 | nein |
| Skala A | 23 | .925 | .086 | 0.4±0.5 | 0.92 | 0.4±1.0 | 0.45 | nein |
| Skala B | 23 | .914 | .049 | 0.5±0.5 | 1.03 | -0.9±1.0 | -0.94 | ja |
| Skala C | 23 | .876 | .008 | 0.1±0.5 | 0.30 | -1.5±1.0 | -1.59 | ja |
| Skala D | 23 | .912 | .045 | 0.5±0.5 | 1.06 | -0.8±1.0 | -0.80 | ja |
| Skala E | 23 | .944 | .218 | -0.5±0.5 | -1.08 | -0.6±1.0 | -0.64 | nein |
| Skala F | 23 | .869 | .006 | -1.1±0.5 | -2.24 | 1.0±1.0 | 1.08 | ja |
| Skala G | 23 | .917 | .058 | 0.4±0.5 | 0.75 | -0.9±1.0 | -1.00 | nein |
| Kognitive Aktivierung$_{2M}$ | 23 | .922 | .073 | 0.1±0.5 | 0.12 | -1.4±1.0 | -1.49 | nein |
| Skala A | 23 | .918 | .061 | 0.6±0.5 | 1.32 | -0.3±1.0 | -0.32 | nein |
| Skala B | 23 | .763 | .000 | 1.4±0.5 | 2.90 | 1.5±1.0 | 1.59 | ja |
| Skala C | 23 | .847 | .002 | 0.4±0.5 | 0.75 | -1.5±1.0 | -1.59 | ja |
| Skala D | 23 | .858 | .004 | 1.2±0.5 | 2.57 | 1.8±1.0 | 1.93 | ja |
| Skala E | 23 | .930 | .109 | -0.5±0.5 | -1.01 | -0.2±1.0 | -0.18 | nein |
| Skala F | 23 | .929 | .104 | -0.5±0.5 | -1.02 | -0.4±1.0 | -0.43 | nein |
| Skala G | 23 | .862 | .005 | 0.2±0.5 | 0.40 | -1.5±1.0 | -1.63 | ja |
| Kognitive Aktivierung$_{1M\&2M}$ | 23 | .941 | .193 | -0.3±0.5 | -0.55 | -1.2±1.0 | -1.24 | nein |
| Skala A | 23 | .970 | .694 | 0.1±0.5 | 0.25 | -0.9±1.0 | -0.99 | nein |
| Skala B | 23 | .927 | .093 | 0.4±0.5 | 0.83 | -0.9±1.0 | -0.95 | nein |
| Skala C | 23 | .938 | .162 | 0.1±0.5 | 0.31 | -1.1±1.0 | -1.16 | nein |
| Skala D | 23 | .933 | .128 | 0.9±0.5 | 1.79 | 0.8±1.0 | 0.89 | nein |
| Skala E | 23 | .961 | .493 | -0.6±0.5 | -1.28 | 0.2±1.0 | 0.24 | nein |
| Skala F | 23 | .960 | .466 | -0.5±0.5 | -1.04 | 0.1±1.0 | 0.15 | nein |
| Skala G | 23 | .943 | .211 | 0.3±0.5 | 0.67 | -1.0±1.0 | -1.10 | nein |
| Klassenführung$_{1M}$ | 23 | .899 | .024 | -0.9±0.5 | -1.77 | -0.2±1.0 | -0.17 | ja |
| Klassenführung$_{2M}$ | 23 | .954 | .361 | -0.6±0.5 | -1.26 | 0.2±1.0 | 0.20 | nein |
| Vernetzheit$_{1M}$ | 23 | .985 | .968 | 0.2±0.5 | 0.36 | 0.6±1.0 | 0.63 | nein |
| **Testwerte SuS** | | | | | | | | |
| Prä | 610 | .992 | .003 | 0.21±0.10 | 2.15 | 0.26±0.20 | 1.31 | ja[1] |
| Post | 610 | .957 | .000 | 0.91±0.10 | 9.20 | 2.57±0.20 | 12.96 | ja[1] |
| Differenz Post-Prä | 610 | .987 | .000 | 0.36±0.10 | 3.68 | 1.14±0.20 | 5.78 | ja[1] |
| KFT | 610 | .991 | .001 | 0.10±0.10 | 1.02 | 0.15±0.20 | 0.78 | ja[1] |
| Sit. Interesse$_{1M}$ | 633 | .989 | .000 | -0.26±0.10 | -2.69 | -0.12±0.20 | -0.60 | ja[1] |
| Sit. Interesse$_{2M}$ | 625 | .984 | .000 | -0.24±0.10 | -2.43 | -0.44±0.20 | -2.25 | ja[1] |
| Sit. Interesse$_{1M\&2M}$ | 600 | .992 | .002 | -0.25±0.10 | -2.51 | -0.17±0.20 | -0.85 | ja[1] |
| **Klassenwerte** | | | | | | | | |
| Post[2] | 23 | .955 | .373 | -0.4±0.5 | -0.76 | -0.8±1.0 | -0.81 | nein |
| Effektiver LZW[3] | 23 | .955 | .373 | -0.3±0.5 | -0.63 | -0.7±1.0 | -0.73 | nein |
| Sit. Interesse$_{1M}$[4] | 23 | .913 | .047 | -0.4±0.5 | -0.82 | -1.2±1.0 | -1.32 | ja |
| Sit. Interesse$_{2M}$[4] | 23 | .947 | .251 | -0.4±0.5 | -0.89 | -0.8±1.0 | -0.89 | nein |
| Unterrichtszeit | 23 | .970 | .678 | 0.4±0.5 | 0.89 | 0.9±1.0 | 0.91 | nein |

*Anmerkung:* Die *z*-Werte der Schiefe und Kurtosis berechnen sich durch Division der unnormierten Werte durch ihre Standardfehler. *z*-Werte > 1.96 ( > 2.58 bzw. > 3.29) zeigen signifikante Abweichungen von der Normalverteilung mit $p < .05$ ($p < .01$ bzw. $p < .001$) an (Field, 2009, S. 139).

[1] Wegen $N > 30$ wurden auf diese Variablen, trotz signifikanter Abweichung von der Normalverteilung, parametrische Testverfahren angewendet und zusätzlich Ergebnisse entsprechender nicht-parametrische Testverfahren berichtet (vergl. Abschnitt 7.4.1 auf Seite 91).

[2] Auf Klassenebene über den Mittelwert aggregierte Post-Testwerte der Lernenden.

[3] In Abschnitt 9.2.2.1 auf Seite 194 eingeführte effektive Leistungszuwächse.

[4] Auf Klassenebene über den Mittelwert aggregierte Maße für das situationale Interesse der Lernenden.

Table B.4.
*Interrater-Übereinstimmung für die PCK- und CK-Aufgaben aus den fachspezifischen Professionswissenstests ($N_{Rater} = 2$)*

| Aufgabe | Beschreibung | $ICC_{\text{2-fakt.,unjust}}$ | $\text{KI}_{95\,\%}$ |
|---|---|---|---|
| PCK_S020 | Warum Experimente | 0.77 | [0.69,0.84] |
| PCK_S230 | Warum Einheiten | 0.85 | [0.79,0.90] |
| PCK_0261 | Lok | 0.96 | [0.94,0.97] |
| PCK_0051 | Diagramm 1 | 0.94 | [0.91,0.96] |
| PCK_0052 | Diagramm 2 | 0.91 | [0.87,0.93] |
| PCK_0151 | Flugbahn 1 | 0.94 | [0.91,0.96] |
| PCK_0152 | Flugbahn 2 | 0.88 | [0.83,0.92] |
| PCK_0071 | Lampe 1 | 0.88 | [0.83,0.91] |
| PCK_0072 | Lampe 2 | 0.91 | [0.87,0.94] |
| PCK_0180 | Schülervorstellungen Geschwindigkeit | 0.96 | [0.95,0.97] |
| PCK_0280 | Wirkung von Kraft | 0.93 | [0.89,0.95] |
| PCK_0320 | Zeichnung Kraft | 0.84 | [0.77,0.88] |
| PCK_0080 | Wasserrakete | 0.95 | [0.93,0.97] |
| PCK_0040[1] | Stundenfortsetzung Experiment | 0.98 | [0.97,0.99] |
| CK_1150 | Rutsche | 1.00 | [0.99,1.00] |
| CK_1160 | Flugzeug Wind | 0.99 | [0.98,0.99] |
| CK_1240 | Puk | 1.00 | [1.00,1.00] |
| CK_1490 | E-Lehre | 0.96 | [0.95,0.97] |
| CK_1450[1] | Hebel | 0.96 | [0.94,0.97] |
| CK_1410 | Ampel | 0.99 | [0.99,0.99] |
| CK_1300 | Beschleunigung | 0.98 | [0.98,0.99] |
| CK_1470 | Looping | 0.98 | [0.98,0.99] |
| CK_1290 | Schaukel | 1.00 | [1.00,1.00] |
| CK_1180 | Kepler | 0.96 | [0.95,0.98] |
| CK_1140 | Pendel | 0.98 | [0.97,0.99] |
| CK_1220 | LKW | 0.98 | [0.96,0.98] |

*Anmerkungen:* Die angegebenen ICCs beziehen sich auf die Skalenmittelwerte der einzelnen Rater und Raterinnen und nicht auf die über alle Rater und Raterinnen gemittelten Skalenmittelwerte. Die Aufgaben PCK_0051/0052, PCK_0151/0152 bzw. PCK_0071/0172 wurden für die Auswertung zusammengefasst (vergl. Abschnitt 7.5.1.4 auf Seite 105)

[1] Aufgabe wurde wegen schlechter Passung ins Rasch-Modell aus den Analysen ausgeschlossen (vergl. Abschnitt 7.5.1.3 auf Seite 104).

Table B.5.
*Korrelationen zwischen den Dimensionen des Professionswissens in der um die ProwiN I-Gymnasiallehrkräfte aus NRW erweiterten Stichprobe der ProwiN II-Lehrkräfte ($N_{Gesamt} = 102$)*

| Korrelierte Merkmale | PCK-CK | PCK-PK | CK-PK |
|---|---|---|---|
| $r_{\text{Pearson}}$ | **.39 ± .09** | **.27 ± .11** | .15 ± .10 |
| $\text{KI}_{95\,\%}$ | [.20,.54] | [.06,.50] | [-.04,.33] |
| $p_{\text{1-seitig}}$ | < .001 | .003 | .065 |
| $r_{\text{Spearman}}$ | **.32 ± .10** | **.34 ± .10** | **.18 ± .10** |
| $\text{KI}_{95\,\%}$ | [.11,.50] | [.14,.52] | [-.02,.36] |
| $p_{\text{1-seitig}}$ | .001 | < .001 | .037 |
| $\tau_{\text{Kendall}}$ | **.23 ± .08** | **.25 ± .07** | **.13 ± .07** |
| $\text{KI}_{95\,\%}$ | [.08,.38] | [.11,.40] | [-.02,.27] |
| $p_{\text{1-seitig}}$ | .001 | < .001 | .037 |

*Anmerkungen.* Signifikante Werte mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Aufgrund signifikanter Abweichungen von der Normalverteilung werden zusätzlich nicht-parametrische Korrelationen berichtet.

Table B.6.
*Korrelationen zwischen den Level-1-Prädiktoren in den Mehrebenenmodellen für die Post-Testwerte der Lernenden im Fachwissen ($N = 610$)*

| Korrel. Prädiktoren | | Prä-Test | KFT | Geschlecht | Sprache |
|---|---|---|---|---|---|
| | $r_{\text{Pearson}}$ | | **.32 ± .04** | **.24 ± .04** | **−.16 ± .04** |
| Prä-Test | $\text{KI}_{95\,\%}$ | | [.25,.39] | [.16,.31] | [-.23,-.08] |
| | $p$ | | < .001 | < .001 | < .001 |
| | $r_{\text{Pearson}}$ | | | .01 ± .05 | −.07 ± .04 |
| KFT | $\text{KI}_{95\,\%}$ | | | [-.07,.09] | [-.15,.01] |
| | $p$ | | | .857 | .067 |
| | $r_{\text{Pearson}}$ [1] | | | | .01 ± .04 |
| Geschlecht | $\text{KI}_{95\,\%}$ | | | | [-.07,.09] |
| | $p$ | | | | .777 |

*Anmerkungen.* Signifikante Werte mit $p < .05$ sind fett gedruckt. Die Prä- und KFT-Testwerte sind nicht normalverteilt, daher sollten zusätzlich nicht-parametrische Korrelationen berichtet werden. Die nicht-parametrischen Korrelationen zwischen Prä- und KFT-Testwerten beträgt $r_{\text{Spearman}} = .28 \pm .04$ ($\text{KI}_{95\,\%} = [.21, .35]$, $p < .001$) bzw. $\tau_{\text{Kendall}} = .197 \pm .026$ ($\text{KI}_{95\,\%} = [.146, .246]$, $p < .001$). Nicht-parametrische Effektstärken für den Zusammenhang zwischen Prä-und KFT-Testwerten und den dichotomen Prädiktoren Geschlecht und Sprache finden sich in Tabelle B.7.

[1] $r_{\text{Pearson}}$ entspricht für dichotome Merkmale dem Korrelationsmaß $\Phi$ (vergl. Bortz & Lienert, 2008, S. 261)

Table B.7.
*Mann-Whitney-Tests für die Zusammenhänge der metrischen Level-1-Prädiktoren mit den dichotomen Level-1-Prädiktoren in den Mehrebenenmodellen für die Post-Testwerte der Lernenden im Fachwissen ($N = 610$)*

| Prädiktoren | | Geschlecht ♀/♂= 267/343 | Sprache deutsch/nicht nur deutsch= 483/127 |
|---|---|---|---|
| Prä-Test | $U$ | $(5.9 \pm 2.2) \cdot 10^3$ | $(2.3 \pm 1.8) \cdot 10^3$ |
| | $z$ | 6.0 | −4.1 |
| | $r_{\text{MW}}$ | **.24** | **−.17** |
| | $p_{\text{asympt.}}$ | < .001 | < .001 |
| KFT | $U$ | $(4.5 \pm 2.2) \cdot 10^3$ | $(2.8 \pm 1.8) \cdot 10^3$ |
| | $z$ | −.41 | −1.7 |
| | $r_{\text{MW}}$ | −.017 | −.069 |
| | $p_{\text{asympt.}}$ | .686 | .088 |

*Anmerkungen.* Signifikante Werte mit $p_{\text{asympt.}} < .05$ sind fett gedruckt. Die Effektstärken wurde über $r_{\text{MW}} = z/\sqrt{N}$ berechnet (vergl. Field, 2009, S. 550).

Table B.8.
*Korrelationen zwischen den Level-2-Prädiktoren in den Mehrebenenmodellen 1a-c (Professionswissensmodelle) und 2.1a 1M, 2M und 1M&2M (Modelle zur kognitiven Aktivierung) für die Post-Testwerte der Lernenden im Fachwissen. Aufgeführt werden die Korrelationen zwischen der Unterrichtszeit (Anzahl der 45-Minuten-Stunden in der Unterrichtseinheit Mechanik) und dem jeweiligen zusätzlichen Level-2-Prädiktor (Modell 1a-c: CK-, PCK- bzw. PK-Testwert der Lehrkräfte; Modell 2.1a$_{1M/2M/1M\&2M}$: Qualitätsmaß für die kognitive Aktivierung (KA) in der 1. (1M) und 2. Unterrichtsstunde (2M) bzw. das über beide Unterrichtsstunden gemittelte Qualitätsmaß (1M&2M)) ($N = 23$)*

| Prädiktor | CK | PCK | PK | $\text{KA}_{1M}$ | $\text{KA}_{2M}$ | $\text{KA}_{1M\&2M}$ |
|---|---|---|---|---|---|---|
| $r_{\text{Pearson}}$ | −.01 ± .24 | .05 ± .23 | .18 ± .21 | .28 ± .20 | −.12 ± .17 | .13 ± .16 |
| $\text{KI}_{95\,\%}$ | [-.42,.49] | [-.37,.48] | [-.25,.53] | [-.11,.64] | [-.40,.23] | [-.17,.42] |
| $p$ | .982 | .828 | .425 | .200 | .602 | .568 |
| $r_{\text{Spearman}}$ | | | | .27 ± .22 | −.05 ± .22 | .16 ± .20 |
| $\text{KI}_{95\,\%}$ | | | | [-.19,.65] | [-.48,.36] | [-.23,.51] |
| $p$ | | | | .220 | .831 | .462 |
| $\tau_{\text{Kendall}}$ | | | | .16 ± .17 | −.03 ± .16 | .11 ± .15 |
| $\text{KI}_{95\,\%}$ | | | | [-.19,.49] | [-.36,.27] | [-.17,.38] |
| $p$ | | | | .302 | .853 | .475 |

*Anmerkung.* Es werden zusätzlich nicht-parametrische Korrelationen berichtet, da das Intervallskalenniveau der Qualitätsmaße zur kognitiven Aktivierung nicht sichergestellt werden kann (vergl. Abschnitt 7.4.1 auf Seite 92 zum Umgang mit Ordinalskalen).

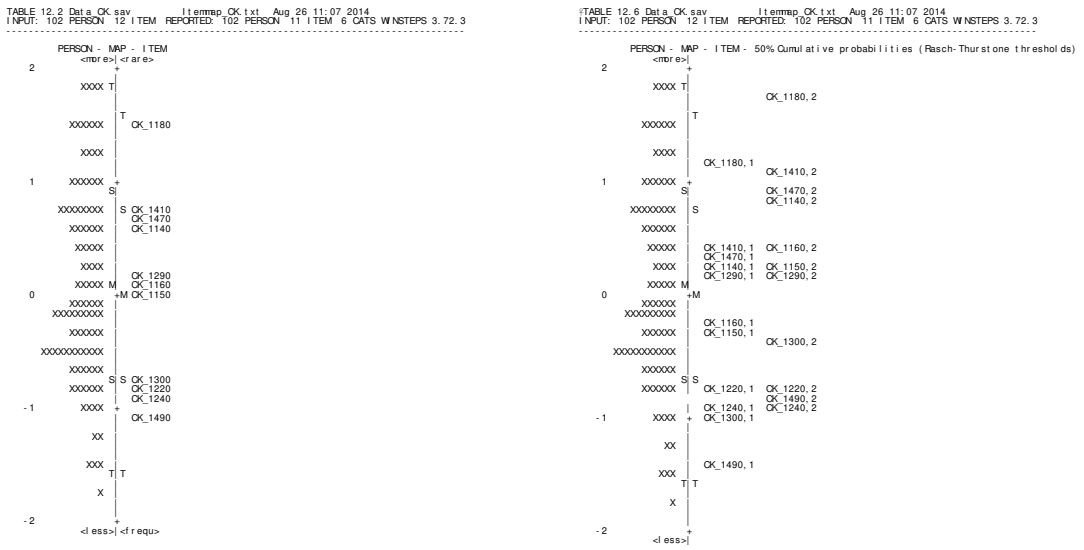
Table B.9.
*Pearson-Korrelationen und gegen Außreißer robuste Rangkorrelationen nach Kendall zwischen den Dimensionen des Professionswissens und den effektiven Leistungszuwächsen (LZW) der Lernenden unter Vernachlässigung der Mehrebenenstruktur*

| Korrelierte Merkmale | CK-LZW | PCK-LZW | PK-LZW |
|---|---|---|---|
| $r_{\text{Pearson}}$ | $.04 \pm .19$ | $-.19 \pm .18$ | $.34 \pm .21$ |
| $\text{KI}_{95\,\%}$ | [-.33,.40] | [-.51,.16] | [-.13,.67] |
| $p_{\text{1-seitig}}$ | .429 | .196 | .056 |
| $\tau_{\text{Kendall}}$ | $.01 \pm .14$ | $-.10 \pm .17$ | $\mathbf{.27 \pm .18}$ |
| $\text{KI}_{95\,\%}$ | [-.27,.27] | [.43,.22] | [-.09,.59] |
| $p_{\text{1-seitig}}$ | .479 | .253 | .036 |

*Anmerkungen.* Signifikante Werte mit $p_{\text{1-seitig}} < .05$ sind fett gedruckt. Die effektiven Leistungszuwächse wurden wie folgt berechnet: Unter Vernachlässigung der Mehrebenenstruktur wurden zunächst residuale Lernzuwächse in einer Regression auf Schülerebene berechnet, in der die durch die Kontrollvariablen (Prä-Testwert, KFT-Testwert, Geschlecht, zuhause gesprochene Sprache) erklärte Varianz aus den Post-Testwerten der Lernenden herausgerechnet wurde. Diese Residuen wurden auf Klassenebene über den Mittelwert aggregiert und anschließend in einer Regression auf Klassenebene um die durch die Unterrichtszeit erklärte Varianz bereinigt. Als effektive Leistungszuwächse werden die Residuen dieser Regression auf Klassenebene bezeichnet.

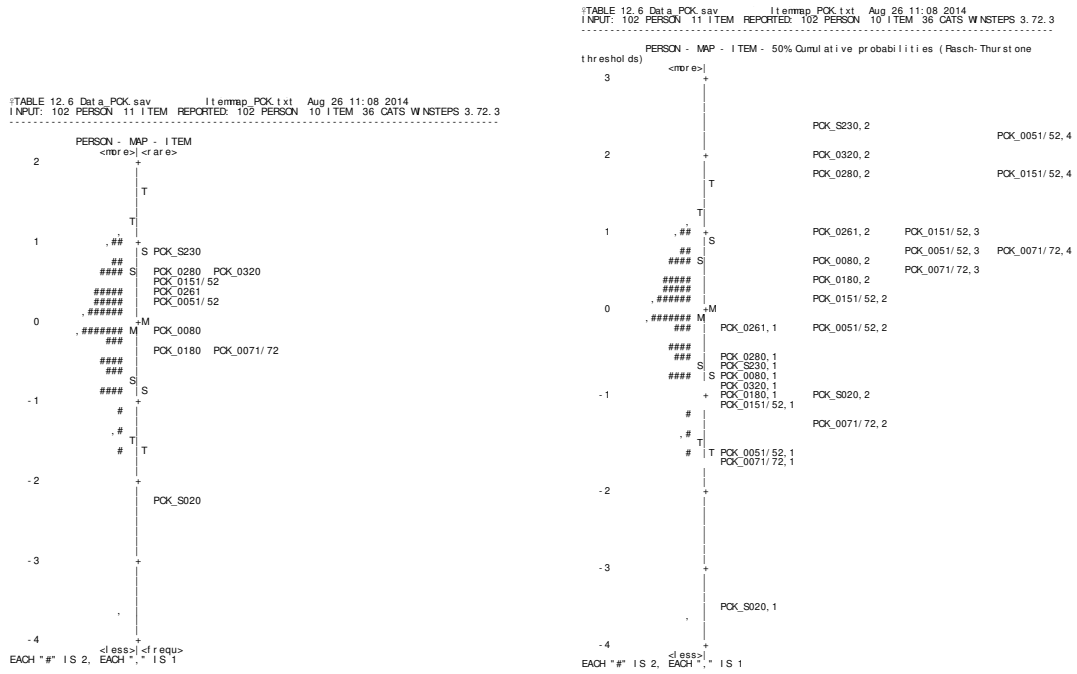
## B.2. Abbildungen

B.1. Wright-Maps für die CK-Aufgaben und deren Schwellenwerte . . . 250
B.2. Wright-Maps für die PCK-Aufgaben und deren Schwellenwerte . . 251
B.3. Wright-Maps für die Aufgaben des Schülerfachwissenstests . . . . 252
B.4. Wright-Map für die Aufgaben des Kognitive Fähigkeitentests . . . 252
B.5. Mplus-Beispielsyntax . . . . . 253



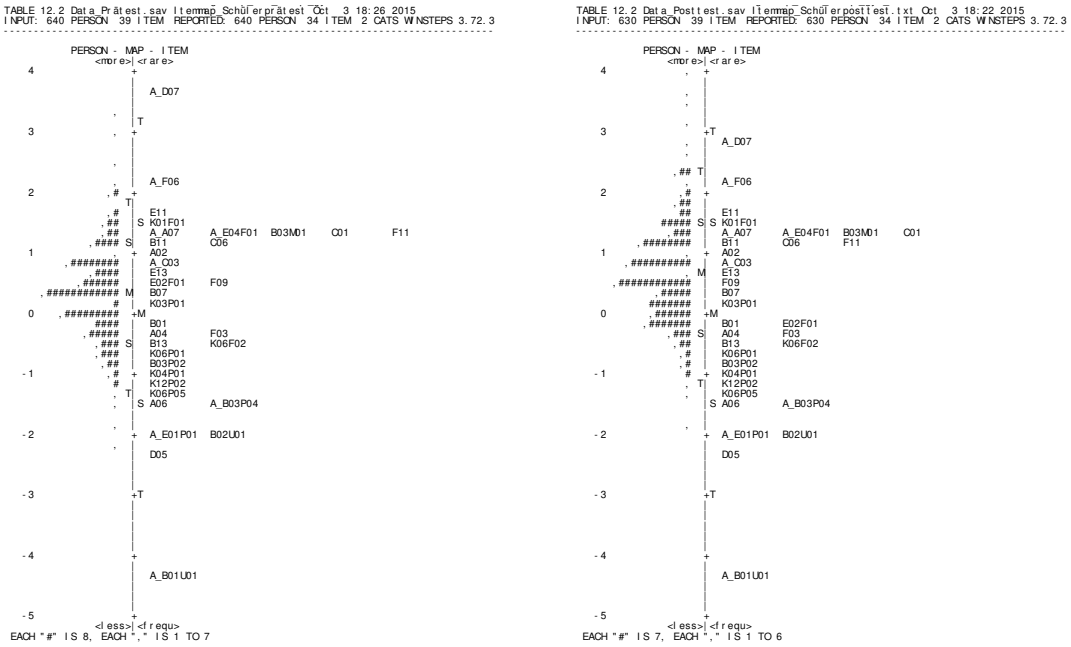

(a) (b)

Figure B.1.
*Wright-Maps für (a) die Aufgaben des CK-Tests und (b) deren Schwellenwerte. Die CK-Aufgaben wurden auf einer zweistufigen Punkteskala bewertet. Bearbeitet eine Person beispielsweise eine Aufgabe, deren Schwellenwert für die Kategorie „1 Punkt“ auf der gemeinsamen Skala auf Höhe ihrer Personenfähigkeit liegt, bedeutet das, dass die Wahrscheinlichkeit dafür, dass die Person in dieser Aufgabe einen Punkt erzielt, bei* **50%** *liegt (vergl. Linacre, 2011, S. 303/330).*




(a) (b)

Figure B.2.
*Wright-Maps für (a) die Aufgaben des PCK-Tests und (b) deren Schwellenwerte. Die PCK-Aufgaben wurden auf einer zweistufigen Punkteskala bewertet. Für die zweiteiligen Aufgaben im PCK-Test (PCK_0051/52, PCK_0151/52, PCK_0071/72) galt dieses Bepunktungsschema für jeden Aufgabenteil. Die Punkte der Teilaufgaben wurden addiert, so dass insgesamt null bis vier Punkte in den zweiteiligen Aufgaben vergeben wurden. Bearbeitet eine Person beispielsweise eine Aufgabe, deren Schwellenwert für die Kategorie „1 Punkt“ auf der gemeinsamen Skala auf Höhe ihrer Personenfähigkeit liegt, bedeutet das, dass die Wahrscheinlichkeit dafür, dass die Person in dieser Aufgabe einen Punkt erzielt, bei* **50%** *liegt (vergl. Linacre, 2011, S. 303/330).*




(a) (b)

Figure B.3.
*Wright-Maps für die Schülerfachwissenstests (a) in der Prä-Testung und (b) in der Post-Testung. Ankeraufgaben, die in Testheft A und Testheft B enthielten waren, sind mit einem der Aufgabenbezeichnung vorangestellten „A“ gekennzeichnet.*

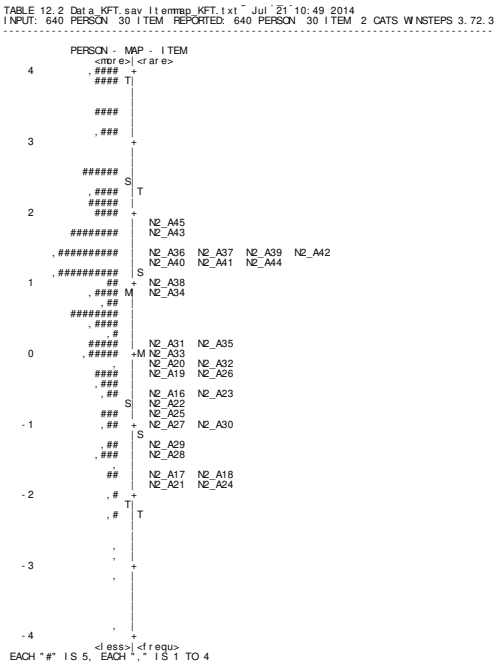



Figure B.4.
*Wright-Map für die Aufgaben des Kognitive Fähigkeitentests.*

```
DATA:       FILE=Daten_Mplus.dat;
Variable:   Names=
                IDKLASSE
                ZPost
                ZPre
                ZstdKFT
                Migrat
                ZTime
                GenderS
                ZCK;
            MISSING = ALL (99);
            Usevar=
                IDKLASSE
                ZPost
                ZPre
                ZstdKFT
                Migrat
                ZTime
                GenderS
                ZCK;
            CLUSTER =  IDKLASSE;
            Within  =  ZPre ZstdKFT Migrat GenderS;
            Between= ZTime ZCK;
ANALYSIS:   Type =Twolevel;
MODEL:        %WITHIN%
                  ZPost on ZPre ZstdKFT Migrat GenderS;
              %BETWEEN%
                  ZPost on ZTime ZCK;
OUTPUT:     sampstat stdyx stdy residual cinterval(symmetric);
```

(a)

```
DATA:       FILE=Daten_Mplus.dat;
Variable:   Names=
                IDKLASSE
                ZFAMin12
                ZCK;
            MISSING = ALL (99);
            Usevar=
                IDKLASSE
                ZFAMin12
                ZCK;
            CLUSTER =  IDKLASSE;
            Between= ZCK;
ANALYSIS:   Type =Twolevel;

MODEL:      %BETWEEN%
                ZFAMin12 on ZCK;
OUTPUT:     sampstat stdyx residual cinterval(symmetric);
```

(b)

Figure B.5.
*Mplus-Beispielsyntax (a) für das Mehrebenenmodell 1a für die z-standardisierten Post-Testwerte (ZPost) der Lernenden im Fachwissen mit den z-standardisierten Prä- (Zpre) und KFT-Testwerten (ZstdKFT), dem Geschlecht (GenderS) und der von den Lernenden zuhause gesprochene Sprache (Migrat) als Prädiktoren auf Schülerebene und der z-standardisierten Unterrichtszeit (ZTime) und dem z-standardisierten CK-Testwert der Lehrkräfte (ZCK) als Prädiktoren auf Klassenebene, und (b) für das Mehrebenenmodell 1d für das z-standardisierte über beide Unterrichtsstunden gemittelte situationale Interesse der Lernenden (ZFAMin12) mit dem CK-Testwert der Lehrkräfte (ZCK) als Prädiktor auf Klassenebene.*

# Literatur


Abel, J. & Faust, G. (2010). Das GLANZ-Projekt–seine Ziele, seine Wirkungen. In *Wirkt Lehrerbildung? Antworten aus der empirischen Forschung* (S. 35–46). Münster u.a.: Waxmann.

Abell, S. K. (2007). Research on science teachers' knowledge. In S. K. Abell & N. G. Lederman (Hrsg.), *Handbook of Research on Science Education* (S. 1105–1149). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Abell, S. K. (2008). Twenty years later: Does pedagogical content knowledge remain a useful idea? *International Journal of Science Education*, *30*(10), 1405–1416.

Bandura, A. (1997). *Self-efficacy: The exercise of control: Diagnose, Evaluation und Verbesserung des Unterrichts* (1. Aufl.). New York: Freeman.

Baumert, J. & Köller, O. (2000). Unterrichtsgestaltung, verständnisvolles Lernen und multiple Zielerreichung im Mathematik- und Physikunterricht der gymnasialen Oberstufe. In J. Baumert, W. Bos & R. Lehmann (Hrsg.), *TIMSS/III. Dritte Internationale Mathematik- und Naturwissenschaftsstudie. Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn. Band. 2: Mathematische und physikalische Kompetenzen am Ende der gymnasialen Oberstufe.* (S. 271–315). Opladen: Leske + Budrich.

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* (S. 29–53). Münster u.a.: Waxmann.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., … Tsai, Y.-M. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *Educational Research Journal*, *47*(1), 133–180.

Baur, N. (2008). Das Ordinalskalenproblem. In N. Baur & S. Fromm (Hrsg.), *Datenanalyse mit SPSS für Fortgeschrittene* (2., überarbeitete und erweiterte Aufl., S. 279–289). Wiesbaden: VS. Zugriff unter http://dx.doi.org/10.1007/978-3-531-91034-5_13

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, *35*(5), 463–482.

Blömeke, S. (2009). Lehrerausbildung. In S. Blömeke, T. Bohl, L. Haag, G. Lang-Wojtasik & W. Sacher (Hrsg.), *Handbuch Schule. Theorie - Organisation - Entwicklung* (S. 483–490). Bad Heilbrunn: Klinkhardt.

Blömeke, S., Kaiser, G., Döhrmann, M. & Lehmann, R. (2010). Mathematisches und mathematikdidaktisches Wissen angehender Sekundarstufen-I-Lehrkräfte im internationalen Vergleich. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.),

*TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 197–238). Münster u.a.: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2008). *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare. Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung.* Münster u.a.: Waxmann.

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2010). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich.* Münster u.a.: Waxmann.

Blömeke, S. & König, J. (2010). Messung des pädagogischen Wissens. Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 239–263). Münster u.a.: Waxmann.

Blömeke, S., Suhl, U., Kaiser, G., Felbrich, A., Schmotz, C. & Lehmann, R. (2010). Lerngelegenheiten und Kompetenzerwerb angehender Mathematiklehrkräfte im internationalen Vergleich. *Unterrichtswissenschaft*, *38*(1), 29–50.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2. Aufl.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Borko, H. & Putnam, R. T. (1996). Learning to Teach. In D. C. Berliner & R. C. Calfee (Hrsg.), *Handbook of Educational Psychology* (S. 673–708). New York: Macmillan.

Borowski, A., Neuhaus, B. J., Tepner, O., Wirth, J., Fischer, H. E., Leutner, D., … Sumfleth, E. (2010). Professionswissen von Lehrkräften in den Naturwissenschaften (ProwiN) - Kurzdarstellung des BMBF-Projekts. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 341–349.

Borowski, A., Olszewski, J. & Fischer, H. E. (2010). Fachdidaktisches Wissen von Physikreferendaren. *Der mathematische und naturwissenschaftliche Unterricht*, *63*(5), 260–263.

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. Aufl.). Heidelberg: Springer Medizin.

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4. Aufl.). Heidelberg: Springer.

Bortz, J. & Lienert, G. A. (2008). *Kurzgefasste Statistik für die klinische Forschung* (3. Aufl.). Heidelberg: Springer.

Bromme, R. (1992). *Der Lehrer als Experte. Zur Psychologie des professionellen Wissens.* Bern u.a.: Huber.

Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F. E. Weinert (Hrsg.), *Handbook of Research on Teaching* (Bd. 3, S. 177–212). 1. Goettingen: Hogrefe.

Bromme, R. (2008). Lehrerexpertise. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie* (S. 159–167). Göttingen: Hogrefe.

Bromme, R. & Rheinberg, F. (2006). Lehrende in den Schulen. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (S. 296–334). Weinheim: Beltz.

Brophy, J. & Good, T. L. (1986). Teacher behavior and student achievement. In M. Wittrock (Hrsg.), *Handbook of Research on Teaching* (S. 328–375). New York: Macmillan.

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*(4), 306–329.

Brunner, M., Kunter, M., Krauss, S., Klusmann, U., Baumert, J., Blum, W., … Tsai, Y.-M. (2006). Die professionelle Kompetenz von Mathematiklehrkräften: Konzeptualisierung, Erfassen und Bedeutung für den Unterricht. Eine Zwischenbilanz des COACTIV-Projekts. In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule* (S. 54–82). Münster u.a.: Waxmann.

Bühner, M. (2006). *Einführung in die Test-und Fragebogenkonstruktion* (2., aktualisierte und erweiterte Aufl.). München: Pearson Studium.

Cardiff. (2011). *TeleForm 10.5.2*. Lüneburg: Electronic Papers.

Carlsen, W. S. (1993). Teacher knowledge and discourse control: Quantitative evidence from novice biology teachers' classrooms. *Journal of Research in Science Teaching*, *30*(5), 471–481. Zugriff unter http://dx.doi.org/10.1002/tea.3660300506

Carpenter, T. P., Fennema, E., Peterson, P. L. & Carey, D. A. (1988). Teachers' Pedagogical Content Knowledge of Students' Problem Solving in Elementary Arithmetic. *Journal for Research in Mathematics Education*, *19*(5).

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P. & Loef, M. (1989). Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study. *American Educational Research Journal*, *26*(4), 499–531.

Carroll, J. B. (1989). The Carroll Model: A 25-Year Retrospective and Prospective View. *Educational Researcher*, *18*(26), 26–31.

Carstensen, C. H. (2000). *Mehrdimensionales Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik*. Kiel: IPN.

Carstensen, C. H. (2006). Technische Grundlagen für die Messwiederholung. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, …U. Schiefele (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres.* (S. 309–323). Münster u.a.: Waxmann.

Cascio, C. (1995). National Board for Professional Teaching Standards: Changing Teaching through Teachers. *The Clearing House*, *68*(4), 211–213.

Cauet, E. (2015). Schülerfachwissenstest Fach Physik. In H. E. Fischer (Hrsg.), *Instrumente fachdidaktischer Unterrichtsforschung, Band I* (S. 4–8). DuePublico, Online-Veröffentlichung Universität Duisburg-Essen. Zugriff unter http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=39374

Chen, W.-H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H. & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot

studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*(2), 485–493. Zugriff unter http://dx.doi.org/10.1007/s11136-013-0487-5

Clausen, M. (2002). *Unterrichtsqualität: Eine Frage der Perspektive? Empirische Analysen zur Übereinstimmung, Konstukt - und Kriteriumsvalidität.* Pädagogische Psychologie und Entwicklungspsychologie. Münster u.a.: Waxmann.

Clausen, M., Reusser, K. & Klieme, E. (2003). Unterrichtsqualität auf der Basis hoch-inferenter Unterrichtsbeurteilungen. Ein Vergleich zwischen Deutschland und der deutschsprachigen Schweiz. *Unterrichtswissenschaft*, *31*(2), 122–141. Zugriff unter http://www.pedocs.de/volltexte/2013/6775/pdf/UnterWiss_2003_2_Clausen_Reusser_Klieme_Unterrichtsqualitaet.pdf

Cochran, K. F., DeRuiter, J. A. & King, R. A. (1993). Pedagogical Content Knowing: An Integrative Model for Teacher Preparation. *Journal of Teacher Education*, *44*(4), 263–272.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences: Eine Einführung für Forschung und Praxis* (3. Aufl.) (H. Sahner, M. Bayer & R. Sackmann, Hrsg.). Mahwah, New Jersey: LEA.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Daniels, Z. (2008). *Entwicklung schulischer Interessen im Jugendalter.* Pädagogische Psychologie und Entwicklungspsychologie. Münster u.a.: Waxmann.

Darling-Hammond, L. (2000). Teacher quality and student achievement. A review of state policy evidence. *Education Policy Analysis Archive*, *8*(1). Zugriff unter http://epaa.asu.edu/ojs/article/download/392/515

De Jong, O. & Van Driel, J. H. (2004). Exploring the Development of Student Teachers' PCK of the Multiple Meanings of Chemistry Topics. *International Journal of Science and Mathematics Education*, *2*(4), 477–491.

De Jong, O., Van Driel, J. H. & Verloop, N. (2005). Preservice teachers' pedagogical content knowledge of using particle models in teaching chemistry. *Journal of Research in Science Teaching*, *42*(8), 947–964.

Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung fuer die Paedagogik. *Zeitschrift für Pädagogik*, *39*(2), 223–238.

Döhrmann, M., Kaiser, G. & Blömeke, S. (2010). Messung des mathematischen und mathematikdidaktischen Wissens: Theoretischer Rahmen und Teststruktur. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 169–196). Münster u.a.: Waxmann.

Dollny, S. (2011). *Entwicklung und Evaluation eines Testinstruments zur Erfassung des fachspezifischen Professionswissens von Chemielehrkräften.* Studien zum Physik- und Chemielernen. Berlin: Logos.

Drechsler, M. & Van Driel, J. H. (2008). Experienced Teachers' Pedagogical Content Knowledge of Teaching Acid–base Chemistry. *Research in Science Education*, *38*(5), 611–631.

Duit, R. & Treagust, D. (2003). Conceptual change: a powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*(6), 671–688.

Edelmann, W. (2003). Intrinsische und extrinsische Motivation. *Grundschule*, *35*(4), 30–32.

Ehmke, T., Blum, W. & Neubrand, M. (2006). Wie verändert sich die mathematische Kompetenz von der neunten zur zehnten Klassenstufe? In M. Prenzel, J. Baumert, R. Blum W.´and Lehmann, D. Leutner, M. Neubrand, R. Pekrun, …U. Schiefele (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres.* (S. 63–85). Münster u.a.: Waxmann.

Ergönenç, J., Neumann, K. & Fischer, H. E. (2014). The Impact of Pedagogical Content Knowledge on Cognitive Activation and Student Learning. In H. E. Fischer, P. Labudde, K. Neumann & J. Viiri (Hrsg.), *Quality of Instruction in Physics: Quality of Instruction in Physics* (S. 13–30). Münster u.a.: Waxmann.

Evertson, C. M. & Emmer, E. T. (1982). Effective management at the beginning of the school year in junior high classes. *Journal of Educational Psychology*, *74*(4), 485–498.

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fend, H. (1980). *Theorie der Schule.* München u.a.: Urban u. Schwarzenberg.

Fenstermacher, G. & Richardson, V. (2005). On making determinations of quality in teaching. *The Teachers College Record*, *107*(1), 186–213.

Fernández-Balboa, J.-M. & Stiehl, J. (1995). The generic nature of pedagogical content knowledge among college professors. *Teaching and Teacher Education*, *11*(3), 293–306. Zugriff unter http://www.sciencedirect.com/science/article/pii/0742051X9400030A

Field, A. (2009). *Discovering Statistics Using SPSS* (3. Aufl.). London u.a.: SAGE Publications Ltd.

Fischer, H. E., Boone, W. J. & Neumann, K. (2014). Quantitative Research Designs and Approaches. In N. G. Lederman & S. K. Abell (Hrsg.), *Handbook of Research on Science Education* (Bd. 2, S. 18–37). New York: Taylor und Francis (Routledge).

Fischer, H. E., Borowski, A. & Tepner, O. (2012). Professional knowledge of science teachers. In B. Fraser, K. Tobin & C. McRobbie (Hrsg.), *Second International Handbook of Science Education* (S. 435–448). New York: Springer.

Fischer, H. E., Labudde, P., Neumann, K. & Viiri, J. (2014a). *Quality of Instruction in Physics: Comparing Finland, Germany and Switzerland.* Münster u.a.: Waxmann.

Fischer, H. E., Labudde, P., Neumann, K. & Viiri, J. (2014b). Theoretical Framework. In H. E. Fischer, P. Labudde, K. Neumann & J. Viiri (Hrsg.), *Quality of Instruction in Physics: Quality of Instruction in Physics* (S. 13–30). Münster u.a.: Waxmann.

Fischler, H. (2008). Physikdidaktisches Wissen und Handlungskompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *14*, 27–49.

Fleischmann, P. (2013). Rundungsregeln in der Metrologie. Zugriff unter https://www.ptb.de/cms/fileadmin/internet/fachabteilungen/abteilung_8/8.4_mathematische_modellierung/268_PTB_SEMINAR/VORTRAEGE/102_Fleischmann_07_Rundungsregeln.pdf

Force Concept Inventory. (1992). Force Concept Inventory: (Deutsche Übersetzung der überarbeiteten englischen Fassung): Original von Hestenes, D.; Wells, M.; Swackhamer, G. (1992). Zugriff unter http://modeling.asu.edu/R%5C&E/Research.html

Förtsch, C., Werner, S., Dorfner, T., von Kotzebue, L. & Neuhaus, B. J. (2015). Kognitive Aktivierung im Biologieunterricht – Wie werden situationales Interesse und Leistung von Lernenden beeinflusst? In *Heterogenität.Wert.Schätzen. Abstractband. 3. Jahrestagung der Gesellschaft für Empirische Bildungsforschung (GEBF). Bochum 11. – 13. März 2015* (S. 648).

Fricke, K. (2015). *Classroom Managment and its Impact on Lesson Outcomes in Physics: A Multi-Perspective Comparison of Teaching Practices in Primary and Secondary Schools* (Unveröffentlichte Dissertation, Universität Duisburg-Essen, Essen).

Fricke, K., van Ackeren, I., Kauertz, A. & Fischer, H. E. (2012). Students' Perceptions of their Teachers' Classroom Management in Elementary and Secondary Science Lessons and the Impact on Student Achievement. In T. Wubbels, J. van Tartwijk, P. den Brok & J. Levy (Hrsg.), *Interpersonal Relationships in Education* (Bd. 3, S. 167–185). Advances in Learning Environments Research. Sense Publishers. Zugriff unter http://dx.doi.org/10.1007/978-94-6091-939-8_11

Geddis, A. N., Onslow, B., Beynon, C. & Oesch, J. (1993). Transforming content knowledge: Learning to teach about isotopes. *Science Education*, *77*(6), 575–591. Zugriff unter http://dx.doi.org/10.1002/sce.3730770603

Geiser, C. (2011). *Datenanalyse mit Mplus: Eine anwendungsorientierte Einführung* (2., durchgesehene Aufl.). Wiesbaden: VS.

Geller, C. (2015). *Lernprozessorientierte Sequenzierung des Physikunterrichts im Zusammenhang mit Fachwissenserwerb: Eine Videostudie in Finnland, Deutschland und der Schweiz*. Studien zum Physik- und Chemielernen. Berlin: Logos.

Geller, C., Neumann, K., Boone, W. J. & Fischer, H. E. (2014). What Makes the Finnish Different in Science? Assessing and Comparing Students' Science Learning in Three Countries. *International Journal of Science Education*, *36*(18), 3042–3066. Zugriff unter http://www.tandfonline.com/doi/pdf/10.1080/09500693.2014.950185

Gess-Newsome, J. (1999). Pedagogical content knowledge: An introduction and orientation. In J. Gess-Newsome & N. G. Lederman (Hrsg.), *Examining pedagogical content knowledge: The construct and its implication for science education* (S. 3–17). Netherlands: Springer. Zugriff unter http://dx.doi.org/10.1007/0-306-47217-1_1

Gess-Newsome, J. (2013). Pedagogical Content Knowledge. In J. Hattie & E. M. Anderman (Hrsg.), *International guide to student achievement* (S. 257–259). Routledge.

Gess-Newsome, J. (2015). A model of meacher professional knowledge and skill including PCK. Results of the thinking from the PCK summit. In A. Berry, P. Friedrichsen & J. Loughran (Hrsg.), *Re-examining Pedagogical Content Knowledge in Science Education* (S. 28–42). New York u.a.: Routledge.

Gess-Newsome, J., Carlson, J., Gardner, A. & Taylor, J. (2010). Impact of Educative Materials and Professional Development on Teachers' Professional Knowledge, Practice, and Student Achievement. Zugriff unter http://bscs.org/primepapers

Gess-Newsome, J. & Lederman, N. G. (1995). Biology teachers' perceptions of subject matter structure and its relationship to classroom practice. *Journal of Research in Science Teaching*, *32*(3), 301–325.

Getzels, J. W. & Jackson, P. W. (1963). The Teacher's Personality and Characteristics. In N. L. Gage (Hrsg.), *Handbook of Research on Teaching* (S. 506–582). Chicago: Rand McNally.

Ghasemi, A. & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489.

Gigl, F., Zander, S., Borowski, A. & Fischer, H. E. (2015). Erfassung des Fachwissens von Lehramtsstudierenden der 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. 112–114). Kiel: IPN. Zugriff unter http://www.gdcp.de/images/tb2015/TB2015_112_Gigl.pdf

Gramzow, J., Y.and Riese & Reinhold, P. (2013). Modellierung fachdidaktischen Wissens angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 7–30. Zugriff unter http://archiv.ipn.uni-kiel.de/zfdn/pdf/19_Gramzow.pdf

Grossman, P. M. (1990). *The making of a teacher: Teacher knowledge and teacher education* (A. Lieberman, Hrsg.). Professional development and practice series. New York: Teachers College Press.

Großschedl, J., Mahler, D., Kleickmann, T. & Harms, U. (2014). Content-Related Knowledge of Biology Teachers from Secondary Schools: Structure and learning opportunities. *International Journal of Science Education*, *36*(14), 2335–2366.

Gruber, H., Mandl, H. & Renkl, A. (2000). Was lernen wir in Schule und Hochschule: Träges Wissen? In H. Mandl & J. Gerstenmeier (Hrsg.), *Die Kluft zwischen Wissen und Handeln: empirische und theoretische Lösungsansätze* (S. 139–156). Göttingen: Hogrefe.

Hammann, M. & Jördens, J. (2014). Offene Aufgaben codieren. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin Heidelberg: Springer.

Hartig, J. & Jude, N. (2007). Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle. In J. Hartig & E. Klieme (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (Bd. 20, S. 17–36). Bildungsforschung. Bonn u.a.: BMBF.

Hartig, J., Jude, N. & Wagner, W. (2008). Methodische Grundlagen der Messung und Erklärung sprachlicher Kompetenzen. In E. Klieme (Hrsg.), *Unterricht und Kompetenzerwerb in Deutsch und Englisch. Ergebnisse der DESI-Studie* (S. 34–54). Weinhein u.a.: Beltz.

Hartig, J. & Kühnbach, O. (2006). Schätzung von Veränderung mit Plausible Values in mehrdimensionalen Rasch-Modellen. In A. Ittel (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 27–44). Wiesbaden: VS.

Hashweh, M. Z. (1987). Effects of subject-matter knowledge in the teaching of biology and physics. *Teaching and Teacher Education*, *3*(2), 109–120.

Hashweh, M. Z. (2005). Teacher pedagogical constructions: a reconfiguration of pedagogical content knowledge. *Teachers and Teaching*, *11*(3), 273–292.

Heller, K. A. & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4.-12. Klassen, Revision (KFT 4-12+ R)*. Göttingen: Hogrefe.

Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität: Diagnose, Evaluation und Verbesserung des Unterrichts* (1. Aufl.). Seelze: Klett-Kallmeyer.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule.* (Bd. 3, S. 71–176). Enzyklopädie der Psychologie. Göttingen u.a.: Hogrefe.

Hetze, P. (2011). *Nachhaltige Hochschulstrategien für mehr MINT-Absolventen.* (2., aktualisierte Aufl.). Essen: Stifterverband für die Deutsche Wissenschaft: Heinz-Nixdorf-Stiftung. Zugriff unter http://www.stifterverband.info/publikationen_und_podcasts/positionen_dokumentationen/mint_hochschulstrategien_2011/mint_hochschulstrategien_2011.pdf

Hill, H. C. & Ball, D. L. (2004). Learning Mathematics for Teaching: Results from California's Mathematics Professional Development Institutes. *Journal for Research in Mathematics Education*, *35*(5), 330–351.

Hill, H. C., Ball, D. L., Blunk, M., Goffney, I. M. & Rowan, B. (2007). Validating the Ecological Assumption: The Relationship of Measure Scores to Classroom Teaching and Student Learning. *Measurement: Interdisciplinary Research and Perspectives*, *5*(2-3), 107–118.

Hill, H. C., Ball, D. L. & Schilling, S. G. (2008). Unpacking "Pedagogical Content Knowledge": Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372–400.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. & Ball, D. L. (2008). Mathematical Knowledge for Teaching and the Mathematical Quality of Instruction: An Exploratory Study. *Cognition and Instruction*, *26*(4), 430–511.

Hill, H. C., Rowan, B. & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*(2), 371–406.

Hill, H. C., Schilling, S. G. & Ball, D. L. (2004). Developing Measures of Teachers' Mathematics Knowledge for Teaching. *The Elementary School Journal*, *105*(1), 11–30.

Hohensinn, C. & Kubinger, K. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, (53), 380–393.

Hugener, I. (2006). Überblick über die Beobachtungsinstrumente. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis", Teil 3: Hugener, I./Pauli, C./Reusser, K.: Videoanalysen.* (Bd. 15, S. 45–54). Materialien zur Bildungsforschung. Frankfurt, Main: DIPF & GFPF. Zugriff unter http://www.pedocs.de/volltexte/2010/3130;%20http://nbn-resolving.de/urn:nbn:de:0111-opus-31304

Hugener, I. (2008). *Inszenierungsmuster im Unterricht und Lernqualität. Sichtstrukturen schweizerischen und deutschen Mathematikunterrichts in ihrer Beziehung zu Schülerwahrnehmung und Lernleistung - eine Videoanalyse.* Pädagogische Psychologie und Entwicklungspsychologie. Münster u.a.: Waxmann.

Hugener, I., Rakoczy, K., Pauli, C. & Reusser, K. (2006). Videobasierte Unterrichtsforschung: Integration verschiedener Methoden der Videoanalyse für eine differenzierte Sicht auf Lehr-Lernprozesse. In S. Rahm, I. Mammes & M. Schratz (Hrsg.), *Schulpädagogische Forschung. Unterrichtsforschung. Perspektiven innovativer Ansätze.* (S. 41–53). Innsbruck: Studien.

IBM Corp. (2012). *IBM SPSS Statistics for Windows, Version 21.0* (3.72.3). Armonk, NY: IBM Corp.

Jansen, M., Schroeders, U. & Stanat, P. (2013). Motivationale Schülermerkmale in Mathematik und den Naturwissenschaften. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 347–365). Münster u.a.: Waxmann.

Jüttner, M. (2013). *Entwicklung, Evaluation und Validierung eines Fachwissenstests und eines fachdidaktischen Wissenstests für die Erfassung des Professionswissens von Biologielehrkräften* (Dissertation, LMU München, München).

Kauertz, A. & Kleickmann, T. (2009). Postersymposium Professionswissen von Lehrkräften, verständnisorientierter naturwissenschaftlicher Unterricht und Zielerreichung im Übergang von der Primar- zur Sekundarstufe (PLUS). In D. Höttecke (Hrsg.), *Chemie- und Physikdidaktik für die Lehramtsausbildung. In Schwäbisch Gmünd 2008.* (S. 395–397). Münster: Lit.

Keller, M. (2011). *Teacher Enthusiasm in Physics Instruction* (Dissertation, Universität Duisburg-Essen, Essen). Zugriff unter http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=25993

Keller, M., Neumann, K. & Fischer, H. E. (2014). Enthusiastic Teaching and its Impact on Students' Interest and Self-Concept: An Investigation of German Physics Classrooms. In H. E. Fischer, P. Labudde, K. Neumann & J. Viiri (Hrsg.), *Quality of Instruction in Physics: Quality of Instruction in Physics* (S. 129–143). Münster u.a.: Waxmann.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R. & Stigler, J. W. (2012). Measuring Usable Knowledge: Teachers' Analyses of Mathematics

Classroom Videos Predict Teaching Quality and Student Learning. *American Educational Research Journal*, *49*(3), 568–589.

Kessler, S. J. (2011). *Mathematisches Fachwissen von gymnasialen Mathematiklehrkräften. Eine empirische Analyse des Konstrukts und dessen Korrelation mit Personen- und Unterrichtsvariablen.* (Dissertation, Technische Universität München, München). Zugriff unter http://nbn-resolving.de/urn:nbn:de:bvb:91-diss-20110802-1071144-1-9

Kirby, K. N. & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior research methods*, *45*(4), 905–927.

Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Studien zum Physik- und Chemielernen. Berlin: Logos.

Kirschner, S., Borowski, A., Fischer, H. E., Gess-Newsome, J. & von Aufschnaiter, C. (in Druck). Developing and Evaluating a Paper-and-Pencil Test to Assess Components of Physics Teachers' Pedagogical Content Knowledge. *International Journal of Science Education*.

Kirschner, S., Sczudlek, M., Tepner, O., Borowski, A., Fischer, H. E., Lenkse, G., … Wirth, J. (in Druck). Professionswissen in den Naturwissenschaften (ProwiN). In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals. Interdisziplinäre Betrachtungen, Befunde und Perspektiven* (S. 113–130). Springer Online.

Kirschner, S., Taylor, J., Rollnick, M., Borowski, A. & Mavhunga, E. (2015). Gathering evidence for the validity of PCK measures: Connecting ideas to analytic approaches. In A. Berry, P. Friedrichsen & J. Loughran (Hrsg.), *Re-examining Pedagogical Content Knowledge in Science Education* (S. 229–242). New York u.a.: Routledge.

Klieme, E. & Clausen, M. (1999). *Identifying facets of problem solving in mathematics instruction. Paper presented at the AERA Annual Meeting, Montreal, 1999*. Berlin: Max-Planck-Institut für Bildungsforschung.

Klieme, E. & Leutner, D. (2006). Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingrichteten Schwerpunktprogramms der DFG. *Zeitschrift für Pädagogik*, *52*(6), 876–903. Zugriff unter http://www.pedocs.de/volltexte/2011/4493;%20http://nbn-resolving.de/urn:nbn:de:0111-opus-44936

Klieme, E., Lipowsky, F., Rakoczy, K. & Ratzka, N. (2006). Qualitätsdimensionen und Wirksamkeit von Mathematikunterricht. Theoretische Grundlagen und ausgewählte Ergebnisse des Projekts "Pythagoras". In M. Prenzel & L. Allolio-Näcke (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms.* (S. 127–146). Münster u.a.: Waxmann.

Klieme, E., Pauli, C. & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Hrsg.), *The power of video studies in investigating teaching and learning in the classroom.* (S. 137–160). Münster u.a.: Waxmann.

Klieme, E. & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik*, *54*(2), 222–237.

Klieme, E., Schümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I. "Aufgabenkultur" und Unterrichtsgestaltung. In *TIMSS - Impulse für Schule und Unterricht.* (S. 43–57). Bonn: Bundesministerium für Bildung u. Forschung.

KMK. (2004). Standards für die Lehrerbildung. Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004. Zugriff unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung.pdf

KMK. (2005a). Bildungsstandards der Kultusministerkonferenz Erläuterungen zur Konzeption und Entwicklung (Am 16.12.2004 von der Kultusministerkonferenz zustimmend zur Kenntnis genommen). Zugriff unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Konzeption-Entwicklung.pdf

KMK. (2005b). Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004. Zugriff unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Physik-Mittleren-SA.pdf

KMK. (2008). Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung (Beschluss der Kultusministerkonferenz vom 16.10.2008 i. d. F. vom 11.12.2014). Zugriff unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2008/2008_10_16-Fachprofile-Lehrerbildungb.pdf

Kolbe, F.-U. (2004). Verhältnis von Wissen und Handeln. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Hrsg.), *Handbuch Lehrerbildung* (S. 206–232). Bad Heilbrunn: Klinkhardt.

Köller, O. & Baumert, J. (2008). Entwicklung schulischer Leistungen. In R. Oerter & L. Montada (Hrsg.), *Entwicklungspsychologie.* (6., vollständig überarbeitete Aufl., S. 735–768). Weinheim u.a.: Beltz.

König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften. Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. *Zeitschrift für Erziehungswissenschaft*, *12*(3), 499–527.

Kounin, J. S. (2006). *Techniken der Klassenführung. Reprint der dt. Ausg. 1976.* Standardwerke aus Psychologie und Pädagogik, Reprints, Band 3. Münster u.a.: Waxmann.

Krapp, A. (1998). Entwicklung und Förderung von Interessen im Unterricht. *Psychologie in Erziehung und Unterricht*, *45*(3), 185–201.

Krapp, A. (2002). An Educational-Psychological Theory of Interest and Its Relation to SDT. In E. L. Deci & R. M. Ryan (Hrsg.), *Handbook of Self-Determination Research* (S. 405–427). University Rochester Press.

Krapp, A. (2003). Die Bedeutung der Lernmotivation für die Optimierung des schulischen Bildungssystems. *Politische Studien*, *54*(3), 91–105.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M. & Jordan, A. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *The Journal of educational psychology*, *100*(3), 716–725.

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 Mathematik-Didaktik*, *29*(3/4), 223–258.

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.

Kromrey, J. D., Coraggio, J. T., Phan, H. T., Romano, J. L., Hess, M. R., Lee, R. S., … Luther, S. L. (2006). The Impact of Measurement Error in Predictor Variables in Multilevel Models: An Empirical Investigation of Statistical Bias and Sampling Error. Paper presented at the annual meeting of the Florida Educational Research Association, 2006, Jacksonville. Zugriff unter http://www.coedu.usf.edu/main/departments/me/documents/theimpactofmeasurementerrorinpredictorvariablesinhierarchicallinearmodelsfera2006.pdf

Kulgemeyer, C., Borowski, A., Fischer, H. E., Gramzow, Y., Reinhold, P., Riese, J., … Walzer, M. (2012). ProfiLe-P – Professionswissen in der Lehramtsausbildung Physik. Vorstellung eines Forschungsprojekts. In V. Nordmeier & H. Grötzebauch (Hrsg.), *PhyDid B, Didaktik der Physik, Beiträge zur DPG-Frühjahrstagung 2012 in Mainz, Berlin.*

Kunter, M. (2005). *Multiple Ziele im Mathematikunterricht.* Pädagogische Psychologie und Entwicklungspsychologie. Münster u.a.: Waxmann.

Kunter, M., Baumert, J. & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, *17*(5), 494–509. Zugriff unter http://www.sciencedirect.com/science/article/pii/S095947520700093X

Kunter, M., Dubberke, T., Baumert, J., Blum, W., Brunner, M., Jordan, A., … Tsai, Y.-M. (2006). Mathematikunterricht in den PISA-Klassen 2004: Rahmenbedingungen, Formen und Lehr-Lernprozesse. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, …U. Schiefele (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres.* (S. 161–194). Münster u.a.: Waxmann.

Kunter, M. & Voss, T. (2011). Das Modell der Unterrichtsqualität in COACTIV: Eine multikriteriale Analyse. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften* (S. 85–113). Münster u.a.: Waxmann.

Lamberti, J. (2001). *Einstieg in die Methoden empirischer Forschung: Planung, Durchführung und Auswertung empirischer Untersuchungen.* Tübingen: dgvt.

Lange, K. (2010). *Zusammenhänge zwischen naturwissenschaftsbezogenem fachspezifisch-pädagogischem Wissen von Grundschullehrkräften und Fortschritten im Verständnis naturwissenschaftlicher Konzepte bei Grundschülerinnen und*

*–schülern*. Münster: Didaktik des Sachunterrichts. Zugriff unter http://nbn-resolving.de/urn:nbn:de:hbz:6-75459654103

Lange, K., Kleickmann, T., Tröbst, S. & Möller, K. (2012). Fachdidaktisches Wissen von Lehrkräften und multiple Ziele im naturwissenschaftlichen Sachunterricht. *Zeitschrift für Erziehungswissenschaft*, *15*, 55–75.

Lange, K., Ohle, A., Kleickmann, T., Kauertz, A., Möller, K. & Fischer, H. E. (2015). Zur Bedeutung von Fachwissen und fachdidaktischem Wissen für Lernfortschritte von Grundschülerinnen und Grundschülern im naturwissenschaftlichen Sachunterricht. *Zeitschrift für Grundschulforschung*, *8*(1), 23–38.

Langer, W. (2009). *Mehrebenenanalyse: Eine Einführung für Forschung und Praxis* (2. Aufl.) (H. Sahner, M. Bayer & R. Sackmann, Hrsg.). Wiesbaden: VS.

Lee, I. A. & Preacher, K. J. (2013a). Calculation for the test of the difference between two dependent correlations with no variable in common. Zugriff unter http://quantpsy.org/corrtest/corrtest3.htm

Lee, I. A. & Preacher, K. J. (2013b). Calculation for the test of the difference between two dependent correlations with one variable in common. Zugriff unter http://quantpsy.org/corrtest/corrtest2.htm

Leitner, E. & Finckh, U. (o.d.). Aufgaben zur Mechanik: Internetportal www.leifiphysik.de. Zugriff unter http://www.leifiphysik.de/teilgebiete/mechanik

Lenske, G., Thillmann, H., Wirth, J., Dicke, T. & Leutner, D. (2015). Pädagogisch-psychologisches Professionswissen von Lehrkräften: Evaluation des ProwiN-Tests. *Zeitschrift für Erziehungswissenschaft*, *18*(2), 225–245. Zugriff unter http://dx.doi.org/10.1007/s11618-015-0627-5

Lenske, G., Wagner, W., Wirth, J., Thillmann, H., Cauet, E. & Leutner, D. (2016). Die Bedeutung des pädagogisch-psychologischen Wissens für die Qualität der Klassenführung und den Lernzuwachs der Schüler/innen im Physikunterricht. *Zeitschrift für Erziehungswissenschaft*, *19*(1), 211–233. Zugriff unter http://dx.doi.org/10.1007/s11618-015-0659-x

Liepertz, S. (2016). *Zusammenhang zwischen dem Professionswissen von Physiklehrkräften, sachstrukturellem Angebot des Unterrichts und Schülerleistung* (Unveröffentlichte Dissertation, Universität Potsdam, Potsdam).

Liepertz, S., Cauet, E., Borowski, A. & Fischer, H. E. (2015). *Influence of Physics Teachers‘ Professional Knowledge on the Interconnectedness of Lessons‘ Content Structure and on Students‘ Outcomes*. Unveröffentlichter Vortrag auf der ESERA Jahrestagung 2015, Helsinki, Finnland.

Linacre, J. M. (2011). *A user's guide to Winsteps ® Ministep: Rasch-model computer programs* (3.72.3). Winsteps.com.

Lipowsky, F. (2006). Auf den Lehrer kommt es an. Empirische Evidenzen für Zusammenhänge zwischen Lehrerkompetenzen, Lehrerhandeln und dem Lernen der Schüler. In C. Allemann-Ghionda & E. Terhart (Hrsg.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern* (S. 47–70). Zeitschrift für Pädagogik. Beiheft. 51. Weinheim u.a.: Beltz.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E. & Reusser, K. (2009). Quality of geometry instruction and its short-term impact on students'

understanding of the Pythagorean Theorem. *Learning and instruction*, *19*(6), 527–537.

Loughran, J., Berry, A. & Mulhall, P. (2012). *Understanding and Developing ScienceTeachers' Pedagogical Content Knowledge.* (2. Aufl.). Rotterdam u.a.: Sense Publishers.

Loughran, J., Mulhall, P. & Berry, A. (2004). In search of pedagogical content knowledge in science: Developing ways of articulating and documenting professional practice. *Journal of Research in Science Teaching*, *41*(4), 370–391.

Löwen, K., Baumert, J., Kunter, M., Krauss, S. & Brunner, M. (2011). Methodische Grundlagen des Forschungsprogramms. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV.* (S. 69–84). Münster u.a.: Waxmann.

Luhmann, N. & Schorr, K. E. (1979). Das Technologiedefizit der Erziehung und die Pädagogik. *Zeitschrift für Pädagogik*, *25*(3), 345–365.

Maas, C. J. M. & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

Maas, C. J. M. & Hox, J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 85–91.

Magnusson, S., Krajcik, J. & Borko, H. (1999). Nature, Sources, and Development of Pedagogical Content Knowledge for Science teacher. In J. Gess-Newsome & N. G. Lederman (Hrsg.), *Examining pedagogical content knowledge: The construct and its implication for science education* (S. 95–132). Netherlands: Springer.

Mahler, D., Großschedl, J. & Harms, U. (2015). *Which Influence does Biology Teachers' Content-related Knowledge has on Students' Conceptual Knowledge Acquisition in Biology Education?* Unveröffentlichter Vortrag auf der NARST Jahrestagung 2015, Chicago, USA.

Malcolm, S. A. & Mavhunga, E. (2015). The Development and Validation of an Instrument Measuring Topic Specific PCK in Stoichiometry. Zugriff unter https://www.researchgate.net/profile/Stephen_Andrew_Malcolm/publication/274015540_The_Development_and_Validation_of_an_Instrument_Measuring_Topic_Specific_PCK_in_Stoichiometry/links/55129a8b0cf20bfdad519148.pdf?origin=publication_detail

Mandl, H., Gruber, H. & Renkl, A. (1993). Das träge Wissen. *Psychologie heute*, *20*(9), 64–69.

Marks, R. (1990). Pedagogical Content Knowledge: From a Mathematical Case to a Modified Conception. *Journal of Teacher Education*, *41*(3), 3–11.

Mechanics Baseline Test. (1992). Mechanics Baseline Test: (Deutsche Übersetzung): Original von Hestenes, D., Wells, M. (1992). Zugriff unter http://modeling.asu.edu/R%5C&E/Research.html

Messick, S. (1987). Validity. *ETS Research Report Series*, *1987*(2).

Meyer, H. (2004). Novice and expert teachers' conceptions of learners' prior knowledge. *Science Education*, *88*(6), 970–983.

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2008). Kernlehrplan für das Gymnasium – Sekundarstufe I in Nordrhein-Westfalen: Physik. Zugriff unter http://www.schulentwicklung.nrw.de/lehrplaene/upload/lehrplaene_download/gymnasium_g8/gym8_physik.pdf

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2011). Kernlehrplan für die Gesamtschule – Sekundarstufe I in Nordrhein-Westfalen: Naturwissenschaften Biologie, Chemie, Physik. Zugriff unter http://www.schulentwicklung.nrw.de/lehrplaene/upload/klp_SI/GE/NW/GE_NW_Bio_Che_Phy_Endfassung.pdf

Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2015). Das Schulwesen in Nordrhein-Westfalen aus quantitativer Sicht: Statistische Übersicht Nr. 386 - 5. Aufl. Zugriff unter http://www.schulministerium.nrw.de/docs/bp/Ministerium/Service/Schulstatistik/Amtliche-Schuldaten/Quantita_2013.pdf

Morris, S. (1989). Teaching Practice: Objectives and Conflicts. *Educational Review*, *21*(2), 120–129.

Muthén, L. K. & Muthén, B. O. (2007). *Mplus User's Guide. Fifth Edition.* (5.21). Los Angeles, CA: Muthén & Muthén.

Narciss, S. & Huth, K. (2004). How to design informative tutoring feedback for multi-media learning. In H. M. Niegemann, D. Leutner & R. Brünken (Hrsg.), *Instructional Design for Multimedia learning* (S. 181–195). Münster, New York: Waxmann.

Ndlovu, M. (2014). *The design of an instrument to measure physical science teachers' topic specific pedagogical content knowledge in electrochemistry* (Dissertation, University of the Witwatersrand, Johannisburg).

Neumann, K., Kauertz, A. & Fischer, H. E. (2012). Quality of Instruction in Science Education. In B. J. Fraser, K. Tobin & C. J. McRobbie (Hrsg.), *Second International Handbook of Science Education* (Bd. 24, S. 247–258). Springer.

Neuweg, G. H. (2002). Lehrerhandeln und Lehrerbildung im Lichte des Konzepts des impliziten Wissens. *Zeitschrift für Pädagogik*, *48*(1), 10–29.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass*, *2*(2), 842–860.

Nezlek, J. B., Schröder-Abé, M. & Schütz, A. (2006). Mehrebenenanalysen in der psychologischen Forschung: Vorteile und Möglichkeiten der Mehrebenenmodellierung mit Zufallskoeffizienten. *Psychologische Rundschau*, *57*(4), 213–223.

Novella, S. (2015). Psychology Journal Bans Significance Testing. Zugriff unter https://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/

OECD. (2012). *Learning beyond Fifteen: Ten Years after PISA* (3.72.3). OECD Publishing. Zugriff unter http://dx.doi.org/10.1787/9789264172104-en

Oevermann, U. (1996). Theoretische Skizze einer revidierten Theorie professionalisierten Handelns. In A. Combe & W. Helsper (Hrsg.), *Pädagogische Pro-*

*fessionalität: Untersuchungen zum Typus pädagogischen Handelns* (1. Aufl., S. 70–182). Frankfurt am Main: Suhrkamp.

Ohle, A. (2010). *Primary school teachers' content knowledge in physics and its impact on teaching and students' achievement*. Studien zum Physik- und Chemielernen. Berlin: Logos.

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.

Olson, L. (1987). An Overview of the Holmes Group. *The Phi Delta Kappan*, *68*(8), 619–621.

Olszewski, J. (2010). *The Impact of Physics Teachers' Pedagogical Content Knowledge on Teacher Actions and Student Outcomes*. Studien zum Physik- und Chemielernen. Berlin.

Oser, F. & Baeriswyl, F. (2001). Choreographies of teaching: Bridging instruction to learning. In V. Richardson (Hrsg.), *Handbook of Research on Teaching* (S. 1031–1065). Washington, DC: American Educational Research Association.

Pant, H. A., Stanat, P., Pöhlmann, C., Hecht, M., Jansen, M., Kampa, A., N.and Lenski, … Ziemke, A. (2013). Der Blick in die Länder. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 159–248). Münster u.a.: Waxmann.

Paris, S. G., Lipson, M. Y. & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, *8*(3), 293–316.

Park, S. & Chen, Y.-C. (2012). Mapping Out the Integration of the Components of Pedagogical Content Knowledge (PCK): Examples From High School Biology Classrooms. *Journal of Research in Science teaching*, *49*(7), 922–941.

Park, S. & Oliver, S. J. (2008). Revisiting the conceptualisation of pedagogical content knowledge (PCK): PCK as a conceptual tool to understand teachers as professionals. *Research in Science Education*, *38*(3), 261–284.

Pauli, C. & Reusser, K. (2003). Unterrichtsskripts im schweizerischen und im deutschen Mathematikunterricht. *Unterrichtswissenschaft*, *31*(3), 238–272.

Peterson, P. L., Carpenter, T. P. & Fennema, E. (1989). Teachers' knowledge of students' knowledge in mathematics problem solving: Correlating and case analysis. *Journal of Educational Psychology*, *81*(4), 558–569.

Phillips, D. C. (2003). The contribution of epistemology to curriculum construction in the sciences. *Zeitschrift für Erziehungswissenschaft*, *6*(3), 421–431. Zugriff unter http://dx.doi.org/10.1007/s11618-003-0043-0

Pöhlmann, C., Haag, N. & Stanat, P. (2013). Zuwanderungsbezogene Disparitäten. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 297–329). Münster u.a.: Waxmann.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K. & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons.

*Learning and Instruction*, *31*, 2–12. Zugriff unter http://www.sciencedirect.com/science/article/pii/S0959475213000832

Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M., … Schiefele, U. (Hrsg.). (2005). *PISA 2003. Der zweite Vergleich der Länder in Deutschland - was wissen und können Jugendliche?* Münster u.a.: Waxmann.

Quesel, C., Möser, G. & Husfeldt, V. (2014). Auswirkungen sozialer Belastungen auf das Schul-, Unterrichts-und Arbeitsklima obligatorischer Schulen in der Schweiz. *Schweizerische Zeitschrift für Bildungswissenschaften*, *36*(2), 283–306. Zugriff unter http://rsse.elearninglab.org/wp-content/uploads/2014/10/SZBW_14.2_Varia_Quesel.pdf

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Zugriff unter http://www.R-project.org

Rakoczy, K. & Pauli, C. (2006). Hoch inferentes Rating: Beurteilung der Qualität unterrichtlicher Prozesse. In E. Klieme, C. Pauli & K. Reusser (Hrsg.), *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis“, Teil 3: Hugener, I./Pauli, C./Reusser, K.: Videoanalysen.* (Bd. 15, S. 206–233). Materialien zur Bildungsforschung. Frankfurt, Main: DIPF & GFPF. Zugriff unter http://www.pedocs.de/volltexte/2010/3130; %20http://nbn-resolving.de/urn:nbn:de:0111-opus-31304

Razali, N. M. & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Reusser, K. (2009). Unterricht. In S. Andresen, R. Casale, T. Gabriel, R. Horlacher, S. Larcher Klee & J. Oelkers (Hrsg.), *Handwörterbuch Erziehungswissenschaft.* (S. 881–896). Weinheim u.a.: Beltz.

Revelle, W. (2015). *psych: Procedures for Psychological, Psychometric, and Personality Research*. R package version 1.5.4. Northwestern University. Evanston, Illinois. Zugriff unter http://CRAN.R-project.org/package=psych

Rheinberg, F., Vollmeyer, R. & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen. *Diagnostica*, *47*(2), 57–66.

Riese, J. (2009). *Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften*. Studien zum Physik- und Chemielernen. Berlin: Logos.

Riese, J., Kulgemeyer, C., Borowski, A., Fischer, H., Gramzow, Y., Reinhold, P., … Zander, S. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik. In S. Blömeke & O. Zlatkin-Troitschanskaia (Hrsg.), *Kompetenzen von Studierenden* (S. 55–79). Zeitschrift für Pädagogik. Beiheft. 61. Weinheim u.a.: Beltz.

Rjosk, C., McElvany, N., Anders, Y. & Becker, M. (2011). Diagnostische Fähigkeiten von Lehrkräften bei der Einschätzung der basalen Lesefähigkeit ihrer Schülerinnen und Schüler. *Psychologie in Erziehung und Unterricht*, *58*, 92–105.

Rollnick, M., Bennett, J., Rhemtula, M., Dharsey, N. & Ndlovu, T. (2008). The Place of Subject Matter Knowledge in Pedagogical Content Knowledge: A

case study of South African teachers teaching the amount of substance and chemical equilibrium. *International Journal of Science Education*, *30*(10), 1365–1387. Zugriff unter http://dx.doi.org/10.1080/09500690802187025

Rollnick, M. & Mavhunga, E. (2014). PCK of teaching electrochemistry in chemistry teachers: A case in Johannesburg, Gauteng Province, South Africa. *Educación Química*, *25*(3), 354–362.

Rosenshine, B. (1983). Teaching Functions in Instructional Programs. *The Elementary School Journal*, *83*(4), 335–351.

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion*. Bern: Huber.

Sadler, P. M., Sonnert, G., Coyle, H. P., Cook-Smith, N. & Miller, J. L. (2013). The Influence of Teachers' Knowledge on Student Learning in Middle School Physical Science Classrooms. *American Educational Research Journal*, *50*(5), 1020–1049.

Sanders, L. R., Borko, H. & Lockard, J. D. (1993). Secondary science teachers' knowledge base when teaching science courses in and out of their area of certification. *Journal of Research in Science Teaching*, *30*(7), 723–736. Zugriff unter http://dx.doi.org/10.1002/tea.3660300710

Schiefele, U. (2008). Lernmotivation und Interesse. In W. Schneider & M. Hasselhorn (Hrsg.), *Handbuch der Pädagogischen Psychologie.* (S. 38–49). Handbuch der Psychologie. 10. Göttingen u.a.: Hogrefe.

Schiefele, U., Krapp, A. & Schreyer, I. (1993). Metaanalyse des Zusammenhangs von Interesse und schulischer Leistung. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie*, *25*(2), 120–148. Zugriff unter https://publish up.uni-potsdam.de/files/3173/schiefele1993_XXV.pdf

Schiefele, U. & Schreyer, I. (1994). Intrinsische Lernmotivation und Lernen. Ein Überblick zu Ergebnissen der Forschung. *Zeitschrift für pädagogische Psychologie*, *8*(1), 1–12.

Schmelzing, S. (2010). *Das fachdidaktische Wissen von Biologielehrkräften: Konzeptionalisierung, Diagnostik, Struktur und Entwicklung im Rahmen der Biologielehrerbildung*. Berlin: Logos.

Schmiemann, P. & Lücken, M. (2014). Validität –Misstmein Test, was er soll? In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung*. Berlin u.a.: Springer.

Schoppmeier, F. (2013). *Physikkompetenz in der gymnasialen Oberstufe: Physikkompetenz in der gymnasialen Oberstufe. Entwicklung und Validierung eines Kompetenzstrukturmodells für den Kompetenzbereich Umgang mit Fachwissen*. Studien zum Physik- und Chemielernen. Berlin: Logos.

Schroeders, U., Penk, C., Jansen, M. & Pant, H. A. (2013). Geschlechtsbezogene Disparitäten. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 249–274). Münster u.a.: Waxmann.

Schroeders, U., Siegle, T., Weirich, S. & Pant, H. A. (2013). Der Einfluss von Kontext-und Schülermerkmalen auf die naturwissenschaftlichen Kompetenzen. In *IQB-Ländervergleich 2012: Mathematische und naturwissenschaftliche*

*Kompetenzen am Ende der Sekundarstufe I* (S. 331–346). Münster u.a.: Waxmann.

Seidel, T. (2003). *Lehr-Lernskripts im Unterricht*. Pädagogische Psychologie und Entwicklungspsychologie. Münster u.a.: Waxmann.

Seidel, T. & Prenzel, M. (2006). Stability of Teaching Patterns in Physics Instruction: Findings from a Video Study. *Learning and instruction*, *16*(3), 228–240.

Seidel, T., Rimmele, R. & Dalehefte, I. M. (2003). Skalendokumentation der Schülerfragebögen. In T. Seidel, M. Prenzel, R. Duit & M. Lehrke (Hrsg.), *Technischer Bericht zur Videostudie „Lehr-Lern-Prozesse im Physikunterricht“* (S. 317–388). Kiel: IPN.

Seidel, T., Rimmele, R. & Prenzel, M. (2003). Gelegenheitsstrukturen beim Klassengespräch und ihre Bedeutung für die Lernmotivation. Videoanalysen in Kombination mit Schülerselbsteinschätzungen. *Unterrichtswissenschaft*, *31*(2), 142–165.

Seidel, T. & Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade. The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research*, *77*(4), 454–499.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*(1), 1–2. Zugriff unter http://pss.sagepub.com/content/8/1/1.short

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, *86*(2), 420–428.

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), 1–22.

Slavin, R. E. (1994). Quality, appropriateness, incentive, and time: A model of instructional effectiveness. *International Journal of Educational Research*, *21*(2), 141–157.

Smith, D. C. & Neale, D. C. (1989). The construction of subject matter knowledge in primary science teaching. *Teaching and Teacher Education*, *5*(1), 1–20.

Spoden, C. & Geller, C. (2014). Uncovering Country Differences in Physics Content Knowledge and their Interrelations with Motivational Outcomes in a Latent Change Analysis. In H. E. Fischer, P. Labudde, K. Neumann & J. Viiri (Hrsg.), *Quality of Instruction in Physics: Quality of Instruction in Physics* (S. 13–30). Münster u.a.: Waxmann.

Stelzl, I. (2006). *Fehler und Fallen der Statistik: für Psychologen, Pädagogen und Sozialwissenschaftler*. Standardwerke aus Psychologie und Pädagogik, Reprints, Band 1. Münster u.a.: Waxmann.

Stender, A., Geller, C., Neumann, K. & Fischer, H. E. (2013). Der Einfluss der Unterrichtstaktung auf die Strukturiertheit und Abgeschlossenheit von Lernprozessen. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 189–202. Zugriff unter http://archiv.ipn.uni-kiel.de/zfdn/pdf/19_Stender.pdf

Strobl, C. (2012). *Das Rasch-Modell: eine verständliche Einführung für Studium und Praxis*. Sozialwissenschaftliche Forschungsmethoden. München und Mering: Hampp.

Tamir, P. (1988). Subject matter and related pedagogical knowledge in teacher education. *Teaching and Teacher Education*, *4*(2), 99–110.

Tatto, M. T., Ingvarson, L., Schwille, J., Peck, R., Senk, S. L. & Rowley, G. (2008). *Teacher Education and Development Study in Mathematics (TEDS-M): Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics. Conceptual Framework.* East Lansing, MI: Teacher Education and Development International Study Center: College of Education, Michigan State University.

Tatto, M. T., Peck, R., Schwille, J., Bankov, K., Senk, S. L., Rodriguez, M., … Rowley, G. (2012). *Policy, Practice, and Readiness to Teach Primary and Secondary Mathematics in 17 Countries: Findings from the IEA Teacher Education and Development Study in Mathematics (TEDS-M).* East Lansing, MI: Teacher Education and Development International Study Center: College of Education, Michigan State University.

Tenorth, H. E. (2006). Professionalität im Lehrerberuf: Ratlosigkeit der Theorie, gelingende Praxis. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 580–597.

Tepner, O., Borowski, A., Dollny, S., Fischer, H. E., Jüttner, M., Kirschner, S., … Wirth, J. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 7–28.

Tesch, M. (2011). *Das Experiment im Physikunterricht - Didaktische Konzepte und Ergebnisse einer Videostudie*. Studien zum Physik- und Chemielernen. Berlin.

TIMSS Assessment. (1995). TIMSS 1995 Science Items: Released Set for Population 2 (Seventh and Eighth Grade). Zugriff unter http://timssandpirls.bc.edu/timss1995i/TIMSSPDF/BSItems.pdf

TIMSS Assessment. (1999). TIMSS 1999 Science Items: Released Set for Eight Grades. Zugriff unter http://timss.bc.edu/timss1999i/pdf/t99science_items.pdf

TIMSS Assessment. (2003). TIMSS 2003 Science Items: Released Set for Eight Grades. Zugriff unter http://timss.bc.edu/PDF/T03_RELEASED_S8.pdf

TIMSS Assessment. (2007). *TIMSS 2007 User Guide for the International database. Released Items: Science- Eighth Grade* (B. C. TIMSS & PIRLS International Study Center Lynch School of Education, Hrsg.). Zugriff unter http://timss.bc.edu/TIMSS2007/PDF/T07_G8_Released_Items_SCI.zip

Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften: Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*. Studien zum Physik- und Chemielernen. Berlin: Logos.

Voss, T. & Kunter, M. (2011). Pädagogisch-psychologisches Wissen von Lehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften* (S. 193–214). Münster u.a.: Waxmann.

Voss, T., Kunter, M. & Baumert, J. (2011a). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology*, *103*(4), 952–969.

Voss, T., Kunter, M. & Baumert, J. (2011b). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *The Journal of Educational Psychology*, *103*(4), 952–969.

Voss, T., Kunter, M., Seiz, J., Hoehne, V. & Baumert, J. (2014). Die Bedeutung des pädagogisch-psychologischen Wissens von angehenden Lehrkräften für die Unterrichtsqualität. *Zeitschrift für Pädagogik*, *60*(2), 184–201.

Walter, O., Senkbeil, M., Rost, J., Carstensen, C. H. & Prenzel, M. (2006). Die Entwicklung der naturwissenschaftlichen Kompetenz von der neunten zur zehnten Klassenstufe: Deskriptive Befunde. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand, …U. Schiefele (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres.* (S. 87–118). Münster u.a.: Waxmann.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.

Weinert, F. E. & Helmke, A. (1996). Der gute Lehrer: Person, Funktion oder Fiktion? In A. Leschinsky (Hrsg.), *Die Institutionalisierung von Lehren und Lernen. Beiträge zu einer Theorie der Schule* (S. 223–233). Zeitschrift für Pädagogik. Beiheft. 34. Weinheim: Beltz.

Widodo, A. & Duit, R. (2004). Konstruktivistische Sichtweisen vom Lehren und Lernen und die Praxis des Physikunterrichts. *Zeitschrift für Didaktik der Naturwissenschaften*, *10*, 233–255.

Wiley, D. E. & Harnischfeger, A. (1974). Explosion of a Myth: Quantity of Schooling and Exposure to Instruction, Major Educational Vehicles. *Educational Researcher*, *3*(7), 7–12.

Wilhelm, O. & Kunina, O. (2009). Pädagogisch-psychologische Diagnostik. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie: mit 27 Tabellen* (S. 307–331). Heidelberg: Springer.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität.* Göttingen: Hogrefe.

Woitkowski, D., Riese, J. & Reinhold, P. (2011). Modellierung fachwissenschaftlicher Kompetenz angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *17*, 289–313. Zugriff unter http://archiv.ipn.uni-kiel.de/zfdn/pdf/17_Woitkowski.pdf

Woodhouse, G., Yang, M., Goldstein, H. & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 201–212.

Yamamoto, K. (1963). Evaluating Teachers' Effectiveness: A Review of Research. *Educational Review*, *21*(2), 120–129.

Zander, S. (2016). *Lehrerfortbildung zu Basismodellen und Zusammenhänge zum Fachwissen.* Studien zum Physik- und Chemielernen. Berlin: Logos.

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

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

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

Das Professionswissen von Lehrkräften wird als wichtige Voraussetzung für gutes und erfolgreiches Unterrichten diskutiert. Professionswissenstests werden daher oft mit dem Ziel eingesetzt, Aussagen über die Wirksamkeit der Lehrerausbildung zu treffen. Die Handlungsrelevanz explizierbaren Wissens ist allerdings nicht empirisch abgesichert, was die Validität solcher Aussagen einschränkt.

Ziel der hier vorgestellten Studie war die Überprüfung der prädiktiven Validität der im Projekt „ProwiN“ entwickelten Tests zur Erfassung des Fachwissens sowie des fachdidaktischen und pädagogischen Wissens von Physiklehrkräften in Bezug auf gutes und erfolgreiches Unterrichten. Hierfür wurden Zusammenhänge zwischen dem Professionswissen von 23 Lehrkräften, der kognitiv aktivierenden Gestaltung ihres Unterrichts und dem Fachwissenserwerb und situationalen Interesse ihrer Schülerinnen und Schüler analysiert.

Die Ergebnisse deuten darauf hin, dass auf Basis der üblichen Validierungsmaßnahmen für Professionswissenstests (Expertenbefragungen, Vergleich bekannter Gruppen, Zusammenhangsanalysen zwischen Professionswissensdimensionen), nicht davon ausgegangen werden kann, dass handlungsrelevantes Wissen für gutes und erfolgreiches Unterrichten erfasst wird. Eine intensive Auseinandersetzung mit den Einschränkungen der vorgestellten Studie macht deutlich, wie wichtig – aber auch wie problematisch – die Untersuchung der Zusammenhänge zwischen Professionswissen, Unterrichtsqualität und Unterrichtserfolg ist.

Logos Verlag Berlin

ISBN 978-3-8325-4276-4