

Vogel, Sebastian

Wie schätzen Lehrkräfte die Lernstandserhebungen in Mathematik ein? Projektevaluation, Typisierung und Ist-Stand zu Einstellungen zu den Lernstandserhebungen Mathematik Klasse 8 in Hessen

Kassel : kassel university press 2020, 414 S. - (Dissertation, Universität Kassel, 2018)



Quellenangabe/ Reference:

Vogel, Sebastian: Wie schätzen Lehrkräfte die Lernstandserhebungen in Mathematik ein? Projektevaluation, Typisierung und Ist-Stand zu Einstellungen zu den Lernstandserhebungen Mathematik Klasse 8 in Hessen. Kassel : kassel university press 2020, 414 S. - (Dissertation, Universität Kassel, 2018) - URN: urn:nbn:de:0111-pedocs-326184 - DOI: 10.25656/01:32618; 10.17170/kobra-202007101439

<https://nbn-resolving.org/urn:nbn:de:0111-pedocs-326184>

<https://doi.org/10.25656/01:32618>

in Kooperation mit / in cooperation with:

kassel
university
press



<http://kup.uni-kassel.de>

Nutzungsbedingungen

Dieses Dokument steht unter folgender Creative Commons-Lizenz: <http://creativecommons.org/licenses/by-sa/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 die daraufhin neu entstandenen Werke bzw. Inhalte nur unter Verwendung von Lizenzbedingungen weitergeben, die mit denen dieses Lizenzvertrags identisch, vergleichbar oder kompatibel sind.

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-sa/4.0/deed.en> - You may copy, distribute and transmit, adapt or exhibit the work or its contents in public and alter, transform, or change this work as long as you attribute the work in the manner specified by the author or licensor. New resulting works or contents must be distributed pursuant to this license or an identical or comparable license.

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

Sebastian Vogel

Wie schätzen Lehrkräfte die Lernstandserhebungen in Mathematik ein?

Projektevaluation, Typisierung und Ist-Stand
zu Einstellungen zu den Lernstandserhebungen
Mathematik Klasse 8 in Hessen

Sebastian Vogel

Wie schätzen Lehrkräfte die Lernstandserhebungen in Mathematik ein?

Projektevaluation, Typisierung und Ist-Stand
zu Einstellungen zu den Lernstandserhebungen
Mathematik Klasse 8 in Hessen

Die vorliegende Arbeit wurde vom Fachbereich Mathematik und Naturwissenschaften der Universität Kassel als Dissertation zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.) angenommen.

Gutachter: Prof. Dr. Werner Blum
Prof. Dr. Frank Lipowsky

Tag der mündlichen Prüfung: 14. Juni 2018



Diese Veröffentlichung – ausgenommen Zitate und anderweitig gekennzeichnete Teile – ist unter der Creative-Commons-Lizenz Namensnennung - Weitergabe unter gleichen Bedingungen International (CC BY-SA 4.0: <https://creativecommons.org/licenses/by-sa/4.0/deed.de>) lizenziert.

 <https://orcid.org/0000-0002-9528-8996> (Sebastian Vogel)

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

Zugl.: Kassel, Univ., Diss. 2018
ISBN 978-3-7376-0844-2 (print)
ISBN 978-3-7376-0845-9 (e-book)
DOI: <https://dx.doi.org/doi:10.17170/kobra-202007101439>

© 2020, kassel university press, Kassel
<http://kup.uni-kassel.de>

Printed in Germany

Für zahlreiche Anmerkungen, Hinweise, Korrekturen und Rückmeldungen,
Unterstützung, Begleitung, Kraft und Ausdauer,
möchte ich mich herzlich bedanken bei

meinem Betreuer und Erstgutachter Prof. Dr. Werner Blum,
meinem Zweitgutachter Prof. Dr. Frank Lipowsky,

meinen Kolleginnen und Kollegen in der Mathematikdidaktik und
der Empirischen Schul- und Unterrichtsforschung in Kassel und

meinen Eltern, Christina, meinen Freunden und Sara.

Inhaltsverzeichnis

Einleitung	1
1 Theorie.....	5
1.1 Bildungsstandards.....	5
1.1.1 Entwicklung zu den Bildungsstandards.....	5
1.1.2 Gesamtstrategie zum Bildungsmonitoring	18
1.1.3 Resultierende Veränderungen für die Schule.....	26
1.2 Schulleistungsmessungen.....	32
1.2.1 Definition	32
1.2.2 Gütekriterien von (Leistungs-)Messung.....	34
1.2.3 Grenzen der Leistungsmessung / Voraussetzungen zur guten Nutzung .	37
1.3 Feedback/Rückmeldung	46
1.3.1 Definition/Einführung	46
1.3.2 Einflüsse auf die Nutzung von Feedback.....	50
1.3.3 Nutzung von Ergebnissen aus Leistungsmessung.....	56
1.4 Evaluation	58
1.4.1 Evaluationsformen	60
1.4.2 Evaluationskreislauf	63
1.4.3 Von der Evaluation zur Innovation.....	67
1.5 Schulentwicklung	77
1.5.1 Definition und Forschungsstand	77
1.5.2 Personalentwicklung	79
1.5.3 Organisationsentwicklung.....	95
1.5.4 Unterrichtsentwicklung.....	96
1.6 Vergleichsarbeiten (VerA) und Lernstandserhebungen	114
1.6.1 Ziele der Vergleichsarbeiten	116
1.6.2 Ablauf VerA.....	127

1.6.3	Exkurs: Abgrenzung VerA.....	132
1.6.4	Bedingungen für die Nutzung von Vergleichsarbeiten.....	135
1.6.5	Wirkung von Lernstandserhebungen	143
1.6.6	Typen von Lehrkräften	147
1.6.7	Kritik an VerA	150
1.7	Fazit und Forschungsfragen.....	156
2	Projektbeschreibung: Verbesserung der Effektivität der Lernstandserhebungen Mathematik Klasse 8 (VELM-8)	161
2.1	Projektskizze	161
2.2	Ziele der Fortbildungsreihe	161
2.3	Fortbildungsbeschreibung.....	162
2.3.1	Informationsveranstaltung – 30.08.2012	162
2.3.2	Fortbildung 1 – 11.09.2012.....	163
2.3.3	Fortbildung 2 – 10.01.2013.....	164
2.3.4	Fortbildung 3 – 13.03.2013.....	165
2.3.5	Fortbildung 4 – 07.05.2013.....	166
2.3.6	Projektausblick.....	167
3	Methode.....	169
3.1	Hypothesen.....	169
3.1.1	VELM-8	170
3.1.2	Hypothesen zur Situation in Hessen	179
3.2	Forschungsdesign	183
3.3	Stichprobe.....	184
3.3.1	VELM-8	184
3.3.2	Online-Erhebung	188
3.4	Instrumente	190
3.4.1	VELM-8	190
3.4.2	Online-Fragebogen	204

3.5	Analysemethoden	207
3.5.1	VELM-8	207
3.5.2	Online-Fragebogen	218
4	Ergebnisse	219
4.1	VELM-8	219
4.1.1	Einstellungsskalen	219
4.1.2	Auswertungen zur Unterrichtsqualität	227
4.1.3	Schülerleistung	236
4.2	Typenfindung	238
4.2.1	Clusteranalyse	238
4.2.2	Zusammenfassende Clusterbeschreibung	247
4.2.3	Konfigurationsfrequenzanalyse	249
4.3	Ist-Stand zu Einstellungen von Mathematiklehrkräften zu den Lernstandserhebungen in Hessen	252
4.3.1	Unterrichts-Validität der Lernstandserhebungen	252
4.3.2	Auswertungszeit	253
4.3.3	Nützlichkeit	254
4.3.4	Besprechung der Ergebnisse und Intensität der Auseinandersetzung mit den Ergebnissen	256
4.3.5	Nutzung der LSE	263
4.3.6	Belastung der Lehrkräfte im Zusammenhang mit Alter sowie Klassengröße	266
4.3.7	Deskriptive Statistik zur Lage in Hessen	267
4.4	Zusammenfassung der Ergebnisse nach Hypothesen	269
5	Diskussion	281
5.1	Reflexion der Ergebnisse	281
5.1.1	VELM-8 Projektevaluation	281
5.1.2	Clusteranalyse/Typenbildung	289

5.1.3	Ist-Stand in Hessen.....	291
5.2	Grenzen der Arbeit	295
5.3	Forschungsdesiderate	301
5.4	Fazit.....	303
6	Literaturverzeichnis	307
7	Anhang	359
7.1	Skalenhandbuch	359
7.1.1	Allgemeine Informationen	359
7.1.2	Skalen zu den Lernstandserhebungen	363
7.1.3	Skalen zur Unterrichtsqualität	375
7.1.4	Online-Fragebogen	390
7.2	Leistungstest.....	399
7.2.1	Raterübereinstimmung.....	399
7.2.2	Q ₃ -Statistiken des Leistungstests in Zusammenfassung.....	402
7.2.3	Itemfit.....	403
7.3	Ergebnisergänzungen	405
7.3.1	Rangtransformierte, messwiederholte Varianzanalyse der Skalen zu Lernstandserhebungen (rmANOVA)	405
7.3.2	Deskriptive Statistiken zur mANOVA für die Skalen der Unterrichtsqualität	408
7.3.3	Deskriptive Statistiken zur rmANOVA für die Skalen der Unterrichtsqualität.....	411
7.3.4	Alternative Clusterlösungen	412
7.3.5	Online-Daten: Regression der Nützlichkeit auf die Nutzung (Einzelitems)	413

Abbildungsverzeichnis

Abbildung 1: Strukturmodell der Bildungsstandards Mathematik für die Allgemeine Hochschulreife (verändert nach KMK, 2012a, S. 12)	10
Abbildung 2: Bildungsstandards als Kern bzw. Konkretisierung von Bildungszielen (A. Hosenfeld, 2010, S. 25)	15
Abbildung 3: Modellierungskreislauf nach Blum und Leiß (2005, S. 19) siehe in Leiß (2007, S. 31)	38
Abbildung 4: Aufgabe „Trainingsanalyse“ (Blum et al., 2012)	42
Abbildung 5: Aufgabe „Abstand auf dem Wasser“	43
Abbildung 6: Feedback-Modell (Hattie & Timperley, 2007, S. 87)	48
Abbildung 7: Merkmale, die (hypothetisch) Datennutzung beeinflussen (Schildkamp & Kuiper, 2010, S. 485)	55
Abbildung 8: Kreislauf der Evaluation durch Lernstandserhebungen (LSA, 2015, S. 24)	63
Abbildung 9: Nutzungskreislauf auf verschiedenen Ebenen des Schulsystems (Stufflebeam, 1967, S. 132)	65
Abbildung 10: Rückmeldungsnutzung nach Helmke (2004, S. 100; vgl. auch I. Hosenfeld et al., 2006)	68
Abbildung 11: Angebots-Nutzungs-Modell der Lehrerfortbildung (vereinfacht nach Lipowsky, 2014, S. 473)	82
Abbildung 12: Angebots-Nutzungs-Modell von Unterricht (Lipowsky, 2009, S. 81)	99
Abbildung 13: Basisdimensionen guten Unterrichts und deren vermutete Wirkung (Klieme et al., 2006, S. 131)	108
Abbildung 14: Wirkung professioneller Kompetenz (nach Löwen, Baumert, Kunter, Krauss & Brunner, 2011, S. 70)	111
Abbildung 15: Gesamtergebnis der Klasse (Sofortbericht)	121
Abbildung 16: Gesamtergebnis in den Teilaufgaben (Sofortbericht)	121
Abbildung 17: Detailliergebnisse der Klasse mit Vergleichswerten	123
Abbildung 18: Ergebnis im Fach Mathematik Klassenstufe 8 (nach Leitideen getrennt)	124

Abbildung 19: Zusammenhänge zwischen auf die Vergleichsarbeiten bezogenen Einschätzungen der Fachlehrkräfte (Bonsen et al., 2006, S. 147)	140
Abbildung 20: Projektablauf VELM-8.....	184
Abbildung 21: Ergebnisse der explorativen Faktorenanalyse zur Prüfung der Eindimensionalität der Testhefte (links: Testheft A; Mitte: Testheft B; rechts: Testheft C)	199
Abbildung 22: Aufgabe „Schultaschen“ aus Testheft 1 (Hauptschule)	200
Abbildung 23: Aufgabe „Durchschnittslinie“ aus Testheft 3 (Gymnasium)	202
Abbildung 24: Aufgabe „Fehler in der Gleichung“ aus Testheft 3 (Gymnasium)	202
Abbildung 25: Verteilung der Aufgaben- und Personenparameter des Leistungstests bei gemeinsamer Skalierung der Messzeitpunkte und Testhefte	203
Abbildung 26: Scree-Plot Nützlichkeit	205
Abbildung 27: Scree-Plot Nutzung	206
Abbildung 28: <i>Allgemeine Akzeptanz</i> in Prä- und Posttest getrennt nach Gruppen (H1a).....	223
Abbildung 29: <i>LSE als Belastung</i> in Prä- und Posttest getrennt nach Gruppen (H3a)	223
Abbildung 30: <i>Förderdiagnostischer Nutzen</i> in Prä- und Posttest nach Gruppen (H4a)	224
Abbildung 31: <i>Unterrichtsplanerischer Nutzen</i> in Prä- und Posttest nach Gruppen (H5a).....	224
Abbildung 32: <i>Selektionsdiagnostischer Nutzen</i> in Prä- und Posttest nach Gruppen (H6).....	224
Abbildung 33: Rangmittelwerte <i>Lehrplanvalidität</i> im Prä- und Posttest nach Gruppen (H2).....	227
Abbildung 34: <i>Insistieren auf Begründungen</i> im Prä- und Posttest nach Gruppen (H7)	230
Abbildung 35: <i>Variation der Aufgabenschwierigkeit</i> im Prä- und Posttest nach Gruppen (H8)	230
Abbildung 36: <i>Kognitiv herausfordernde Aufgaben</i> im Prä- und Posttest nach Gruppen (H9).....	230
Abbildung 37: <i>Kognitiv herausforderndes Üben</i> im Prä- und Posttest nach Gruppen (H10).....	230
Abbildung 38: <i>Einschleifendes Üben</i> im Prä- und Posttest Gruppen (H11a).....	230

Abbildung 39: <i>Tutorielle Unterstützung in Lernpartnerschaften</i> und Gruppen im Prä- und Posttest nach Gruppen (H18).....	230
Abbildung 40: Rangmittelwerte <i>Umgang mit Schülerbeiträgen</i> im Prä- und Posttest nach Gruppen (H12a).....	233
Abbildung 41: Rangmittelwerte <i>Fehlervermeidende Erarbeitung</i> im Prä- und Posttest nach Gruppen (H13a).....	233
Abbildung 42: Rangmittelwerte <i>Fürsorgliche Anleitung und Kontrolle</i> im Prä- und Posttest nach Gruppen (H14a).....	233
Abbildung 43: Rangmittelwerte <i>Unterstützung kognitiver Selbstständigkeit</i> im Prä- und Posttest nach Gruppen (H15a).....	233
Abbildung 44: Rangmittelwerte <i>Soziale Konstruktion von Verständnis</i> im Prä- und Posttest nach Gruppen (H17).....	234
Abbildung 45: Rangmittelwerte <i>Persönliches Vertrauen</i> im Prä- und Posttest nach Gruppen (H19a).....	234
Abbildung 46: Rangmittelwerte <i>Diagnose im sozialen Bereich</i> im Prä- und Posttest nach Gruppen (H20).....	234
Abbildung 47: Leistung der Schülerinnen und Schüler in Prä- und Posttest nach Gruppen (H22a).....	237
Abbildung 48: Elbow-Plot der Verschmelzungshöhen.....	239
Abbildung 49: Dendrogramm Clusteranalyse (Ward-Verfahren; Manhattan-Metrik)	239
Abbildung 50: Mittelwerte der Skalen zu Lernstandserhebungen nach Clustern (Ward-Verfahren mit Manhattan-Metrik)	240
Abbildung 51: Rangmittelwerte <i>Allgemeine Akzeptanz</i> (H1 _{nonpara})	407
Abbildung 52: Rangmittelwerte <i>LSE als Belastung</i> (H3 _{nonpara})	407
Abbildung 53: Rangmittelwerte <i>förderdiagnostischer Nutzen</i> (H4 _{nonpara})	407
Abbildung 54: Rangmittelwerte <i>unterrichtsplanerischer Nutzen</i> (H5 _{nonpara})	407
Abbildung 55: Rangmittelwerte <i>selektionsdiagnostischer Nutzen</i> (H6 _{nonpara})	407
Abbildung 56: Rangmittelwerte <i>Insistieren auf Begründungen</i> (H7 _{nonpara}).....	410
Abbildung 57: Rangmittelwerte <i>Variation Aufgabenschwierigkeit</i> (H8 _{nonpara})	410
Abbildung 58: Rangmittelwerte <i>Kognitiv herausfordernde Aufgaben</i> (H9 _{nonpara}) ...	410
Abbildung 59: Rangmittelwerte <i>Kognitiv herausforderndes Üben</i> (H10 _{nonpara})	410
Abbildung 60: Rangmittelwerte <i>Repetitives, einschleifendes Üben</i> (H11 _{nonpara})	410
Abbildung 61: Rangmittelwerte <i>Lernpartnerschaften</i> (H18 _{nonpara})	410

Abbildung 62: Dendrogramm Complete-Linkage (Manhattan-Metrik)	412
Abbildung 63: Dendrogramm Complete-Linkage (euklidische Metrik).....	412
Abbildung 64: Dendrogramm Average-Linkage (Manhattan-Metrik)	412
Abbildung 65: Dendrogramm Average-Linkage (euklidische Metrik)	412

Tabellenverzeichnis

Tabelle 1: Übersicht zu den Testzyklen (grau hinterlegt sind Testungen; verändert nach KMK, 2015, S. 25).....	21
Tabelle 2: Übersicht zu den Testzyklen (grau hinterlegt sind Testungen; verändert nach KMK, 2015, S. 25).....	24
Tabelle 3: Aufgabenkennwerte „Trainingsanalyse“ und „Abstand auf dem Wasser“	39
Tabelle 4: Zielsetzung Kompetenzdiagnostik, Vergleichsarbeiten und Bildungsmonitoring (verändert nach Leutner et al., 2008, S. 151).....	116
Tabelle 5: Vergleich vers. Leistungsmessungen in Abgrenzung der Lernstandserhebungen (zusammenfassend nach Bonsen et al., 2006; Helmke & I. Hosenfeld, 2003; Hessisches Kultusministerium, 2009; I. Hosenfeld et al., 2006; KMK & IQB, 2006; KMK, 2015; Koch, 2011; LSA, 2015; Peek et al., 2006; Schwippert, 2005; A. Voss & Blatt, 2010).....	134
Tabelle 6: Einschätzung der Kompetenztests im Hinblick auf Nützlichkeit (Mittelwertangaben; Nachtigall, 2014, S. 89).....	141
Tabelle 7: Qualitativer Überblick über die Ergebnisse von Groß-Ophoff et al. (2007) im Vergleich der einzelnen Cluster zueinander	148
Tabelle 8: Qualitative Unterscheidung zwischen Typen in A. Hosenfeld (2010, S. 161) bzgl. schriftlicher Rückmeldungen	149
Tabelle 9: t-Test für unabhängige Stichproben im Querschnitt sowie Wald-Test für abhängige Stichproben im Längsschnitt: T ₁ vs. T ₂ (Wacker & Kramer, 2012, 697ff.)	155
Tabelle 10: Übersicht der Hypothesen zu Einstellungen und Nutzen der Lernstandserhebungen	172
Tabelle 11: Übersicht der Hypothesen zum Bereich Unterrichtsqualität und Leistung der Schülerinnen und Schüler	178
Tabelle 12: Übersicht der allgemeinen Hypothesen zur Lage in Hessen.....	182
Tabelle 13: Stichprobengrößen für Leistungstests der Schülerinnen und Schüler zu beiden Messzeitpunkten.....	186
Tabelle 14: Stichprobenbeschreibungen	187

Tabelle 15: Stichprobengröße differenziert nach Fächern bei Berücksichtigung der Teilnahme an den LSE 2013	188
Tabelle 16: Übersicht Stichproben-Hypothesen-Zuordnung.....	189
Tabelle 17: Reliabilitäten (Cronbachs α) der Skalen zu den Lernstandserhebungen bzgl. der verschiedenen Stichproben	190
Tabelle 18: Einstellungen zu den Lernstandserhebungen: Skalenmittelwerte, Standardabweichung getrennt nach Stichproben.....	191
Tabelle 19: Übersicht (marginal) signifikant von der Normalverteilung abweichender Skalen (nach Kolmogorow-Smirnow-Lilliefors-Test)	192
Tabelle 20: Reliabilitäten der Skalen zur Unterrichtsqualität im Vergleich zu Baumert et al. (2008) für Stichprobe 1	193
Tabelle 21: Itemformulierungen Skala „Kognitiv herausfordernder Umgang mit Schülerbeiträgen (Ideen/Fehler)“	194
Tabelle 22: Itemanalyse „Kognitiv herausfordernder Umgang mit Schülerbeiträgen“	194
Tabelle 23: Itemformulierungen Skala „Persönliches Vertrauen“	195
Tabelle 24: Itemanalyse Skala „Persönliches Vertrauen“	195
Tabelle 25: Itemformulierungen Skala „Lob nach individueller Bezugsnorm“	195
Tabelle 26: Itemformulierungen Skala „Diagnose im Leistungsbereich“	196
Tabelle 27: Itemanalyse Skala „Diagnose im Leistungsbereich“	196
Tabelle 28: Korrelationen zwischen Einzelitems der Skala „Diagnose im Leistungsbereich“	196
Tabelle 29: Reliabilität und deskriptive Statistiken der Leistungsdaten der Schülerinnen und Schüler	197
Tabelle 30: Prozentuale Übereinstimmung und Cohens Kappa der Doppelkodierung im Leistungstest	198
Tabelle 31: Auffällige Q ₃ -Werte (vgl. ausführliche Darstellung im Anhang 7.2.2) ..	199
Tabelle 32: Itemanalyse Nützlichkeit.....	204
Tabelle 33: Items Nützlichkeit der Lernstandserhebung.....	205
Tabelle 34: Itemanalyse Schulentwicklung/Nutzung	206
Tabelle 35: Übersicht der Auswertungsverfahren nach Hypothesen unter Prüfung von Prätestunterschieden und Voraussetzungen der parametrischen Verfahren	210
Tabelle 36: Mittelwerte, t-Test und Cohens d für Einstellungsskalen (Stichprobe 1) im Prätest	212

Tabelle 37: Überprüfung der Voraussetzungen für <i>t</i> -Test und messwiederholte Varianzanalyse.....	213
Tabelle 38: Ergebnisse der Suche nach Prätestunterschieden für Unterrichtsqualität mittels <i>t</i> -Test und Welch-Test.....	214
Tabelle 39: Ergebnisse der Suche nach Vortestunterschieden mittels Mann-Whitney-Test	214
Tabelle 40: Prätestunterschiede der Schülerleistung Welch-Test.....	216
Tabelle 41: Deskriptive Statistiken der Skalen zu beiden Messzeitpunkten (Stichprobe 1)	221
Tabelle 42: Ergebnisse mANOVA Skalen zu Lernstandserhebungen (H1a bis H6) ..	223
Tabelle 43: ANCOVA für <i>Allgemeine Akzeptanz</i> (H1b).....	225
Tabelle 44: ANCOVA für <i>LSE als Belastung</i> (H3b)	225
Tabelle 45: ANCOVA für <i>förderdiagnostischen Nutzen</i> (H4b).....	225
Tabelle 46: ANCOVA für <i>unterrichtsplanerischer Nutzen</i> (H5b)	226
Tabelle 47: Deskriptive Beschreibung der Rangtransformation der Skala <i>Lehrplanvalidität</i>	226
Tabelle 48: Ergebnisse der rangtransformierten, messwiederholten Varianzanalyse für die Skala <i>Lehrplanvalidität</i> (H2).....	227
Tabelle 49: Ergebnisse der mANOVA zu Skalen der Unterrichtsqualität (Hypothesen 7, 8, 9, 10, 11a und 18).....	228
Tabelle 50: ANCOVA für <i>Einschleifendes repetitives Üben</i> (H11b)	229
Tabelle 51: Ergebnisse der rangtransformierten, messwiederholten Varianzanalyse	232
Tabelle 52: rANCOVA für <i>Schülerbeiträge (Ideen/Fehler) als Ausgangspunkt für weiteren Unterricht</i> (H12b)	235
Tabelle 53: rANCOVA für <i>Fehlervermeidende Erarbeitung durch den Lehrer selbst</i> (H13b).....	235
Tabelle 54: rANCOVA für <i>Fürsorgliche Anleitung und Kontrolle</i> (H14b).....	235
Tabelle 55: rANCOVA für <i>Unterstützung kognitiver Selbstständigkeit</i> (H15b)	236
Tabelle 56: rANCOVA für <i>Persönliches Vertrauen</i> (H19b).....	236
Tabelle 57: Deskriptive Statistiken der Skalen zu beiden Messzeitpunkten mit reduzierter Stichprobe 1	236
Tabelle 58: Ergebnisse mANOVA (Hypothese 22a).....	237
Tabelle 59: ANCOVA Schülerleistung (H22b)	237

Tabelle 60: Mittelwerte, Standardabweichung und ANOVA für LSE-Skalen nach Clustern	239
Tabelle 61: t-Test auf Abweichungen der Mittelwerte der Cluster vom semantischen Median von 3	241
Tabelle 62: <i>p</i> -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für <i>Akzeptanz der LSE</i>	241
Tabelle 63: <i>p</i> -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für <i>LSE als Belastung</i>	242
Tabelle 64: <i>p</i> -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für <i>Lehrplanvalidität</i>	242
Tabelle 65: <i>p</i> -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für <i>förderdiagnostischen Nutzen</i>	243
Tabelle 66: <i>p</i> -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für <i>unterrichtsplanerischen Nutzen</i>	243
Tabelle 67: <i>p</i> -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für <i>selektionsdiagnostischen Nutzen</i>	244
Tabelle 68: „Dichotomisierte“ Beschreibungen der Abweichungen der Cluster-Skalenmittelwerte aufbauend auf <i>t</i> -Tests (vgl. Tabelle 61)	244
Tabelle 69: Deskriptive Statistiken nach Clustern getrennt für Unterrichtserfahrung der Lehrkräfte	246
Tabelle 70: Häufigkeitstabelle nach Clustern getrennt für Geschlecht.....	246
Tabelle 71: Häufigkeitstabelle nach Cluster getrennt für Projektteilnahme	246
Tabelle 72: Deskriptive Statistiken nach Clustern getrennt für LSE-Vorerfahrungen (als Summenscore der Jahrgänge 2009, 2010, 2011, 2012)	246
Tabelle 73: Häufigkeitstabelle nach Clustern getrennt für aktuelle Schulform	247
Tabelle 74: Häufigkeitstabelle nach Clustern getrennt für Ausbildung	247
Tabelle 75: Ergebnisse der Konfigurationsfrequenzanalyse für die Wahrnehmungsskalen <i>Akzeptanz</i> , <i>Belastung</i> und <i>Lehrplanvalidität</i> (Mediansplit)	249
Tabelle 76: Ergebnisse der Konfigurationsfrequenzanalyse für die Nutzungsskalen <i>Förderung</i> , <i>Planung</i> und <i>Selektion</i> (Mediansplit)	249
Tabelle 77: Ergebnisse der Konfigurationsfrequenzanalyse für die Wahrnehmungsskalen (Split am semantischen Median).....	251
Tabelle 78: Ergebnisse der Konfigurationsfrequenzanalyse für die Nutzungsskalen (Split am semantischen Median)	251

Tabelle 79: Ergebnisse der t -Tests für einzelne Stichproben der Unterrichtsvalidität (operationalisiert durch die Ungewohntheit der Aufgaben) zwischen Mathematik und Deutsch sowie Englisch (bei Kontrolle der Teilnahme).....	252
Tabelle 80: Ergebnisse der t -Tests der Auswertungszeit (in Minuten) zwischen Mathematik und Deutsch sowie Englisch	253
Tabelle 81: Angemessenheit der Auswertungszeit im Fächervergleich (mit Kontrolle der Teilnahme)	254
Tabelle 82: Regressionsanalyse der Auswertungszeit (allgemein bzw. mathematikbezogen) auf die Einschätzung der Nützlichkeit bzgl. der Lernstandsdiagnose, Unterrichts- und Schulentwicklung	255
Tabelle 83: Einschätzung der Nützlichkeit nach Fächern (mit Kontrolle der Teilnahme)	255
Tabelle 84: Nützlichkeit der LSE nach Fächern (mit Kontrolle der Teilnahme)	255
Tabelle 85: Durchsicht der Rückmeldungen	258
Tabelle 86: Besprechung der Ergebnisse der Lernstandserhebungen differenziert nach Fächern	259
Tabelle 87: Regression der Nützlichkeit auf die Durchsicht der verfügbaren Berichte	259
Tabelle 88: Regression der Nützlichkeit auf die Durchsicht der verfügbaren Berichte (nur Mathematik)	260
Tabelle 89: Ergebnisse Regressionsanalyse Durchsicht auf Besprechung.....	262
Tabelle 90: Ergebnisse Regressionsanalyse Durchsicht auf Besprechung nur Mathematiklehrkräfte	262
Tabelle 91: Regressionsanalyse der Nützlichkeit auf die Gesamtskala „Nutzung“ .	263
Tabelle 92: Regressionsanalyse der Durchsicht bzw. Besprechung auf die Nutzung (als Skala)	264
Tabelle 93: Ergebnisse zur Nutzung der Lernstandserhebungen differenziert nach Fächern	265
Tabelle 94: Korrelationen zwischen Klassengröße und der Belastung von Mathematik-Lehrkräften der Stichprobe 1 in Prä- und Posttest	266
Tabelle 95: Korrelationen zwischen Unterrichtserfahrung und Einstellungen zu den Lernstandserhebungen (getrennt nach Stichproben)	267
Tabelle 96: Deskriptive Statistik zur Frage: „Kennen Sie das didaktische Material?“	268

Tabelle 97: Deskriptive Statistik zur Bewertung des didaktischen Materials mit Schulnoten	268
Tabelle 98: Deskriptive Statistik zur Frage: „Wie haben Sie mit dem didaktischen Material gearbeitet?“	268
Tabelle 99: Deskriptive Statistik zur Frage: „Die Zusatzinformationen zur Aufgabe helfen mir die Klassenergebnisse richtig einzuordnen.“	268
Tabelle 100: Deskriptive Statistik zur Frage: „Die Anregungen zur Weiterarbeit finde ich hilfreich für meine eigene Arbeit.“ ¹⁰¹	268
Tabelle 101: Deskriptive Statistik zur Frage: „Ich nutze die didaktischen Materialien zur Information über didaktische Konzepte und Entwicklungen.“	269
Tabelle 102: Deskriptive Ergebnisse zur Frage: „Ich nutze die aktuellen didaktischen Materialien als Übungsmaterial nach dem Lernstand.“ ¹⁰²	269
Tabelle 103: Hypothesenübersicht H1 - H6	277
Tabelle 104: Hypothesenübersicht H7 - H13	278
Tabelle 105: Hypothesenübersicht H14 - H22 ¹⁰⁷	279
Tabelle 106: Vergleich der Itemparameter aus der freien Skalierung und Skalierung mit vorgegebenen Itemparametern aus der Pilotierung VerA2010/2011	289
Tabelle 107: maximale Q ₃ -Werte je Items im Leistungstest	402
Tabelle 108: Itemfit-Statistiken des Leistungstests (Teil 1)	403
Tabelle 109: Itemfit-Statistiken des Leistungstests (Teil 2)	404
Tabelle 110: Deskriptive Beschreibung der Rangtransformation für Lernstandserhebungsskalen– Verfahren rmANOVA	405
Tabelle 111: Ergebnisse der rmANOVA für Lernstandserhebungsskalen	406
Tabelle 112: Deskriptive Ergebnisse der Unterrichtsqualitätsskalen – Verfahren mANOVA	408
Tabelle 113: Ergebnisse der rmANOVA für Unterrichtsqualitätsskalen	409
Tabelle 114: Deskriptive Ergebnisse der Rangtransformation für Unterrichtsqualitätsskalen– Verfahren rmANOVA	411
Tabelle 115: Regression der Nützlichkeitsdimensionen sowie der Nützlichkeits-Skala auf die Nutzung (Einzelitems)	413
Tabelle 116: Regression der Nützlichkeitsdimensionen sowie der Nützlichkeits-Skala auf die Nutzung (Einzelitems; nur Mathematik-LPS)	414

Abkürzungsverzeichnis

ANCOVA	Kovarianzanalyse
COACTIV	Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz
Co ² CA	Conditions and Consequences of Classroom Assessment
DISUM	Didaktische Interventionsformen für einen selbstständigkeitsgesteuerten Unterricht am Beispiel Mathematik
IGLU	Internationale Grundschul-Lese-Untersuchung
IQB	Institut für Qualitätsentwicklung im Bildungswesen
IQB-LV	Ländervergleich des IQB (mittlerweile Bildungstrend)
kA	Kognitive Aktivierung
KMK	Kultusministerkonferenz
LSA	Landesschulamt und Lehrkräfteakademie
LSE	Lernstandserhebung
mANOVA	Messwiederholte Varianzanalyse
OECD	Organisation for Economic Coordination and Development
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
rANCOVA	Rangtransformierte Kovarianzanalyse
rmANOVA	Rangtransformierte messwiederholte Varianzanalyse
SINUS	Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts
TEDS(-M)	Teacher Education and Development Study in Mathematics
TIMSS	Trends in International Mathematics and Science Study
VerA	Vergleichsarbeiten

Einleitung

„VerA oder DiaNA“ lautete der Titel eines Vortrages von Prof. Dr. Werner Blum im Juni 2010 am Institut für Qualitätsentwicklung (IQB) in Berlin in Bezug auf die Vergleichsarbeiten in Mathematik im Jahrgang 8. Die **Vergleichs-Arbeiten**¹ sollten demnach eher als **Diagnostisch-Nutzbare-Arbeiten** betitelt sein, um den Zielen und Funktionen des Verfahrens gerecht zu werden, Lehrkräfte nicht zu verwirren und so letztlich das Verfahren korrekt darzustellen. Seit dem Schuljahr 2008/2009 werden in Deutschland die Vergleichsarbeiten durchgeführt und zielen darauf ab, Schulentwicklung anzustoßen. Bei vielen Lehrkräften fällt der Name „Vergleichs“-Arbeit jedoch negativ auf, da das Verfahren zentral entwickelt und von den jeweiligen Landesinstituten bereitgestellt wird. Lehrkräfte fragen sich also: „Wer wird hier verglichen?“ und „Wer erfährt meine Ergebnisse?“. Diesen Missverständnissen hätte mit dem Titel DiaNA vermutlich vorgebeugt werden können, da dies Verfahren insbesondere keinen Vergleich zwischen Lehrkräften anstrebt bzw. deren „Leistung“ bewertet werden soll und die Ergebnisse – zumindest in Hessen – nicht an die Bildungsadministration weitergegeben werden.

Offensichtlich wurden diese Schwierigkeiten bei der Einführung der Vergleichsarbeiten nicht oder nur unzureichend vorab berücksichtigt. Im Schuljahr 2012/2013 wurde das Verfahren bereits zum fünften Mal durchgeführt und für Hessen lagen, abseits vereinzelter persönlicher Rückmeldungen, bis dato keine systematischen Untersuchungen über die Wahrnehmung des Verfahrens, dessen Nutzung sowie die Auswirkungen auf den Unterricht vor. Die vorliegende Arbeit versucht, diese Forschungslücke ansatzweise zu schließen. Dabei soll, ausgehend von einer theoretischen Erläuterung der Grundlagen, das Projekt VELM-8 (**V**erbesserung der **E**ffektivität der **L**ernstandserhebungen **M**athematik Klasse **8**) vorgestellt werden, in das diese Arbeit eingebettet ist. Wesentlicher Inhalt dieser Arbeit ist die Erläuterung der Vergleichsarbeiten bzw. Lernstandserhebungen, worauf die ersten Abschnitte des Theorie-Teils mit entsprechenden Hintergrundthemen hinarbeiten: Zunächst wird mit den Bildungs-

¹ Das Instrument Vergleichsarbeiten (VerA) wird in verschiedenen Bundesländern mit anderen Bezeichnungen versehen. Für Hessen wird das Instrument unter dem Namen „Lernstandserhebungen“ geführt. Der Begriff „Vergleichsarbeit“ birgt im Wortlaut, wie eingangs erwähnt, die Gefahr von Missverständnissen. Im Verlauf der Arbeit werden die Begriffe weitestgehend synonym verwendet, wobei die Bezeichnung „Lernstandserhebungen“ im Speziellen die Umsetzung in Hessen meint.

standards in Kapitel 1.1 die seit nunmehr gut zehn Jahren verbindliche Grundlage des deutschen Mathematikunterrichts sowie der Vergleichsarbeiten vorgestellt. Dabei dienen die Bildungsstandards als kriteriale Bezugsnorm für jegliche Vergleiche und als Referenz für die Leistungen der Schülerinnen und Schüler in Schulleistungsmessungen (Kapitel 1.2). Allgemein gesprochen lassen sich einige Anforderungen an Leistungsmessung artikulieren, die Voraussetzung sind, informative Daten gewinnen zu können. Erhobene Daten – sei es aus international vergleichenden Leistungsmessungen oder auch Klassenarbeiten – müssen in jedem Falle an die Beteiligten zurückgemeldet werden. Es muss also ein Feedback (Kapitel 1.3) erfolgen, welches für die jeweilig im Fokus stehende Gruppe (z. B. Schülerinnen und Schüler, Lehrkräfte oder Bildungspolitik) informativ gestaltet ist. Aus diesem bisher eher linearen Prozess wird durch das Selbstverständnis bzw. die Definition von Evaluation (Kapitel 1.4) ein Kreislauf. Maßnahmen und Entwicklungsrichtungen, die aus dem Feedback abgeleitet werden, gilt es umzusetzen und in einem nächsten Schritt wiederum einer kritischen Prüfung zu unterziehen, um dauerhaft verlässliche Daten zu erhalten und den Entwicklungsprozess zu begleiten. Im Falle schulischer Evaluation richten sich die Maßnahmen auf Schulentwicklung (Kapitel 1.5), sodass die einzelnen Schulen ihre Möglichkeiten im Rahmen der inhaltlichen und rechtlichen Vorgaben nutzen können. In diesem Kapitel soll auch näher auf den Aspekt der persönlichen Weiterqualifikation durch Fortbildungen eingegangen werden, um die spätere Vorstellung des Projekts VELM-8 sowie dessen Evaluation vorzubereiten. Die Thematisierung des Verfahrens VerA bildet den Abschluss des Theorieteils. Die Vergleichsarbeiten stellen eine konkrete Umsetzung der angesprochenen Aspekte dar. In Kapitel 1.6 soll das Instrument mit Zielen und Funktionen sowie die Entwicklung näher vorgestellt, die verfügbaren Rückmeldungen präsentiert und etwaige Entwicklungsrichtungen aufgezeigt werden. Während diesen Darstellungen zunächst vor allem bildungsbehördliche Texte zu Grunde liegen, wird im Anschluss die tatsächliche Akzeptanz, Wahrnehmung und Nutzung des Verfahrens anhand empirischer Forschungen erläutert.

Im Teil 2 der Arbeit wird das Projekt VELM-8 detailliert vorgestellt. Damit wird die Struktur einer empirischen Arbeit bewusst durchbrochen: Das Projekt VELM-8 war in der Anlage mehr ein Fortbildungs- als ein Forschungsprojekt. Daher stellt die Konzeption der Fortbildungen an sich einen wichtigen Inhalt dar, dem durch die Platzierung eine höhere Bedeutung gegeben werden soll. In Teil 0 folgen mit dem Methodenteil die Aufstellung der zu untersuchenden Hypothesen, die Vorstellung der

zugrundeliegenden Stichproben, die Prüfung der genutzten Skalen sowie die Vorstellung der genutzten Auswertungsverfahren. Daran anschließend werden in Teil 4 die Ergebnisse präsentiert. Die Arbeit schließt mit einer kritischen Diskussion der Ergebnisse vor dem Hintergrund der bestehenden Forschungsergebnisse und den Grenzen dieser Arbeit (Teil 1).

Einzelne Teile der Arbeit wurden bereits auf Tagungen der Gesellschaft für Didaktik der Mathematik (GDM), der Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF) sowie des Arbeitskreises Empiriegestützte Schulentwicklung (EMSE) vorgestellt und sind Inhalt von Veröffentlichungen, die alle unter Erstautorenschaft des Autors erfolgt sind (Vogel, Achmetli, Krawitz & Blum, 2013; Vogel, Blum, Achmetli & Krawitz, 2013; Vogel, Achmetli, Krawitz & Blum, 2014; Vogel, Blum, Achmetli & Krawitz, 2016).

1 Theorie

1.1 Bildungsstandards

1.1.1 Entwicklung zu den Bildungsstandards

In vielen anderen Veröffentlichungen stehen die Bildungsstandards jedoch lediglich als pars pro toto für sämtliche neue Verfahren zur Lenkung des Schulwesens. (Uhl, 2006, S. 3)

Die öffentliche Ergebnisrezeption um TIMSS (Trends in International Mathematics and Science Study) und PISA 2000 (Programme for International Student Assessment) stellte nach langer Zeit erstmals eine Situation dar, in der öffentliches Interesse und Aufmerksamkeit auf die Schule, die Schülerinnen und Schüler sowie deren Leistungen gerichtet waren. Dabei standen folglich auch die Lehrkräfte im Fokus. Die öffentliche Debatte fiel dabei umso dramatischer aus, da die Ergebnisse scheinbar einen dringenden – wenn nicht überfälligen – Handlungsbedarf aufzeigten (Blum et al., 2006; I. Hosenfeld, Groß-Ophoff & Bittins, 2006; Maritzen, 2014; Rösner, 2005). TIMSS und PISA stellten jedoch Ende der 1990er und Anfang der 2000er Jahre keinesfalls die ersten Leistungsvergleichsmessungen ihrer Art dar: Bereits in den 70er Jahren wurde im OECD-Report dem deutschen Bildungssystem ein schlechtes Zeugnis und dringender Reformbedarf attestiert (Lührig, 1973). Auch die Teilnahmen an FIMS (First International Mathematics Study) bzw. FISS (First International Science Study) Anfang der 60er Jahre zeigten ein Abrutschen junger deutscher Schülerinnen und Schüler gegenüber ihren älteren Schulkameradinnen und -kameraden (van Ackeren & Klemm, 2011). Weitestgehend abseits dieser internationalen Vergleichsuntersuchungen wurde von Picht (1964) die deutsche Bildungskatastrophe proklamiert, dessen Fokus jedoch weniger auf der reinen Leistung der Schülerinnen und Schüler als auf der nicht vorhandenen Chancengleichheit u. a. zwischen städtischen und ländlichen Regionen sowie einer besseren Ausstattung von Schulen lag. Selbstverständlich haben sich die Institution Schule und die Rahmenbedingungen schulischer Bildung seither deutlich geändert. Dennoch wurden bereits etwa 40 Jahre vor den „erschreckenden“ und „aufrüttelnden“ Ergebnissen von TIMSS und PISA Mangelerscheinungen im deutschen Bildungssystem offengelegt, ohne jedoch, dass tragfähige Maßnahmen ergriffen worden wären, diese zu beheben.

Seit Mitte der 1990er Jahre lassen Beschlüsse der Kultusministerkonferenz sowie diverse Projekte auf regionaler Ebene erahnen, dass das bisher nicht in Frage gestellte „Aushängeschild“ Deutschlands – das Bildungssystem – nun kritisch beleuchtet wird. Bereits 1997 verabschiedete die Kultusministerkonferenz in den „Konstanzer Beschlüssen“ die Notwendigkeit der Qualitätssicherung hinsichtlich der Bereiche (1) Gleichwertigkeit von schulischer Ausbildung, (2) Vergleichbarkeit von Schulabschlüssen sowie (3) Durchlässigkeit innerhalb des Bildungssystems. Neben der normativen Vorgabe sollten entsprechende Evaluationsinstrumente entwickelt und erprobt werden sowie in internationalen Vergleichsuntersuchungen regelmäßig komparativ der Leistungsstand deutscher Schülerinnen und Schüler gemessen werden. Diese internationalen Verfahren sollen dabei durch (bundes-)länderbezogene Verfahren ergänzt werden (KMK, 1997). Diese Entwicklung entspricht den Forderungen bzw. kritischen Betrachtungen seitens der Bildungsforschung, wie beispielsweise Klieme und Baumert (2001), die das Fehlen einer „Qualitätssicherung, das heißt die kritische Bilanzierung durch breite und regelmäßige empirische Bewährungskontrolle“, anmahnen. Darüber hinaus versuchten im Anschluss an die Veröffentlichung der TIMSS- und PISA-Ergebnisse diverse national und regional angelegte Studien das schlechte Abschneiden zu erklären sowie Ansatzpunkte für Verbesserungen aufzuzeigen (I. Hosenfeld et al., 2006; KMK, 2001, 2002c).

Neben diesem „Bildungsschock“ und der Diskussion um Qualitätssicherung im Bildungswesen bedeutete die Leistungs-Operationalisierung internationaler Vergleichsuntersuchungen einen weiteren Diskussionsansatz. Durch TIMSS und PISA wurden normative Leistungs- und Kompetenzmaße aufgestellt, die nur bedingt anschlussfähig an die eher input- bzw. an Lehrplänen orientierte Bildungssteuerung Deutschlands waren. Eine möglichst direkte Nutzung von Ergebnissen aus Vergleichsuntersuchungen jedoch bedarf der Möglichkeit, die Leistungsmaße von Untersuchung und Bildungssystem miteinander in Verbindung zu bringen (Peek, Pallack, Döbelstein, Fleischer & Leutner, 2006; Rolff, 2001). Resultierend aus dieser Entwicklung wurde auf der 298. Sitzung der Kultusministerkonferenz in Eisenach entschieden, u. a. Bildungsstandards einzuführen (KMK, 2002a).

1.1.1.1 Definition Bildungsstandards

Angestoßen vom schlechten Abschneiden in den internationalen Schulvergleichsuntersuchungen wurde eine Expertengruppe unter der Leitung von Eckard Klieme damit

beauftragt, ein Konzept zur Entwicklung nationaler Bildungsstandards vorzulegen. Diese sogenannte Klieme-Expertise umfasst neben den Empfehlungen für die Umsetzung von Bildungsstandards und deren Eigenschaften auch einen internationalen Blick in Nationen, die bereits über Erfahrungen mit Output-Standards sowie deren Evaluation haben (Klieme et al., 2003, 2009).

Die Problematik einer Inputsteuerung, wie sie vor Einführung der Bildungsstandards in Deutschland vorhanden war, besteht darin, dass eine Veränderung des Inputs nicht zwingend eine Veränderung des Outputs nach sich zieht (Hofmann-Göttig, Eschmann & Daumen, 2005; I. Hosenfeld et al., 2006; Oelkers & Reusser, 2008). Dies lässt sich mit einem Augenzwinkern wie folgt illustrieren:

Max: „Ich habe meinem Hund Spot beigebracht zu pfeifen!“
Luisa: „Ich kann Spot aber gar nicht pfeifen hören.“
Max: „Ich habe es ihm ja auch nur beigebracht,
ich habe nicht gesagt, dass er es gelernt hat.“

Anders als Lehrpläne formulieren Bildungsstandards „gegenüber Lehrkräften, Eltern, Schülerinnen und Schülern, Bildungsadministration und Öffentlichkeit verbindliche Erwartungen in bestimmten inhaltlichen Lernbereichen, die auch überprüft werden sollen“ (Oelkers & Reusser, 2008, S. 7). Klieme und Kollegen definieren Bildungsstandards als „Anforderungen an das Lehren und Lernen in der Schule“, die den „Bildungsauftrag, den allgemeinbildende Schulen zu erfüllen haben“, konkretisieren (Klieme et al., 2003, S. 19).

Bevor jedoch Ausführungen zu den inhaltlichen Ausgestaltungen von Standards beginnen, gilt es einige Vorüberlegungen zu berücksichtigen. Dabei sollten nach der Klieme-Expertise (2003) die Leitfragen nach (1) dem Bezug der Standards, (2) der Ebene der Standards und (3) dem Zielniveau der Standards geklärt sein. Handelt es sich um Input-, Prozess- oder Outputstandards und sollen diese auf der Ebene allgemeiner Bildungsziele, bereichsspezifischer Kompetenzen oder fachspezifischer Inhalte konkretisiert werden? Zentral ist auch die Frage nach dem Niveau: Handelt es sich um Mindestanforderungen (Minimalstandards²), um Höchstanforderungen (Maximalstandards) oder soll ein durchschnittlicher Standard, sog. Regelstandard, angestrebt werden? Zumindest beim zweitgenannten Maximalstandard lässt sich schnell

² vgl. zur Diskussion um Mindeststandards u. a. Drüke-Noe (2011, 2012), Linneweber-Lammerskitten & Wälti (2007), Reiss (2007) und Sill (2005, 2010).

Teil 1 – Theorie Bildungsstandards

erkennen, dass es inhaltlich äußerst schwerfallen dürfte und aus pädagogischer Sicht sehr zweifelhaft wäre, einen Standard zu setzen, der nicht übertroffen werden kann/darf. Die Empfehlung an die Bildungspolitik lautet im Rahmen dieser drei Leitfragen: ergebnisbezogene, bereichsspezifische Mindeststandards zu setzen, die an allgemeinen Bildungszielen orientiert nach „oben“ und „unten“ (an die jeweils angrenzenden Bildungsabschnitte) anschlussfähig sind und sich in Aufgaben operationalisieren lassen.

Aus diesen Antworten auf die genannten Leitfragen lässt sich ein grundlegender Unterschied der bildungsstandardsgeprägten Outputorientierung und der lehrplangeprägten Inputorientierung kennzeichnen: Während die Lehrpläne mit einer Liste von *Lehrinhalten* gefüllt sind, enthalten Bildungsstandards *Kompetenzanforderungen*. Dabei liegt den Bildungsstandards die vielfach zitierte Kompetenzdefinition von Weinert (2001, S. 27f.) zugrunde: Kompetenzen sind demnach „die bei Individuen verfügbaren oder von ihnen erlernbaren kognitiven Fähigkeiten und Fertigkeiten, bestimmte Probleme zu lösen sowie die damit verbundenen motivationalen, volitionalen und sozialen Bereitschaften und Fähigkeiten, die Problemlösungen in variablen Situationen erfolgreich und verantwortungsvoll nutzen zu können.“ Damit zeigen Kompetenzen neben dem reinen *Wissen* weitere Facetten (Klieme et al., 2003, S. 73; vgl. auch Blum & Leiß, 2007b):

- Fähigkeit
- Verstehen
- Können
- Handeln
- Erfahrung
- Motivation

Aus dieser Definition resultiert, dass sich die in den Bildungsstandards formulierten Kompetenzanforderungen in Aufgaben, also in konkreten Anforderungssituationen, veranschaulichen lassen: Zur Bewältigung dieser Anforderungen bedarf es der *Fähigkeit*, das Wissen (*Verstehen*) und *Können* in *Handeln* umzusetzen und dabei unter Rückgriff auf die bisherigen *Erfahrungen* anzuwenden, wobei explizit „die damit verbundenen *motivationalen*, *volitionalen* und sozialen Bereitschaften“ angesprochen werden (Weinert, 2001, S. 27f.).

In der Ausprägung als bereichsspezifische Kompetenzanforderungen stellen Bildungsstandards in der von der Kultusministerkonferenz beschlossenen Form (2003: Mittlerer Bildungsabschluss; KMK, 2004a, 2004b: Primarbereich, Hauptschulabschluss; KMK, 2012a: Allgemeine Hochschulreife)³ fachliche Standards dar, die sich auf Kernbereiche des Faches fokussieren (van Ackeren & Klemm, 2011). Für die Fächer Deutsch und Mathematik liegen mittlerweile Bildungsstandards vom Primarbereich bis zur Allgemeinen Hochschulreife vor, die einen kumulativen Kompetenzaufbau ermöglichen und einfordern. Neben diesen Merkmalen „guter“ Bildungsstandards (*Fachlichkeit, Fokussiertheit, Kumulativität*) benennt die Expertengruppe um Eckhard Klieme noch *Verbindlichkeit für alle, Differenziertheit* (vgl. Anforderungsniveaus in der Sekundarstufe 2 und Bildungsgänge in der Sekundarstufe 1), *Verständlichkeit* und *Realisierbarkeit* (Klieme et al., 2003).

Im folgenden Abschnitt sollen die Bildungsstandards für das Fach Mathematik konkreter vorgestellt werden.

1.1.1.2 Bildungsstandards Mathematik

Im Sinne des Merkmals der *Kumulativität* bauen die Bildungsstandards im Fach Mathematik vom Primarbereich bis zur Allgemeinen Hochschulreife aufeinander auf. Deutlich wird dies sehr schnell, wenn die Struktur der Bildungsstandards betrachtet wird.

Beginnend am Ende der Schullaufbahn liegt für die Allgemeine Hochschulreife (KMK, 2012a) das folgende Modell⁴ zugrunde, welches sich in drei Dimensionen aufgliedert:

1. Leitideen
2. Allgemeine mathematische Kompetenzen
3. Anforderungsbereiche

³ Diese Jahresangaben beziehen sich auf die Bildungsstandards im Fach Mathematik. Die Veröffentlichungsjahre der Bildungsstandards für andere Fächer können u. U. abweichen.

⁴ Zu unterscheiden ist vom Kompetenzstrukturmodell das Kompetenzstufenmodell. Das Kompetenzstufenmodell ist eine nachträgliche empirische Setzung von Schwierigkeitsstufen (KMK, 2012b).

Teil 1 – Theorie
Bildungsstandards

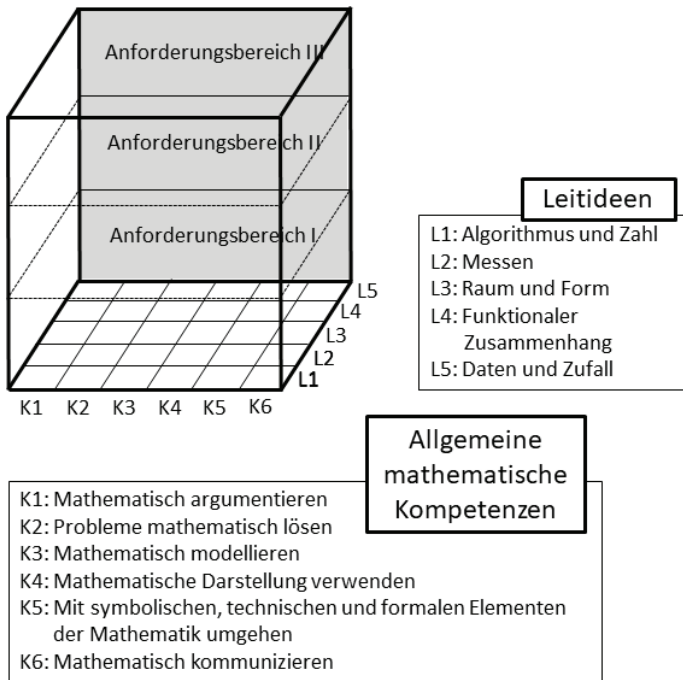


Abbildung 1: Strukturmodell der Bildungsstandards Mathematik für die Allgemeine Hochschulreife (verändert nach KMK, 2012a, S. 12)

Die Leitideen sind in den Bildungsstandards für den Hauptschulabschluss (KMK, 2004a), den Mittleren Bildungsabschluss (KMK, 2003) und die Allgemeine Hochschulreife nahezu identisch und gliedern sich in die Bereiche:

1. (Algorithmus und) Zahl
2. Messen
3. Raum und Form
4. Funktionaler Zusammenhang
5. Daten und Zufall

Für die gymnasiale Oberstufe ist die Leitidee 1 gegenüber den Standards für die Sekundarstufe I um den Aspekt von *Algorithmen*, „die prinzipiell automatisierbar und damit einer Rechnernutzung zugänglich sind“, erweitert.

Die allgemeinen mathematischen Kompetenzen stellen typische mathematische Handlungsweisen dar, die sich zwischen Sekundarstufe I und II nicht unterscheiden. Konkret benannt werden diese sechs Handlungsweisen:

1. Mathematisch argumentieren
2. Probleme mathematisch lösen
3. Mathematisch modellieren
4. Mathematische Darstellung verwenden
5. Mit symbolischen, formalen und technischen Elementen der Mathematik umgehen
6. Mathematisch kommunizieren

Die Anforderungsbereiche gliedern sich in drei Ebenen kognitiven Anspruchs, die für alle Bildungsabschnitte identisch sind. Dabei sind sie für die Standards zur Allgemeinen Hochschulreife von den Anforderungsniveaus zu unterscheiden. Letztere beschreiben gemäß den Vereinbarungen zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II unterschiedliche Unterrichtsgestaltungen (vergleichbar mit der früheren Unterscheidung in Grund- und Leistungskurs; KMK, 2016b). Aufgaben oder Aktivitäten im Mathematikunterricht sind demgegenüber mit unterschiedlichem kognitivem Anspruch verbunden, was in Anforderungsbereichen ausgedrückt wird: Einfaches Ausführen von Routinen oder Nutzung von direkt zugänglichem Wissen fällt in den Anforderungsbereich 1. Im zweiten Anforderungsbereich ist es notwendig, Verknüpfungen zwischen verschiedenen Aspekten herzustellen, mehrschrittige Lösungswege zu planen und durchzuführen oder Routinen auf andere Zusammenhänge zu übertragen. Der Anforderungsbereich 3 beschreibt Aufgaben/Tätigkeiten, die vielschrittige Lösungswege beinhalten, eine kritische Stellungnahme erfordern oder etwa wegen der Komplexität hohe Anforderungen bei der Planung und Durchführung der Bearbeitung stellen.

Das vorgestellte Kompetenzstrukturmodell aus den Sekundarstufen ist dabei direkt anschlussfähig an jenes aus dem Primarbereich und differenziert dies weiter aus. Die Bildungsstandards Mathematik für den Primarbereich nennen die folgenden allgemeinen mathematischen Kompetenzen (KMK, 2004b): Argumentieren, Problemlösen, Modellieren, Darstellen von Mathematik, Kommunizieren.

Teil 1 – Theorie Bildungsstandards

Dementsprechend wird in der Sekundarstufe lediglich das symbolische, technische und formelle Arbeiten mit Mathematik ergänzt. Die inhaltsbezogenen mathematischen Leitideen lauten: (L1) Zahlen und Operationen, (L2) Größen und Messen, (L3) Raum und Form, (L4) Muster und Strukturen, (L5) Daten, Häufigkeit und Wahrscheinlichkeit. Dabei weicht namentlich vor allem die vierte Leitidee in der Sekundarstufe ab: *Muster und Strukturen* (Primarstufe) gegenüber *funktionaler Zusammenhang* (Sekundarstufe I und II). Betrachtet man allerdings die genauere inhaltliche Beschreibung dieser Labels (S. 10f.), so fällt auf, dass es in der Leitidee für den Primarbereich darum geht, „Gesetzmäßigkeiten“ sowie „funktionale Beziehungen erkennen, beschreiben und darstellen“ zu können. Unter der Leitidee *funktionaler Zusammenhang* aus dem Sekundarbereich werden Aktivitäten wie „beschreiben und interpretieren funktionaler Zusammenhänge und ihre Darstellungen in Alltagssituationen“, die Verwendung verschiedener Darstellungsformen für funktionale Zusammenhänge und die Nutzung und Erkennung von proportionalen/antiproportionalen Zuordnungen sowie die Nutzung der Prozentrechnung genannt (S. 10). Daraus wird deutlich, dass es sich in diesem Falle trotz unterschiedlichen Namens um im Kern ähnliche, jedoch in der Sekundarstufe ausdifferenziertere inhaltliche Bereiche im Umfeld von Zuordnungen und Funktionen handelt.

Entsprechend der Definition von Bildungsstandards lassen sich diese drei Dimensionen direkt in Aufgaben konkretisieren. Viele Beispiele hierzu sind in den Veröffentlichungen der Kultusministerkonferenz mit entsprechender Einordnung in das Kompetenzstrukturmodell zu finden. Darüber hinaus gibt es für die einzelnen Bildungsabschnitte begleitende und erläuternde Literatur, in der sich neben genaueren Ausführungen zu den Kompetenzdimensionen und bildungsstandardsorientierter Unterrichtspraxis eine Vielzahl von klassifizierten Aufgaben finden (vgl. Primarbereich: Walther, van den Heuvel-Panhuizen, Granzer & Köller, 2012; Sekundarstufe I: Blum, Drüke-Noe, Hartung & Köller, 2012; Sekundarstufe II: Blum, Vogel, Drüke-Noe & Ropelt, 2015). Daher soll hier nun darauf verzichtet werden, einzelne Aufgaben konkret auf ihre Bildungsstandardsklassifikation zu untersuchen.

1.1.1.3 Funktionen von Bildungsstandards

Mit der Entwicklung und Einführung von länderübergreifenden Bildungsstandards haben die Länder nicht nur Grundlagen für eine wirksame Qualitätssicherung im Bildungswesen, sondern auch für die Entwicklung eines an Kompetenzen orientierten Unterrichts gelegt. (KMK, 2010, S. 7)

Als Konkretisierung von (bereichsspezifischen) Bildungszielen, die durch (Kompetenz)-Anforderungen operationalisiert werden und dabei in konkrete Aufgabenstellungen bzw. Verfahren zur Kompetenzmessung umgesetzt werden können, übernehmen die Bildungsstandards zwei zentrale Funktionen im Bildungssystem (Grabensberger, Freudenthaler & Specht, 2008; Klieme et al., 2003):

1. Orientierungsfunktion
2. Rückmeldefunktion

Die Bildungsstandards geben vor, welche Kompetenzen zum Ende eines Bildungsabschnittes von den Schülerinnen und Schülern beherrscht werden müssen, und stellen damit „Leitplanken“ für den Unterricht bereit, in denen sich das konkrete Unterrichtsgeschehen „frei“ bewegen kann. Insofern bieten die Bildungsstandards den Lehrkräften bei der Planung, Durchführung und Reflexion von Unterricht *Orientierung*. Ein kompetenzorientierter Unterricht bedeutet in diesem Sinne, dass

jede einzelne Unterrichtsstunde und jede Unterrichtseinheit (muss) sich daran messen lassen [muss, S.V.], inwieweit sie zur Förderung und Weiterentwicklung inhaltsbezogener und allgemeiner Schüler-Kompetenzen beiträgt, und der Unterricht über längere Zeiträume hinweg muss so konzipiert sein, dass der Aufbau von Kompetenzen im Zentrum steht. Die wichtigste Frage ist nicht ‚Was haben wir durchgenommen?‘, sondern ‚Welche Vorstellungen, Fähigkeiten und Einstellungen sind entwickelt worden?‘. (Blum, 2012, S. 17)

Greifbare Realisierung einer solchen Orientierungsfunktion ist beispielsweise die Umsetzung in zentralen Vergleichsarbeiten (VerA; vgl. Kapitel 1.6). Hierbei handelt es sich um deutschlandweite Testverfahren, die die Leistung der Schülerinnen und Schüler zu einem bestimmten Zeitpunkt messen. Dabei orientieren sich die Aufgaben an

den abschlussbezogenen Standards des jeweiligen Bildungsabschnitts. Durch die Nutzung von kompetenzorientierten Aufgaben anstelle von technischen und verfahrensorientierten Aufgaben, wie sie beispielsweise häufig in Klassenarbeiten eingesetzt werden (Drücke-Noe, 2014), bieten die Vergleichsarbeiten Lehrkräften Einblicke in eine *neue Aufgabenkultur* und unterstützen damit den „Paradigmenwechsel von der inhaltlichen Orientierung des Unterrichts zur Kompetenzorientierung“ (Posch, 2009, S. 125; Hessisches Kultusministerium, 2009). Damit legen die Bildungsstandards die Grundlage für die Unterrichtsentwicklung und Schulentwicklung, deren Basis diagnostische Informationen sind. Mithilfe von konkreten Erwartungen, was Schülerinnen und Schüler am Ende eines Bildungsabschnittes zu leisten im Stande sein sollen, ist es möglich, den Lehrkräften auf Basis dieser Referenzwerte Rückmeldungen zu geben (Klieme et al., 2003).

Bildungsstandards haben also einen anderen Anspruch als bisherige Lehrpläne. Sie sollen diese aber keineswegs überflüssig zu machen. Vielmehr legen Bildungsstandards „fachliche ‚Essentials‘ sowie die innere Gliederung und das Verhältnis zu anderen Fächern“ fest (Oelkers & Reusser, 2008, S. 17). Die klaren inhaltlichen und prozessualen Vorgaben des *was-wann-wie-wo*-Unterrichtens aus den Lehrplänen sollten nach Empfehlung von Klieme et al. (2003) in Richtung eines Kerncurriculums umstrukturiert werden. Diese Kerncurricula bilden dann die Grundlage für die Entwicklung von Curricula auf unterschiedlichen Ebenen (Schule, Fach etc.), in denen etwa genauere Absprachen zur Unterrichtsgestaltung festgehalten sind und so die aus den Bildungsstandards resultierenden Freiheiten genutzt werden können.

Bildungsstandards, deren Implementation sowie die direkt mit den Bildungsstandards verbundenen Verfahren sind somit letzten Endes Elemente der Qualitätsentwicklung (u. a. von Schulen). In diesem Sinne haben sie vorrangig eine Schul(-system-)Entwicklung und eine individuelle Förderung im Blick; nicht jedoch eine Selektion von Schülerinnen und Schülern (Klieme et al., 2003; KMK, 2010). Gegen eine Selektion oder Bewertung spricht dabei, dass Bildungsstandards lediglich „Leitplanken“ vorgeben und damit nur den „Kern“ des Curriculums abbilden (vgl. Abbildung 2). Dies gilt umso stärker für die Vergleichsarbeiten, da diese einen noch deutlich reduzierten Fokus legen. Des Weiteren sollte eine bewusste Trennung zwischen Evaluation bzw. Monitoring auf der einen und Bewertung bzw. Selektion auf der anderen Seite deutlich werden, da eine Benotung eines einzelnen Schülers oder einer einzelnen

Schülerin professionelle, pädagogische Entscheidungskompetenz der *zuständigen* Lehrkraft bedarf (Klieme et al., 2003).

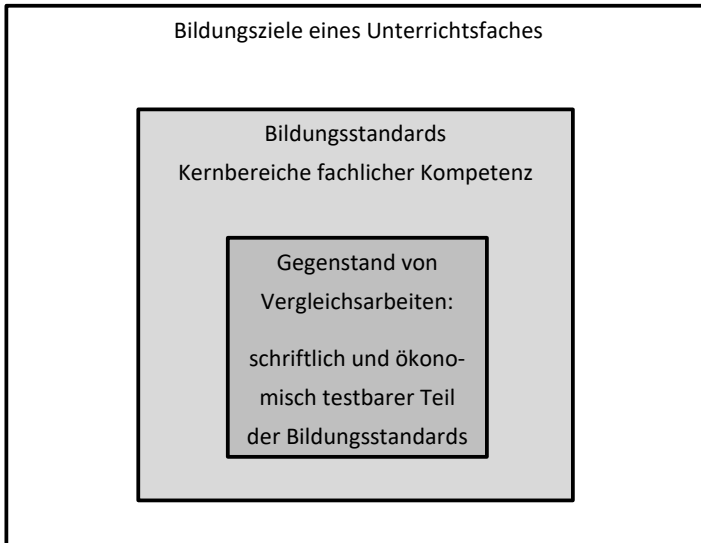


Abbildung 2: Bildungsstandards als Kern bzw. Konkretisierung von Bildungszielen (A. Hosenfeld, 2010, S. 25)

Dennoch sind die Verfahren zur Testung der Bildungsstandards bzw. deren Erreichung nach der normativ-theoretischen Bestimmung von Bildungsstandards notwendig. Mit der Benennung von zentralen Bildungszielen und damit verbundenen Kompetenzen ist eine Entwicklung von entsprechenden Testverfahren sogar erst möglich. Diese Testverfahren dienen dabei zum einen der empirischen Prüfung der Bildungsstandards bzw. der darauf aufbauend entwickelten Aufgaben zur Testung im Sinne einer Validierung der Standards und zum anderen zur Evaluation von Unterricht und individuellen Förderung. Damit Einzelschulen sowie Lehrkräfte aus diesen Testverfahren entsprechende Schlüsse ziehen können, bedarf es allerdings zusätzlich eines Unterstützungssystems sowie der Betonung von Schulen als zentralen Handlungseinheiten des Bildungssystems (Klieme et al., 2003; Maag Merki, 2010; vgl. Kapitel 1.5).

1.1.1.4 Kritik an den Bildungsstandards

Die Einführung der Bildungsstandards wurde von vielen Seiten als eine notwendige und überfällige Entwicklung sehr begrüßt. Trotz all dieser positiven Meinungen und der Notwendigkeit von Reformen nach den aufrüttelnden Ergebnissen aus TIMSS und PISA gibt es vielfach auch Kritik bzw. Risiken, die mit den Bildungsstandards

verbunden werden. Bereits in der Klieme-Expertise (Klieme et al., 2003, S. 56) werden Kritikpunkte antizipiert:

1. Reduktionismusvorwurf
2. Vereinheitlichung und Nivellierung
3. Legitimationsdefizit

Die Bildungsstandards sind als Regelstandards konzipiert und decken dabei nur die inhaltlichen *Kernbereiche* ab. Durch diese Gestaltung besteht jedoch die Gefahr, dass eine übermäßige Fokussierung auf die in den Standards genannten Inhalte und Kompetenzen erfolgt (*teaching-to-the-test* bzw. fachbezogenes *narrowing-the-curriculum*). Kritiker werfen den Bildungsstandards weiter vor, dass *Bildung* nicht nur durch *fachliche* Kompetenzen zu erfassen ist (Baumert, 2001b; Brügelmann, 2005a; Uhl, 2006). In eine ähnliche Richtung geht die Kritik, die den Bildungsstandards eine unverhältnismäßige Fokussierung auf Kernfächer unterstellt. Der schulische Fächerkanon werde durch die Einführung von Bildungsstandards in den Fächern Mathematik, Deutsch und den Fremdsprachen sowie den Naturwissenschaften unverhältnismäßig auf diese fokussiert und so resultierend das schulische Curriculum verengt (*narrowing-the-curriculum*; Bartnitzky, 2005).

Der Begriff der Standardisierung und Vereinheitlichung bringt vielfach die Assoziation von maschineller Produktion der Industrialisierung hervor. Das plakativste Beispiel einer standardisierten Produktion ist das Maschinengewehr 08/15 aus dem Ersten Weltkrieg. In der Redewendung ist „08/15“ mittlerweile als „umgangssprachlich abwertend“, „mit fehlender Originalität“ sowie „Überdruß erzeugend“ beschrieben (Duden, 2016). Eine solche Standardisierung ist mit den Bildungsstandards selbstverständlich nicht intendiert: Daher muss bei aller Vereinheitlichung durch Bildungsstandards bewusst gemacht werden, dass diese lediglich *Kernbereiche* und zentrale Bildungsziele der Fächer abdecken, nicht jedoch einen allumfänglichen Anspruch haben (Baumert, 2001b; Herzog, 2006). Nicht zuletzt stellen auch Lehrpläne oder Schulbücher Elemente einer Standardisierung dar, indem durch diese der Unterricht in verschiedenen Klassen, Schulen, Städten etc. in Bezug auf Inputvorgaben bzw. die genutzten Aufgaben gleich oder zumindest ähnlich ausfällt.

Der letzte Kritikpunkt, der von der Experten-Gruppe um Eckhard Klieme bereits vorhergesehen wurde, betrifft die Legitimation der Standards. Eingeführt wurden die

Bildungsstandards vermeintlich mit Bezug auf den in PISA und TIMSS genutzten Begriff *literacy*, also einem Grundbildungsgedanken. Die zu messenden Kompetenzen bzw. Leistungen in PISA bezogen (und beziehen) sich nicht auf den gesamten Bildungsauftrag der Schule im betreffenden Fach/Domäne, sondern lediglich auf zentrale Kernbereiche. Mit der Orientierung der Bildungsstandards an dieser Konzeption würden bewusst Aspekte von schulischer Bildung ausgeklammert, so die Kritik. Darüber hinaus orientiere sich das Bildungssystem damit an einer Studie, die von der OECD – einer internationalen und vor allem wirtschaftlich orientierten Organisation – durchgeführt wird (vgl. u. a. Jahnke & Meyerhöfer, 2008). Tatsächlich nehmen die Bildungsstandards Mathematik in der vorliegenden Ausgestaltung hingegen Bezug auf das dänische KOM-Projekt (Competency and the Learning of Mathematics). In diesem Projekt wurden analog zur Entwicklung durch die Bildungsstandards versucht, die Syllabi aus inhaltlichen Themen durch „mathematische Kompetenz“ zu ersetzen. Die im Projekt definierten Kompetenzen deuten bereits auf den intensiven Bezug hin (Niss, 2003):

1. Thinking mathematically
2. Posing and solving mathematical problems
3. Modelling mathematically
4. Reasoning mathematically
5. Representing mathematically entities
6. Handling mathematical symbols and formalisms
7. Communicating in, with, and about mathematics
8. Making use of aids and tools

In diese drei Hauptkritikpunkte eingliedern lassen sich weitere, später artikulierte Kritikpunkte, die beispielsweise ein zu einfaches „Ursache-Wirkungs-Denken“ unterstellen, welches den Bildungsstandards zu Grunde läge. Nur durch die Änderung von Rahmenbedingungen ließen sich noch keine Veränderungen für Unterricht direkt erreichen, so die Kritik von Uhl (2006). Die Bildungsstandards als Output-Vorgabe legen jedoch das Ziel fest, auf das hingearbeitet wird. Vor diesem Hintergrund müssen die Fragen nach dem Nutzen kritisch betrachtet werden und dabei zwischen den unterschiedlichen Ebenen des Schulsystems unterschieden werden (Herzog, 2006).

Ebenfalls kritisch angemerkt wird fehlende Transparenz und Akzeptanz: Die Bildungsstandards wurden weitestgehend ohne breite Mitbestimmung der zentralen Akteure in der Schule normativ-theoretisch gesetzt und erst im Nachhinein empirisch geprüft (vgl. Kompetenzstufenmodell; KMK, 2012b). Derartige *top-down*-Vorgehensweisen bergen Risiken, da die Schülerinnen und Schüler sowie die Lehrerinnen und Lehrer als zentrale Akteure von Schule lediglich ein Ergebnis präsentiert bekommen, mit dem sie umgehen müssen (Kuhn, 2014). Letzten Endes sitzen die Experten für den alltäglichen Unterricht in den Klassenzimmern. In diesem Zusammenhang artikuliert Herzog (2006, S. 28): „Wo Experten sind, entstehen auch Laien. Das Schlimmste, was als Konsequenz der Einführung von Bildungsstandards geschehen könnte, wäre daher, dass die pädagogische Expertise von den Lehrerinnen und Lehrern abwandern und sich in Gremien der Bildungsverwaltung einnisten würde.“ Die Lehrkräfte stehen dabei nun mit eingeführten Bildungsstandards und damit verbundener Outputorientierung vor regelmäßigen Evaluationen, deren Akzeptanz bei den Lehrkräften vorausgesetzt wird, ohne dass die entscheidenden Ebenen in Politik oder Schulamt sich ähnlichen Evaluationen unterziehen würden (Böttcher & Rürup, 2010).

Mit den Bildungsstandards sind enorme Erwartungen seitens der Politik verbunden. Allerdings entsprechend die Bildungsstandards in ihrer umgesetzten Ausprägung nicht in allen Belangen den Empfehlungen der Experten-Kommission um Eckhard Klieme, sodass beispielsweise keine Mindest-, sondern Regelstandards eingeführt wurden (Demmer & Schweitzer, 2005). Die Bildungsstandards sind dabei jedoch nur ein Teil umfangreicher Veränderungen, die in der Gesamtstrategie zum Bildungsmonitoring (KMK, 2015) übergreifend orchestriert werden.

1.1.2 Gesamtstrategie zum Bildungsmonitoring

Mit der Gesamtstrategie hat die KMK den Versuch unternommen, zum einen die bis 2006 zum Teil unverbunden entstandenen Aktivitäten quasi nachholend zu einem in sich schlüssigen Konzept zu bündeln, zum anderen die Selbstverpflichtung auf eine gesamtstaatliche, langfristige Strategie zur Transparenz- und Qualitätssicherung festzuschreiben. (Maritzen, 2014, S. 401)

Mit der Einführung der Bildungsstandards wurden verbindliche abschlussbezogene Rahmenvorgaben für bestimmte Bildungsabschnitte eingeführt, die explizit eine

Evaluation ermöglichen sollen. Ein hochkomplexes Gefüge, wie es das Bildungssystem von Bundesebene bis zur Ebene der einzelnen Schülerinnen und Schüler ist, lässt sich jedoch nicht durch ein einzelnes Verfahren oder Instrument gänzlich evaluieren; mehr noch: „Damit die verschiedenen Maßnahmen die notwendigen Impulse zur Verbesserung des Bildungswesens auch tatsächlich auslösen, ist es erforderlich, Prozesse der Qualitätsentwicklung und Standardsicherung auf allen Ebenen, von der einzelnen Schule bis zum gesamten Bildungssystem, systematisch umzusetzen und miteinander zu verbinden“ (KMK, 2015, S. 6). Die Bildungsstandards liefern dazu den verbindlichen Rahmen, sodass im Sinne Rolffs (2001) die einzelnen Verfahren auf verschiedenen Ebenen ineinandergreifen und sich verzahnen (Maritzen, 2014; van Ackeren & Klemm, 2011).

Die Kultusministerkonferenz hat im Jahr 2006 (aktualisiert im Jahr 2015) mit der *Gesamtstrategie zum Bildungsmonitoring* ein Konzept vorgestellt, mit dem die verschiedenen Verfahren zur Qualitätssicherung und -entwicklung des Bildungswesens miteinander in Bezug gestellt werden. Diese Verfahren lassen sich in drei Ebenen gliedern (Abschnitt 1.2; Hessisches Kultusministerium, 2009; I. Hosenfeld et al., 2006; KMK & IQB, 2006; KMK, 2015; Landesschulamt und Lehrkräfteakademie, 2015): (1) internationale Verfahren, (2) nationale Verfahren und (3) bundeslandinterne Verfahren.

Grundsätzlich dienen alle Verfahren dazu, systematisch und miteinander in Verbindung zu setzende Informationen über das Bildungssystem zu generieren und diese zur Unterrichts- und Qualitätsentwicklung sowie zur Unterstützung von Schulen einzusetzen (Maritzen, 2014; KMK, 2015). Damit verfolgt die Gesamtstrategie neben der Qualitätsentwicklung im Unterricht auch eine Weiterentwicklung des gesamten Bildungssystems sowie dessen regelmäßige und wiederholte umfassende Beobachtung sowie Überwachung: In diesem Sinne fordert die Gesamtstrategie ein *Systemmonitoring* (KMK, 2010; van Ackeren & Klemm, 2011; Oelkers & Reusser, 2008). Wacker und Kramer (2012, S. 684) betonen, dass für die Sicherstellung von Erträgen aus dem Bildungssystem auf den verschiedenen Ebenen, unterschiedliche Instrumente der *Rechenschaftslegung* (*accountability*) herangezogen werden: Auf der Makroebene des Bildungssystems wird beim Bildungsmonitoring/Systemmonitoring durch internationale Vergleichsuntersuchungen die Leistungsfähigkeit des deutschen Bildungssystems mit andern Systemen verglichen (KMK, 2015). Das Ziel dieser Vergleiche liegt darin, mögliche Fehlentwicklungen oder Entwicklungsnotwendigkeiten frühzeitig

erkennen zu können und Basisinformationen für Entscheidungen zu erhalten (I. Hosenfeld et al., 2006; Klieme & Baumert, 2001; Kohler & Schrader, 2004). Die Mesoebene wird beispielsweise durch den IQB-Bildungstrend in den Blick genommen. Auf der Mikroebene letztendlich geht es um den Unterricht vor Ort, dem sich bundeslandinterne Verfahren wie beispielsweise die Vergleichsarbeiten widmen (vgl. Kapitel 1.6; Oelkers & Reusser, 2008).

Ergänzt werden diese Verfahren durch den Bildungsbericht (KMK, 2015, S. 9). Es handelt sich bei allen diesen Varianzen weitestgehend um *externe* Evaluationen (vgl. Abschnitt 1.4.1), bei denen das Erreichen von außen gesetzter Kriterien durch unabhängige Forschungseinrichtungen untersucht wird (Kohler & Schrader, 2004).

1.1.2.1 Internationale Verfahren

Das wohl bekannteste internationale Verfahren zur Feststellung der Leistungsfähigkeit von Schülerinnen und Schülern bzw. daraus abgeleitet des Bildungssystems ist PISA (vgl. Organisation for Economic Co-operation and Development [OECD], 2014a, 2016; Prenzel, Baumert, Blum, Lehmann, Leutner, Neubrand, Pekrun, Rolff, Rost & Schiefele, 2004a). Dieses Verfahren wird seit 2000 im dreijährigen Rhythmus mit unterschiedlichen fachlichen Schwerpunkten in Deutschland durchgeführt und gilt als „der zentrale Teil eines umfassenden Indikatorensystems der OECD“ (Prenzel, Drechsel, Carstensen & Ramm, 2004, S. 14). Neben dieser prominenten Vertretung sind mit PIRLS (**P**rogress in **I**nternational **R**eadin**G** **L**iteracy **S**tudy) bzw. IGLU (**I**nternationale **G**rundschul-**L**ese-**U**ntersuchung) im fünfjährigen und TIMSS im vierjährigen Rhythmus drei weitere regelmäßig durchgeführte, internationale Vergleichsuntersuchungen vorgesehen (KMK, 1997; 2002b; vgl. Tabelle 1). Ziel der Verfahren ist dabei eine wiederholte, begleitende wissenschaftliche Überprüfung der Leistungsfähigkeit des deutschen Bildungssystems aufgrund von repräsentativen Stichproben deutscher Schülerinnen und Schüler (Hessisches Kultusministerium, 2009; I. Hosenfeld et al., 2006; Prenzel, Baumert, Blum, Lehmann, Leutner, Neubrand, Pekrun, Rolff, Rost & Schiefele, 2004b; KMK, 2015).

Tabelle 1: Übersicht zu den Testzyklen (grau hinterlegt sind Testungen; verändert nach KMK, 2015, S. 25)

Jahr	PIRLS	IGLU	PISA
	Grundschule		Sek. 1
2009			
2010			
2011	x	x	
2012			x
2013			
2014			
2015		x	x
2016	x		
2017			
2018			x
2019		x	
2020			
2021	x		x

Mit den drei genannten Verfahren wird die Möglichkeit geschaffen, Informationen über das Bildungssystem zu erhalten, da neben den reinen leistungsbezogenen Merkmalen Hintergrundvariablen wie etwa die soziale Herkunft oder der sozioökonomische Status miterhoben werden und so repräsentative Informationen auf Ebene der Nationalstaaten über Zusammenhänge dieser Merkmale generiert werden. Dies spricht jedoch bereits eine Einschränkung in der Nutzung der entstehenden Daten und Informationen an, auf die vielfach hingewiesen wird: „So informativ und anregend das *Programme for International Student Assessment* der OECD ist, handelt es sich doch vom Design her letztlich um eine große Überblicksstudie, die auch mit ihren Korrelationsbefunden in erster Linie Beschreibungswissen liefert“ (Prenzel, 2006, S. 15; vgl. auch I. Hosenfeld et al., 2006). Damit erlaubt PISA insbesondere keine kausalen Interpretationen, sondern lediglich Trendaussagen auf Basis von deskriptiven Zusammenhangsbeschreibungen (Bischof, Hochweber, Hartig & Klieme, 2013; Rolff, 2001, 2002). Diese Aussagen können im Anschluss als Ausgangspunkt einer Weiterentwicklung dienen (I. Hosenfeld et al., 2006; Prenzel, Drechsel et al., 2004). Um jedoch kausal interpretierbare Ergebnisse zu erhalten, bedarf es weiterer Untersuchungen, die mit ausreichend großen Stichproben dem Anspruch auf Repräsentativität gerecht werdend und aussagekräftigen Designs *Handlungswissen* anstelle von *Beschreibungswissen* liefern (Maritzen, 2014).

Die Interpretierbarkeit der Ergebnisse von internationalen Vergleichsuntersuchungen ebenfalls einschränkend lässt sich anführen, dass diese Untersuchungen mit der Leistung von Schülerinnen und Schülern in grundlegenden Kompetenzen den Output messen, ohne dabei Merkmale des Unterrichts zu berücksichtigen. Somit bleiben

veränderbare Größen der Unterrichtsgestaltung unberührt, was bei der Nutzung der Ergebnisse für Unterrichts- und Schulentwicklung zu bedenken bleibt (Ditton, 2000). Entsprechend resultiert als Kritik die wesentliche Frage nach der Nützlichkeit bzw. dem Nutzen v. a. für die einzelne Schule. Zwar stellen die durchführenden Organe der internationalen Vergleichsuntersuchungen⁵ nach Abschluss der Untersuchung mit einiger zeitlicher Verzögerung Daten und Ergebnisse zur Verfügung, jedoch liegt gerade bei internationalen Verfahren meist ein gehöriger zeitlicher Abstand zwischen Erhebung und Ergebnisverfügbarkeit. Außerdem ist eine unzulässige Interpretation auf Basis veröffentlichter Daten ebenso wenig zu verhindern, wie eine bewusste „Täuschung“ durch statistische Berechnungen (Stamm, 2003). Bei der Frage nach dem Nutzen von PISA o. ä. für die einzelne Schule stellt jedoch Baumert (2001b, S. 30) heraus, dass es sich bei internationale Vergleichsuntersuchungen „nicht um die Evaluation von Einzelschulen, sondern um Untersuchungen, die dem Systemmonitoring dienen, und in erster Linie Wissen über Systemzusammenhänge erzeugen wollen“ handelt (vgl. auch Rolff, 2002) und die Nutzung solcher Daten in der „professionellen Verantwortung“ von Schulen liegt (vgl. Abschnitt 1.5).

1.1.2.2 Nationale Verfahren

Während internationale Verfahren den Fokus auf den Vergleich und die Analyse von gesamtstaatlichen Bildungssystemen legen, haben nationale Verfahren die „zentrale Überprüfung des Erreichens der nationalen Bildungsstandards im Ländervergleich“ als Ziel (Hessisches Kultusministerium, 2009, S. 8). Dabei liegt der Fokus auf den Bildungssystemen der einzelnen Bundesländer (I. Hosenfeld et al., 2006). PISA und TIMSS können verständlicherweise keine nationalen Standards als Referenz anlegen, da es unmittelbar zur Benachteiligung anderer Nationalstaaten führen würde. Den internationalen Verfahren liegen somit speziell für diese Verfahren entwickelte Standards/Referenzrahmen zugrunde. Die nationalen Verfahren streben jedoch einen Vergleich zwischen Bundesländern an, in denen die nationalen Bildungsstandards gültig sind. Dem nationalen Ländervergleich, der vom Institut für Qualitätsentwicklung im Bildungswesen (IQB) mit Sitz in Berlin administriert wird, liegen somit Standards zugrunde, die eine Vergleichbarkeit von verschiedenen Bildungsgängen in verschiedenen Bundesländern möglich machen. Zusätzlich sind die Ergebnisse des

⁵ Vielfach wird jedoch auch bei anderen Verfahren und Forschungsprojekten bereits so verfahren, dass nach Abschluss der Auswertungen die Daten offengelegt werden.

Ländervergleichs damit anschlussfähig an Verfahren auf bundeslandinterner Ebene (KMK, 2015). Damit adressiert der IQB-Bildungstrend (ehemals IQB-Ländervergleich) „nicht primär Schulen, sondern die Bildungspolitik der Länder. Es liegt bei ihnen, die Konsequenzen aus den Ergebnissen abzuleiten“ (Petersen, 2013, S. 30) und ergänzt die zuvor genannten Verfahren, was auch durch die zeitliche Verzahnung (Primarstufe & Sekundarstufe) und die fachliche Aufteilung (Mathematik & Naturwissenschaften und Deutsch & erste Fremdsprache) ersichtlich wird (Tabelle 2).

1.1.2.3 Bundeslandinterne Verfahren

Nach den internationalen und nationalen Verfahren schließen sich auf der „untersten“ Ebene bundeslandinterne Verfahren an. Diese Verfahren haben vorrangig weder ein Systemmonitoring oder dezidierte Forschungsinteressen im Blick, sondern sollen eine „schulinterne, datenbasierte Standortbestimmung“ ermöglichen und so „weitere Entwicklungen im Hinblick auf das Erreichen der abschlussbezogenen Bildungsstandards“ anstoßen (Hessisches Kultusministerium, 2009, S. 8; I. Hosenfeld et al., 2006). Damit zielen diese Verfahren auf die Qualitätsentwicklung in Schule und Unterricht. Die Nutzung von zentral entwickelten Vergleichs- und Orientierungsarbeiten wurde dabei bereits deutlich vor der Einführung der Gesamtstrategie festgelegt (KMK, 2002b). Die jährlich durchgeführten Vergleichsarbeiten in Klasse 3 und 8 lassen sich als konkretes Instrument hierfür nennen (vgl. Kapitel 1.6). Darüber hinaus fallen auch weniger breit angelegte Verfahren in diese Kategorie wie beispielsweise lokale lerngruppenübergreifende Klassenarbeiten oder Parallelarbeiten.

Tabelle 2: Übersicht zu den Testzyklen (grau hinterlegt sind Testungen; verändert nach KMK, 2015, S. 25)

Jahr	PIRLS	IGLU	PISA	IQB-Ländervergleich ⁶		
	Grundschule		Sek. 1	Grundsch.	Sekundarstufe 1	
					Mat./Nat. ⁷	Deu, Eng/Fr ⁸
2009						x
2010						
2011	x	x		x		
2012			x		x	
2013						
2014						
2015		x	x			x
2016	x			x		
2017						
2018			x		x	
2019		x				
2020						
2021	x		x	x		x

1.1.2.4 Fazit und Kritik

Wie das einführende Zitat von Maritzen (2014) deutlich macht, liegt mit der Gesamtstrategie zum Bildungsmonitoring ein konkreter Vorschlag zur Verzahnung der bis dahin unverbunden nebeneinanderstehenden Verfahren vor. Maritzen (2014) nimmt damit Bezug auf Rolff (2001, S. 342), der für verschiedene Untersuchungen bzw. Rückmeldungen drei unterschiedliche Bezugsmöglichkeiten nennt:

- Verschiedene Untersuchungen können vollkommen unverbunden nebeneinander stehen bleiben und dabei keinerlei Bezug aufeinander nehmen (*Unverbundenheit*).
- Verschiedene Rückmeldungen bzw. deren zugrundeliegenden Messungen können sich in der Struktur widersprechen und entsprechend gegeneinanderstehen. Bezogen auf eine einzelne Rückmeldung steht die Rückmeldung nicht im Einklang mit den vorherigen Erfahrungen (*Gegentakt*).
- Optimal verhalten sich verschiedene Messungen/Rückmeldungen und deren Nutzung in *Passung/Verzahnung* zueinander. Die grundlegende Struktur sowie die Operationalisierungen der Messungen stimmen überein bzw. die Ergebnisse werden akzeptiert. Die Ergebnisse ergänzen sich dabei entweder

⁶ Da der erste Zyklus des IQB-Ländervergleichs mittlerweile abgeschlossen ist, wird die Untersuchung nun unter dem Namen „Bildungstrend“ geführt (<https://www.iqb.hu-berlin.de/bt>; Zugriff am 19.09.2017).

⁷ Mathematik & Naturwissenschaften

⁸ Deutsch, Englisch & Französisch

gegenseitig oder ergänzen das vorhandene Bild über den eigenen Unterricht bzw. die eigene Schule oder die einzelnen Schülerinnen und Schüler.

Dennoch wurden und werden der Gesamtstrategie, insbesondere den vielfältigen Testverfahren, einige Kritikpunkte entgegengebracht (u. a. Klein & Jahnke, 2012; Meyerhöfer, 2003; Rindermann, 2006 Uhl, 2006):

- Die Nutzung von internationalen Vergleichsstudien wird von vielen Autoren generell sehr kritisch gesehen. Dabei zielt die Kritik häufig darauf ab, dass internationale Vergleichsstudien die lokalen Gegebenheiten nicht berücksichtigen würden, eine „Kompetenzorientierung“ zu Lasten des Wissens von Schülerinnen und Schülern gehe und aus den internationalen Vergleichen keine Schlüsse abgeleitet werden könnten, diese Studien jedoch eine beträchtliche Menge Geld kosteten.
- Ein weiterer Kritikpunkt liegt in der zunehmenden Testung von Schülerinnen und Schülern, die zweifelsfrei mit der Einführung der Bildungsstandards eingesetzt hat. Da jedoch die Verfahren gegenseitigen Bezug nehmen und bewusst verschiedene Aspekte von Bildungsmonitoring berücksichtigen, kommt das Anführen der sog. „Testeritis“ einer Pauschalkritik gleich. Seitens der durchführenden Institutionen wird durch die zeitliche Verzahnung und gleichzeitige Erhebung beispielsweise des Sekundarstufen-Ländervergleichs und PISA versucht, die Test-Belastung für die einzelnen Schülerinnen und Schüler im Mittel gering zu halten.
- Ebenfalls häufig kritisiert werden die Auswirkungen externer Testverfahren auf den Unterricht und ein damit befürchtetes *teaching-to-the-test*; also ein unverhältnismäßiges Fokussieren der unterrichtlichen Handlungen auf die im Test zu leistenden Inhalte bzw. Kompetenzen.

Hier sei jedoch erwähnt, dass den nationalen und bundeslandinternen Verfahren die für den Unterricht verbindlichen Bildungsstandards zugrunde liegen. Die international zugrundeliegenden Kriterien sind Orientierungshilfe für die Bildungsstandardskonzeption gewesen, sodass eine Fokussierung auf Testinhalte v. a. der nationalen und bundeslandinternen Verfahren letzten Endes zu einer verstärkten Implementation der Bildungsstandards in der Schule und im Unterricht führen dürfte (KMK, 2010). Ein solches *teaching-to-the-test* ist also vollkommen legitim.

Grundlage vieler kritischer Anmerkungen ist dabei, dass jede Leistungsmessung zunächst einmal Daten liefert, die eine Erstellung von Rankings einzelner Schülerinnen und Schüler, Lehrkräfte, Schulen, Bundesländer oder nationaler Bildungssysteme ermöglicht. Dies ist ein Aspekt, bei dem das jeweilige Ziel der einzelnen eingesetzten Verfahren kritisch geprüft werden muss: Liefert das Verfahren verlässliche Informationen für die Erstellung von Rankings? Und: Auf welcher Ebene sind Rankings aus diesem Verfahren sinnvoll zu interpretieren? Zumindest für die bundeslandinternen Verfahren lässt sich eindeutig sagen, dass sowohl die Ziele als auch die pädagogische Intention, beispielsweise der Vergleichsarbeiten in Klasse 8, keine Rankings zulassen (vgl. auch Ziele der LSE in Abschnitt 1.6.1). Bei all der Kritik werden jedoch selten konstruktive Gegenvorschläge geäußert, sodass der Schluss nicht lauten kann, wegen der Kritik die Outputorientierung inkl. entsprechender Evaluation von Schulleistungen etc. schlicht abzuschaffen, da sie eine „wahrscheinliche Ursache für den Erfolg von Bildungssystemen“ (van Ackeren & Brauckmann, 2010, S. 41) darstellen.

1.1.3 Resultierende Veränderungen für die Schule

Mit der Einführung von Bildungsstandards, dem Paradigmenwechsel von Input- zur Outputsteuerung, der regelmäßigen Teilnahme an internationalen Vergleichsuntersuchungen sowie nationalen und bundeslandinternen Verfahren ist nur ein kleiner Teil an Veränderungen angesprochen, die unmittelbar das Schulsystem und damit auch (wenn auch in bestimmten Bereichen eher mittelbar) die Institution Schule beeinflussen, da die Schulen durch die mit diesen Veränderungen einhergehende *Dezentralisierung* als operative Einheiten des Bildungssystems in den Fokus gestellt sind (Kuper & Hartung, 2007).

Seit Einführung der Bildungsstandards ist es Aufgabe der Lehrkräfte in Deutschland, für Schülerinnen und Schüler Lerngelegenheiten zu schaffen, in denen die genannten Kompetenzanforderungen erworben werden können und Schülerinnen und Schülern bei deren Aneignung zu unterstützen (Klieme et al., 2003). Der Schule kommen dabei viele neue Freiheiten zu: Sie erhalten beispielsweise Handlungsspielraum in der Ausgestaltung der nationalen Vorgaben (Bildungsstandards) im eigenen Schulcurriculum bzw. Schulprogramm (Hofmann-Göttig et al., 2005; KMK, 2010). Diese Entwicklung ist eine logische Konsequenz des Paradigmenwechsels von der Input- zu Outputsteuerung (Oelkers & Reusser, 2008). Mit der *Dezentralisierung* geht nach Kuper und Hartung (2007, S. 215) eine Verlagerung von Entscheidungskompetenzen auf die

„operativen Einheiten“ des Bildungssystems – die Schulen – einher. Darin wird wiederum die *Orientierungsfunktion* der Bildungsstandards deutlich, der Schule „Leitplanken“ zu setzen, innerhalb derer eigene Ideen bearbeitet werden können (Blum et al., 2006). Damit wird die Schule zur zentralen, pädagogischen Handlungseinheit, wie dies in der pädagogischen Diskussion seit mehreren Jahrzehnten gefordert wird (Blum et al., 2006; Fend, 1986; Gruschka et al., 2005; Holtappels, 2014b; Korngiebel, 2009; O'Day, 2002; Rolff, 2001; van Ackeren & Klemm, 2011).

Einher mit der verringerten Vorgabe im Input geht eine erhöhte Selbst- und Fremdkontrolle des Outputs, da die Steuerung von außen zurückgenommen wird (Holtappels & Rolff, 2010; vgl. auch Ditton, 2000; Oelkers & Reusser, 2008). Hierbei wird von erhöhter *Rechenschaftslegungspflicht* oder auch *accountability* der Schulen gesprochen (Hofmann-Göttig et al., 2005; O'Day, 2002). Die *Rechenschaftsablegung* der Schulen erfolgt dabei beispielsweise durch die bereits angesprochenen Vergleichsarbeiten. Weitere Instrumente können Schulinspektionen aber auch zentrale Abschlussprüfungen sein (Landwehr, 2011). Durch die Messung eines Outputs und der entsprechenden Rechenschaftslegung wird die Aufmerksamkeit innerhalb des Schul- und Bildungssystems auf die Ergebnisse von schulischen Prozessen und deren datengestützte Reflexion respektive resultierende Entwicklungen gerichtet (vgl. Abschnitt 1.3). Damit stehen Schulentwicklung und *Rechenschaftsablegung* in einem „geschwisterlichen Verhältnis“ (Baumert, 2001b, S. 16): So wird eine sinnvolle, den Bedürfnissen entsprechende Ressourcenverteilung ermöglicht und eine Professionalisierung in Gang gebracht (O'Day, 2002). Peek und Dobbelstein (2006b) bestätigen mit PISA-Daten empirisch, dass ein solches *rechenschaftsbasiertes* Schulsystem höhere Leistungen erreiche.

Dies bedeutet neben tief greifenden Veränderungen in der Rolle der Einzelschule und für Lehrkräfte (Kuper & Hartung, 2007) auch Anpassungen in der Funktion von Schulaufsicht: Zu den neuen Aufgaben gehören nun das Anstoßen, die Begleitung und die Reflexion von Qualitätsentwicklungs- und -sicherungsprozessen auf der Einzelschulebene. Um trotz des gewonnenen Handlungsspielraumes der Einzelschule die Vergleichbarkeit verschiedener Schulen sowie insbesondere der Bildungsabschlüsse zu gewährleisten (vgl. KMK, 1997), sind entsprechende Evaluationsverfahren nötig sowie die Unterstützung von Schulen beim Aufbau, der Aufrechterhaltung und Weiterentwicklung von internen Qualitätsmanagementprozessen (Hofmann-Göttig et al.,

2005; Holtappels & Rolff, 2010; Maag Merki, 2010). Rolff (2007) ergänzt in diesem Zusammenhang neben den genannten Punkten die Aufgabe der Gesetzgeber – wie es in Deutschland zumindest für Kernfächer sowie die Naturwissenschaften erfolgt ist – verbindliche Standards, also klare Vorgaben mit ausreichend Spielraum zur Selbstentwicklung zu geben. Ohne näher darauf eingehen zu wollen, kommt der Finanzierung eine wichtige Rolle zu, sowohl für den Gesetzgeber als auch die Behörden in den Bundesländern (Baumert, 2001b; Holtappels & Rolff, 2010; Rolff, 2007; van Ackeren & Klemm, 2011).

Ziel sollte es sein, aus den Ergebnissen der *Rechenschafts*ablegung eine interne Qualitätsentwicklung der Schulen anzustoßen, um somit „von den – ohnehin stattfindenden – Leistungsmessungen wirklich zu profitieren“ (Helmke & I. Hosenfeld, 2005, S. 3). Dazu muss die „Messung von Lernergebnissen [als, S.V.] eine Chance für die Schulen [erkannt werden, S.V.], sich der Ergebnisse der eigenen Arbeit zu vergewissern und darauf professionell zu reagieren“ (Klieme et al., 2003, S. 53; Oelkers & Reusser, 2008). Eine entsprechende Nutzung der Testergebnisse bzw. Testmaterialien sieht auch die Kultusministerkonferenz vor (KMK, 2010). Allerdings muss hierbei vor allem wegen der enormen Heterogenität zwischen Schulen Berücksichtigung finden, für welche Ergebnisse eine Verantwortung vorhanden ist bzw. wer eine Verantwortung für diese Ergebnisse trägt (Reimers, Ramm & Klein, 2009; Schildkamp & Vischer, 2009); also letztlich gilt es zu prüfen, „wer für welche Lernergebnisse [...] ‚accountable‘ ist“ (Maag Merki, 2010, S. 165; Lortie, 1977).

Ausgehend von einer Diagnose der eigenen Schule/des schuleigenen Profils sollte eine Qualitätsentwicklung beginnen. Die Entwicklungsrichtung im Sinne eines *bottom-up*-Effektes ist schnell nachvollziehbar:

To begin with macrolevel reform without compelling evidence about effective practice at the point of context between teachers and students would be like designing airports before practical airplanes existed. (Slavin, 1996, S. 261)

Die Notwendigkeit von *bottom-up*-Entwicklungen wird an diesem Zitat auf ironische Weise deutlich, welches auf Deutsch sinngemäß aussagt: „Ohne innere Schulreform ist die äußere Schulreform nichts“ (Hofmann-Göttig et al., 2005, S. 34). Grundsätzlich ergeben sich vor allem bei *top-down*-Verordnungen Probleme, da die Vorgaben von

außen ohne Kenntnis der „systeminneren Logik“ nur bedingt Einfluss nehmen können (O'Day, 2002, S. 295; vgl. auch Cuban, 1984; Kuhn, 2014). Damit determinieren letzten Endes bei *top-down*-Entscheidungen die „Einzelsysteme, also die Schulen selbst, ob und wie sie diese Interventionen verarbeiten“ (Holtappels & Rolff, 2010, S. 75; Rolff, 2007; Tillmann, 2005). Die darin implizit angesprochene Diversität der Bildungslandschaft und die damit von Behörden kaum zu überblickende Vielfalt ist ein deutliches Kennzeichen, dass Einheitslösungen zentraler Steuerung (*one size fits all*) nur wenig wirkungsvoll erscheinen. Ditton und Arnoldt (2004) verstehen den Widerstand von Lehrkräften gegen ein solches Vorgehen ergänzend als Zeichen dafür, dass Lehrkräfte sich durch zentrale Vorgaben in ihrer Professionalität angegriffen fühlen.

Um jedoch umgedreht *bottom-up*-Entwicklungen anstoßen zu können, bedarf es m. E. zusätzlicher Kompetenzen in der Bildungsverwaltung und v. a. auch bei Lehrkräften wie beispielsweise in den Bereichen Statistik, Methodenlehre aber auch Forschungsdesign (Kuhn, 2014; vgl. auch Abschnitt 1.4 zu Evaluation). *Bottom-Up*-Entwicklungen erfordern zwar ohne Zweifel Freiheiten der Akteure, allerdings sind in der Institution Schule zwei wesentliche Besonderheiten zu bedenken:

- (1) Schulen als „lose gekoppelte Systeme“
- (2) „Endemische Unsicherheit“ von Lehrkräften

Weick (1976) prägte den Begriff *loosely coupled system* in der Diskussion um Organisationsentwicklung (vgl. auch das Modell des *garbage can* nach M. Cohen, March & Olsen, 1972). Dabei meint er Systeme, die zwar an einem gemeinsamen Ziel arbeiten, deren Teilsysteme jedoch weitestgehend unabhängig und unverbunden bleiben. Dies wird beispielsweise dadurch deutlich, dass „Fifth, if there is a breakdown in one portion of a loosely coupled system then this breakdown is sealed off and does not affect other portions of the organization“ (Weick, 1976, S. 7). Übertragen auf das System Schule ergibt sich ein Bild von lose verbundenen Unterrichtsstunden bzw. Unterrichtseinheiten sowie Klassen bzw. Lerngemeinschaften, aber auch Lehrkräften bzw. Fach-/Jahrgangsteams: beispielsweise haben alle Lehrkräfte zwar das gleiche Ziel einer möglichst guten schulischen Bildung der Schülerinnen und Schüler, jedoch ist jede Unterrichtsstunde bzw. -einheit von anderen weitestgehend abgetrennt und die verantwortliche Person für die Unterrichtsstunde oder -einheit ist eine einzelne Lehrperson. Die Lehrkräfte genießen hierin somit eine große Autonomie bei gleichzeitig „eher schwacher Machtposition“ der Schulleitung (Fend, 1986, S. 275). Erreicht eine

Unterrichtsstunde bzw. -einheit bei einer einzelnen Schülerin oder einem einzelnen Schüler nicht ihr Ziel, beeinflusst dies andere Unterrichtseinheiten, Unterrichtsfächer oder Lerngruppen vermutlich im Grunde gar nicht.

Diese strukturelle Besonderheit von Schulen gegenüber beispielsweise deutlich hierarchischer gegliederten Unternehmen verringert jedoch die Wahrscheinlichkeit, unmittelbar auf eine Veränderung in der Systemumwelt reagieren zu können und zu müssen (Weick, 1976), da jedes einzelne „Event“ (im schulischen Fall also Unterrichtsstunden oder -einheiten) „preserves its own identity and some evidence of its physical or logical separateness“ (Weick, 1976, S. 3). Für eine Veränderung im Unterricht ist dabei zunächst die Lehrperson selbst verantwortlich (Tillmann, 2005). Wie Rolff (2001, S. 338) anmerkt, stellen jedoch „lernende Subjekte“ (in diesem Falle also Lehrkräfte, die Veränderungen im Unterricht anstreben) „geschlossene Systeme [dar], die auf sich selbst bezogen sind“. Eine identische Umsetzung einer zentralen Vorgabe durch mehrere Lehrkräfte ist somit nicht erwartbar: Auch dies unterstreicht nochmals die Probleme von *top-down*-Entwicklungen oder -Vorgaben im Stile von *one size fits all*.

Der Begriff der *endemischen Unsicherheit* wurde von Lortie (1977) eingeführt und beschreibt die Ungewissheit von Lehrkräften in der Institution Schule, „gut genug zu sein“. Der Großteil des Unterrichts verläuft hinter geschlossener Tür: Laut TALIS-GEW erhalten beispielsweise etwa 15% der Lehrkräfte in Deutschland nie Rückmeldung seitens der Schulleitung über ihre Arbeit (Barfknecht & Saldern, 2010, S. 97). Hieraus resultiert das Phänomen des Einzelkämpfertums. Lehrkräfte stehen zwar (lose) zusammen und geben an, sich zu unterstützen (Lortie, 1977). Sie sind jedoch grundsätzlich unsicher, ob ihre Art des Unterrichtens erfolgreich ist, die Schülerinnen und Schüler profitieren und der eigene Kenntnis- und Fähigkeitsstand angemessen ist (Lortie, 1977; Rolff, 2001).

Sowohl die lose Kopplung als auch die *endemische Unsicherheit* führt zu dem „Egalitäts-Autonomie-Syndrom“ (Rolff, 2001, S. 339), welches mit dem Sprichwort beschrieben werden kann: „Leben und leben lassen.“ Die Autonomie des „Einzelkämpfertums“ sorgt dafür, dass die lose Kopplung zwischen den Lehrkräften aufrechterhalten wird, die Lehrkräfte wiederum wenig Kenntnis davon haben, wie „erfolgreich“ sie sind und damit nicht in die Gefahr kommen, angeprangert zu werden (beispielsweise erkennen nur etwa 50% der Lehrkräfte in Deutschland Feedback als Hilfe zur

Weiterentwicklung von Unterricht an; Institut für Sozialwissenschaftliche Studien (ISS) e.V. & Gewerkschaft Erziehung und Wissenschaft (GEW), 2009, S. 16; vgl. auch Altrichter, 1998). So wird allerdings auch eine Weiterentwicklung erschwert (Krainz-Dürr, 1999). Vor diesem Hintergrund rückt *vergleichende* Leistungsmessung für Lehrkräfte in ein anderes Licht und es gilt Lehrkräften und Schulen Entwicklungsrahmen zu geben, um selbst gestaltend sich als Organisation gemeinsam zu entwickeln (Kuper & Diemer, 2012; O'Day, 2002; Rolff, 2001).

1.2 Schulleistungsmessungen

Bereits im Vorfeld der viel diskutierten Ergebnisse von PISA wurde von der KMK (1997) beschlossen, regelmäßig an länderübergreifenden, internationalen Vergleichsuntersuchungen teilzunehmen und zusätzlich länderbezogene Qualitätssicherungsmaßnahmen einzuführen bzw. bestehende Maßnahmen zu koordinieren. Diese bereits 1997 angestoßene Entwicklung mündete 2006 in die angesprochene *Gesamtstrategie zum Bildungsmonitoring*. Darin wird deutlich dargelegt, auf welche Weise verschiedene Verfahren gezielt ineinandergreifen sollen, um so die Leistungsfähigkeit des deutschen Bildungssystems im nationalen und internationalen Vergleich sicherzustellen und lokale Qualitätsentwicklungs- und -sicherungsmaßnahmen zu ermöglichen, da diese Verfahren „nicht nur für ein Bildungssystem, sondern auch für die einzelne Schule als ‚Schulevaluation‘ bedeutsam“ sind (G. Orth, 2005, S. 64; KMK, 2015). Im Fokus steht dabei sehr häufig die *Leistung*.

1.2.1 Definition

1.2.1.1 Leistung/Schulleistung

Im schulischen Alltag kommt es an verschiedensten Stellen zu einer Messung von Leistung. Dabei kann *Leistung* je nach Situation unterschiedlich verstanden werden. Jeder Messung liegt immer eine vorherige, notwendigerweise theoriebasierte Definition der Messgröße zugrunde (*Operationalisierung des Messgegenstandes*; Ingenkamp, 1985; Ingenkamp & Lissmann, 2008; Pellegrino, Chudowsky & Glaser, 2001). Eine genaue Definition von *Leistung*, die für alle Kontexte gültig wäre, ist daher kaum möglich, was sich auch darin widerspiegelt, dass sich in einschlägigen pädagogischen und psychologischen Handbüchern keine definierenden Einträge in Glossaren finden lassen (A. Krapp & Weidenmann, 2005; A. Krapp & Seidel, 2014; Rost, 2007; E. Wild & Möller, 2015). Gage, Berliner und Bach (1996, S. 70) grenzen Leistung lediglich diskriminierend von Intelligenz ab: „Lange Zeit hat man *Intelligenz* als etwas betrachtet, das von angeborenen Faktoren bestimmt ist, während *Leistung* etwas ist, das man sich aneignet oder lernt. [...] Leistung auf der anderen Seite ist das Resultat von Unterricht und von Lerntätigkeiten.“ Bereits der Lexikoneintrag von Spieler (1932, S. 255) zeugt beim Definitionsversuch von Leistung ansatzweise von der mittlerweile umgesetzten Kompetenzorientierung: „Die L. beruht auf dem Können, nicht auf dem Wissen. Doch gibt es auf geistigem Gebiete ohne Wissen kein Können.“

Es geht also im schulischen Kontext nicht um *die eine* Leistung, sondern Leistung ist in einzelnen Situationen unterschiedlich operationalisiert: In einem Referat beispielsweise können Aspekte einer guten Präsentation und Verständlichkeit neben inhaltlichen Belangen ebenfalls leistungsrelevant sein. Geht es um ein Diktat, so zählt ggf. ausschließlich die absolute Fehleranzahl. Grundsätzlich ist jedoch mit Leistung eine Anforderungssituation verbunden, in der *Leistung* gezeigt werden muss.

1.2.1.2 Messung und Beurteilung

Prinzipiell gilt es zwischen den beiden Schritten der *Leistungsmessung* und der *Leistungsbeurteilung* zu unterscheiden (Ingenkamp & Lissmann, 2008; Jäger, 1995; B. Orth, 1995; Rheinberg, 2001). Während die reine *Messung* die Überführung einer Eigenschaft des Messgegenstandes in eine Zahl bedeutet, liegt der *Beurteilung* die Messung zugrunde und das Messergebnis wird im Kontext mit einem Standard/einer Norm verglichen (Ingenkamp, 1995; Ingenkamp & Lissmann, 2008; Jäger, 1995; Schrader & Helmke, 2001): Stellt sich eine Person auf die Waage und erhält als Messergebnis 75kg, so ist der Vorgang des Wiegens zunächst lediglich eine *Messung*. Eine Beurteilung der Masse benötigt jedoch den Vergleich mit einem Standard, etwa einem Durchschnittsgewicht, und einer Interpretation. Im schulischen Kontext werden üblicherweise verschiedene Bezugsnormen (Vergleichsmaßstäbe) angelegt, die jeweils ihre Vor- und Nachteile haben: Während bei der *individuellen Bezugsnorm* vorherige Leistungen des Probanden oder der Probandin selbst als Referenz betrachtet werden, wird bei der *sozialen Bezugsnorm* das individuelle Ergebnis in Relation zu einer sozialen Gruppe (z. B. der Klasse) gesetzt. Die *sachliche Bezugsnorm* legt ein inhaltliches Kriterium an, etwa die Weite von 40 Metern beim Weitwurf zu erreichen oder nicht. Nähere Ausführungen zu Bezugsnormen und deren Nutzung sowie Vor- und Nachteile finden sich u. a. in Rheinberg (1980, 1982, 2001) sowie in diversen Handbüchern (Ingenkamp, 1985; Ingenkamp & Lissmann, 2008; A. Krapp & Weidenmann, 2005; Rost, 2010; E. Wild & Möller, 2009, 2015). In der Schule sollte diese Unterscheidung von Messung und Beurteilung bewusst getroffen werden. Hat ein Schüler beispielsweise von 20 Aufgaben in einem Test 15 korrekt bearbeitet, so ist dies zunächst *nur* ein Messergebnis, welches erst in einem weiteren Schritt beurteilt werden muss.

1.2.2 Gütekriterien von (Leistungs-)Messung

Nachdem eine Messgröße operationalisiert wurde, sollte eine Messung möglichst gut sein. *Gut* wird in diesem Zusammenhang u. a. über drei Hauptgütekriterien definiert und meint konkret eine unabhängige (*objektiv*), zuverlässige (*reliabel*) und korrekte (*valide*) Messung (Bortz, 2005; Bortz & Döring, 2006; Eid, Gollwitzer & Schmitt, 2011; Ingenkamp, 1985; Ingenkamp & Lissmann, 2008; Schnell, Hill & Esser, 2014).

1.2.2.1 Objektivität

Der Anspruch an eine Messung ist, dass diese personenunabhängig zum selben Ergebnis führt. Dabei kann personeller Einfluss an verschiedenen Stellen des Messprozesses wirken; entsprechend untergliedert man die *Objektivität* in:

- **Durchführungsobjektivität:** Die Durchführung einer Messung sollte nicht von der durchführenden Person bzw. der Testleitung abhängig sein. Im Rahmen großer Vergleichsuntersuchungen wird dies etwa durch gezielte Schulungen von Testleitenden und vorgegebenen Regiebüchern für den Testablauf gewährleistet.
- **Auswertungsobjektivität:** Darunter ist zu verstehen, dass beispielsweise eine einzelne Mathematikaufgabe von zwei unabhängigen Korrigierenden einheitlich als richtig bzw. falsch bewertet werden sollte.
- **Interpretationsobjektivität:** Dieser Punkt geht über die reine Messung hinaus, da es um die Interpretation/Beurteilung eines Ergebnisses geht und damit um den an die Messung anschließenden Prozess. Auch hierbei sollten sich keine personenbezogenen Unterschiede ergeben.

1.2.2.2 Reliabilität

Die *Reliabilität* einer Messung gibt deren Zuverlässigkeit an, also etwa ob eine Messung bei wiederholter Durchführung zum identischen Ergebnis kommt. Hierbei steht anders als bei der *Objektivität* nicht das durchführende Personal, sondern das Messinstrument selbst im Fokus. Misst ein Fieberthermometer bei mehrfacher Messung in kurzem zeitlichen Abstand (also ohne dass eine Veränderung der Körpertemperatur erwartbar wäre) die gleiche oder zumindest eine ähnliche Temperatur? Ist dies nicht der Fall, ist intuitiv klar, dass dieses Messinstrument wenig geeignet ist, um zuverlässige Informationen zu generieren.

In der Literatur (vgl. Bortz & Döring, 2006; Bühner, 2011; Moosbrugger & Kelava, 2012a, 2012b) werden als Maße der Reliabilität (1) die Retestreliabilität, (2) die Paralleltestreliabilität, (3) die Testhalbierungsreliabilität und (4) die interne Konsistenz angegeben. Vielfach wird stellvertretend die interne Konsistenz, meist als *Cronbachs Alpha* (Cortina, 1993; Schermelleh-Engel & Werner, 2012), angegeben. Bei der *Testhalbierungsreliabilität* werden lediglich zwei mögliche „Testhälften“ miteinander verglichen, während *Cronbachs Alpha* als Durchschnittswert aller potentiellen „Testhälftenvergleiche“ zu verstehen ist (Baur & Blasius, 2014; Cortina, 1993; Cronbach, 1951; Schmitt, 1996).

Ergänzend wird mit der Interraterreliabilität die Übereinstimmung von mehreren Kodierenden und Kodierern angegeben. Die Interraterreliabilität ist damit jedoch genau genommen eher als Teil der *Objektivität* zu verstehen, da die personenunabhängig identische Kodierung überprüft wird. Im erweiterten Sinne werden so allerdings die *Zuverlässigkeit* und Eindeutigkeit des Messinstrumentes evaluiert. Häufige Maße für die Interraterreliabilität sind die prozentuale Übereinstimmung sowie Cohens K (J. Cohen, 1960; Fleiss & Cohen, 1973; Wirtz & Caspar, 2002).

1.2.2.3 Validität

Mit der *Validität* einer Messung ist deren Gültigkeit angesprochen. Eine *valide* Messung liefert Ergebnisse im untersuchten Bereich. Plakativ heißt das: Es ist zweifelsfrei möglich, mit einem Fieberthermometer *objektiv* und *reliabel* zu messen, nur ist diese Messung nicht *valide*, wenn es darum geht, das Körpergewicht zu erfassen. Dieses Beispiel macht bereits deutlich, dass die Gütekriterien aufeinander aufbauen. Dabei kann kein Instrument oder keine Messung *valide* sein, solange sie nicht *objektiv* und *reliabel* ist. Insbesondere spielt bei der *Validität* eine Rolle, dass bei der Messung lediglich die interessierende Messgröße festgestellt wird und die Messung nicht von anderen Aspekten beeinflusst wird: Beispielsweise bei einer Mathematikaufgabe, die die technische Anwendung eines Verfahrens (etwa Satz des Pythagoras) prüfen soll, jedoch zunächst einen halbseitigen, komplizierten Text zur Einleitung hat, wird die zu messende Größe (die technische Anwendung eines Verfahrens) durch die Lesekompetenz der Schülerin/des Schülers beeinflusst. Damit stellt eine solche Aufgabe kein *valides* Instrument dar, um die Verfahrensanwendung zu testen.

Teil 1 – Theorie

Schulleistungsmessungen

Während die Objektivität und Reliabilität durch recht direkte statistische Analysen einer Prüfung zugänglich sind, fällt dies für die Validität etwas schwieriger aus (Bortz & Döring, 2006). Es werden grundsätzlich drei Arten von Validität unterschieden:

- Inhaltsvalidität
- Kriteriumsvalidität
- Konstruktvalidität

Im oben angeführten Beispiel der Pythagoras-Aufgabe mit komplexem Einführungstext wäre die *Inhaltsvalidität* kritisch zu hinterfragen; *Inhaltsvalidität* beansprucht also die inhaltliche Passung zwischen Testitem(s) und zu messendem Konstrukt. Eine numerische Bestimmung ist hierfür jedoch nicht möglich. Es wird hier vielmehr auf subjektive Einschätzungen von Experten zurückgegriffen.

Eine quantitative Einschätzung ermöglicht hingegen die *Kriteriumsvalidität*. Hierin wird geprüft, inwieweit die Messung einer Eigenschaft durch ein Instrument mit einem externen Kriterium zusammenhängt: Wie hoch hängen beispielsweise die Messung des Berufserfolges einmal gemessen mit einem validierten, erprobten Instrument und zum anderen gemessen mit einem neu entwickelten Instrument zusammen? Je nach der Höhe der Korrelation lässt sich ein Urteil über die *Kriteriumsvalidität* des neu entwickelten Testinstrumentes fällen.

Problematisch bleibt, dass bei der *Inhaltsvalidität* kein objektiver Messwert vorhanden ist und eine Prüfung der *Kriteriumsvalidität* externe Kriterien voraussetzt. Als Alternative zur Prüfung der Validität bietet sich die *Konstruktvalidität* an. Hierbei wird anhand von ableitbaren Hypothesen die Gültigkeit der Messung geprüft. Bortz und Döring (2006, 201f.) geben als Beispiel einen Fragebogen zur Überprüfung der subjektiven Einsamkeit. Theoretisch erwartbar wäre ein Zusammenhang von Einsamkeit mit geringem Selbstwertgefühl und sozialer Ängstlichkeit. Eine Prüfung dieser Hypothesen dient der *Konstruktvalidierung*. Einzuschränkend bleibt zu diesem Weg anzumerken, dass eine *Konstruktvalidierung* selbstverständlich nur dann erfolgsversprechend ist, wenn die anderen Messungen (im Beispiel Selbstwertgefühl und soziale Ängstlichkeit) mit etablierten, bereits validierten Instrumenten erfolgt sind. Zu unterscheiden ist beim Vorgehen der *Konstruktvalidierung*, dass dies auf *diskriminante* oder *konvergente* Art geschehen kann. Es lassen sich also theoriegestützte Hypothesen

zu positiven Zusammenhängen (*konvergente Validität*) und zu negativen Zusammenhängen (*diskriminante Validität*) aufstellen und prüfen (Baur & Blasius, 2014).

1.2.3 Grenzen der Leistungsmessung / Voraussetzungen zur guten Nutzung

Trotz aller wissenschaftlich fundierten Gütekriterien für Leistungsmessung, ergeben sich im schulischen und alltäglichen Kontext zwangsläufig Grenzen: Ausschnitthaft und exemplarisch soll in gebotener Kürze mit entsprechenden Verweisen näher auf die folgenden kritisch zu betrachtenden Punkte eingegangen werden:

- Testbarkeit von Inhalten, Kompetenzen etc.
- Genauigkeit von Messungen
- Interpretation von Messergebnissen
- Nützlichkeit von reinen „Leistungs“-Messungen / Fokus von Schule

1.2.3.1 Testbarkeit von Inhalten/Kompetenzen etc.

Wie bereits oben angeführt, sollte einer Messung eine genaue Operationalisierung des Messgegenstandes vorangehen. Bezogen auf eine Klassenarbeit im schulischen Kontext zum Thema „Satz des Pythagoras“ heißt das konkret, dass sich die betreffende Lehrkraft u. a. mit den Fragen auseinandersetzen muss: Welche Inhalte möchte ich prüfen? Ist auch die Umkehrung des Satzes prüfungsrelevant? Prüfe ich lediglich die technische Anwendung oder interessiert mich auch die Modellierung? Analog bedeutet die vorangehende Operationalisierung für vergleichende Leistungsmessung, dass die zu messende Fähigkeit von Schülerinnen und Schülern genau umrissen werden muss. Beispielsweise hat der IQB-Ländervergleich den Anspruch, die mathematische Grundbildung in der neunten Jahrgangsstufe zu messen, wobei kein konkreter Unterrichtsinhalt, sondern die bildungsstandardskonformen Kompetenzen seit der fünften Klasse zur Disposition stehen (Pant et al., 2013; Stanat, Pant, Böhme & Richter, 2012). Prinzipiell gibt es nach Baumert (2001b, S. 24) den Basiseinwand, dass zentral entwickelte und administrierte Leistungsmessungen im Widerspruch zu einem ganzheitlichen Bildungsprozess stünden, da die getesteten/gemessenen Inhalte notwendigerweise ausschnitthaft bleiben müssen. Darüber hinaus wird angeführt, dass aufgrund von fehlenden Operationalisierungen von Bildung (vgl. Abschnitt 1.1.1.4) die Gefahr bestünde, dass *Bildung* mit Testerfolg gleichgesetzt würde. Dies könnte langfristig zur Einschränkung des Curriculums auf die Testinhalte (*narrowing-*

the-curriculum) führen. Außerdem fehle bei dieser Sichtweise eine Berücksichtigung „proximaler und veränderbarer Größen im Unterricht“ (Ditton, 2000, S. 73).

Konkreter auf Ebene der einzelnen Aufgaben liegen die Grenzen von Leistungsmessung jedoch u. a. darin, dass bestimmte Kompetenzen bzw. Fähigkeiten und Fertigkeiten durch schriftliche Tests nicht (oder nur unzureichend) abgebildet werden können. Insbesondere bedeutet dies bei der recht knappen zur Verfügung stehenden Test-/Messzeit, dass durch Aufgaben selektiv messbare Aspekte fokussiert sowie notwendigerweise ausschnitthaft Fachinhalte getestet werden (Baumert, 2001b). Betrachtet man exemplarisch hier zwei Modellierungsaufgaben aus dem Aufgabenpool der Vergleichsarbeiten und der Aufgabensammlung des die Bildungsstandards illustrierenden Buchs von Blum et al. (2012), so fällt auf, dass der Anspruch der verlangten Modellierungskompetenz in der „Nicht-Test-Aufgabe“ um ein Vielfaches höher liegt. Eine solche Unterscheidung zwischen „Test-Aufgaben“ und „Lern-Aufgaben“ fordern auch Büchter und Leuders (2005a, S. 326).

Um den Anspruch dieser Aufgaben besser einschätzen zu können, wird der Modellierungskreislauf nach Leiß (2007) herangezogen (vgl. auch Leiß & Blum, 2012; Blum & Leiß, 2006, 2007a, 2007b; Blum, 2015; Kaiser & Stender, 2015; Leuders & Leiß, 2012).

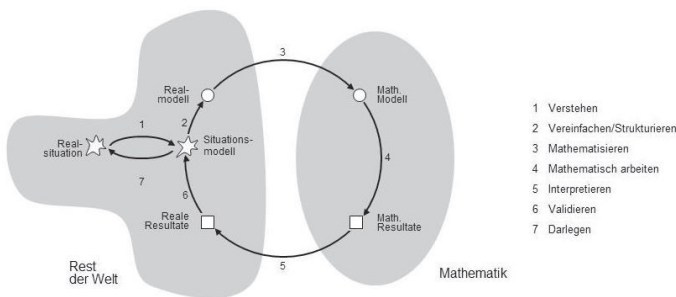


Abbildung 3: Modellierungskreislauf nach Blum und Leiß (2005, S. 19) siehe in Leiß (2007, S. 31)

Exemplarisch sollen hier die beiden Aufgaben „Trainingsanalyse“ aus Blum et al. (2012) und „Abstand auf dem Wasser“ aus den Vergleichsarbeiten (VerA) verglichen werden.

Tabelle 3: Aufgabenkennwerte „Trainingsanalyse“ und „Abstand auf dem Wasser“

	Trainingsanalyse	Abstand auf dem Wasser ⁹
Leitidee	4. Funkt. Zusammenhang	4. Funkt. Zusammenhang
Kompetenzen ¹⁰	K2, K3, K4	K2, K3, K4
Anforderungsbereich	I-III	II
Kompetenzstufe	--	4

In beiden Fällen muss die Aufgabe selbstverständlich zunächst verstanden werden. Bei der Aufgabe „Trainingsanalyse“ (Abbildung 4) muss in den ersten drei Teilaufgaben im Schritt zwei des Modellierungskreislaufs die komplexe grafische Darstellung auf die relevanten Teile reduziert werden. Dies erfordert bereits in beträchtlichem Maße die Kompetenz „Mathematische Darstellung verwenden“. Die Übersetzung in die Mathematik ist für die ersten drei Teilaufgaben im Grunde bereits geschehen, da der Graph vorliegt. Innermathematisch (Schritt 4 im Modellierungskreislauf) müssen nun die korrekten Abschnitte der Graphen erkannt und abgelesen werden: Für Teilaufgabe a) entsprechend auf der x-Achse die Strecke ablesen, die Anzahl der lokalen Extremstellen zwischen etwa 53 und 61 km abzählen (Teilaufgabe b) und die Nullstellen des Geschwindigkeits-Plots identifizieren (Teilaufgabe c). Diese Informationen müssen in Schritt fünf aus der mathematischen Welt zurück in die Aufgabenrealität übersetzt werden. Entsprechend werden die Anzahl der lokalen Extremstellen in die Anzahl der Serpentinien übersetzt und die Anzahl der Nullstellen des Geschwindigkeits-Plots in Pausen. Die Schritte 6 (Validieren) und 7 (Darlegen) im Modellierungskreislauf fallen für diese Teilaufgaben eher knapp aus. Anders sieht es für die letzten beiden Teilaufgaben der Aufgabe aus:

In einem ersten Schritt muss nach dem Verständnis der Aufgabe für Teilaufgabe d) vereinfacht werden und der Fokus auf den Geschwindigkeits-Plot gelegt werden. Zusätzlich muss ein Plan erdacht werden, mit dem eine Durchschnittsgeschwindigkeit gefunden wird. Dies könnte beispielsweise dadurch geschehen, dass besonders schnelle Passagen langsamere Abschnitte ausgleichen. Dieses Vorhaben wird nun in die Mathematik übersetzt, in der es darum geht, mathematisch eine Ausgleichsgerade zu finden. Dies ist für Schülerinnen und Schüler der achten Klasse ein tatsächliches Problem ohne „Rezept“, da Mittel der Integralrechnung noch nicht zur

⁹ Zugriff unter: <https://www.iqb.hu-berlin.de/vera/aufgaben/ma1/system/taskpool/getTaskFile?fileid=5537> [07.02.2017]

¹⁰ Die Abkürzungen orientieren sich an den gängigen Kurzfassungen: „K2“ steht für „Probleme mathematisch lösen“, „K3“ steht für „Mathematisch modellieren“, „K4“ steht für „Mathematische Darstellung verwenden“.

Teil 1 – Theorie

Schulleistungsmessungen

Verfügung stehen. Möglich ist die Suche nach einer Ausgleichsgeraden beispielsweise durch das Interpolieren des Graphs durch einen Polygonzug. Verbunden mit einer Parallelen zur x-Achse entstehen so Trapeze. Diese Parallele wird nun explorativ so gewählt, dass die Fläche der Trapeze „oberhalb“ und „unterhalb“ der Ausgleichs-„Parallelen“ nahezu gleich sind. Nachdem so eine Ausgleichsgerade gefunden wurde, die etwa bei 30 bis 35 liegen sollte, werden die Ergebnisse in die Realität zurückübersetzt. Hier wird nun zunächst die entsprechende Einheit, nämlich Kilometer pro Stunde, ergänzt und im nächsten Schritt das Ergebnis validiert. Hierzu sollten sowohl Informationen der Aufgabe selbst, also beispielsweise die maximale und minimale Geschwindigkeit, aber auch aufgabenexterne Informationen, wie etwa das eigene Fahrradfahrt tempo, Wissen über Geschwindigkeiten bei Radrennen etc. herangezogen werden. Daraus sollte zu dem Schluss gekommen werden, dass diese Durchschnittsgeschwindigkeit (30-35 km/h) im Bereich des Erwartbaren liegt und das Ergebnis kann dargelegt werden.

In der letzten Teilaufgabe muss der Modellierungskreis ebenfalls durchlaufen werden, wenngleich keine Aufgabe direkt gestellt ist. Für die Konstruktion von Aufgaben ist es dabei jedoch genauso erforderlich, die notwendigen Informationen herauszufinden und zu vereinfachen, worauf eine Aufgabe abzielen soll. Hieraus muss sich ein mathematisches Verfahren ableiten lassen, welches abschließend nochmals validiert wird, ob die erdachte Aufgabe sinnvoll und bearbeitbar ist.

Dabei stellt insbesondere die Teilaufgabe e) aus „Trainingsanalyse“ ein sehr offenes Aufgabenformat dar, welches in dieser Form in einem Test nur schwer *objektiv* auswertbar wäre. Entsprechend fällt die Analyse der zu vergleichenden Testaufgabe aus VerA deutlich kürzer aus: Ein Verständnis der Aufgabenstellung und der Gesamtsituation ist vorausgesetzt und die Vereinfachung kann vernachlässigt werden, da keine zusätzlichen oder ablenkenden Informationen präsentiert werden. Die Übersetzung in die Mathematik ist durch die Graphen im Multiple-Choice-Format ebenfalls bereits erfolgt, sodass die Schülerinnen und Schüler lediglich innermathematisch erkennen müssen, welcher Graph den Abstand beschreibt. Da der Abstand offensichtlich zu einem Zeitpunkt minimal ist und zuvor wie auch danach größer war/sein wird, fallen beide Graphen mit Verlauf durch den Ursprung heraus. Der zeitliche Horizont (markiert durch die Punkte auf der Bewegungsrichtung in der Aufgabenstellung) zeigt an, dass der minimale Abstand nach etwa 15 Minuten erreicht ist. Diese Information

führt zur Lösung unten links. Vermutlich läuft die tatsächliche Lösung von Schülerinnen und Schülern in diesem Falle eher als ein mehrfaches Durchlaufen des Modellierungskreislaufes ab. In einem ersten Durchlauf wird die Tatsache des sinkenden und später wieder wachsenden Abstandes zum Turm mathematisiert und als Schluss nach erfolgter Übersetzung in die Realität die Graphen oben links und unten rechts ausgeschlossen. Zur Entscheidung zwischen den verbleibenden Graphen wird in einem erneuten Durchlauf der Zeitpunkt des geringsten Abstandes als lokales Minimum des Graphen mathematisiert und damit nach Rückübersetzung in die Realität Graph 3 (unten links) als die korrekte Lösung identifiziert.

Aus diesen Ausführungen wird deutlich, dass Testaufgaben noch stärker als Aufgaben allgemein die Aufmerksamkeit von Schülerinnen und Schülern lenken. Dabei werden insbesondere in Testaufgaben eher geschlossene Formate (u. a. Multiple-Choice) gewählt, da dies mit Blick auf die Gütekriterien einen lohnenswerten Kompromiss darstellt.

Trainingsanalyse

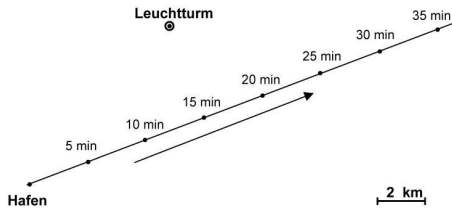
Das unten stehende Diagramm zeigt einen Ausschnitt aus einer Trainingsaufzeichnung eines Radrennfahrers.



- Wie lang ist die Abfahrt vom ersten Berg?
- Wie viele Serpentinaen (enge Kurven) kamen auf der Abfahrt vom ersten Berg vor?
- Wie oft hat der Radrennfahrer angehalten?
- Wie groß war die ungefähre Durchschnittsgeschwindigkeit des Radrennfahrers?
- Überlege dir noch eine weitere interessante Aufgabe, die man anhand dieses Graphen beantworten kann, und löse sie.

Abstand auf dem Wasser

Ein Schiff fährt mit gleichbleibender Geschwindigkeit aus einem Hafen an einem Leuchtturm vorbei. Der nachstehenden Abbildung kann man entnehmen, wie viele Minuten es vom Ablegen bis zu den markierten Positionen braucht.



Welcher Graph stellt den Abstand des Schiffes zum Leuchtturm als Funktion der Zeit dar?

Kreuze an.

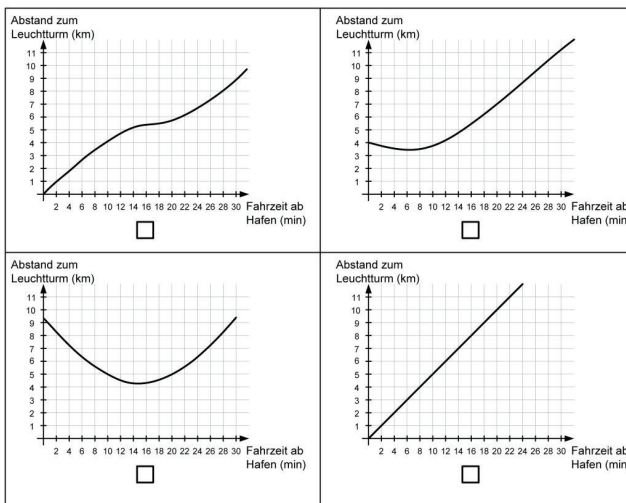


Abbildung 5: Aufgabe „Abstand auf dem Wasser“

1.2.3.2 Genauigkeit von Messungen / Gütekriterien / Verlässlichkeit auf unterschiedlichen Ebenen

Der erste Punkt von mangelnder Messbarkeit bestimmter Kompetenzen geht fließend in diesen zweiten Punkt über: Mit der Genauigkeit von Messungen wird ein zentrales Problem angesprochen. Jede Messung sollte nach Möglichkeit *reliabel* sein, doch gilt es die Reliabilität auf verschiedenen Ebenen zu beurteilen: Nimmt man eine einzelne Aufgabe aus einer zentralen Leistungsmessung, so wird in der Entwicklung dieser Aufgabe (und des zugehörigen Kodiermanuals mit Blick auf die *Objektivität*)

darauf geachtet, dass die mit dieser Aufgabe durchgeführte Messung reliabel ist. Allein diese Aufgabe kann jedoch noch weder ein *zuverlässiges* noch *valides* Bild von mathematischer Leistung oder mathematischer Grundbildung darstellen. Dafür benötigt es eine größere Stichprobe von Aufgaben: Die Tests bzw. die Testentwickelnden stehen damit vor dem Dilemma, ein breites Bild des Curriculums in einer sehr begrenzten Zeit zu überprüfen (Arnold, 1999; Nachtigall & Kröhne, 2006). Verlässliche Aussagen über die mathematische (Grund-)Bildung einer Schülerin oder eines Schülers sind also nur für einen gesamten Test hinreichend *valid*. Umgedreht bedeutet das, dass „je höher die Aggregationsebene [ist, S.V.], desto präziser die Messung. Oder anders herum ausgedrückt: Die Leistungsbestimmung ist auf der Ebene einzelner Schülerinnen und Schüler deutlich am ungenauesten“ (I. Hosenfeld et al., 2006, S. 21).

Weitere Brisanz bringt der Aspekt der Grundgesamtheit und der geprüften Stichprobe in diese Thematik: In Leistungsmessungen, die wie PISA oder der IQB-Ländervergleich beispielsweise dem *Systemmonitoring* dienen, wird keine Vollerhebung durchgeführt, sondern lediglich einzelne Schülerinnen und Schüler aus einzelnen Klassen getestet, „sodass sie keine generalisierten Aussagen über die gesamte Schule erlauben“ (Baumert, 2001b, S. 31; Kohler & Schrader, 2004).

1.2.3.3 Interpretation der / Fehlerhafte Schlüsse aus Ergebnissen sowie Nützlichkeit von Rückmeldungen ohne konkreten Handlungsvorschlag

Die zuvor angesprochenen Grenzen der Leistungsmessung beziehen sich vorwiegend auf die technischen Grenzen von Leistungsmessung. Mit Blick auf eine fehlerhafte Interpretation von Leistungsmessungen wird der Blick eher auf die durchführenden Personen gerichtet. Weiterhin unterscheiden sich die beiden folgenden Punkte von den zuerst genannten darin, dass die Rückmeldung die Ergebnisse der Leistungsmessung in den Fokus rückt und deren fehlerhafte Nutzung bzw. deren Nützlichkeit thematisiert wird. Daher soll an dieser Stelle auf die entsprechenden Teile im Kapitel zu *Feedback und Rückmeldung* verwiesen sein (vgl. Abschnitt 1.3). Prinzipiell gilt jedoch das bekannte Sprichwort: „Vom Wiegen wird die Sau nicht fett!“; also übertragen auf die Schule: „Vom Testen allein lernen Schülerinnen und Schüler nicht mehr!“.

1.2.3.4 Unklare Zieldefinition von Schule

Für Unternehmen ist die Zielgröße vergleichsweise einfach zu bestimmen: Ziel ist es, bei geringstmöglichem Einsatz ein qualitativ und quantitativ bestmögliches Produkt

zu schaffen; konkret bedeutet das also, bei gleichem Personal- und Kosteneinsatz beispielsweise die Arbeitsabläufe so umzustrukturieren, dass die Anzahl der produzierten Autos bei gleichbleibender Qualität gesteigert werden kann. Im System Schule ist jedoch die Zielgröße weniger klar zu greifen: In der Schule geht es u. a. darum, in der zur Verfügung stehenden Unterrichtszeit die bestmögliche Leistung der Schülerinnen und Schülern zu erreichen. Dazu gehört jedoch, ein tiefgreifendes Verständnis der Inhalte und, dass neben reinem Faktenwissen auch Routinen eingeübt werden. Neben diesen fachlichen, leistungsrelevanten Aspekten hat Schule jedoch zweifelsfrei den Auftrag, soziale Kompetenzen bei Schülerinnen und Schülern zu fördern und zu fordern (Baumert, 2001b), welche sich jedoch nur bedingt durch schriftliche Tests messen lassen.

Oelkers und Reusser (2008, S. 17) beschreiben sehr treffend: „Vor dem Hintergrund wird sich der Aufwand langfristig dann lohnen, wenn es gelingt, die Ergebnisse von Tests und Schulevaluation nicht nur als neue Form der Überprüfung, etwa durch die Benennung von Stärken und Schwächen, sondern auch zur Ingangsetzung empiriegestützter Prozesse der Systementwicklung und Weiterbildung zu nutzen.“ Auf der Grundlage gemeinsamer Standards werden mithilfe von Leistungsmessungen Informationen generiert, die in Bezug gesetzt Entwicklungsmöglichkeiten offenbaren. Wesentlicher Bestandteil von Leistungsmessung sollte die Rückmeldung der Ergebnisse an die beteiligten Personen sein, sodass diese nach der Rezeption und Reflexion der Ergebnisse Entwicklungsmöglichkeiten erkennen und nutzen können. Bevor jedoch der Blick auf die Evaluation als Ganzes gehen darf, soll im kommenden Kapitel zunächst der Prozess der Rückmeldung näher betrachtet werden, um anschließend in Kapitel 1.4 den Bogen mit *Evaluation* umfassend zu spannen.

1.3 Feedback/Rückmeldung

Dabei wird zunehmend klar: Die bloße Bereitstellung von Informationen im Rahmen von Ergebnissrückmeldungen allein führt noch nicht zu den gewünschten Veränderungen. (Kohler & Schrader, 2004, S. 10)

1.3.1 Definition/Einführung

Als wesentlicher Punkt wurde bisher immer wieder betont, dass nach der Einführung der Bildungsstandards – als *pars pro toto* der Veränderungen (Uhl, 2006, S. 3) – das Bildungswesen anhand des an diesen Standards orientierten Outputs „gesteuert“ werden soll. Der Output wird beispielsweise durch Leistungsmessung bestimmt. Diese Messergebnisse müssen jedoch in einem nächsten Schritt nutzbar gemacht und dazu vor allem rückgemeldet werden. Diese als *Feedback* zurückgemeldeten Informationen sind damit ein essentieller Teil der neuen Steuerungsphilosophie. Zunächst handelt es sich bei Rückmeldungen jedoch oft um eine rein analytisch-deskriptive Aussage zu einem Messergebnis (oder zu Beobachtungen), auf Basis derer Hypothesen zu möglichen Ursachen entwickelt werden können (van Ackeren & Klemm, 2011).

Feedback wird definiert als „...actions taken by (an) external agent(s) to provide information regarding some aspect(s) of one's task performance“ (Kluger & DeNisi, 1996, S. 255). *Feedback* kommt insofern eine zentrale Rolle zu, da es die Aufmerksamkeit von Adressaten lenkt (etwa auf die eigene Leistung und nicht auf den Vergleich zu anderen Personen; Ditton & Arnoldt, 2004; Visscher & Coe, 2003). Im Kontext schulischer Leistungsmessung ist *Feedback* sogar in zweifacher Hinsicht von Bedeutung (vgl. auch Sleeter, 2014): Zum einen spielt *Feedback* für die Lehrkräfte, Schulen, Schulämter, Kultusministerien etc. im Sinne der Steuerung des Bildungssystems eine wesentliche Rolle. Zum anderen spielt *Feedback* eine zentrale Rolle in Unterrichtsprozessen: zum einen also *Feedback* an die „Anbietenden“¹¹ von Unterricht, zum anderen *Feedback* als Rückmeldung (von Lehrkräften) an Schülerinnen und Schüler als „Teilnehmende“ an Unterricht. Während Letzteres mittlerweile vielfach untersucht, beinahe unbestritten wirksam ist und vielfach eingesetzt wird (vgl. $d = 0.73$ in Hattie, 2009, S. 173; Hattie & Timperley, 2007; Hattie, Beywl & Zierer,

¹¹ Im weitesten Sinne betrachtet von den Verantwortlichen für behördliche Vorgabe, personelle, materielle und finanzielle Ressourcenverteilung etc. bis hin zur konkreten Unterrichtsplanung/-durchführung durch die einzelne Lehrkraft.

2013; Kluger & DeNisi, 1996), „erhalten [in Deutschland, S.V.] 15 Prozent der Lehrerinnen und Lehrer laut TALIS GEW nie Feedback über ihre Arbeit an der Schule“ (Barfknecht & Saldern, 2010, S. 97). Rückmeldung ist jedoch äußerst wirksam für den Lernerfolg von Schülerinnen und Schülern: Ergänzend bestätigt Hattie (2009, S. 173) die Relevanz von Feedback mit der großen Effektstärke von $d = 0.90$ für die formativen Evaluation von Unterricht (vgl. auch Hattie et al., 2013).

Zunächst soll jedoch ein allgemeines theoretisches Grundgerüst zum Thema *Feedback* aufgebaut werden, bevor die Nutzung von Feedback im Unterricht bzw. zur Schul-/Unterrichtsentwicklung näher beleuchtet wird. Wie bereits aus der Definition von Kluger und DeNisi (1996) deutlich wird, handelt es sich bei Feedback zunächst nur um reine Information (Ditton & Arnoldt, 2004), die dazu beitragen sollte, dass die/der Adressat/in die gestellte Anforderung besser absolvieren kann (im schulischen Kontext also Schülerinnen und Schüler eine höhere Lernleistung erbringen können; Visscher & Coe, 2003). Die im Feedback transportierte Information sollte nach Hattie und Timperley (2007, S. 86) „three major questions“ beantworten können:

1. Where am I going? (What are the goals?) – *Feed Up*
2. How am I going? (What progress is being made toward the goal?) – *Feed back*
3. Where to next? (What activities need to be undertaken to make better progress?) – *Feed forward*

Feedback, welches Informationen zu diesen drei Fragen bietet, sollte in Analogie von Ist-Soll-Regelkreisen eine Reduktion der Diskrepanz zwischen aktuellem Ist-Stand und anvisiertem Soll-Zustand ermöglichen. Neben diesen Fragen, die den Informationsgehalt des Feedbacks betreffen, nennen Hattie und Timperley (2007) in ihrem Modell des Weiteren vier Ebenen, auf denen Feedback wirken kann (Abbildung 6).

Teil 1 – Theorie

Feedback/Rückmeldung

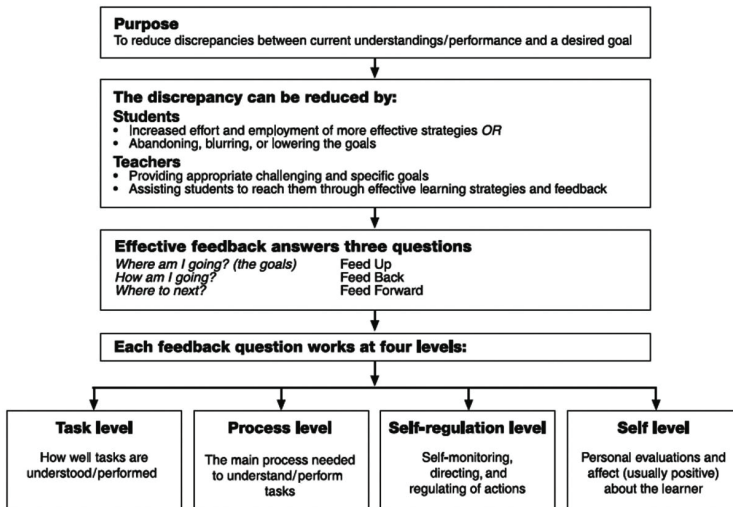


Abbildung 6: Feedback-Modell (Hattie & Timperley, 2007, S. 87)

Dieses Modell ist für Feedback von Lehrkräften an Schülerinnen und Schüler gedacht, lässt sich allerdings in ähnlicher Form auch auf Feedback übertragen, welches Lehrkräfte beispielsweise aus vergleichender Schulleistungsmessung, Abschlussprüfungen oder auch Schulinspektionen erhalten: Schnell wird so ersichtlich (S. Krapp, 2006), dass Feedback aus solchen Verfahren mehr Informationen beinhalten muss als ausschließlich deskriptive Leistungswerte.¹² Feedback zielt darauf ab, die Diskrepanz zwischen einem *gemessenen* Ist-Zustand und einem anvisierten Soll-Zustand (Standard) zu verringern. Diese Diskrepanzreduzierung kann vonseiten der Lernenden durch erhöhten Einsatz oder durch Resignation bzw. Abkehr vom anvisierten Ziel-Zustand erreicht werden; Lehrkräfte haben zumindest beim ersten Lösungsansatz der Schülerinnen und Schüler die Möglichkeit zu unterstützen (etwa durch Feedback-Informationen).

¹² Um näher an der Originalquelle zu bleiben und zugunsten der einfacheren Lesbarkeit, wird sich darauf beschränkt, die Ebenen von Feedback anhand des von Lehrkräften an Schülerinnen und Schüler gegebenen Feedbacks zu verdeutlichen.

Angemerkt werden sollte an dieser Stelle jedoch auch, dass der „Lerngegenstand“ von Lehrkräften bei der Entwicklung von Unterricht ungleich komplexer ist als beim Feedback an Schülerinnen und Schüler. Daher ist das Modell nur bedingt übertragbar bzw. zielt nur auf kleine Anpassungen im Unterricht durch die Lehrkräfte.

Feedback adressiert mindestens eines der vier Level/eine der vier Ebenen. Hattie und Timperley (2007) nehmen mit den genannten Ebenen Bezug auf eine für Feedback zentrale Meta-Analyse von Kluger und DeNisi (1996). Letztere unterscheiden drei Ebenen von Feedback: (1) task-learning-process, (2) task-motivation-process und (3) meta-task-process. Dabei beschreiben „task-learning-process“ und „task level“ ähnliche Bereiche. Solches Feedback dreht sich um Informationen, die mit dem Verständnis des Ziels/der Aufgabe zusammenhängen. Wird beispielsweise eine Aufgabe bearbeitet, so wäre etwa ein Feedback auf „task level“ eine Rückmeldung zu Aufgabendetails, die bisher nicht verstanden oder nicht berücksichtigt wurden. Geht es um die Aufgabebearbeitung, könnte ein Feedback auf „process-level“ bzw. „task-motivation-process“ beispielsweise betonen, dass die Reihenfolge der Schritte des Modellierungskreislaufes von Schülerinnen und Schülern beachtet werden sollte.

Im Sinne der Selbstregulation (A. Krapp & Weidenmann, 2005; K.-P. Wild & Schiefele, 1994; E. Wild & Möller, 2015) fokussiert ein Feedback auf „self-regulation“-Ebene interne Kontroll- und Überwachungsstrategien. Analog zu durch Feedback bereitgestellten Informationen liefern interne Monitoringprozesse Informationen über den Grad der Zielerreichung und beeinflussen so die Aufgabebearbeitung oder den Prozess der Zielerreichung in Planung, Durchführung oder Reflexion (vgl. Prozessmodell der Selbstregulation u. a. in Landmann, Perels, Otto & Schmitz, 2009).

Die vierte Ebene von Feedback nach Hattie und Timperley (2007) ist die Ebene „self“. Hierbei handelt es sich um Feedback zur Person selbst: Ein Teil von Feedback dieser Ebene ist Lob. Zu unterscheiden ist hierbei von „aufmerksamkeitsablenkendem“ Feedback/Lob zur Person und „aufmerksamkeitssteuerndem“ Feedback zu persönlichem Aufwand oder Selbstregulation (vgl. Ditton & Arnoldt, 2004, S. 117). Während einzelne Studien einen geringen positiven Einfluss von Lob auf die Leistung finden (vgl. $d = 0.12$ in Wilkinson, 1981 nach Hattie & Timperley, 2007), zeigen Brummelman, Thomaes, Orobio de Castro, Overbeek und Bushman (2014), dass Lob abhängig vom Selbstwertgefühl differenzielle Wirkungen haben kann. Die Autoren konnten nachweisen, dass Kinder mit geringem Selbstwertgefühl nach überschwänglichem Lob weniger Herausforderungen suchten einerseits als Kinder mit hohem Selbstwertgefühl bei ebenfalls überschwänglichem Lob und andererseits insbesondere als Kinder mit gleichsam geringem Selbstwertgefühl bei nicht überschwänglichem Lob. Vor

diesem Hintergrund muss der Einsatz von Lob differenziert und abhängig von der zu lobenden Person betrachtet werden.

Kluger und DeNisi (1996) unterscheiden im Unterschied zu Hattie und Timperley (2007) nur drei Prozesse, sodass unter „meta-task-process“ die beiden letztgenannten Ebenen nach Hattie und Timperley zu fassen sind. Deutlich wird durch die Betrachtung der Ebenen und der drei Fragen, dass Feedback vor allem mit der individuellen Bezugsnorm verbunden werden sollte, da so die Aufmerksamkeit auf die eigene Person und damit auf selbstregulatorisch veränderbare Aspekte gelenkt wird (Visscher & Coe, 2003).

1.3.2 Einflüsse auf die Nutzung von Feedback

Unter Feedback wird also eine informationshaltige Rückmeldung verstanden, die sich auf verschiedene Ebenen beziehen kann. Diese theoretischen Annahmen über die Klassifikation und den Informationsgehalt von Feedback werden durch empirische Modelle ergänzt. Dabei sind neben dem reinen Informationsgehalt und der Ebene von Feedback weitere Faktoren bzw. Aspekte zu beachten, wie u. a. Schildkamp und Kuiper (2010) oder Visscher und Coe (2003) zeigen. Bevor es darum gehen kann, wie die konkrete Nutzung von Feedback aussehen könnte, gilt es Voraussetzungen (a) für die Gestaltung des Feedbacks, (b) in den persönlichen Ressourcen und der Persönlichkeit des Feedback-Erhaltenden und (c) des Kontextes, in den das Feedback eingebettet ist, zu beachten. Diese drei Bereiche sollen im Weiteren zunächst erläutert werden, bevor die Nutzung von Feedback in den Fokus rückt.

1.3.2.1 Gestaltung von Feedback

In der Gestaltung von Feedback lassen sich neben den theoretisch bereits aufgeführten drei Fragen von Feedback v. a. folgende Punkte herausstellen:

1. Zeitpunkt von Feedback
2. Auflösung und Zielgruppe des Feedbacks
3. Inhalt des Feedbacks
4. Verständlichkeit des Feedbacks

1.3.2.1.1 Zeitpunkt von Feedback

In der Feedback-Forschung wurde unter dem *delay retention effect* einige Zeit propagiert, dass verspätetes Feedback einen besonders lernwirksamen Effekt habe (Brackbill, Bravos & Starr, 1962; Brackbill & Lienz, 1967). Der vermeintlich positive

Effekt wurde von Kulhavy und Anderson (1972) widerlegt und u. a. als Einfluss der längeren Verarbeitung von Feedback identifiziert (vgl. auch Kulhavy, 1977; Kulik & Kulik, 1988). In der jüngeren Feedback-Forschung wird entsprechend die positive Wirkung von unmittelbarem Feedback betont (vgl. u. a. Narciss, 2006; zsf. mit kleinem Effekt von $d = 0.16$; Hattie, 2009, S. 183; Hattie et al., 2013).

1.3.2.1.2 Auflösungsgrad und Zielgruppe von Feedback

Bei der Bewertung von Feedback gilt es immer im Blick zu behalten, welchen Auflösungsgrad das Feedback hat und an wen sich Feedback richtet. Um ausreichend Informationen liefern zu können, muss die zugehörige Messung zunächst den Gütekriterien genügen und entsprechend detaillierte Informationen auf dem geforderten Auflösungsgrad liefern (Popham, 2003): Betrachtet man dies exemplarisch vor dem Hintergrund internationaler Vergleichsstudien, so stellen viele Autoren fest, dass stichprobenbasierte Untersuchungen – also insbesondere Untersuchungen, die an einzelnen Schulen lediglich einen Teil der Schülerinnen und Schüler befragen – „keine generalisierten Aussagen über die gesamte Schule erlauben“ (Baumert, 2001b, S. 31; Kohler & Schrader, 2004; Leuders, 2006; Leutner, Fleischer, Spoden & Wirth, 2008). Mit diesen stichprobenbasierten Untersuchungen werden damit vorrangig Informationen (frühestens) auf der nächst höheren Aggregationsebene erzeugt (Kohler & Schrader, 2004): Konkret bedeutet dies, dass standardisierte Rückmeldungen internationaler Vergleichsstudien keine individuellen und auch keine einzelschulbezogenen Schlüsse zulassen.

1.3.2.1.3 Inhalt

Die in Feedback enthaltenen Informationen sollten die drei oben genannten Fragen beantworten können und für Lehrkräfte vor allem unterrichtsrelevante Aspekte fokussieren (Ditton, 2000; Popham, 2003; Schildkamp & Kuiper, 2010; Visscher, 2008). Mit Blick auf eine vorangegangene Messung von Kompetenz oder Leistung, gilt es zu berücksichtigen, dass die alleinige Messung/Beschreibung von Kompetenz oder Leistung bzw. deren Rückmeldung nicht unmittelbar zu einer Verbesserung/Veränderung der Unterrichtspraxis führt (Fitz-Gibbon, 1996; Netzwerk Empiriegestützte Schulentwicklung [EMSE], 2008). Dies ist mit dem Sprichwort „Vom Wiegen wird die Sau nicht fett!“ zu vergleichen. Ebenfalls kritisch zu bemerken ist, dass insbesondere internationale Vergleichsuntersuchungen lediglich Beschreibungswissen liefern, ohne jedoch direkt ableitbares Handlungswissen vorzuschlagen (oder vorschlagen zu können;

Maritzen, 2014; Rolff, 2001, 2002). Aus wissenschaftlicher Sicht wird hierbei wiederholt gefordert, die Messung als Grundlage von Feedback um weitere Aspekte, wie beispielsweise persönliche oder soziale Hintergrundvariablen, zu erweitern. Allein durch eine analytische Berücksichtigung von Drittvariablen ergeben sich zwar weiterhin keine kausalen Zusammenhänge, aber differenziertere Analysen vervollständigen das Bild von Leistung bzw. Kompetenz (Terhart, 2002). Über diese Ergänzung hinaus ist es notwendig, wenn wirklich kausale Schlüsse gezogen werden sollen, dass durch längsschnittliche Designs die geforderten psychometrisch-methodischen Anforderungen erfüllt werden (Bortz & Döring, 2006; Rolff, 2002).

Eine beliebte Art der Rückmeldung sind Rankings, also eine Rangfolge basierend auf bestimmten ausgewerteten Merkmalen wie etwa der Leistung. Allerdings sind diese Rankings sowohl aus methodischer Sicht und im Besonderen für Schulen aus pädagogischer Sicht kritisch zu prüfen (Rolff, 2001). Methodisch zweifelhaft sind Rankings, da sichergestellt sein muss, dass Ausprägungen des identischen Merkmals miteinander in Bezug gesetzt werden. Betrachtet man hierzu beispielsweise die Leistungsrankings in PISA, so wird im Rahmen der Entwicklung sehr genau darauf geachtet, dass es keine länder- und sprachspezifischen Einflüsse, sog. *differential-item-functioning*, gibt. Trotz sehr genauer Kontrollen in diesem Bereich lässt sich nicht ausschließen, dass die Pilotierungsstichprobe von der Erhebungsstichprobe abweicht und so dennoch systematische Unterschiede auftreten (vgl. u. a. Krawitz, Achmetli, Blum, Vogel & Besser, 2016; OECD, 2014b). Einen pädagogisch zweifelhaften Stand haben Rankings zusätzlich daher, weil sie ausschließlich Informationen über einen Ist-Stand transportieren, ohne dabei weitere Informationen zu liefern. Legt man die drei Fragen von Hattie und Timperley (2007) an, so beantworten Rankings lediglich die Frage „How am I going? (*Feed back*)“. Informationen über mögliche Ziele („Where am I going?“ – *Feed up*) sind (wenn überhaupt) nur implizit sozial Bezug nehmend am Ranking anführenden enthalten. Die Frage, wie als nächstes vorgegangen werden soll („Where to next?“ – *Feed forward*), lassen Rankings jedoch gänzlich unbeantwortet und bieten entsprechend wenig pädagogisch wertvolle Informationen (Rolff, 2001).

Absolute Leistungskennwerte an sich liefern dabei ähnlich wenig Information wie relative Leistungswerte, die implizit in Rankings abzulesen sind. Watermann und Stanat (2008, S. 47) fordern, die „Abweichungen des beobachteten Wertes vom regressionsanalytisch bestimmten Erwartungswert“ zu betrachten (Visscher, 2008). In diesen

regressionsanalytischen Untersuchungen können leistungsrelevante Merkmale, wie beispielsweise der sozioökonomische Status Berücksichtigung finden. Dieser erweiterte Leistungskennwert liefert in zweierlei Hinsicht zusätzlichen Benefit:

- Zum einen steigert die Rückmeldung eines Leistungskennwertes, der weitere leistungsrelevante Merkmale berücksichtigt, die Akzeptanz (vgl. Abschnitt 1.6.4) der Lehrkräfte, da keine „Äpfel mit Birnen“ verglichen werden (Arnold, 1999; Visscher, 2008).
- Zum anderen liefert ein erweiterter Leistungskennwert für die Lehrkräfte eine Orientierung anhand eines objektiven Leistungsstands. Im Vergleich mit anderen Schulen oder Ländern können Lehrkräfte in der sozialen Bezugsnorm die Leistung ihrer Klassen/Lerngruppen nach Berücksichtigung bestimmter Voraussetzungen einschätzen. Liefert die Rückmeldung beispielsweise auch Kompetenzstufen oder lässt ähnliche Aussagen über zu erwartende Leistungen zu, so stehen Informationen in der kriterialen Bezugsnorm zur Verfügung und können als Ausgangspunkt für Entwicklungen durch die Lehrkräfte genutzt werden (A. Hosenfeld, 2010; Rolff, 2001; Visscher, 2008; Watermann & Stanat, 2008).

Lehrkräfte benötigen also inhaltlich bei Rückmeldungen aus Leistungsmessungen insbesondere mehr als nur Rangfolgen reiner Leistungskennwerte. Darüber hinaus wäre kausal-interpretierbares Handlungswissen ein tatsächlicher Gewinn gegenüber dem Beschreibungswissen. Viele Lehrkräfte wünschen sich Informationen, die vor allem auf der Klassenebene nützlich sind, zur Umsetzung neuer Curricula und zur Wahrnehmung des Unterrichts der Schülerinnen und Schüler sowie zu deren Leistung (Schildkamp & Kuiper, 2010).

1.3.2.1.4 Verständlichkeit

Eine grundlegende Voraussetzung für die Wirkung von Feedback ist, dass dieses genutzt wird. Jedoch ist es intuitiv verständlich, dass Feedback nur dann genutzt werden kann, wenn die enthaltenen Informationen auch verstanden und korrekt dekodiert werden (zu Kommunikationsmodellen vgl. u. a. Frindte, 2001; Nünning & Zierold, 2008; Röhner & Schütz, 2016; Six, Gleich & Gimmler, 2007):

So gesehen wundert es doch, dass die „zweite empirische Wendung in der Schulpädagogik“ bereits ausgerufen worden ist,

denn allein im Bereich der Evaluationsdiagnostik zu hinreichend verstehender Expertise zu gelangen, erfordert weit mehr als ein Lehramtsstudium. Eine ärgerliche Nebenwirkung des neuen, allseitigen Interesses an empirischen Daten muss deshalb auch thematisiert und angegangen werden: Vieles, was jetzt in Schulen diagnostisch und evaluativ geschieht, ähnelt mehr einem naiven, neo-positivistischen Aktionismus als einer kompetenten, sich der eigenen Grenzen bewussten, d. h. wissenschaftsgeleiteten Berufspraxis. (Arnold, 2007, S. 456)

Dieser Schwierigkeit muss begegnet werden, soll sichergestellt sein, dass Rezipienten von Feedback dieses bewusst verarbeiten und als Ausgangspunkt von Veränderung anerkennen (Klieme & Baumert, 2001; Visscher, 2008). Um ein umfassendes Verständnis der Rückmeldung zu erreichen, ist eine „ausführliche Erklärung der eingesetzten Skalen“ bzw. Instrumente ebenfalls sinnvoll (A. Hosenfeld, 2010, S. 43). Negativ ausgelegt warnt bereits Rolff (2001, S. 343) davor, dass „hochkomplexe Rezeptionsprobleme“ auftreten können und prinzipiell „mit Skepsis, Missverständnissen, Fehlinterpretationen und unzulässigen Verallgemeinerungen“ gerechnet werden muss. Hierauf rekurrierend empfehlen Watermann und Stanat (2008, S. 50) die deutliche Kennzeichnung von *bedeutsamen* Werten bzw. Abweichungen, um die/den Lesenden zu lenken. Geschieht dies nicht, so kann es schnell zu einem „information overload“ kommen (Schildkamp & Kuiper, 2010, S. 493f.), der darin mündet, die Nutzung der Rückmeldung zu unterlassen.

1.3.2.2 Rahmenbedingungen bei der Nutzung

Bereits bei der Gestaltung von Rückmeldungen an sich ist also auf eine hier nur angerissene Vielfalt an Aspekten zu achten, um im weiteren Nutzungsprozess so wenige Probleme wie möglich entstehen zu lassen. Während Hattie und Timperley (2007) ein Modell für die Gestaltung von Feedback geliefert haben, betten Schildkamp und Kuiper (2010) die Feedback- bzw. Datennutzung in einen größeren Zusammenhang ein (vgl. Abbildung 7).

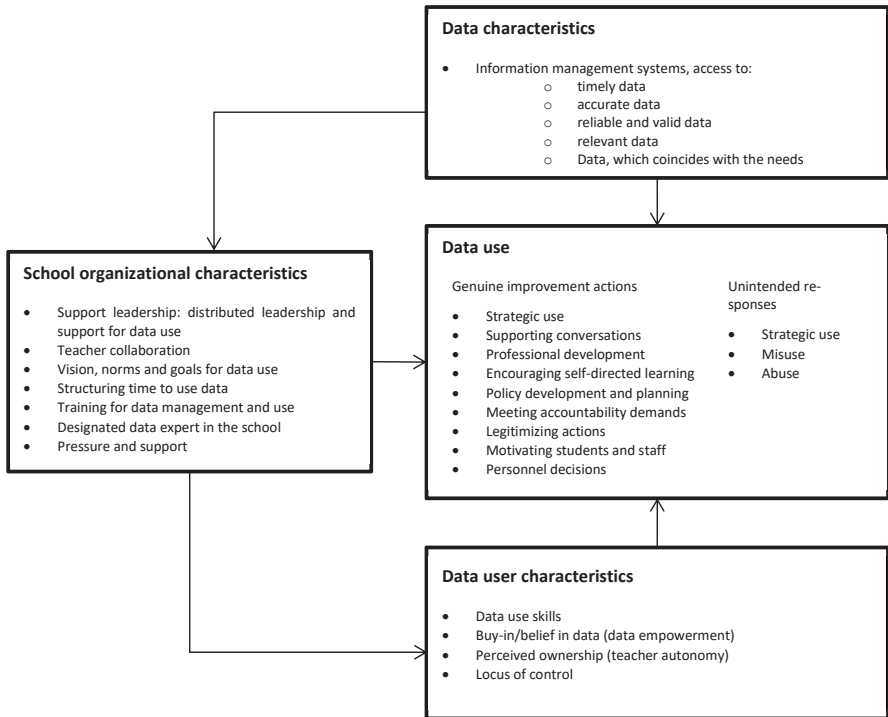


Abbildung 7: Merkmale, die (hypothetisch) Datennutzung beeinflussen (Schildkamp & Kuiper, 2010, S. 485)

Grundlage für Feedback- bzw. Datennutzung sind dabei Daten, die einer gewissen Qualität entsprechen: Daten sollten dabei nach Schildkamp und Kuiper (2010) unmittelbar verfügbar, zur Entwicklung passend (vgl. *Feed forward*), verlässlich sowie aussagekräftig, valide und nicht zuletzt relevant sein (Punkt *Data characteristics* in Abbildung 7). Mit dem Feedback sollte die Basis für eine *self-evaluation* gelegt werden, also aus dem Feedback eine entsprechende (Re-)Aktion erfolgen, um die Diskrepanz zwischen Ist- und Soll-Zustand zu minimieren (Visscher & Coe, 2003, S. 323). Neben Schildkamp und Kuiper (2010) stellen Visscher und Coe (2003) ein ähnliches Modell für die schulische Feedbacksituation vor, das sich in den Merkmalen von Feedbackgestaltung weitestgehend deckt (vgl. auch Ehren & Visscher, 2006; Visscher, 2008). In diesem Modell werden jedoch bereits Organisationsmerkmale wie etwa die Unterstützung zur Problemlösung als Eingangsmerkmal genannt.

Einen wichtigen Einfluss auf die Nutzung von Rückmeldungen haben die Charakteristika der nutzenden Person selbst. Dabei werden von Schildkamp und Kuiper (2010)

etwa Fähigkeiten zur Datennutzung und der *locus of control* genannt. An dieser Stelle sei jedoch, um Dopplungen zu vermeiden, auf den Abschnitt 1.4.3.1 zu individuellen Bedingungen bei der Evaluations-Nutzung verwiesen. Ebenfalls mit Einfluss auf die Nutzung von Feedback stellen sich Merkmale der Organisation dar. Wiederum sei hier auf die Ausführungen im Evaluationskapitel verwiesen (Abschnitt 1.4.3.2).

1.3.3 Nutzung von Ergebnissen aus Leistungsmessung

Der zentrale Punkt des Modells (vgl. Abbildung 7) ist die eigentliche Nutzung, der Umgang mit den zur Verfügung stehenden Daten und deren Rückmeldung. Denn wie von vielen Autoren angeführt bringt eine Rückmeldung *per se* keine Entwicklung mit sich (repräsentativ beispielsweise EMSE, 2008; Schildkamp & Visscher, 2009; Visscher & Coe, 2003; Visscher, 2008). Schildkamp und Kuiper (2010) unterscheiden drei Arten von *unintendierter* sowie weitere drei Arten *intendierter Nutzung* (Visscher & Coe, 2003; Visscher, 2008):

Unintendierter Nutzen

- strategic use
- misuse
- abuse

Intendierter Nutzen

- direct/instrumental use
- conceptual use
- convincing/symbolic use

Erstgenannter *strategic use* meint die fälschliche (aber weit verbreitete) Nutzung „only easy to use data to change, and ignore data that involve more complicated long term improvement trajectories“ (Schildkamp & Kuiper, 2010, S. 484). Es werden also lediglich direkt zugängliche Daten/Rückmeldungen genutzt ohne eine tiefere Verarbeitung oder Ursachensuche. In diesem Sinne handelt es sich um eine oberflächliche Nutzung. Dies kann an nicht vorhandener Kompetenz liegen oder Kennzeichen mutwilliger Ignoranz gegenüber einer umfassenden Nutzung sein. Dies ist besonders unter dem Gesichtspunkt außerordentlich kritisch, dass die meisten Informationen aus Rückmeldungen von Leistungsmessungen lediglich querschnittliche, nicht kausal interpretierbare Informationen liefern und daher, wie Terhart (2002, S. 92) fordert, „Leistungsvergleiche von tiefer gehenden, differenzierten Analysen begleitet und vervollständigt werden“ müssen. Es sind also gerade die komplexeren Analysen, die entscheidende Informationen enthalten.

Die strategische Nutzung von Datenrückmeldungen steht im Grunde in der Mitte eines „Nutzungskontinuums“. *Misuse* und *abuse* schließen sich auf der negativen Seite an und beschreiben fehlerhafte Interpretationen der Rückmeldung. *Abuse* stellt eine mutwillige Falschnutzung dar: Schildkamp und Kuiper (2010, 484f.) berichten

beispielsweise von einer „Gruppeneinteilung“ im Anschluss an Rückmeldungen, so dass der folgende Unterricht weder auf die „safe cases“ noch auf die „hopeless cases“, sondern ausschließlich auf jene Schülerinnen und Schüler ausgerichtet war, bei denen eine zeitnahe Zielerreichung aufgrund des Testergebnisses erwartet werden konnte. Um Missverständnissen vorzubeugen, muss erwähnt werden, dass es sich um Tests im anglo-amerikanischen Raum handelte, aus denen als *high-stakes*-Tests Konsequenzen für die Lehrkräfte abgeleitet wurden. *Misuse* hingegen stellt die (im Endeffekt nicht minder problematische) unabsichtliche Fehlnutzung von Feedback dar. Bereits früh warnt Rolff (2001, S. 343) vor „hochkomplexen Rezeptionsproblemen“ und Arnold (2007, S. 456) stellt fest, dass die „zweite empirische Wendung in der Schulpädagogik bereits ausgerufen worden“ sei, obwohl die Ausbildungsstrukturen sich an diese neuen Herausforderungen noch nicht angepasst hätten.

An einen strategischen Nutzen schließen sich auf der anderen Seite die intendierten Nutzungen an. Dabei versteht sich *convincing/symbolic use* als Nutzung des Feedbacks zur Unterstützung der eigenen Position. Durch das Feedback wird damit kein „blinder“ Fleck behoben oder neue Aspekte in den Entwicklungsfokus gerückt. Unter *conceptual use* wird die indirekte Beeinflussung der Akteure gefasst, während *direct/instrumental use* dann eintritt, wenn direkt aus den Daten/dem Feedback abgeleitet Entwicklungen angestoßen werden.

Diese verschiedenen Nutzungsarten lassen sich auf die im Modell von Schildkamp und Kuiper (2010) angeführten Nutzungszwecke anwenden (linke Spalte im Punkte *Data use*, Abbildung 7). Beispielsweise ist die Nutzung von Rückmeldungen in *symbolischer* Art möglich, um die bestehende Kommunikation und Kooperation zu bestärken oder in *direkter/instrumenteller* Art diese zu optimieren. Rückgemeldete Daten können ebenfalls genutzt werden, um die Motivation von Schülerinnen und Schülern zu steigern, zu individueller Fortbildung (professional development) anzuregen und Personalentscheidungen zu treffen. Insbesondere der letztgenannte Punkt macht deutlich, dass die administrative, höhere Hierarchieebene mit einbezogen wird.

Zu welcher Art der Nutzung es kommt, lässt sich nach Ditton und Arnoldt (2004, S. 118) anhand von vier Faktoren einschätzen: „(a) Merkmale des Feedbacks, (b) Merkmale der Aufgabe, (c) Merkmale der Person und (d) Merkmale der Handlungssituation.“ Eine nähere Betrachtung schließt sich im folgenden Abschnitt zur Evaluation an.

1.4 Evaluation

Nur wer fortlaufend überprüft, wo er steht, was er oder sie erreicht hat und was nicht, kann sein Lernen selber steuern, bleibt auf Dauer überhaupt lernfähig. (Rolff, 2007, S. 145)

Mit Evaluation ist nach den Richtlinien der Deutschen Gesellschaft für Evaluation (DeGEval, 2002, S. 15) „die systematische Untersuchung des Nutzens oder Wertes eines Gegenstandes“ gemeint. „Die erzielten Ergebnisse, Schlussfolgerungen oder Empfehlungen müssen nachvollziehbar auf empirisch gewonnenen qualitativen und/oder quantitativen Daten beruhen.“ Evaluation stellt damit also eine Diagnose bzw. Messung dar, an die sich eine bewusste Nutzung etwa durch Ableitung sowie Durchführung von Maßnahmen anschließt (Stufflebeam, 1967). In diesem Verständnis erweitert der Begriff *Evaluation* die bisherigen Beschreibungen zu Messung (vgl. Kapitel 1.2) und Rückmeldung (vgl. Kapitel 1.3).

Seitens der DeGEval (2002) wurden für Evaluation Standards aufgestellt: Nach diesen Vorgaben sollte Evaluation den folgenden vier Eigenschaften genügen, welche prinzipiell als Referenzrahmen zur Beurteilung von Evaluationsmaßnahmen dienen können:

1. Nützlichkeit
2. Durchführbarkeit
3. Fairness
4. Genauigkeit

Eine Evaluation sollte also dem evaluierten Personenkreis *nützen*. Im Detail bedeutet dies, dass betroffene bzw. beteiligte Personengruppen identifiziert werden und es eine transparente Bezugsgröße/Bezugsnorm/Standard gibt. Mit der Nützlichkeit indirekt verbunden ist ein zeitlicher und personeller Horizont: Stufflebeam (1967, S. 129) weist darauf hin, dass Evaluationsergebnisse über die notwendige Gültigkeit (*valid*) verfügen sowie zeitnah (*timely*) und dabei allen betroffenen Personengruppen (*pervasive*) zugänglich sein müssen. A. Hosenfeld (2010) betont darüber hinaus, dass eine Evaluation ziel- und zweckgebunden ist und eine Bewertung von Handlungsalternativen erfolgt. Dabei liefern anspruchsvolle und komplexe Evaluationsdesigns in vielen Fällen die nützlichsten Ergebnisse, weshalb Evaluationsverfahren – soweit möglich – dem aktuellen Entwicklungs- und Forschungsstand entsprechen sollten.

Allerdings wird insbesondere im schulischen Umfeld „methodische Perfektion letztlich die Abnehmerbedürfnisse nur schwer befriedigen“ und damit u. a. die *Durchführbarkeit* der Evaluation in Frage stellen (Stamm, 2003, S. 185). Weitere Schwierigkeiten dürften bei der Interpretation statistisch und methodisch anspruchsvoller Evaluationsdesigns seitens dafür nur bedingt ausgebildeten Personals auftreten (Arnold, 2007; Rolff, 2001; Stamm, 2003; Visscher, 2008; Watermann & Stanat, 2008).

Entsprechenden Schwierigkeiten beugt der Standard *Durchführbarkeit* der DeGEval (2002) vor: Eine Evaluation sollte in Bezug auf den Ablauf und die entstehenden Kosten *durchführbar* sein. Aus Evaluationsverfahren resultierende Kosten stellen einen kritischen Punkt dar, weil die Finanzierung im schulischen Fall aus öffentlicher Hand erfolgt. Durch die veränderte Steuerungsphilosophie von Schule, welche mit den Bildungsstandards (vgl. Kapitel 1.1) und der damit wachsenden Eigenverantwortung von Schule einhergegangen ist, ist auch der Anspruch der Rechenschaftsablegung (*accountability*) – vor allem für die Einzelschule – gestiegen (Holtappels, 2014a; Korngiebel, 2009; van Ackeren & Klemm, 2011).

Mit der Durchführbarkeit im schulischen Kontext eng verbunden ist der Aspekt der Partizipation durch die Lehrkräfte: Dabei ist zum einen eine Mitwirkung der Betroffenen bei der Entwicklung des Evaluationsverfahrens und dessen Durchführung gemeint (vgl. u. a. Rolff, 2007; Watermann & Stanat, 2008), zum anderen aber auch die Möglichkeit zur Anknüpfung der Evaluationsergebnisse an bereits bestehende Informationen (Kühle & Peek, 2007). Durch dieses Vorgehen wird bei den Lehrkräften Transparenz erzeugt und so die wahrgenommene *Fairness* gesteigert. Die Lehrkräfte schätzen die Ergebnisse/Verfahren als glaubwürdig ein (*credible* bei Stufflebeam, 1967, S. 129). Eine Evaluation sollte sich dabei u. a. an formale Vorgaben halten, individuelle Rechte schützen und vor allem Ergebnisse offenlegen. Bei vielen standardisierten Evaluationsverfahren (beispielsweise PISA, TIMSS oder Vergleichsarbeiten) werden Hintergrundvariablen von Leistung berücksichtigt, um Lehrpersonen „faire Vergleiche“ anbieten zu können (vgl. vertiefende Informationen am Beispiel VerA in Kapitel 1.6).

Der Standard der *Genauigkeit* schließt sich direkt an das Gütekriterium *Reliabilität* an (*reliable* bei Stufflebeam, 1967, S. 129; vgl. Abschnitt 1.2.2.2). Die Evaluation sollte eine möglichst zielgenaue Beschreibung ermöglichen und dabei die genutzten

Informationsquellen (beispielsweise die eingesetzten Evaluationsverfahren) angeben (A. Hosenfeld, 2010).

Analog zu den Standards der DeGEval wird auch für das Bildungswesen davon gesprochen, dass „Evaluation ohne ernsthafte Bemühungen, ihren Ertrag für die Unterrichtsentwicklung und die Förderung von Schülerinnen und Schülern zu nutzen, (wäre) sinnlos, ja zynisch [wäre, S.V.]“ (Isaac, Halt, Hosenfeld, Helmke & Groß-Ophoff, 2006, S. 107; O'Day, 2002). Dabei geht Baumert (2001b) noch einen Schritt weiter und spricht von einer regelmäßigen Evaluation, also insbesondere mehr als eine *Baseline*-Messung (vgl. auch A. Hosenfeld, 2010). Wichtig ist also, dass Evaluation hier nicht nur als eine einmalige Bestandsaufnahme erfolgt, sondern mit dem Ziel, das schulische Lernen zu verbessern und damit letzten Endes die Leistung der Schülerinnen und Schüler gezielt zu steigern (vgl. Abbildung 12; Lipowsky, 2009; Stryck, 2005); damit geht Evaluation über Feedback an sich hinaus.

1.4.1 Evaluationsformen

Vielfach wird zwischen verschiedenen Evaluationsformen und -arten unterschieden (Baumert, 2001b; Klieme et al., 2003; van Ackeren & Klemm, 2011). Bereits in den vorherigen Kapiteln wurden verschiedene Verfahren von Evaluation z. T. explizit angesprochen. Die bisher genannten Verfahren lassen sich beispielsweise von Klassenarbeiten deutlich abgrenzen und vorrangig als *externe* Evaluationsverfahren einordnen. Folgerichtig lassen sich davon *interne* Evaluationsverfahren differenzieren: Dabei meint *interne* Evaluation nach Ramm (2014, S. 45) „diejenigen Untersuchungen [...], die die Schulgemeinschaft selbstverantwortlich durchführt“, während *externe* Evaluation entsprechend ohne Selbstverantwortung der Schulgemeinschaft zu fassen ist. *Interne* und *externe* Evaluation stehen jedoch keinesfalls unverbunden nebeneinander, sondern sollten im Sinne Rolffs (2001, S. 342) identische (oder zumindest ähnliche) Fokusse haben und sich in ihren Ergebnissen ergänzen – also letztlich ineinander verzahnt sein und zueinander passen (Ditton & Arnoldt, 2004). Den Gedanken von Evaluationen und Veränderungen weiterdenkend, stellen Hofmann-Göttig et al. (2005, S. 34) fest: „Ohne innere Schulreform ist die äußere Schulreform nichts“, *interne* und *externe* Verfahren müssen folglich zusammenspielen und gegenseitigen Bezug haben. Während die Verpflichtung zur *internen* Evaluation im Hessischen Schulgesetz fest verankert ist (HSchG, §127b, Abs. 2; Hessisches Kultusministerium, 2016; Ramm, 2014), stellen die diversen Beschlüsse der Kulturministerkonferenz den

Rahmen für *externe* Evaluation (u. a. van Ackeren & Klemm, 2011). Beide Formen der Evaluation sind nach Baumert (2001b) vor dem Hintergrund der höheren Selbstständigkeit und Eigenverantwortung von Schule (vgl. Schule als *pädagogische Handlungseinheit*) an sich unstrittig.

Betrachtet man die verschiedenen *internen* Evaluationsmaßnahmen in ihrer Funktion, so lassen sich nach Baumert (2001b, S. 19) Evaluationen zur (1) individuellen Bewertung und Zertifizierung von Schülerinnen und Schülern von (2) schulinterner Evaluation zur Optimierung von Schulentwicklung und (3) Prüfung externer Programme/Projekte abgrenzen. Die Funktionen *externer* Evaluationsmaßnahmen unterscheidet Baumert (2001b, S. 20) in

- individuelle Zertifizierung durch zentrale Tests oder Prüfungen mit berechtigender Wirkung,
- flächendeckende Evaluation von Einzelschulen bzw. ausgewählten Jahrgängen an Einzelschulen,
- Systemmonitoring auf Stichprobenbasis und
- internationale Vergleichsstudien auf Stichprobenbasis (vgl. auch van Ackeren & Klemm, 2011).

Auf die beiden letztgenannten Funktionen *externer* Evaluation wurde in den vorangegangenen Kapiteln bereits eingegangen (vgl. Kapitel 1.1 und 1.2). In der Klieme-Expertise (Klieme et al., 2003) wird zu einer deutlichen Trennung zwischen *interner* Evaluation zur individuellen Bewertung und der individuellen Zertifizierung durch *externe* Evaluation geraten. Eine externe Zertifizierung anhand von Verfahren, die eher *intern* durchgeführt werden, birgt einige Risiken der Verfälschung: Vor allem aus dem anglo-amerikanischen gibt es viele Erfahrungen zum Einfluss von *high-stakes*-Tests, die beispielsweise für Angst bei Lehrkräften und in Klassenräumen sorgen (Wall, 2000; eine gute Übersicht über positive und negative Effekte auf verschiedenen Ebenen von *high-stakes*-Tests liefert Maag Merki, 2010). Gemeint sind damit Tests, die für weitreichende Entscheidungen genutzt werden, gegenüber *low-stakes*-Tests, deren Ziel es „nur“ ist, schulische Leistung zu messen und Diagnoseinformationen zu liefern (Abbott, 2014). Je bedeutsamer die auf Basis eines Tests getroffenen Entscheidungen für eine Personengruppe sind, umso wahrscheinlicher wird ein „Schummeln“ bei der Durchführung, Auswertung oder Nutzung (Koch, 2011). Diese „Gesetzmäßigkeit“ wurde bereits in den 80er Jahren von Campbell (1979) u. a. bei der

Teil 1 – Theorie Evaluation

Unterstützung von Schülerinnen und Schülern durch Nachhilfelehrkräfte nachgewiesen, die nach der Leistung der Schülerinnen und Schüler im Abschlusstest bezahlt wurden. Im Bildungsbereich kann *campbell's law* u. a. zu einer Fokussierung auf Testinhalte (*teaching-to-the-test*) einhergehend mit einer Verengung des Curriculums (*narrowing-the-curriculum*) führen (Arnold, 2007; Helmke & I. Hosenfeld, 2003; I. Hosenfeld, Koch, Groß-Ophoff & Scherthan, 2008; Landwehr, 2011; Richter, Böhme, Becker, Pant & Stanat, 2014; Spoden, Fleischer & Leutner, 2014; Wacker & Kramer, 2012; Wall, 2000).

Nichtsdestotrotz ist *externe* Evaluation wichtig, insbesondere da letztlich nur durch *externe*, nicht von der Schulgemeinschaft selbstverantwortlich durchgeführte Evaluation „an Vergleichbarkeit ausgerichtete Messungen [...] einen objektiven Eindruck von [den] nachhaltigen Lernergebnissen schaffen“ können (Leuders, 2006, S. 79). Landwehr (2011, S. 41) nennt leicht abweichende Funktionen von *externer* Evaluation, wobei der Fokus eher auf der Einzelschule liegt (McDowall, Cameron, Dingle, Gilmore & MacGibbon, 2007):

1. Wissensgewinnung
2. Rechenschaftsablegung
3. Schulentwicklung
4. Normendurchsetzung

Diese vier Wirkungsbereiche greifen jeweils ineinander, da beispielsweise neu erzeugtes Wissen direkt der Schulentwicklung zu Gute kommt, ggf. die Grundlage der Rechenschaftsablegung verändert und damit auch Normen beeinflussen kann bzw. durch die Prüfung von Normen diese wiederum vertreten werden. In Bezug auf die oben genannten Funktionen lässt sich die Rechenschaftsablegung sowie die Schulentwicklung als einzelne Teile des zweiten Punktes bei Baumert (2001b), flächendeckende Evaluation, fassen. Der Aspekt der Normendurchsetzung korrespondiert mit der individuellen Zertifizierung, da Prüfungen automatisch inhaltliche Rückwirkungen auf den Unterricht haben (*washback*-Effekt; Anderson & Wall, 1993; Wall, 2000). Landwehr (2011) unterstreicht ebenfalls, dass *externe* Evaluation die größte Wirkung erzielt, wenn eine Bereicherung und ein Bezug zur *internen* Evaluation bestehen. Auf diese Weise wird auch einer Ablehnung von Seiten der Lehrkräfte begegnet, die sich womöglich in ihrer Kompetenz angegriffen fühlen (Koch, 2011; vgl. auch individuelle Bedingungen weiter unten). Insbesondere der Einfluss der hohen Sachkenntnis der

Lehrkräfte ist in „ihrem“ Feld bei der *internen* Evaluation nutzbar, wenn auch eine daraus resultierende geringere Unabhängigkeit von den Evaluierenden sowie das Risiko einer gewissen Betriebsblindheit gegeben ist (vgl. Abschnitt 1.2.2; Koch, 2011).

1.4.2 Evaluationskreislauf

Aus den bisherigen Ausführungen wird deutlich, dass eine Evaluation keine einmalige Messung von Ergebnissen darstellen sollte, sondern ein sich wiederholender Prozess ist. In der Literatur finden sich diverse exemplarische Prozessbeschreibungen für Evaluation (u. a. Ditton & Arnoldt, 2004; Helmke, 2004; Hessisches Kultusministerium, 2009; LSA, 2015; Ramm, 2014; Stufflebeam, 1967; van Ackeren & Klemm, 2011). Exemplarisch zeigen Abbildung 8 den von der Hessischen Lehrkräfteakademie in den Handreichungen vorgeschlagenen Evaluationskreislauf und Abbildung 9 einen theoretisch vorgestellten Rahmen für die Evaluation im US-amerikanischen Schulsystem, bei dem bewusst der Fokus auf die verschiedenen Ebenen des Schulsystems gelegt wurde:

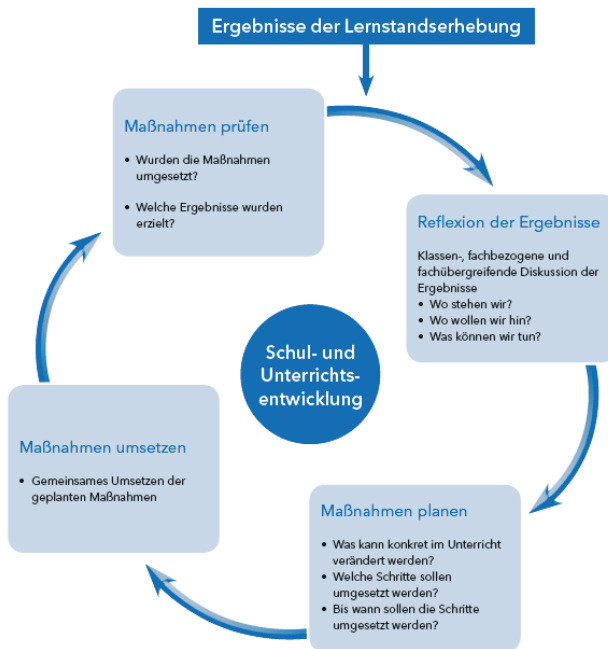


Abbildung 8: Kreislauf der Evaluation durch Lernstandserhebungen (LSA, 2015, S. 24)

Teil 1 – Theorie Evaluation

Das Modell von Stufflebeam (1967) ist dabei allgemeiner gehalten, da es zum einen mehrere Ebenen des Schulsystems einbezieht und zum anderen die Organisation/Planung der Evaluation selbst als eigenen Punkt anführt. Bei den Lernstandserhebungen in Hessen hingegen (vgl. Kapitel 1.6) liegt die Planung und Konzeption des Evaluationsinstruments beim Institut für Qualitätsentwicklung im Bildungswesen (IQB). Stufflebeams (1967) Kreislauf ist zwischen den einzelnen Ebenen in den Schritten deckungsgleich, sodass hier auf eine der Ebenen näher eingegangen werden kann.¹³ Im Folgenden sollen die einzelnen Schritte der Kreisläufe etwas detaillierter betrachtet werden (vgl. auch DeGEval, 2002; LSA, 2015; Ramm, 2014; Rolff, 2002; Stufflebeam, 1967).

Operationalisierung und Planung der Evaluation

Im ersten Schritt geht es – analog zum exemplarischen Ablauf eines Forschungsprozesses mit einer genauen Formulierung von statistisch falsifizierbaren Hypothesen (Bortz & Döring, 2006; Bühner, 2011; Eid et al., 2011) – darum, Ziel und Zweck (A. Hosenfeld, 2010) der Evaluation zu bestimmen und damit einhergehend auch den eigentlichen Messgegenstand zu operationalisieren. Selbstverständlich ist, dass eine Operationalisierung und Zieldefinition so genau wie möglich erfolgen sollte, um im weiteren Verlauf aus den Evaluationsergebnissen möglichst konkrete Handlungen ableiten zu können. Je nach Partizipation in der Entwicklung von Evaluationsinstrumenten bzw. -verfahren sind Lehrkräfte in diesem Schritt u. U. nicht beteiligt. Nach der Operationalisierung ist die Auswahl eines geeigneten Messinstruments bzw. deren Entwicklung entscheidend. Ebenfalls vor Beginn der tatsächlichen Evaluation gilt es natürlich auch, Ablauf und Organisation der Evaluation zu planen und realistische, wenn auch anspruchsvolle Ziele zu formulieren (I. Hosenfeld et al., 2006).

Messung

Im Anschluss an die Planung und Operationalisierung der Evaluation erfolgt die gezielte Messung (Schildkamp & Kuiper, 2010); hierbei kann es sich beispielsweise rein um die Leistung in einem ausgewählten Fach handeln (vgl. VerA), um einzelne Ausschnitte einer Lerneinheit (vgl. Klassenarbeit; Drücke-Noe, 2014), aber auch um eine Vielzahl verschiedener Messungen, etwa um zu erfahren, wie der Unterricht von Schülerinnen und Schülern erlebt wird (beispielsweise bei Schulinspektionen mit der

¹³ Damit hier der direkte Bezug zum Kreislauf des LSA (2015) besteht, wird hierbei die Unterrichtsebene aufgegriffen, also im Konkreten die Schritte 1 bis 5 in Abbildung 9.

Beobachtung von Unterricht, Interviews und Dokumentenprüfung; Böttcher & Keune, 2011; Ditton & Arnoldt, 2004; Landwehr, 2011; Nieder & Frühauf, 2012).

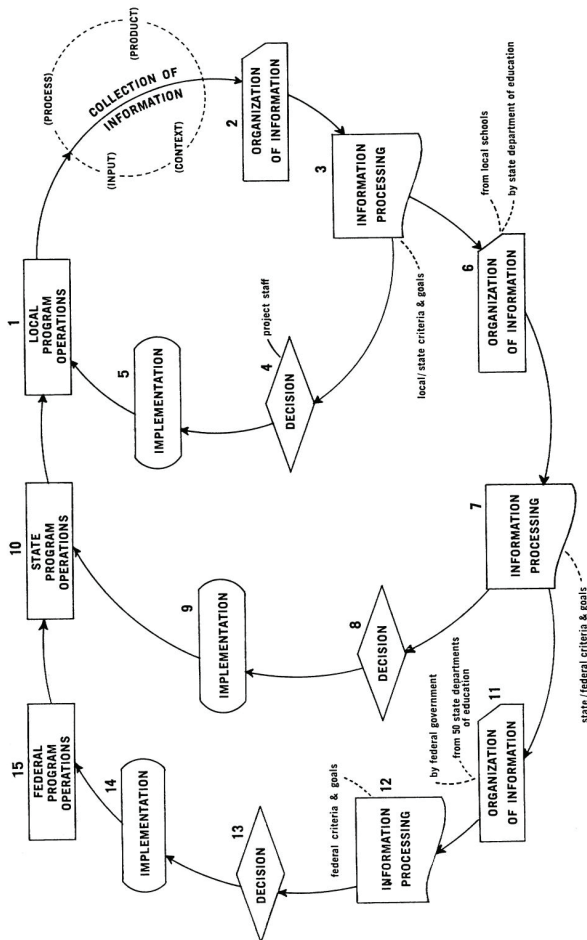


Abbildung 9: Nutzungskreislauf auf verschiedenen Ebenen des Schulsystems (Stufflebeam, 1967, S. 132)

Rezeption der Ergebnisse/Rückmeldung

Mit dem nächsten Schritt „beginnt“ der Kreislauf des LSA (2015), nämlich die *Rezeption* der Evaluationsergebnisse, wobei intuitiv klar ist, dass eine Rückmeldung der Ergebnisse diesem Schritt vorausgehen muss. Wie bereits bei den Standards zur

Teil 1 – Theorie Evaluation

Evaluation (DeGEval, 2002) angedeutet, ist es wichtig, dass diese Ergebnisse an alle verantwortlichen Ebenen des Bildungssystems zurückgemeldet werden (Rolff, 2001). Eine umfassende Rückmeldung der Ergebnisse beeinflusst dabei neben der Nutzung der Ergebnisse auch motivationale Aspekte, da eine professionelle Rückmeldung eine Wertschätzung darstellt, die den Lehrkräften sowie den Schülerinnen und Schülern etc. entgegengebracht wird (Arnold, 1999; Bonsen, Büchter & Peek, 2006).

Reflexion der Ergebnisse/Suche nach Ursachen

Im Anschluss an eine Rezeption der Rückmeldung gilt es, die Ergebnisse in den jeweiligen Rezeptions-Gruppen zu *reflektieren*. Dabei muss darauf geachtet werden, inwiefern die Ergebnisse für die jeweilige Gruppe aussagekräftig sind. Ausgehend von diesem zurückgemeldeten Ist-Stand gilt es im kriterialen, sozialen oder individuellem Vergleich, Verbesserungspotenzial auszumachen (van Ackeren & Klemm, 2011). Dabei kann handlungsleitend beispielsweise ein Abgleich von Erwartung und tatsächlichem Ergebnis sowie die Suche nach Ursachen hierfür sein (Helmke, I. Hosenfeld & Schrader, 2003; Schrader & Helmke, 2005). Die Reflexion wie auch die nächsten Schritte des Evaluationskreislaufs können nicht nur von der einzelnen Lehrkraft, sondern auch von Teams, Fachkollegien oder Gesamtkollegien gemeinsam angegangen werden.

Planung von Maßnahmen

Aus den Ergebnissen und deren Reflexion werden im letzten Schritt mögliche Maßnahmen abgeleitet. Dementsprechend werden die Ergebnisse direkt für die Schulentwicklung genutzt (vgl. Kapitel 1.5; Visscher & Coe, 2003). Im Fokus von Evaluationsmaßnahmen sollte nicht primär die Selektion stehen (Kuper & Diemer, 2012). Mögliche Handlungsfelder sind von Schildkamp und Kuiper (2010) bereits in Abbildung 7 genannt: Die Ergebnisse können vielmehr in Maßnahmen zu Personalentscheidungen bzw. -entwicklung münden, Möglichkeiten für eine veränderte Unterrichtspraxis bieten (vgl. Personalentwicklung und Unterrichtsentwicklung in Kapitel 1.5), aber auch zur Rechenschaftsablegung nach außen (*accountability*) genutzt werden. Bei der Planung der Maßnahmen sollte neben der theoretischen Passung zum vorher identifizierten Bereich auch darauf geachtet werden, dass realistische, aber anspruchsvolle Ziele artikuliert werden, da so die Bereitschaft zur Veränderung gesteigert wird (Gruschka et al., 2005). Eine möglichst präzise Zieldefinition dient dazu, nach Abschluss der entsprechenden Maßnahme deren Erfolg eindeutig messen zu können, also eine „erneute“ Evaluation möglich zu machen (I. Hosenfeld et al., 2006). Die

rückgemeldeten Daten und deren Auswertungen eröffnen zusätzlich die Option, die vorherige Operationalisierung zu konkretisieren bzw. neu zu fassen, Optimierungen am Messinstrument durchzuführen und so auch zu verdeutlichen, welche Ziele für die Zukunft ausgegeben werden (Klieme et al., 2003; Ramm, 2014).

Durchführung von Maßnahmen

Im dritten Schritt (des LSA-Kreislaufes) geht es um die Durchführung von Maßnahmen. Bei diesem und dem vorherigen Schritt gilt es, sich dessen bewusst zu sein, dass es grundsätzlich eine Entscheidung zwischen Handlungsalternativen/Alternativmaßnahmen ist (Gruschka et al., 2005).

Erneute Evaluation und „Schließen des Kreislaufes“

Den Abschluss des ersten Kreislaufes bildet eine Evaluation bzw. Kontrolle der durchgeführten Maßnahmen. Alle angestrebten und umgesetzten Veränderungen müssen sich einer Kontrolle der Wirksamkeit stellen. Die Ergebnisse dieser Evaluation stellen wiederum die Grundlage eines neuen Kreis-Laufes dar, der wiederum mit der Reflexion dieser Ergebnisse beginnt (Rolff, 2007). Damit gehen Schulentwicklung und Evaluation im Grund ineinander über und – ergänzend zur intuitiven Annahme, dass Evaluation und Datennutzung als Voraussetzung von Schulentwicklung stehen – gilt: „Letztlich (ist) Schulentwicklung sogar Voraussetzung von wirksamer Datennutzung“ (Rolff, 2002, S. 93).

1.4.3 Von der Evaluation zur Innovation

Die beiden vorgestellten Kreisläufe konzentrieren sich auf den Ablauf von Evaluationen. Allerdings vernachlässigen sie dabei die Rahmenbedingungen, die die Nutzung beeinflussen. Helmke (2004) stellt ein vielzitiertes Rahmenmodell im Zusammenhang mit der Einführung von Vergleichsarbeiten in Klasse 3 in Rheinland-Pfalz vor. Kritisch bei diesem Modell anzumerken ist, dass der kreisförmige Charakter nicht direkt ersichtlich wird, sondern dieses Modell eher nach einem linearen, und damit nach *einem* Durchlauf abgeschlossenen Prozess anmutet (Abbildung 10).

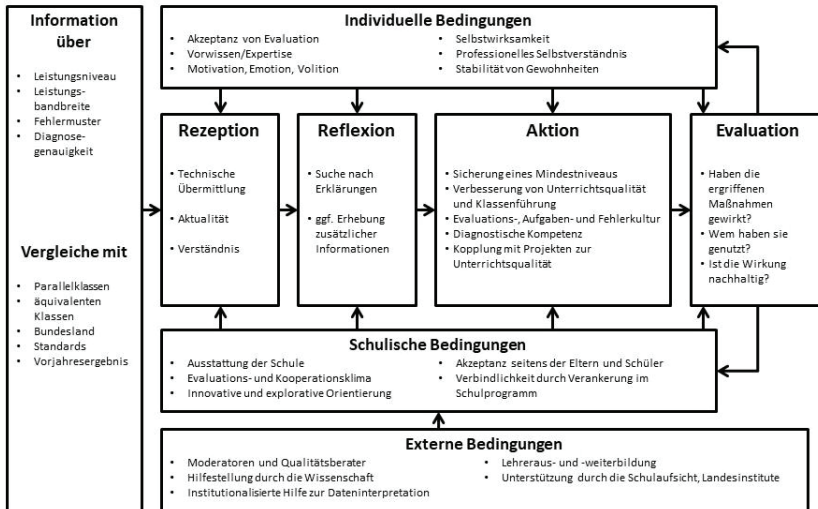


Abbildung 10: Rückmeldungs- und Evaluationsnutzung nach Helmke (2004, S. 100; vgl. auch I. Hosenfeld et al., 2006)

Analog zu den vorgestellten Kreisläufen (Abbildung 8 und Abbildung 9) stehen die Schritte *Reflexion/Rezeption*, *Aktion* (Maßnahmen planen und durchführen) und erneute *Evaluation* im Zentrum, werden aber in Ergänzung zu den bisherigen Beschreibungen durch (1) individuelle, (2) schulische und (3) externe Bedingungen gerahmt. Der Fokus dieses Modells liegt vorrangig auf der Nutzung innerhalb der Einzelschule, ist also beschränkt auf den innersten Kreis im Modell von Stufflebeam (1967). Die Rahmenbedingungen sollen im Folgenden näher beschrieben werden.

1.4.3.1 Individuelle Bedingungen

Breite Einigkeit besteht in der Literatur darüber, dass die *Akzeptanz* der Lehrkräfte gegenüber einer Evaluation wichtige Voraussetzung für deren Nutzung ist (Bonsen et al., 2006; Holtappels, 2014b; Rolff, 2001; van Ackeren & Klemm, 2011; Watermann & Klieme, 2006). Altrichter (1998, 2010) ergänzt, dass anschlussfähig an die Ausführungen zur Evaluation eingangs dieses Kapitels Daten für Lehrkräfte *relevant* sein müssen und die für die Nutzung – also insbesondere auch für die Entwicklung und Durchführung von Maßnahmen – notwendigen persönlichen und materiellen Ressourcen vorhanden sein müssen (Schildkamp & Visscher, 2009; vgl. zu persönlichen Ressourcen u. a. Klusmann, 2011; Schaarschmidt, 2005, 2007). Ditton und Arnoldt (2004, S. 133) weisen zwischen der Belastung als Indikator für fehlende persönliche Ressourcen und

der Kompetenz von Schülerinnen und Schüler einen signifikanten mittleren negativen Zusammenhang nach ($r = -.32, p < .05$); einen entsprechenden negativen Einfluss von Belastung im Sinne der von Schaarschmidt (2005, 2007) gefundenen vier Muster zeigt sich auch in COACTIV (Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung mathematischer Kompetenz; Kunter et al., 2011) mit signifikanten Unterschieden zwischen den Belastungsmustern in den Kategorien Unterrichtstempo ($F(3,312) = 6.37, p < .05$), kognitiver Aktivierung ($F(3,312) = 3.19, p < .05$) und soziale Unterstützung ($F(3,312) = 8.57, p < .05$). Als Indikator für vorhandene persönliche Ressourcen findet sich für die Berufszufriedenheit ein signifikanter positiver Zusammenhang mit der Verwendbarkeit von Rückmeldungen ($r = .33, p < .05$). Belastete Lehrkräfte zeigen sowohl höhere Werte bei Indikatoren für fehlende/mangelnde Unterrichtsqualität als auch geringere Werte bei Indikatoren für Unterrichtsqualität als ihre weniger belasteten Kolleginnen und Kollegen.

Eine weitere zentrale Voraussetzung für die Nutzung von Evaluations- und Rückmeldeinformationen stellt das *Vorwissen* dar. Mit Bezug auf COACTIV (Baumert & Kunter, 2011; Kunter et al., 2011) sollte hier jedoch nicht nur von „Wissen“ gesprochen werden, sondern der Blick erweitert werden auf die gesamte *professionelle Kompetenz* der Lehrkräfte, also ergänzend zum reinen Wissen auch auf Überzeugungen/Werthaltungen/Ziele (*professionelles Selbstverständnis* und *Stabilität der Gewohnheiten* in Abbildung 10 respektive *beliefs* in Abbildung 7; vgl. auch *sokratischer Eid* in Hentig, 1994, 258f.), motivationale Orientierungen (*Motivation, Emotion, Volition* in Abbildung 10) und Aspekte der Selbstregulation. Ein Teil selbstregulatorischer Fähigkeiten wurde gerade unter Verweis auf Schaarschmidt (2005, 2007) angesprochen. Rolff (2001, S. 343) deutet aufbauend auf den Ausführungen von Altrichter (1998, 2010) an: „Die *Motivation der Betroffenen zur Weiterentwicklung* ist im Allgemeinen höher, wenn sie an den Evaluationsprozessen mitwirken konnten“ (vgl. auch Koch, 2011; Watermann & Stanat, 2008). Als wesentlicher Einflussfaktor auf die Motivation muss nach dem Erwartungs-Wert-Modell die *Bedeutung* bzw. *Relevanz* der Rückmeldungen/Evaluation für die Lehrkräfte genannt werden (Altrichter, 2010; Atkinson, 1957; Rolff, 2001; Wigfield & Eccles, 2000; vgl. auch Abschnitt 1.3.2.1 zur Gestaltung von Feedback), wobei vermutet werden kann, dass durch die Mitwirkung von Lehrkräften für diese ein erhöhter *Wert* entsteht.

Teil 1 – Theorie Evaluation

Zwar ist die Motivation zur Weiterentwicklung hiernach variabel und beeinflussbar. Generell gilt jedoch nach Bensen und Rolff (2006, S. 168) unter Bezug auf DuFour und Eaker (1998), dass „die Notwendigkeit, die eigenen Unterrichtspraxis fortlaufend zu reflektieren und gegebenenfalls zu verändern, [...] im professionellen Anspruch der Lehrertätigkeit“ – also im *professionellen Selbstverständnis* – begründet liegt, letztlich eigentlich eine Selbstverständlichkeit darstellen sollte. Der Fokus von Evaluationen für/durch Lehrkräfte sollte entsprechend auf individuelle Denk- und Verstehensprozesse gelegt werden (Sjuts, 2008), wobei Maag Merki (2010, S. 165) einschränkend anmerkt, dass bei allen Evaluationsrückmeldungen zunächst geprüft werden muss, „wer für welche Lernergebnisse [...] ‚accountable‘“ also verantwortlich ist (vgl. auch Bensen & Frey, 2014; Koch, Groß-Ophoff, Hosenfeld & Helmke, 2006; Posch, 2009; Rolff, 2002).

Im Kompetenzmodell von COACTIV ist das Professionswissen als Teil von *professioneller Kompetenz* wiederum untergliedert in die Bereiche (1) Fachwissen, (2) fachdidaktisches Wissen, (3) pädagogisch-psychologisches Wissen, (4) Organisationswissen und (5) Beratungswissen. Für die Nutzung von rückgemeldeten Evaluationsergebnissen kommt vor allem dem pädagogisch-psychologischen und auch dem fachdidaktischen Wissen eine besondere Rolle zu: Einerseits benötigen Lehrkräfte zur Nutzung von Evaluationsinformationen das notwendige Wissen über Leistungsbeurteilungen und -messungen (Posch, 2009; Schildkamp & Kuiper, 2010; van Ackeren & Klemm, 2011; vgl. *data use skills* in Abbildung 7), andererseits gibt es „Kein Stricken ohne Wolle“, sodass jede Rückmeldung im konkreten fachlich-didaktischen Licht interpretiert werden muss, um so aus den Informationen Handlungsalternativen ableiten zu können (Altrichter, 2010). Insbesondere fehlt es jedoch Lehrkräften häufig am Wissen über Leistungsmessungen/-beurteilung und der Auswertung von statistischen Daten (Arnold, 2007; Ditton & Arnoldt, 2004).

Bisher unbeachtet, aber durchaus nachvollziehbar führt Helmke (2004) als weitere individuelle Bedingung die *Selbstwirksamkeit*¹⁴ an. Bandura (1977, S. 191) definiert, dass „expectations of personal efficacy determine whether coping behavior will be

¹⁴ Für weitere Informationen zur Selbstwirksamkeit vgl. u. a. Bandura (1982), Jerusalem und Hopf (2002), E. Wild, Hofer und Pekrun (2005), Harackiewicz, Barron, Tauer und Elliot (2002) und B. Zimmerman (2000). Dieses pädagogisch-psychologische Konstrukt ist vom Selbstkonzept zu unterscheiden (für eine Übersicht vgl. Bong & Skaalvik, 2003).

initiated, how much effort will be expended, and how long it will be sustained in the face of obstacles and aversive experiences.“ Die Überzeugung eine Anforderungssituation – hier die Rückmeldung von Evaluationsinformationen, deren Interpretation, das Ableiten von Maßnahmen mit deren Durchführung sowie die erneute Überprüfung – meistern zu können, beeinflusst den Umgang mit den Daten. In Abbildung 7 wird dies mit dem *locus of control* bereits gestreift: Damit ist nicht nur entscheidend, ob die Lehrkräfte im Sinne der Selbstwirksamkeit sich die Bewältigung der Rückmelde-Situation zutrauen, sondern auch, ob sie die Situation als veränderbar einschätzen und somit die Ursachen *intern* statt *extern* lokalisieren (vgl. Weiner, 1975, 1986; E. Wild, Hofer & Pekrun, 2005). Kuper und Diemer (2012, S. 238) weisen beispielsweise darauf hin, dass es „durchaus als ein praktikabler Weg der Unsicherheitsbekämpfung angesehen werden [kann, S.V.], wenn Ursachenzuschreibungen gefunden werden, die nicht die Kompetenz der Lehrkraft und die Qualität des Unterrichts betreffen, sondern äußere Umstände der Unterrichtsorganisation, wie z. B. Unterrichtsausfall“ (vgl. auch Rolff, 2002; Visscher, 2008). Ditton und Arnoldt (2004, S. 135) berichten ähnlich von einer bedeutsamen Korrelation zwischen der Akzeptanz von Erhebungs- und Rückmeldeverfahren sowie dem erreichten Ergebnis, also dass bei abnehmender Leistung die Zweifel und Skepsis an der Erhebung bzw. der Rückmeldung zunehmen.

Ein Aspekt an der Schnittstelle zwischen individuellen und schulischen Bedingungen wird bereits in der Nutzung von Feedback vorgeschlagen (vgl. Abbildung 7 in Abschnitt 1.3.2.2): *level of performance*. Je nach Leistung der Schülerinnen und Schüler fällt der Handlungs- und Erklärungsbedarf von Lehrkräften unterschiedlich aus, sodass nach Kuper und Diemer (2012) vermeintlich insbesondere bei schwächeren Leistungen erhöhter Bedarf besteht.

1.4.3.2 Schulische Bedingungen

Im Abschnitt zu individuellen Bedingungen wurde Altrichter (2010, S. 174) bereits zitiert: Er weist auf die „persönlichen und materiellen Ressourcen“ zur Nutzung von Evaluationsrückmeldungen hin. Während die persönlichen Ressourcen oben bereits näher betrachtet wurden, spielen auch die (materiellen) Ressourcen im schulischen Umfeld etwa in Form der schulischen Ausstattung eine Rolle (vgl. auch Posch, 2009; Schildkamp & Visscher, 2009; Krainz-Dürr, 1999; Lortie, 1977; O'Day, 2002; Rolff, 2001; Visscher & Coe, 2003; Weick, 1976; vgl. *school resources* in Abbildung 7).

Zu einem entscheidenden Teil beeinflusst das *Evaluations- und Kooperationsklima* an der Schule respektive innerhalb des Kollegiums sowie die *innovative und explorative Orientierung* die Nutzung von Rückmeldungen (Ditton, 2000; Helmke, 2004; Holtapfels, 2014b; van Ackeren & Klemm, 2011; vgl. *innovation attitude* in Abbildung 7). Vielfach fühlen sich Lehrkräfte als Einzelkämpfer und haben keine oder nur wenig Kenntnis darüber, was ihre Kolleginnen und Kollegen im Unterricht machen. Dies erzeugt die von Rolff (2007, S. 145) sogenannte *endemische Unsicherheit*. Unter Rückgriff auf Erhebungsergebnisse aus TALIS-GEW zeigt sich, dass über die bereits erwähnten etwa 15% der Lehrkräfte in Deutschland, die nie Feedback seitens der Schulleitung über Ihre Arbeit erhalten, der Anteil der Lehrkräfte, die von externen Institutionen nie Rückmeldung erhalten, sogar bei über 50% liegt (Barfknecht & Saldern, 2010, S. 97). In Kombination mit dem Gefühl keine Fehler machen zu dürfen (Angabe von 69% der befragten Lehrkräfte in Vodafone Stiftung Deutschland, 2017, S. 6), führt dies dazu, dass Evaluation „oft [...] mit Kontrolle und Fremdbestimmung assoziiert“ wird (Ditton & Arnoldt, 2004, S. 134) und Lehrkräfte diese nicht als Anlass zur Verbesserung und Reflexion wahrnehmen (O'Day, 2002, S. 296; Rolff, 2001). Der *endemischen Unsicherheit* kann eine gute Kommunikations- und Kooperationskultur an der Schule vorbeugen. Diese Kommunikation über den eigenen Unterricht und Hilfestellungen in der Unterrichtsplanung werden an vielen Stellen in der Literatur als hilfreich konstatiert. Rolff (2007, S. 148) fasst es noch deutlicher: „Ohne dass Lehrkräfte wie reflektierende Praktiker miteinander reden, sich gegenseitig beobachten und auch helfen, gibt es keine Schulentwicklung.“ Anzuführen wären hierbei der positive Einfluss von Kooperation auf das Belastungserleben von Lehrkräften (Schaarschmidt, 2005, 2007) und auf die Unterrichtsqualität beispielsweise in COACTIV (Klusmann, 2011; Kunter et al., 2011).

Bereits in TIMSS (Blum, 2001; Bundesministerium für Bildung und Forschung [BMBF], 2001; Klieme & Baumert, 2001) wurden Unterschiede zwischen den Verläufen japanischer, deutscher und amerikanischer Unterrichtsstunden festgehalten. In Japan fest etabliert ist die sogenannte *lesson study* (Bonsen & Frey, 2014). Eine ähnliche Kooperation regt das Konzept der *Professionellen Lerngemeinschaften* an (PLG). Nach Hord (2005) zeigt sich in Schulen mit PLGs erfolgreicherer Lernen seitens der Schülerinnen und Schüler sowie der Lehrkräfte, in Form von erhöhter Wahrscheinlichkeit von Fortbildungen, einer „signifikant höhere[n, S.V.] Kapazität, sich auf die besonderen Bedürfnisse der Schüler einzustellen und den eigenen Unterricht entsprechend

zu adaptieren“ (Bonsen & Rolff, 2006, S. 168). Analog zu Schaarschmidt (2005, 2007) geht vermehrte Kooperation mit einer höheren Berufszufriedenheit sowie höherer Motivation einher (vgl. *teacher collaboration* in Abbildung 7).

In Bezug auf die Nutzung von Evaluationsrückmeldungen kommt dem „Einzelkämpfertum“ insofern eine besonders kritische Rolle zu – und damit umgekehrt der Kooperation und Kommunikation innerhalb des Kollegiums sowie mit der Schulleitung eine besonders wichtige Rolle – da nach O'Day (2002, S. 295) die Akteure von möglichen Innovationen bzw. Maßnahmen aus der Evaluation Lehrkräfte sind, das Feld der Innovationen jedoch die Schule. Durch eine Kooperation können Teams von Lehrkräften zum „Motor der Entwicklung“ werden (Rolff, 2007, S. 27). Während eine einzelne Lehrkraft im *loosely coupled system* (Weick, 1976) Schule wenig Möglichkeiten zur umfassenden Veränderung hat, können Teams oder Professionelle Lerngemeinschaften *bottom-up*-Entwicklungen anstoßen (Blum et al., 2006; Hofmann-Göttig et al., 2005; Holtappels, 2014b; van Ackeren & Klemm, 2011).

Sinnvoll ist eine Verankerung einer kooperativen und kommunikativen Ausrichtung des Kollegiums im Schulprogramm, welches nach Helmke (2004) auch die Evaluation verbindlich machen sollte (vgl. *innovation capacity* und *innovation attitude* in Abbildung 7). Das Hessische Schulgesetz (HSchG) schreibt entsprechend in §127b Abs. 2 neben der Entwicklung eines Schulprogrammes auch die regelmäßige *interne* Evaluation vor (Hessisches Kultusministerium, 2016). Allerdings ist es grundsätzlich kritisch zu betrachten, wenn anstelle der geforderten *internen* Evaluation, die *bottom-up*-Entwicklungen anstoßen soll, externe Vorgaben gemacht werden: Im Falle einer solchen Intervention „entscheiden die Einzelsysteme, also die Schulen selbst, ob und wie sie diese Interventionen verarbeiten“ (Rolff, 2007, S. 22; Hofmann-Göttig et al., 2005), da u. a. kein oder nur bedingter Einblick in eine system- bzw. schulinnere Logik besteht (O'Day, 2002). Erschwerend kommt hinzu, dass Lehrkräfte nach Schildkamp und Kuiper (2010, S. 484) Unterricht als ihre Haupttätigkeit und -aufgabe ansehen und eine Nutzung und Durchführung von Evaluation scheinbar nicht selbstverständlich darin enthalten ist. Vor diesem Hintergrund erscheint die partizipatorische und demokratische Gestaltung von Evaluationen und Schulprogrammen umso wichtiger, da einerseits wegen der föderal bedingten, andererseits lokal bedingten (etwa unterschiedliche Schülerinnen und Schüler, anderes regionales Umfeld, andere Ausstattung, andere Lehrkräfte) Diversität von Schulen, *top-down*-Einheitslösungen (*one size*

fits all) von zentralen Behörden nur wenig geeignet erscheinen, der „Vielfalt und Differenziertheit der Schullandschaft“ gerecht werden zu können (Holtappels & Rolff, 2010, S. 75; Baumert, 2001a; Cuban, 1984; Desimone, 2006; Reimers et al., 2009; Schildkamp & Visscher, 2009).

Der Aspekt der Schulprogrammarbeit und ein im Modell von Helmke (2004) zumindest nicht explizit berücksichtigter Aspekt stehen sich jedoch sehr nahe: Die *Rolle der Schulleitung* prägt sämtliche schulischen Bedingungen selbstverständlich in enormem Maße (Korngiebel, 2009; Meyer & Demuth, 2009; Tillmann, 2005; Visscher & Coe, 2003). Trotz der nur losen Verkopplung hat die Schulleitung einen großen Einfluss darauf, ob und wie Evaluationsergebnisse genutzt werden. Dabei kann sie direkt, beispielsweise durch die Thematisierung von Evaluation auf Konferenzen, oder indirekt durch entsprechende Schulentwicklungsmaßnahmen (vgl. Abschnitt 1.5) Einfluss nehmen (Bonsen & Rolff, 2006; Bonsen, 2010; Burmeister et al., 2005; Holtappels, 2014a; Korngiebel, 2009; Meyer & Demuth, 2009).

Während Lehrkräfte sowie das Personal der Schulleitung direkt zur Schule gehören, spielen in der Nutzung von Evaluationsrückmeldungen Eltern und Schülerinnen und Schüler eher eine „externe“ Rolle, die jedoch nicht minder wichtig ist. Das Modell von Schildkamp und Kuiper (2010) zur Nutzung von Feedback (Abbildung 7) greift in diesem Punkt kürzer als Helmkes (2004) Modell. Rolff (2001, S. 347) weist jedoch bereits darauf hin, dass zur Erreichung einer wirksamen Schulentwicklung bzw. Unterrichtsentwicklung „die Befunde zur Mathematikleistung und deren Kontext allen zur Verfügung stehen [müssen, S.V.], die für die Schulentwicklung verantwortlich sind. [...] (1) die Mathematik-Lehrkräfte, (2) die Schulleitungen und (3) die Schulaufsicht sowie (4) die Eltern.“

1.4.3.3 Externe Bedingungen

Über die schulischen und individuellen Bedingungen hinaus beeinflussen einige externe Bedingungen die Nutzung von Evaluationsrückmeldungen. Dabei lassen sich im Grunde alle von Helmke (2004) in Abbildung 10 aufgeführten Aspekte unter dem Begriff *externe Unterstützung* subsumieren. Bereits bei den individuellen Bedingungen wurde angeführt, dass für das Verständnis von Rückmeldungen ein gewisses Vorwissen pädagogisch-psychologischer sowie statistischer Aspekte notwendig ist. Betrachtet man die Standards für die Lehrerbildung (KMK, 2004c, S. 6), so zeigt sich in den geforderten Kompetenzbereichen für die Aus- und Fortbildung von Lehrkräften, dass

Entsprechendes vorgesehen ist: Beispielsweise sollen Lehrkräfte Kompetenzen im Bereich „Diagnostik, Beurteilung und Beratung“ erlangen, also insbesondere Fertigkeiten und Fähigkeiten im Bereich der Leistungsmessung und -beurteilung. Dies ist sicherlich notwendige Voraussetzung für Klassenarbeiten, umfasst aber ebenso statistisch anspruchsvollere Evaluationsergebnisse: Lehrkräfte sollen mit Methoden der Bildungsforschung sowie der Interpretation der Ergebnisse entsprechender statistischer Berechnungen umgehen können, was letztlich direkt auch auf Evaluationsverfahren, wie beispielsweise PISA, den IQB-Ländervergleich oder die Vergleichsarbeiten, zutrifft. Diese geforderten Kompetenzen sind jedoch vielfach bei Lehrkräften nicht vorhanden und kommen in der Lehrerbildung auch nicht vor (Blum, 2001, S. 79), obwohl etwa statistische Grundbegriffe für Evaluationsdiagnostik notwendig sind (Arnold, 2007), was schlussendlich zu den bereits angesprochenen „hochkomplexen Rezeptionsproblemen“ führen kann (Rolff, 2001, S. 343).

Trotz dieser Berücksichtigung in den Standards für die Lehrerbildung ist „eine externe Unterstützung [...] bei der Gewinnung einer verlässlichen, objektiven Datenbasis nötig, da in der Schule selbst oft nicht die Zeit, das erforderliche Wissen und die Methodenkompetenz dafür“ vorhanden sind (A. Hosenfeld, 2010, S. 19; Ditton & Arnoldt, 2004). Diese Unterstützung kann dabei sowohl durch Moderierende und Qualitätsberatende wie beispielsweise Unterrichtsentwicklungsberatung in Hessen¹⁵ erfolgen oder durch entsprechende Unterstützung seitens der Wissenschaft oder Schulaufsicht bzw. Landesinstitute (Posch, 2009).

Solche theoretischen Rahmenmodelle und auch exemplarischen Kreisläufe können für schulische Evaluationen sicherlich nur bedingt zur direkten Nutzung geeignet sein. Je komplexer ein Kreislauf aufgebaut ist, umso unwahrscheinlicher dürfte die Nutzung im Schulalltag vermutlich ausfallen. Ähnlich dem Modellierungskreislauf dienen komplexere Modelle (vgl. etwa Abbildung 9 oder Abbildung 10) jedoch dem Aufstellen von Hypothesen und dem Beforschen von Schulentwicklung, während einfacher gehaltene Modelle für den alltäglichen Einsatz in der Schule geeignet sind (vgl. etwa Abbildung 8; I. Hosenfeld et al., 2006).

¹⁵ vgl. Webauftritt der Schulämter Hessen (beispielsweise <https://schulaemter.hessen.de/schulqualitaet/unterrichtsentwicklungsberatung>; Zugriff am 14.03.2017)

Teil 1 – Theorie

Evaluation

Mit den Entwicklungen der letzten Jahre seit PISA und TIMSS ist die Schule als eigenverantwortliche Institution immer wichtiger geworden. „Immerhin hat sich inzwischen die Einsicht gefestigt, daß mehr Autonomie für die einzelnen Schulen nur in Verbindung mit begleitenden Evaluationsprogrammen verantwortbar ist, wenn quasi im Sinne einer *Nachweispflicht* eine Überprüfung vorgenommen wird“ (Ditton, 2000, S. 74). Damit einher geht jedoch neben der angesprochenen *Nachweispflicht* auch die Anforderung einer stetigen Evaluation, welche in Maßnahmen zur Schulentwicklung mündet und diese gleichsam kritisch bilanzierend prüft.

1.5 Schulentwicklung

Das geschwisterliche Verhältnis von Schulentwicklung und Rechenschaftsablegung war vermutlich vielen Schulentwicklern in der Anfangsphase nicht klar. (Baumert, 2001b, S. 16)

oder

*Gelingende Schulentwicklung hat zum ultimativen Ziel, die Lern-
gelegenheiten der Schülerinnen und Schüler zu verbessern.
(Rolff, 2010, S. 35)*

1.5.1 Definition und Forschungsstand

Die Lerngelegenheiten von Schülerinnen und Schülern sind nach Rolff (2010) der Gradmesser für Schulentwicklung. Damit wird der Fokus der Entwicklungsprozesse – auch für andere bildungspolitische Akteure als Lehrkräfte – auf den „Output“ von Bildungsprozessen gelegt. Aus den vorherigen Ausführungen ist deutlich geworden, dass ein Paradigmenwechsel in der Steuerung von Bildungsprozessen vorgenommen wurde bzw. anders formuliert: „von zentraler Bildungsplanung zur Einzelschule als Gestaltungseinheit“ (Rolff, 2001, S. 338). Den obigen Ausführungen entsprechend funktioniert „eine gezielte und gesteuerte Schulentwicklung [...] nur, wenn Rückmeldungen über den Stand eingeholt, darüber reflektiert und Steuerungsmaßnahmen ergriffen werden“ (Ramm, 2014, S. 46), wobei diese Maßnahmen als Einzelschulsteuerung und nicht als *one-size-fits-all*-Lösung respektive *top-down*-Entwicklungen zu sehen sind (vgl. Abschnitt 1.1.3; Blum et al., 2006; Landwehr, 2011; Meyer & Demuth, 2009; Oelkers & Reusser, 2008; Reimers et al., 2009; van Ackeren & Klemm, 2011). Durch diesen Wechsel werden jedoch die Dinge nicht unbedingt leichter: Mit einer klaren Vorgabe, *was* im Unterricht *wann* gemacht werden muss, und der Sicherheit an der Struktur der Schule keine Mitverantwortung zu tragen, ist die Aufgabe von Lehrkräften, den Unterricht für Schülerinnen und Schüler zu gestalten, sicherlich weniger komplex als bei umfänglicher (Mit-)Verantwortung für die Institution im Sinne einer *eigenverantwortlichen Schule* sowie der Vorgabe von verbindlichen Kompetenzanforderungen für Schülerinnen und Schüler nach Abschluss eines bestimmten Bildungsabschnittes (O'Day, 2002; Baumert, 2001b). Das Kerngeschäft von Lehrkräften ist der Unterricht: Dies ist der primäre Fokus, den Lehrkräfte haben (Rolff, 2010; Schildkamp & Kuiper, 2010), und genau aus diesem Grund geht mit den Veränderungen im Bildungswesen und einer erhöhten Autonomie von Einzelschulen die

Teil 1 – Theorie Schulentwicklung

Notwendigkeit einher, *Schulentwicklung* als „neue“ Aufgabe für Lehrkräfte zu beleuchten. Wenngleich Lehrkräfte und Schulleiter identische Informationen aus Evaluationen o. ä. zurückgemeldet bekommen, nutzen sie diese nach Schildkamp und Kuiper (2010, S. 494) in unterschiedlicher Weise: Lehrkräfte nutzen Informationen „nicht überraschend“ (Übersetzung durch Autor) zur Leistungsstandsdiagnose, zur Optimierung von Unterricht, Reflexion des eigenen Unterrichts und zur Unterstützung in Elterngesprächen, während Schulleiterinnen und Schulleiter Daten zur Überwachung von Entwicklungs-/Fortschrittsbereichen, Planung und Entwicklung der Institution, der Evaluation von Lehrkräften, öffentlichen Beziehungen und zur Beantwortung von *accountability*-Belangen (beispielsweise aus Schulinspektionen) nutzen. Insbesondere wegen dieser verschiedenen Fokusse ist jedoch ein gemeinsames Miteinander für die Entwicklung der Institution *Schule* unerlässlich (Bonsen & Rolff, 2006; Kuper & Hartung, 2007; Rolff, 2007).

Klar geregelt ist die Verantwortlichkeit für die bauliche Substanz einer Schule, die Bezahlung des nicht lehrenden sowie lehrenden Personals etc. Die Zuständigkeiten hierfür liegt bei Gemeinden oder Ländern (van Ackeren & Klemm, 2011). Die Entwicklung einer einzelnen Schule führt jedoch aufgrund der enormen Komplexität mit verschiedensten Personengruppen, inhaltlicher Breite etc. (Baumert, 2001b; O'Day, 2002; Weick, 1976) zu einer *Problemüberlast* für einen isolierten Personenkreis, der in sich zu homogen ist: Bereits 1986 sprach Fend (S. 284) davon, dass die Wahrnehmung einer *Problemüberlast* gute von schlechten Schulen differenziert. Daher empfiehlt es sich, das Thema *Schulentwicklung* anhand einer theoretischen Grundlage „handhabbar“ zu machen, also das Innovations- und Entwicklungsfeld zu reduzieren bzw. zu unterteilen (Baumert, 2001b; Klippert, 1997; Rolff, 2007).

Im Sinne dieser „Handhabbarmachung“ von Schulentwicklung (SE) wird von Rolff (2007, 2010) diese als Trias aus *Personal-* (PE), *Organisations-* (OE) und *Unterrichtsentwicklung* (UE) umgeben vom zugehörigen schulischen/organisationalen Umfeld definiert. Essenziell bei diesem Modell ist die Tatsache, dass die einzelnen Teilbereiche direkt miteinander in Bezug stehen und eine Entwicklung lediglich in einem der drei Teilbereiche somit nicht denkbar ist (Bischof et al., 2013). Anzumerken ist dabei, dass dieses Modell sich vorrangig auf die Entwicklung einer Einzelschule bezieht und SE im weiteren Sinne eine Systementwicklung außer Acht lässt. Veränderungen auf

höheren Ebenen müssen jedoch mit Einzelschulentwicklung einhergehen (Rolff, 2010).

Im Folgenden sollen die einzelnen Teile der Schulentwicklung näher vorgestellt werden. Dabei wird vor allem auf die beiden Aspekte *Lehrerfortbildung* (als Teil der PE) und *Unterrichtsqualität* detaillierter eingegangen, während die Organisationsentwicklung lediglich angerissen wird.

Zwar ist SE als Ganzes nicht von einzelnen Lehrkräften leistbar, doch kann von jeder Lehrkraft individuell der Anstoß zur Veränderung und Entwicklung gegeben werden, sodass als Ausgangspunkt aller Teilaspekte von SE hier die PE gewählt wird (Rolff, 2007).

1.5.2 Personalentwicklung

Nach Baumert und Kunter (2011, S. 32) unterteilt sich die *professionelle Kompetenz* von Lehrkräften in die Aspekte (1) Überzeugungen/Werthaltungen/Ziele, (2) Motivationale Überzeugungen, (3) Selbstregulation und (4) das Professionswissen (vgl. auch Abschnitt 1.4.3.1). Im Rahmen von COACTIV wurden diese Bereiche professioneller Kompetenz bei Mathematiklehrkräften und angehenden Mathematiklehrkräften untersucht (Baumert et al., 2009; Blum, Krauss & Neubrand, 2008; Dubberke, Kunter, McElvany, Brunner & Baumert, 2008; Krauss et al., 2008; Kunter et al., 2011). In ähnlicher Weise untersuchen auch die verschiedenen TEDS-Studien (**T**eacher **E**ducation and **D**evelopment **S**tudy in Mathematics) den Einfluss von Professionalität auf den Lernerfolg (vgl. u. a. Blömeke, Kaiser & Lehmann, 2010). Beide Studien sprechen nicht ausschließlich von Fachwissen als notwendigem Teil von Professionalität, sondern bewusst von Kompetenz, also der Fähigkeit, „unter Bezugnahme auf wissenschaftliches Wissen zur kompetenten Bewältigung von alltagspraktischen Problemen beizutragen“ (Kuper & Hartung, 2007, S. 217; vgl. auch Ball, Thames & Phelps, 2008; Shulman, 1986). Mit „alltagspraktischen Problemen“ ist für Lehrkräfte zweifelsfrei vor allem ihr „Kerngeschäft“, das Unterrichten (Schildkamp & Kuiper, 2010, S. 484), gemeint. In diesem Sinne stellt eine *professionelle Kompetenz* eine Grundvoraussetzung von gutem Unterricht dar.

Bezugnehmend auf Weinert (2001, 27f.) stellen *Kompetenzen* „die bei Individuen verfügbaren oder von ihnen erlernbaren kognitiven Fähigkeiten und Fertigkeiten“ dar. Damit wird deutlich, dass *Kompetenz* als etwas Veränderbares und mit der

Möglichkeit zur Entwicklung angesehen werden kann. Folglich ist eine gelingende Unterrichtsgestaltung durch Fortbildungen, durch den Aufbau und die Förderung neuer bzw. die Erweiterung vorhandener Kompetenzen erreichbar (zumindest im Sinne des Angebots guten Unterrichts; vgl. Abbildung 12; Lipowsky, 2009, S. 81; Helmke, 2009). Damit wird die direkte Verbindung von UE und PE im Sinne einer persönlichen, individuellen Weiterqualifikation deutlich (Rolff, 2007).

Es ist also Aufgabe von Lehrerinnen und Lehrern (Hoffmann & Richter, 2016), eine stetige Weiterqualifizierung anzustreben (vgl. *professionelles Selbstverständnis*). Darüber hinaus bedarf es aufgrund der Eigenverantwortlichkeit, Outputsteuerung und ähnlicher bereits beschriebener Veränderungen im Bildungswesen sowie der sich wandelnden Gesellschaft (beispielsweise die gestiegene Relevanz von Medienbildung und der Nutzung von Medien im Fachunterricht; KMK, 2012c, 2016a) einer veränderten Lehrerrolle (Klieme, Schlümer & Knoll, 2001, S. 52). Ebenfalls notwendig sind neue Kompetenzen, die über die Inhalte eines reinen Lehramtsstudiums hinausgehen und im Rahmen der Personalentwicklung, konkret der Lehrerfortbildung, erworben werden können (Arnold, 2007, S. 456; Kuhn, 2014, S. 422).

1.5.2.1 Fortbildung

Fortbildung ist also ein wesentlicher Teil von PE und damit von SE. Allerdings gilt nach Rolff (2010, S. 32) dabei lediglich die Schlussfolgerung, dass SE immer Lehrerfortbildung bedeutet. Die Umkehrung dieser Aussagen ist dabei nicht gültig, da eine individuelle Fortbildung nicht zwingend Auswirkungen auf OE und UE hat. Die Lehrkräfte in Deutschland werden im internationalen Vergleich zwar recht lange aus-, jedoch nur wenig fortgebildet (Fortbildungstage in Deutschland: neun Tage/Schuljahr; International: 15 Tage/Schuljahr; Institut für Sozialwissenschaftliche Studien (ISS) e.V. & Gewerkschaft Erziehung und Wissenschaft (GEW), 2009, S. 10; Goddar, 2010). Von der Vodafone Stiftung Deutschland (2017, S. 1) werden Ergebnisse einer Befragung dahingehend interpretiert, „dass das System Schule die Weiterbildungsbemühungen von Lehrern nicht ausreichend unterstützt und anerkennt.“

Es soll nun vor allem auf die Wirksamkeit von Fortbildungen eingegangen werden. Dabei ist jedoch zunächst eine Unterscheidung von Lerngelegenheiten (als weitergefasster Begriff der Fortbildung) vorzunehmen: Generell stellt das tägliche Arbeitsumfeld bereits eine Lerngelegenheit dar. Hierbei spricht man nach Richter (2011, S. 317) von *informellen Lerngelegenheiten*. Demgegenüber handelt es sich bei *nonformalen*

Lerngelegenheiten um bewusste, intentionale Fortbildungen, die jedoch außerhalb klassischer Bildungseinrichtungen erfolgen (etwa in Lerngemeinschaften im Kollegium oder individueller Fachliteratur- bzw. Internetrecherche). Unter *formalen Lerngelegenheiten* werden stereotype Fortbildungen beispielsweise an der Universität verstanden (vgl. Desimone, 2009).

Wenngleich *nonformale* und *informelle Lerngelegenheiten* ein äußerst interessantes Feld für Forschung bilden, soll sich im Weiteren auf die *formellen Lerngelegenheiten* konzentriert und der Frage nachgegangen werden, wie wirksam diese sind.

Bevor detaillierter auf die Wirksamkeit und die empirischen Ergebnisse hierzu eingegangen wird, sollen zunächst zwei unterschiedliche Hypothesen zum Fortbildungsbesuch von Lehrkräften vorgestellt werden (Cookson, 1986; Desimone, 2006; Richter, 2011):

1. Neigungshypothese (*catalytic function*)
2. Kompensationshypothese (*educative function*)

Während die *Neigungshypothese* davon ausgeht, dass Lehrkräfte vor allem Fortbildungen zu Themen besuchen, welche sie interessieren und/oder sie bereits auf eine gute Wissens- und Kompetenzbasis zurückgreifen können, besagt die *Kompensationshypothese*, dass Lehrkräfte eher zu Themen Fortbildungen suchen, bei denen sie sich unsicher fühlen bzw. Entwicklungsnot erkennen. Erstgenannte Hypothese spiegelt den Matthäus-Effekt („Wer hat, dem wird gegeben“) wider.

Bei diesen Hypothesen liegt die Annahme zugrunde, dass die Entscheidung zu einer Fortbildung individuell und ausschließlich von der betreffenden Lehrkraft allein getroffen wird. Diese Annahme ist vor dem Hintergrund von komplexeren Personalstrukturen und häufiger werdenden Fach- oder Jahrgangsteams jedoch sicherlich nicht immer haltbar (Ahlring & Brömer, 1999; Tuschel, 2006). Es scheint daher unwahrscheinlich, alle Fortbildungsentscheidungen monokausal begründen zu können; vielmehr dürfte ein komplexes System dahinterstehen, wodurch wiederum der interne Bezug der PE zur OE und UE als „Summanden“ der SE deutlich wird.

Teil 1 – Theorie Schulentwicklung

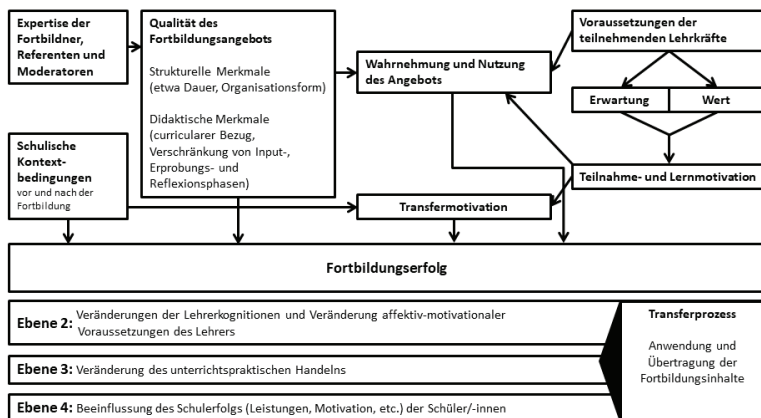


Abbildung 11: Angebots-Nutzungs-Modell der Lehrerfortbildung (vereinfacht nach Lipowsky, 2014, S. 473)

Lipowsky (2014) verknüpft in diesem Modell ähnlich dem Angebots-Nutzungs-Modell für Unterricht (Abbildung 12) verschiedene Bereiche und Theorie (u. a. Erwartungs-Wert-Theorie; Atkinson, 1957; Wigfield & Eccles, 2000) zu einem holistischen Erklärungsgefüge für Fortbildungserfolg. Dieser lässt sich dabei auf vier Ebenen von direkten Reaktionen/Emotionen der teilnehmenden Lehrkräfte zur Fortbildungsveranstaltung an sich (in der Abbildung nicht ausgewiesen), über veränderte Kognitionen sowie unterrichtspraktisches Handeln bis zur veränderten Leistung von Schülerinnen und Schülern ablesen.

Betrachtet man das Modell von links beginnend, fallen zunächst die Angebotsqualität in Abhängigkeit der Expertise der Fortbildner in den Blick. In Anlehnung an den Sozialkonstruktivismus (Vygotskii, Cole, Stein & Sekula, 1978) beeinflussen die Merkmale der fortbildenden Person(en) die Fortbildungsnutzung/den Fortbildungserfolg. Analog zur Situation in der Klasse nehmen in der Fortbildungssituation Lehrkräfte die Position der Schülerinnen und Schüler ein. Daher lassen sich die Ansprüche an die *professionelle Kompetenz* von Lehrkräften nahezu eins zu eins auf die *professionelle Kompetenz* von Fortbildenden übertragen. Die Fortbildenden entscheiden über die Konzeption und Gestaltung von Fortbildungen, stellen somit ein Angebot. Im weniger detaillierten, linearen Modell von Cookson (1986) zur Erwachsenenbildung wird dieser Aspekt unter *external context* als Eingangsbedingung gefasst. Desimone (2006, S. 207) fordert in diesem Zusammenhang die Fortbildenden auf: „stop providing low-quality professional development“. Diese Forderung erhält durch die Ergebnisse von

TALIS-GEW Unterstützung, da vielfach von Lehrkräften angegeben wird, „keine passende Fortbildung“ (über 80% der Angaben bei der Frage nach Hinderungsgründen) gefunden zu haben (Institut für Sozialwissenschaftliche Studien (ISS) e.V. & Gewerkschaft Erziehung und Wissenschaft (GEW), 2009, S. 10; Kast, 2010). Von knapp 50% der Lehrkräfte im IQB-Ländervergleich wird korrespondierend angegeben, dass aus Fortbildungen wenig praktischer Nutzen gezogen werden könne, sie wenig Neues bieten würden (47.0%) oder die Teilnahme nicht zur Bewältigung beruflicher Aufgaben beitrage (41.0%; Richter, Kuhl, Haag & Pant, 2013).

Neben den persönlichen Merkmalen der Fortbildenden nehmen im Sinne eines Angebots-Nutzungs-Verständnisses (Abbildung 11) selbstverständlich insbesondere auch die persönlichen Eigenschaften der teilnehmenden Lehrkräfte als Nutzende erheblichen Einfluss auf die Wirkung von Fortbildungen. Lipowsky (2014) untergliedert dabei *motivationale, kognitive, volitionale, berufsbiografische/private Voraussetzungen* sowie *Persönlichkeit*. Die Motivation von Lehrkräften an einer Fortbildung teilzunehmen lässt sich zunächst grundlegend auf die Selbstbestimmungstheorie nach Deci und Ryan (1993) beziehen, der zu Folge bei (1) *Autonomieerleben*, (2) *Kompetenzerleben* und (3) *Erleben sozialer Eingebundenheit* eine höhere Motivation vorliegt (vgl. auch Deci, 1972; Deci & Ryan, 2004; Ryan & Deci, 2000). Dies gilt ebenso für Fortbildungssituationen wie auch im Unterrichtskontext für die Motivation von Schülerinnen und Schülern. Im unterrichtlichen Zusammenhang ebenfalls vielfach untersucht ist das Erwartungs-Wert-Modell (Atkinson, 1957; Wigfield & Eccles, 2000). Nach diesem Modell steigert sich die Motivation, wenn *erwartet* wird mit einer Situation erfolgreich umgehen zu können (vgl. Selbstwirksamkeit und Selbstkonzept¹⁶) und dabei dieser Situation einen *Wert* zugeschrieben wird, etwa durch persönliches Interesse, durch Karriereorientierung bzw. externe Erwartungen oder beispielsweise die Aussicht auf soziale Interaktion (Kao, Wu & Tsai, 2011; Rzejak et al., 2014).

Neben der Motivation spielen die *kognitiven Voraussetzungen* eine große Rolle. Im Umgang mit Rückmeldungen (als *nonformale* Lerngelegenheiten einzustufen) zeigte sich bei A. Hosenfeld (2010, S. 44) beispielsweise, dass insbesondere die Personen mit Rückmeldungen umgehen, die wenig Anlass zu Veränderungen geben und

¹⁶ vgl. u. a. Bong und Skaalvik (2003); Marsh (1986); Marsh, Byrne und Shavelson (1988); Shavelson, Hubner und Stanton (1976); B. Zimmerman (2000)

„ohnehin schon neue Wege beschreiten“. Dieser Befund unterstreicht die *Neigungshypothese*. Ähnliches zeigt sich bei Desimone (2006): Mehr Fachwissen führt zu einer größeren Fortbildungsbereitschaft und -nutzung. Das Ausmaß an Fortbildungsveranstaltungen korreliert hoch signifikant und negativ mit dem Nichtvorhandensein eines höheren Abschlusses ($r = -.23, p < .01$) sowie mit dem Fokus von Mathematik in der Ausbildung ($r = .67, p < .01$). Ebenfalls hoch signifikante Korrelationen finden sich zwischen der Fortbildungsnutzung und der Abschlussnote in Mathematik ($r = .16, p < .01$) sowie dem Gefühl der *preparedness* ($r = .27, p < .01$). Diese Korrelationen unterstützen den Ansatz der *Neigungshypothese*, da vor allem gut ausgebildete Lehrkräfte scheinbar mehr Fortbildungen besuchen. Als persönliche Hinderungsgründe werden von vielen Lehrkräften familiäre Gründe angeführt, sodass etwa Fortbildungszeiten mit der Familie einschränken würden (41.2%) oder keine Kinderbetreuung gefunden wurde (17.2%; Richter et al., 2013).

1.5.2.1.1 Ebenen von Fortbildungserfolg

Bevor Merkmale wirksamer Fortbildungen näher betrachtet werden, stellt sich zunächst die Frage, ob Fortbildungen *per se* einen Einfluss auf eine der vier genannten Ebenen haben. In diversen Meta-Analysen wird der Fortbildung grundsätzlich ein positiver Effekt (wenn auch in unterschiedlichem Ausmaß) zugesprochen (zsf. Lipowsky & Rzejak, 2012; Lipowsky, 2014). Wade (1984, 1984/1985) berichtet von einem großen Einfluss ($d = 0.90$) von *inservice teacher training* auf das Lernen von Lehrkräften. Dieser Effekt wird jedoch bei anderen Zieldimensionen etwas geringer: $d = 0.60$ auf das Verhalten von Lehrkräften und $d = 0.37$ auf die Leistungen von Schülerinnen und Schüler. Diese rückläufigen Effektstärken sind vor dem Hintergrund des Angebots-Nutzungs-Modells von Unterricht (Abbildung 12) verständlich. Je indirekter der Einfluss der Fortbildung auf die jeweilige Untersuchungsgröße ausfällt, desto kleiner sind die zu erwartenden Effekte. Im Modell der Fortbildungsnutzung (Abbildung 11) deutet sich diese Tatsache durch die verschiedenen Ebenen an. Im Bereich mittlerer Effekte bewegen sich auch viele der jüngeren Meta-Analysen: $d = 0.54$ für den Leistungszuwachs bei Schülerinnen und Schülern (Yoon, Duncan, Lee, Silvia Wen-Yu, Scarloss & Shapley, 2007), $d = 0.66$ allgemein und $d = 0.50$ für Mathematik bei Timperley, Wilson, Barrar und Fung (2007) sowie $d = 0.62$ in der Meta-Analyse von Hattie (2009).

Bereits in den kurzen Ausführungen zu vorhandenen Meta-Analysen lässt sich erkennen, dass Fortbildungserfolg oder -wirksamkeit auf verschiedenen Ebenen gemessen werden kann. Analog zu den in Abbildung 11 genannten Ebenen differenzieren auch Yoon et al. (2007) und Desimone (2009) die Veränderungen im Wissen und den Fähigkeiten von Lehrkräften, also deren *professionelle Kompetenz*, vom tatsächlichen Unterrichtsablauf und den daraus resultierenden Leistungen von Schülerinnen und Schülern. Diese Differenzierung ist im Sinne des Angebots-Nutzungs-Modells für den Unterricht schlüssig nachvollziehbar (vgl. Abschnitt 1.5.4; Abbildung 12). Im Folgenden sollen die Effekte von Fortbildungen auf den einzelnen Ebenen differenziert betrachtet werden.

Ebene 1: Reaktionen und Einschätzungen der teilnehmenden Lehrpersonen

Direkt im Anschluss an eine Fortbildungsveranstaltung werden vielfach Instrumente eingesetzt, um die direkte Zufriedenheit und Wahrnehmung der Lehrkräfte zu erfassen. Diese Evaluationsmethoden sind dabei nicht immer elaborierten wissenschaftlichen Standards entsprechend; es zeigt sich bei korrelativer Analyse kein Zusammenhang zwischen diesem unmittelbar erhobenen, sehr kontextspezifischen Merkmal und dem Wissenszuwachs oder dem unterrichtlichen Handeln von Lehrkräften (zsf. Lipowsky & Rzejak, 2012; vgl. auch Colquitt, LePine & Noe, 2000). Generell wird jedoch eine hohe Akzeptanz der Lehrkräfte als notwendige Bedingung für Wirkungen auf höheren Ebenen angesehen und es zeigt sich entsprechend der Selbstbestimmungstheorie, dass Partizipationsmöglichkeiten – als Merkmal einer erhöhten *Autonomie* – die Zufriedenheit der Lehrkräfte mit Fortbildungen steigert (Lipowsky, 2010; Nir & Bogler, 2008).

In einer Studie zur Nutzung von Rückmeldungen in schriftlicher gegenüber videobasierter Form als eine Art non- oder informelle Lerngelegenheiten¹⁷ findet A. Hosenfeld (2010, S. 149) eine signifikante Korrelation zwischen den erfüllten Erwartungen an das Rückmeldevideo und der Reflexion der Ergebnisse sowie der Ableitung von Aktionen beim zugrundeliegenden Modell der Evaluation nach Helmke (Abbildung 10; $r = .54, p < .001$, $r = .49, p < .001$). Die Zufriedenheit mit dem Video korreliert danach ebenfalls signifikant mit der Rezeption der Rückmeldung ($r = .32, p < .05$). Auch für das schriftliche Feedback finden sich signifikante Zusammenhänge der

¹⁷ Je nach schulischem Kontext, in denen die Rückmeldungen eingebunden sind, ist eine genauere Bestimmung hier nicht möglich.

Zufriedenheit mit der Rezeption ($r = .56, p < .001$) und der Reflexion ($r = .25, p < .05$). Kritisch anzumerken ist, dass die Konstrukte durchweg mittels Fragebögen als Selbstauskünfte der Lehrkräfte erhoben wurden.

Ebene 2: Erweiterung der Lehrerkognitionen

Nicht mehr unmittelbar erfassbar sind durch Fortbildungen induzierte Veränderungen in den Kognitionen der Lehrkräfte bzw. deren *professionellen Kompetenz*. Die Wirksamkeit auf dieser Ebene ist dennoch in vielen Studien belegt und bereits oben erwähnt worden (u. a. Wade, 1984, 1984/1985). Durch Fortbildungen lässt sich das Fachwissen von Lehrkräften signifikant steigern: exemplarisch sei hier etwa die Fortbildung von Mathematiklehrkräften zum Wissen über Addition und Subtraktion sowie das Denken der Schülerinnen und Schüler hierüber angeführt. In der Begleitforschung dieser Fortbildung zeigen Carpenter, Fennema, Peterson, Chiang und Looef (1989, S. 521) u. a. signifikante Mittelwertunterschiede zwischen Projekt- und Kontrollgruppe im Wissen über number fact strategies ($t(38) = 3.51, p < .05$) und problem-solving strategies ($t(38) = 3.56, p < .05$) zugunsten erstgenannter Projektgruppe auf. Ähnliche Effekte lassen sich bei anderen Autoren auch für weitere Domänen nachweisen, so etwa bei Gersten, Dimino, Jayanthi, Kim und Santoro (2010) für Lehrkräfte der ersten Klasse im Bereich Leseverständnis und Vokabular sowie für das neuseeländische Literacy-Projekt in McDowall et al. (2007) und Parr, Timperley, Reddish, Jesson und Adams (2007). Wie eingangs beschrieben, gehören zur *professionellen Kompetenz* von Lehrkräften in der Operationalisierung nach COACTIV neben dem reinen Wissen auch Beliefs und Sichtweisen von Lehrkräften. In der bereits oben angesprochenen Fortbildung zur Addition und Subtraktion zeigen sich varianzanalytisch auf zwei Sub-Skalen¹⁸ sowie für die Gesamtskala der Beliefs signifikante Interaktionen zwischen Zeit und Treatment zugunsten der Projektgruppe, sodass „CGI teachers were significantly more cognitively guided in their beliefs than were control teachers“ (CGI steht für „cognitive guided instruction“; $F_{\text{Skala2}} = 9.80^{19}, p < .01$; $F_{\text{Skala4}} = 5.26, p < .05$; $F_{\text{Gesamt}} = 9.16, p < .01$; Carpenter et al., 1989, S. 523). In einem schuljahresbegleitenden Mathematikprojekt mit zehn Lehrkräften der zweiten Jahrgangsstufe zu

¹⁸ Skala 2 beschreibt die Zusammenhänge zwischen Fähigkeiten, Verständnis und Problemlösen. Skala 4 adressiert die Rolle der Lehrperson. Skala 1 stellt die Rolle des Lernenden in den Fokus und Skala 3 richtet sich auf Unterrichtsabfolgen in Bezug auf Addition und Subtraktion.

¹⁹ Leider fehlen die Hinweise auf Freiheitsgrade in der Originalquelle.

soziokonstruktivistischen Lerntheorien und damaligen Empfehlungen der NCTM²⁰ zeigt sich mittels *t*-Test eine verstärkte soziokonstruktivistische Sichtweise der Lehrkräfte am Ende des Projektes: $t(17) = 36.66, p < .001$ (Cobb et al., 1991).

Mit Verweis auf das Angebots-Nutzungs-Modell (Abbildung 12) von Unterricht muss hier mahndend ergänzt werden, dass eine Veränderung der Kognitionen von Lehrkräften keinesfalls gleichbedeutend mit der Veränderung von Unterrichtspraktiken oder gar dem Leistungs-/Lerngewinn von Schülerinnen und Schülern ist (Carpenter et al., 1989; Lipowsky, 2009; Timperley et al., 2007). Ebenfalls einschränkend bzw. als Forschungsdesiderat zu nennen, ist die bisher weitgehende Vernachlässigung der Untersuchung affektiv-motivationaler sowie selbstregulatorischer Outcomes von Fortbildungen auf Seiten der Lehrkräfte als weitere Aspekte von *professioneller Kompetenz* (Baumert & Kunter, 2011; Lipowsky & Rzejak, 2012).

Ebene 3: Unterrichtspraktisches Handeln

Wünschenswerte Konsequenzen aus veränderter *professioneller Kompetenz* von Lehrkräften stellen Veränderungen im Unterricht dar. Wie erwartbar, zeigen sich auch für diese Ebene positive Effekte von Fortbildungen, wobei jedoch angemerkt werden muss, dass mit steigender Ebene die direkte Zugänglichkeit der Messung erschwert wird. Insbesondere die Messung von Veränderungen im Unterricht gestaltet sich sowohl theoretisch als auch personell kompliziert, während die Erfassung von *professioneller Kompetenz* mittels Fragebögen, Wissenstest oder Vignetten noch vergleichsweise einfach zugänglich ist (Lipowsky & Rzejak, 2012). Nach Durchlaufen einer Fortbildung beobachteten geschulte Ratende im Unterricht eine häufigere Nutzung mehrerer Strategien ($t(38) = 3.76, p < .01$), häufiges Zuhören beim Lösungsprozess der Schülerinnen und Schüler ($t(38) = 3.54, p < .01$) oder ein häufigeres Stellen von Probleme ($t(38) = 2.45, p < .05$; Carpenter et al., 1989; vgl. zsf. Desimone, 2006; Lipowsky & Rzejak, 2012; Lipowsky, 2014; Gersten et al., 2010). Prinzipiell zeichnet sich insbesondere für die aktuellen MINT-Lehrkräfte ein positives Bild, da der Großteil der Lehrkräfte angibt, Fortbildungsinhalte im Unterricht umzusetzen (24% der Lehrkräfte geben an Inhalte umfassend, 57% ein bisschen angewendet zu haben; wobei hier lediglich die Frage *ob* und nicht *wie* thematisiert wird; forsa Politik- und Sozialforschung GmbH, 2017).

²⁰ National Council of Teachers of Mathematics

Ebene 4: Effekte auf Schüler/innen

Distales Ziel von Bemühungen im schulischen Kontext sollte grundsätzlich die Verbesserung des Lernens von Schülerinnen und Schülern sowie deren Leistung sein: Konsequenterweise wird dies als vierte Ebene im Modell von Lipowsky (2014) ebenfalls aufgenommen. Erfreulicherweise lassen sich auch auf dieser Ebene Effekte von Fortbildungen finden, wenngleich die Schülerinnen und Schüler nur mittelbar von der Fortbildung betroffen sind: Für Lehrkräfte stellt die Fortbildung eine zum Unterricht für Schülerinnen und Schüler analoge Lernsituation dar; insofern sind für Effekte auf Schülerebene zwei Lernsituationen erfolgreich zu gestalten (Sleeter, 2014; zsf. Lipowsky & Rzejak, 2012; Lipowsky, 2014; Timperley et al., 2007). Dass dies gelingen kann, zeigen verschiedene Projekte und Forschungsergebnisse: Auch hier lässt sich das neuseeländische *Literacy*-Projekt mit großen Effekten anführen, die vor allem für schwächere Schülerinnen und Schüler nachzuweisen sind (McDowall et al., 2007; Parr et al., 2007). In der bereits für Ebene 2 herangezogene Untersuchung von Carpenter et al. (1989) wird von positiven Effekten in ausgewählten Bereichen berichtet: Schülerinnen und Schüler der Projektgruppe erreichen höhere Mittelwerte für die Bereiche *number facts* mit $t(37) = 2.23, p < .05$ sowie für *complex addition/subtraction* mit $t(37) = 1.99, p < .05$ (bei einseitiger Testung) und im Interview $t(37) = 2.51, p < .05$. Da die Schülerinnen und Schüler sich jedoch in einer standardisierten Testleistung nicht unterscheiden, lässt sich das Ergebnis eher als vertieftes Verständnis interpretieren. Ein ähnliches Ergebnis zeigt sich bei Cobb et al. (1991) in qualitativen Auswertungen mit einem höheren konzeptuellen Verständnis der Schülerinnen und Schüler trotz fehlender Leistungsunterschiede. Weitere Belege für Effekte von Fortbildungen auf Ebene der Schülerinnen und Schüler lassen sich u. a. bei Cardelle-Elawar (1995) und Saxe, Gearhart und Nasir (2001) sowie Wood und Sellers (1996) finden.

1.5.2.1.2 Merkmale von Fortbildungen

Aus der Vielzahl an Studien mit positiven Effekten auf den verschiedenen Ebenen lassen sich Merkmale ableiten, die als *notwendige* Voraussetzungen wirksamer Fortbildungen betrachtet werden können (Lipowsky & Rzejak, 2012; Lipowsky, 2013; Timperley et al., 2007). Dabei unterscheiden sich verschiedene Analysen in Ihren Merkmalen. Im Folgenden werden die wesentlichen Aspekte genannt und kurz beschrieben. Saxe et al. (2001, S. 55) stellen dabei fest, dass Fortbildungsinhalt und Fortbildungsgestaltung idealerweise aufeinander abgestimmt sein sollten (Desimone, 2009).

Dauer

Die Dauer einer Fortbildung ist ein entscheidendes Merkmal und dies gleich in zweierlei Hinsicht (Desimone, 2006, 2009; Lipowsky, 2010, 2013, 2014; Mayr & Müller, 2010; Nir & Bogler, 2008; Oelkers & Reusser, 2008; Timperley et al., 2007): Zum einen bedeutet eine längere Fortbildungsdauer, eine höhere Wahrscheinlichkeit von Einflüssen auf den Ebenen 2 bis 4 (Carpenter et al., 1989). Zum anderen sind *one-shot*-Fortbildungen, also Fortbildungen, die nur einmal stattfinden, wenig geeignet, um Veränderungen herbeizuführen (Penuel, Fishman, Yamaguchi & Gallagher, 2007). Vielfach belegt ist, dass es einige Zeit bedarf, damit Lehrkräfte eingefahrene Routinen hinterfragen (vgl. hierzu auch die Fortbildungsmerkmale „Wirkung des eigenen Handelns erfahrbar machen“ und „Verschränkung von Input-, Erprobungs- und Reflexionsphasen“) und eine längere Fortbildungsdauer sich auch positiv auf die Wahrnehmung der Fortbildung auswirkt (Gersten et al., 2010; Lipowsky & Rzejak, 2012). Yoon et al. (2007) spricht von einem „Schwellenwert“ von etwa 14 Stunden, was nach Timperley et al. (2007) tendenziell zu niedrig angesetzt ist, unterhalb dessen keine Effekte auf Ebene 4 – also Veränderungen bei den Schülerinnen und Schülern – zu erwarten wären (Lipowsky, 2014).

Wirkung des eigenen Handelns erfahrbar machen

Eine gewisse Dauer einer Fortbildung ist notwendige Voraussetzung, Lehrkräften die Wirkungen ihres Handelns deutlich aufzuzeigen. Durch die Möglichkeit im Rahmen von fiktiven Unterrichtsversuchen innerhalb der Fortbildung, Beobachtungen von Unterricht, durch die Fortbildung begleiteten Unterrichtsversuchen etc., das eigene unterrichtliche Handeln bewusst zu erfahren und mögliche Veränderungen zu vergegenwärtigen, wird das unterrichtliche Handeln nachhaltig verändert, Neues gelernt und letztlich eingeschliffene Muster aufgebrochen. Garett et al. (2008) beschreiben eine schuljahresbegleitende Fortbildung samt reflexivem Coaching des Unterrichts, welches zu deutlichen, signifikanten Verbesserungen in den Leistungen der Schülerinnen und Schüler sowie zu signifikanten Unterrichtsveränderungen beigetragen hat (vgl. Desimone, 2006, 2009; Gersten et al., 2010; Horster & Rolff, 2001; Lipowsky, 2010; Lipowsky & Rzejak, 2012; Lipowsky, 2014; Piaget, 1976; Timperley et al., 2007). Dabei kann das eigene Handeln auch Grundlage für kollegialen Austausch sein (Oelkers & Reusser, 2008; Vygotskij et al., 1978). Die Vergegenwärtigung von Unterrichtsroutinen bzw. die Wirkung des Aufbrechens solcher Routinen kann auch durch Beobachtung und Antizipation von Unterrichtssituationen geschehen, so wie es von

Lipowsky, Rzejak und Dorst (2011) beschrieben wird: Die Präsentation einer video-graphierten Unterrichtssequenz wird unterbrochen und die Lehrkräfte in der Fortbildung um erwartbare Schülerantworten zu einer Frage gebeten. Die Schülerinnen und Schüler sollten eine Geschichte zu einem Bild erzählen. Diese antizipierten Antworten wurden von den Lehrkräften in der Fortbildung gesammelt, wobei auffiel, dass vorrangig deskriptive Aussagen erwartet wurden. Als alternatives Vorgehen wurde den Lehrkräften in der Fortbildung mit dem Video die Möglichkeit aufgezeigt, Schülerinnen und Schüler aufzufordern, sich in eine der Personen auf dem Bild hineinzuversetzen und *die eigene* Geschichte zu erzählen. Dabei zeigten sich zum Erstaunen der Lehrkräfte die Antworten der Schülerinnen und Schüler im Video deutlich narrativer, kreativer und umfassender als erwartet. Auf diesem Weg wurden die üblichen Routinen/Erwartungen verdeutlicht und die Wirkung einer veränderten Handlung erfahrbar.

Verschränkung von Input-, Erprobungs- und Reflexionsphasen

Grundlage einer Erfahrbarmachung eigenen Handelns kann eine fiktive Situation und deren Antizipation durch Lehrkräfte innerhalb der Fortbildung sein oder aber, wie bei Garet et al. (2008) beschrieben, durch die Verschränkung von Inputphasen (also *formalen* Lerngelegenheiten) mit Erprobungs- und Reflexionsphasen (eher *informelle* Lerngelegenheiten). Eine solche Verschränkung von Phasen gilt ebenfalls als wirkungsvoll und sollte insbesondere in den Reflexionsphasen durch Kooperation und Kommunikation zwischen Lehrkräften angereichert werden (Blömeke, 2013; Desimone, 2006, 2009; Lipowsky, 2010; Lipowsky & Rzejak, 2012; Lipowsky, 2013; Oelkers & Reusser, 2008; Timperley et al., 2007). In diesem Sinne kann Fortbildung als „kollaborativ-reflexives Modell“ verstanden werden, wie es beispielsweise in der *lesson study* und den *Professionellen Lerngemeinschaften* praktiziert wird (vgl. Abschnitt 1.5.2.2; Bensen & Frey, 2014, S. 13).

Feedback an LPS

Die Bedeutung von Feedback für Lernprozesse wurde bereits oben thematisiert (vgl. Abschnitt 1.3). Dies gilt selbstverständlich uneingeschränkt auch für Lehrpersonen, die in Fortbildungen *lernen* sollen. Für umfangreiche Coachingmaßnahmen, wie beispielsweise bei Garet et al. (2008) oder Cobb et al. (1991), werden Effekte auf Ebene der Leistungen von Schülerinnen und Schülern berichtet (vgl. Lipowsky, 2010, 2014). Daneben ist die Rückmeldung von Leistungsmaßen vielfach untersucht. So geschehen

etwa im *Literacy*-Projekt in Neuseeland oder den Vergleichsarbeiten in Deutschland (vgl. Kapitel 1.6). Für die Leistungsrückmeldungen kann dabei ebenfalls ein positives Fazit auf Ebene der Schülerinnen und Schüler gezogen werden. Auch Parr et al. (2007, S. 1) berichten von u. a. großen Effekten auf das Schreiben (1.28) und das Lesen (0.46 bzw. 0.49)²¹, wobei diese Effekte kausal nicht ausschließlich mit Feedback verbunden werden können (vgl. auch McDowall et al., 2007).

Erweiterung professionellen Wissens und inhaltliche Relevanz

Unabhängig von den zuvor beschriebenen Ebenen sollte eine Fortbildung darauf abzielen, die *professionelle Kompetenz* der Lehrkräfte in den Fokus zu nehmen. Durch die Erfahrung der Wirksamkeit des eigenen Handelns und ein daraus resultierender „conceptual change“ soll ein Lernen der Lehrkräfte erreicht werden (Lipowsky, 2014). Die Erweiterung des *professionellen Wissens* von Lehrkräften kann entsprechend dem Kompetenzmodell von COACTIV sowohl am Fachwissen als auch am pädagogisch-psychologischen Wissen, dem fachdidaktischen Wissen etc. ansetzen und ist nach Stand der Forschung Voraussetzung für verändertes Unterrichtshandeln (vgl. Fortbildung bei Penuel et al., 2007). Daher sprechen viele Quellen auch von einem „content focus“, den Fortbildungen haben sollten, um an den Voraussetzungen der Lehrkräfte anzuknüpfen und von den Lehrpersonen als kohärent zum *Fach*-Unterricht erlebt zu werden (Blömeke, 2013; Carpenter et al., 1989; Desimone, 2006; Lipowsky & Rzejak, 2012; Lipowsky, 2013; Oelkers & Reusser, 2008).

Fokus auf domänenspezifische Lern- und Verstehensprozesse von Schülerinnen und Schülern

Fachdidaktisches sowie pädagogisch-psychologisches Wissen im COACTIV-Kompetenzmodell ist direkt auf das Lernen und Verstehen von Schülerinnen und Schülern ausgerichtet. Ist das Ziel von Fortbildungen also diese Aspekte *professioneller Kompetenz* zu verändern, so müssen diese Aspekte auch im Kern der Fortbildung selbst stehen und den Lehrkräften diesbezüglich aktive Lerngelegenheiten ermöglicht werden. Besondere Relevanz erhält der Fokus auf *domänenspezifische Lern- und Verstehensprozesse*, da vielfach belegt wurde, dass Transferleistungen nicht eintreten (Carpenter et al., 1989; Cobb et al., 1991; Desimone, 2006, 2009; Lipowsky, 2010;

²¹ Im Bericht ist leider nicht eindeutig erkennbar, um welche Effektstärke es sich handelt, da jedoch Mittelwerte verglichen werden, liegt *Cohens d* nahe. In der Interpretation ist jedoch die Wahl der Effektstärke weitgehend irrelevant und führt hier nicht zu Bedeutungsveränderungen.

Lipowsky & Rzejak, 2012; Lipowsky, 2013, 2014; Nir & Bogler, 2008; Wood & Sellers, 1996). Entsprechend dem Erwartungs-Wert-Modell nach Wigfield und Eccles (2000) bzw. Atkinson (1957) sollte die Fortbildungsmotivation der Lehrkräfte insbesondere bei fachlichem respektive domänenspezifischem Fokus steigen, da so die Unterrichtsnähe, direkte Verwertbarkeit und damit die persönliche Relevanz, der *Wert*, höher ausfällt (Blömeke, 2013; Lipowsky, 2014; Oelkers & Reusser, 2008).

Orientierung an evidenzbasierten Merkmalen erfolgreichen Unterrichts

Ein wesentlicher Teil des fachdidaktischen und pädagogisch-psychologischen Wissens von Lehrkräften ist es, Lern- und Verstehensprozesse von Schülerinnen und Schülern anzuregen. Jenseits der individuellen Perspektive ist jedoch auch integraler Bestandteil dieser Facetten *professioneller Kompetenz*, Unterricht erfolgreich gestalten zu können (vgl. Abschnitt 1.5.4.2). Um Lehrkräften das hierfür notwendige Wissen zur Verfügung zu stellen, sollten Input-, Erprobungs- und Reflexionsphasen sowie das Feedback an die Lehrkräfte etc. sich an evidenzbasierten Merkmalen erfolgreichen Unterrichtens orientieren (Lipowsky, 2009, 2010; Lipowsky & Rzejak, 2012; Lipowsky, 2013). Ebenfalls zentral ist, dass die Fortbildungsinhalte kohärent mit dem aktuellen Curriculum sind (Desimone, 2006, 2009; Timperley et al., 2007). Für Hessen liegt beispielsweise der Hessische Referenzrahmen Schulqualität für Schulen zugrunde, deren Operationalisierung von Schul- und Unterrichtsqualität ebenfalls Bezug zur aktuellen Forschung nimmt (Diel, 2011; Hessisches Kultusministerium & Institut für Qualitätsentwicklung [IQ], 2011). Abschließend bleibt anzumerken, dass die Berücksichtigung der genannten Merkmale keinesfalls eine Garantie für Fortbildungswirksamkeit darstellt. Letztlich ist eine Fortbildung eine Lehr-Lern-Situation wie Unterricht, in der für Lehrkräfte ein Angebot geschaffen wird, welches in Abhängigkeit vieler Faktoren genutzt wird (vgl. Abbildung 14 auf Fortbildungsebene und Abbildung 15 für die Unterrichtsebene).

1.5.2.2 Kooperation

Überhaupt scheint die Lehrerverkooperation eine unabdingbare Voraussetzung oder ganz wesentliche Schubkraft für Schulentwicklung zu bilden, wie zahlreiche Schulentwicklungsstudien zeigen. (Holtappels, 2014b, S. 34)

Zweifelsfrei spielt die individuelle Weiterqualifikation und Fortbildung eine wesentliche Rolle für die PE einer Schule und ist u. U. sogar Ausgangspunkt von weiteren

Entwicklungen auf Unterrichts- und Organisationsebene. Nichtsdestotrotz erfordert eine Entwicklung von Schule insbesondere ein Zusammenspiel und eine „Handlungsabstimmung zwischen den Akteuren“ (Ditton, 2000, S. 73). Auch Rolff (2007, S. 27) weist darauf hin, dass SE „gerade nicht von einzelnen ‚einsamen‘ Lehrern aus(zu)gehen [kann, SV.], sondern miteinander kooperierende und untereinander vernetzte Teams als Motor der Entwicklung“ erfordert. Diese Sichtweise von Kooperation im Kollegium gemeinsam mit kontinuierlicher PE (also insbesondere auch Fortbildung) wird dabei in vielen Schulqualitätsstudien als wichtige Einflussgröße für Schulqualität angesehen (Holtappels, 2014b).

Was genau unter Kooperation zu verstehen ist, definiert Spieß (2004, S. 199) folgendermaßen: „Kooperation ist gekennzeichnet durch den Bezug auf andere, auf gemeinsam zu erreichende Ziele bzw. Aufgaben, sie ist intentional, kommunikativ und bedarf des Vertrauens. Sie setzt eine gewisse Autonomie voraus und ist der Norm von Reziprozität verpflichtet“ (vgl. auch Gräsel, Fußangel & Pröbstel, 2006, S. 206). Aus dieser Definition lässt sich ablesen, dass Kooperation zielgerichtet und auf vorher definierte Ziele ausgerichtet zu verstehen ist.

Im Zusammenhang mit einer veränderten Unterrichtskultur wird vielfach von einer Fehlertoleranz gesprochen und davon, Fehler als Lerngelegenheit anzuerkennen (Blum et al., 2012; Blum et al., 2015; Bensen & Rolff, 2006; Ramm, 2014). Dies gilt jedoch nicht nur für den Klassenraum: Die Stimmung innerhalb des Kollegiums muss ebenfalls vertrauensvoll und fehlertolerant sein. Vor diesem Hintergrund bedeutet PE an der Schnittstelle zur OE vor allem auch die Veränderung von Strukturen. Wie oben beschrieben, sind viele Lehrkräfte und Kollegien vom „Einzelkämpfertum“ (l. Hosenfeld, 2005; Rolff, 2001, 2007) gekennzeichnet und lassen sich als *loosely coupled systems* (O'Day, 2002; Weick, 1976) betrachten. Krainz-Dürr (1999) konnte in Fallanalysen zeigen, dass solch lose strukturierte Kollegien Schwierigkeiten bei der Identifizierung und selbstständigen Lösung von Problemen hatten (vgl. auch Rolff, 2001). Mit Blick auf vergleichende Leistungsmessungen bekommt das „Einzelkämpfertum“ und die Fehlertoleranz (Bensen & Rolff, 2006) vor dem Hintergrund der Attribution (Weiner, 1975, 1986) weitere Relevanz: Bei fehlender Akzeptanz und Toleranz kommt es schnell zu Abwehrreaktionen und externer Attribution von Misserfolgen bzw. unerwünschten Ergebnissen. Besteht diese Gefahr, bleiben Ergebnisse vergleichender Leistungsmessungen eher „Privatsache“ der einzelnen Lehrkräfte und das

Potenzial, gemeinsame Entwicklungen und Unterstützungen anzustoßen, verpufft (Kuper & Diemer, 2012; Rolff & Buchen, 2016).

Die Kooperation zwischen den einzelnen Lehrkräften und die Arbeitsatmosphäre sowie das Evaluationsklima an einer konkreten Schule wird in wesentlichem Maße von den jeweiligen Lehrkräften im Zusammenspiel mit der Schulleitung geprägt. Im Rahmen der zunehmenden Eigenverantwortung von Schulen ist es insbesondere die Aufgabe der Schulleitung für ein wertschätzendes Miteinander zu sorgen: Hierfür müssen Raum und Zeit eingeräumt werden, um etwaige Probleme und Entwicklungslinien zu identifizieren und daran zu arbeiten, eine adäquate Kommunikationskultur etabliert werden und die einzelnen Gremien der Schule (etwa Fachkonferenzen) mit ausreichendem Spielraum für Entscheidungen ausgestattet werden (Burmeister et al., 2005; Maier, 2010; Reimers et al., 2009; Rolff, 2001).

Eine Möglichkeit der institutionalisierten Kooperation an Schulen sind *Professionelle Lerngemeinschaften* (PLG). „Ohne dass Lehrkräfte wie reflektierende Praktiker miteinander reden, sich gegenseitig beobachten und auch helfen, gibt es keine Schulentwicklung“ (Rolff, 2007, S. 148). In diesem Zitat von Rolff (2007) wird klar, dass zwischen den Lehrkräften eine Kooperation entstehen sollte, in der über den eigenen Unterricht gesprochen wird – und damit insbesondere auch über Fehler im eigenen Unterricht. Im japanischen Unterricht wurden in der TIMSS-Videostudie wie auch in nachfolgenden Untersuchungen u. a. auf die *lesson study* als besonderes Merkmal der teamorientierten Unterrichtsplanung, -durchführung und -auswertung hingewiesen (Baumert, Klieme & Bos, 2001; Blum & Neubrand, 1998; Blum, 2001; Bensen & Frey, 2014; Wiegand, 1999). Eine solche Kooperation ist in Deutschland eher die Ausnahme (Institut für Sozialwissenschaftliche Studien (ISS) e.V. & Gewerkschaft Erziehung und Wissenschaft (GEW), 2009, 12ff.). Im deutschsprachigen Raum wird hingegen von *Professionellen Lerngemeinschaften* (PLG) gesprochen. PLGs zeichnen sich dabei durch folgende Merkmale aus (Bensen & Rolff, 2006; Bensen & Frey, 2014; Lipowsky & Rzejak, 2012; Posch, 2009; Rolff, 2007):

- Gemeinsame Normen, Werte und Ziele
- Kooperation
- De-Privatisierung von Unterricht
- Reflexiver Dialog
- Fokus auf das Schülerlernen

Die Lehrkräfte einer *PLG* beraten sich gegenseitig in einer konstruktiven und fehler-toleranten, weil „de-privatisierten“ Atmosphäre. Dabei teilen sie gemeinsame Ziele und fokussieren letztlich ihre Handlungen und Entwicklungen auf die Optimierung des Lernens durch Schülerinnen und Schüler (vgl. Abbildung 12).

Die Initiierung und Unterstützung solcher *PLGs* ist dabei auch Aufgabe der Schulleitung. Damit lässt sich der Aufbau von *PLGs* als eine Verbindung von PE und OE ansehen; deren Aufrechterhaltung und inhaltliche Arbeit steht jedoch an der Schnittstelle von PE und UE (Bonsen & Rolff, 2006, S. 181; Holtappels, 2014b). Damit wird nochmals deutlich, wie eng die drei Teile der SE verzahnt sind. Einen guten Überblick über *PLGs* wird in den bereits genannten Quellen gegeben (Bonsen & Rolff, 2006; Bonsen & Frey, 2014; Rolff, 2007).

1.5.3 Organisationsentwicklung

Letztendlich ist die Schule aufgrund der sich wandelnden Gesellschaft dazu verpflichtet, sich auch als Organisation zu entwickeln. Durch die Schaffung von Kooperation werden unmittelbar Strukturen innerhalb der Organisation *Schule* verändert, beispielsweise durch die Einrichtung von *PLGs* und damit einhergehender Ressourcen zur Arbeit in diesen wie etwa gemeinsame Unterrichtsbeobachtung, gemeinsame Beratungsgespräche etc. Eine besondere Rolle kommen bei Entwicklungen, die die Schule als Ganzes betreffen, der Schulleitung als Vorbild und Vorreiter zu (Holtappels, 2014b; Schildkamp & Kuiper, 2010). Dabei sollte die Schulleitung und das Kollegium (ähnlich einer großen *PLG*) ein gemeines Ziel oder eine Vision haben, von der Schulleitung ein innovationsfreundliches und anerkennendes Klima aufgebaut und Transparenz in Entscheidungen gewahrt werden (Meyer & Demuth, 2009). Eine gerechte Verteilung von Ressourcen sowie Belastungen bzw. Verantwortungen beispielsweise auf Fachkonferenzen (oder andere Instanzen/Steuergruppen innerhalb der Schule) ist sowohl unmittelbar als auch mittelbar (in der Wirkung auf das Organisationsklima) entscheidend (Burmeister et al., 2005; Meyer & Demuth, 2009; Schaarschmidt, 2005, 2007). Ziel sollte es schlussendlich sein, die Verantwortung für die Entwicklung von Schule auf die gesamte Organisation zu verteilen und dabei alle Akteure mit einzubeziehen (also u. a. auch Eltern sowie Schülerinnen und Schüler; Burmeister et al., 2005).

An dieser Stelle soll auf diesen Aspekt der SE nicht weiter eingegangen werden. Für weitere Informationen vergleiche u. a. Holtappels und Feldhoff (2010), Krainz-Dürr (1999), Rolff (1993, 2001, 2007, 2010), van Ackeren und Klemm (2011).

1.5.4 Unterrichtsentwicklung

Den Unterricht zu entwickeln gehört zu den schwierigsten Reformaufgaben überhaupt, gilt es doch, den eigenen Arbeitsplatz zu verändern und sich selbst auch, seine Vorstellungen, seine Einstellungen und sein Verhalten. (Horster & Rolff, 2001, S. 9)

oder

Wer Unterricht verändern will, muss mehr als den Unterricht verändern. (Rolff, 2007, S. 29)

1.5.4.1 Zieldimension von UE

Jede einzelne Unterrichtsstunde und jede Unterrichtseinheit muss sich daran messen lassen, inwieweit sie zur Förderung und Weiterentwicklung inhaltsbezogener und allgemeiner Schüler-Kompetenzen beiträgt, und der Unterricht über längere Zeiträume hinweg muss so konzipiert sein, dass der Aufbau von Kompetenzen im Zentrum steht. Die wichtigste Frage ist nicht ‚Was haben wir durchgenommen?, sondern ‚Welche Vorstellungen, Fähigkeiten und Einstellungen sind entwickelt worden?‘. (Blum, 2012, S. 17)

Mit diesem einführenden Zitat von Werner Blum lassen sich zwei essentielle Aspekte von UE konkret benennen: Es geht darum, dass Unterricht *reflektiert* werden muss und dabei der Fokus auf das *Lernen von Schülerinnen und Schülern* gelegt wird (vgl. Outputorientierung; O'Day, 2002). Das Kerngeschäft von Lehrkräften und ihr primärer Fokus ist der Unterricht (Schildkamp, 2007; Schildkamp & Kuiper, 2010), jedoch gehört andererseits eine klare Analyse bzw. Reflexion des Ist-Standes zwingend dazu, um hiervon ausgehend geeignete Entwicklungen anzustoßen (vgl. Abschnitt 1.4; Rolff, 2001; Stryck, 2005). In der allgemeinen Pädagogik ohne besonderen Fachbezug werden in der UE vielfach diejenigen Maßnahmen angeführt, die die „Wirksamkeit“ des eigenen Unterrichts steigern bzw. das Lernen der Schülerinnen und Schüler effektivieren (Horster & Rolff, 2001; A. Hosenfeld, 2010), wobei unter „Wirksamkeit“

ein hoher Lernfortschritt und verringerte Leistungsheterogenität verstanden wird (Helmke & Weinert, 1989; van Ackeren & Klemm, 2011). Meyer und Demuth (2009, S. 10) gehen dabei noch einen Schritt weiter und fordern, dass zur UE Möglichkeiten geschaffen werden müssen, „in denen die Lehrer und Schüler die Chance haben, ihre fest verinnerlichten Unterrichtsbilder ‚im Hinterkopfe‘ wieder zu verflüssigen und ihnen neue Facetten hinzuzufügen“. Dies deckt sich mit Befunden über eine sehr hohe Stabilität von Beliefs/Alltagstheorien bei Lehrkräften (vgl. u. a. A. Hosenfeld, 2010; Kunter et al., 2011; Rolff, 2007; T. Voss, Kleickmann, Kunter & Hachfeld, 2011).

Die Anforderungen zur Reflexion und Weiterentwicklung der alltäglichen Unterrichtspraxis durch die Lehrkräfte selbst stellt einen erheblichen Anspruch an diese und erfordert insbesondere vor dem Hintergrund sich wandelnder Aufgaben bzw. dem sich wandelnden Rollenverständnis von Lehrkräften (Lehrkräfte als *Organisatoren/Modérateuren*, als *Erziehende* und als *Vertretende des jeweiligen Faches*; Klieme et al., 2001, S. 52) externe Unterstützung und Fortbildung (I. Hosenfeld et al., 2006; Rolff, 2007).

Im Unterricht geht es darum, Lerngelegenheiten für Schülerinnen und Schüler zu schaffen, die deren aktuellem Lernstand entsprechen (Hofmann-Göttig et al., 2005; O'Day, 2002; Reimers et al., 2009). Daher sollten UE-Maßnahmen darauf ausgelegt sein, das *Lernen von Schülerinnen und Schülern* zu verbessern. Entsprechend sollten Lehrkräfte nach Sjuts (2008, S. 58) den Fokus regelmäßig (auch unabhängig von expliziten Leistungskontrollsituationen wie Klassenarbeiten; *Diagnose-on-the-fly*; Rolff, 2007) auf „das individuelle Denken und Verstehen der Schülerinnen und Schüler“ lenken und damit sicherstellen, dass Schülerinnen und Schüler nicht nur unterrichtet werden, sondern etwas lernen (Bonsen & Frey, 2014, S. 15; Isaac et al., 2006). Explizite Leistungskontrollsituationen und davon ausgehende Evaluationen sind immer zusätzlich mit dem Ziel verbunden, Personal und Unterricht zu verbessern (Helmke et al., 2007; Isaac et al., 2006).

Vor dem Hintergrund der vergleichenden Leistungsmessungen um die Jahrtausendwende wurde vielfach über die Leistungsfähigkeit deutscher Schülerinnen und Schüler²² diskutiert: Zusammenfassend lässt sich sagen, dass der deutsche Unterricht (als

²² Sicherlich ist die Leistungsfähigkeit nur ein Teil des Lernerfolgs von Schülerinnen und Schülern, aber für die empirische Bildungsforschung vermutlich einer der am leichtesten zugänglichen Teile. Ob das Ziel von Schule und Unterricht ausschließlich in der Fokussierung und Optimierung der Leistungsfähigkeit von Schülerinnen und Schülern besteht, soll an dieser Stelle nicht diskutiert werden.

Spiegel der Leistungen von Schülerinnen und Schüler) vor allem „didaktisch und methodisch einförmig“ und auf „kurzfristige Behaltensleistungen und Routinen“ ausgerichtet (Prenzel, 2001, S. 59; Baumert et al., 2001) sowie wenig inner- und außermathematische Vernetzung, wohl aber von Kalkülen und Verfahren (im Sinne der Bildungsstandards also K5 „Mit symbolischen, formalen und technischen Elementen der Mathematik umgehen“) geprägt war (Blum, 2001; Wiegand, 1999, 2001). Insbesondere mit dem bereits mehrfach angesprochenen Paradigmenwechsel von der *Input-* zur *Outputsteuerung* (vgl. Abschnitt 1.1) geht es umso mehr darum, wie Bildungsstandards Einzug in den alltäglichen Unterricht erhalten, also letztlich der alltägliche Unterricht unter dem Blickwinkel von *Outputorientierung* gesehen werden kann (KMK, 2010, S. 7). Dabei forcieren die Bildungsstandards mit dem Perspektivenwechsel, „fachbezogene (und fachübergreifende) Lehr- und Lernprozesse in Hinblick auf klare Zielstellungen neu zu durchdenken, das Wissen und Können der Schülerinnen und Schüler differenziert wahrzunehmen, eine realistische Vorstellung von der eigenen Wirksamkeit aufzubauen und an der Weiterentwicklung der eigenen Professionalität und der der eigenen Schule (mit)zuarbeiten“ (Klieme et al., 2009, S. 51). Ein Unterricht im Sinne der Bildungsstandards (und damit „kompetenzorientiert“) erfordert daher bei der Aufgabenbearbeitung mehr als das reine Fachwissen und technische Fähigkeiten, sondern insbesondere die Fähigkeit dieses Wissen in konkreter Anforderungssituation einsetzen zu können (Blum & Leiß, 2007b; Weinert, 2001). Der Einzug von Bildungsstandards verändert dabei jedoch nicht ausschließlich die Rolle der Lehrenden, sondern auch der Lernenden, die selbstständiger in durch Lehrkräfte vorbereiteten Lernumgebungen für ihre Leistungen verantwortlich sind (Bauch, Maitzen & Katzenbach, 2011; Oelkers & Reusser, 2008).

Dies bedeutet für die Veränderung von Unterricht, dass sich prinzipielle Strukturen ändern müssen: Blum (2001) fasst dabei – in Einklang mit dem Eingangszitat dieses Kapitels – eine neue Unterrichtskultur als Summe von neuer Aufgabenkultur²³ und neuer Kommunikationskultur (sowohl zwischen Lehrkräften als auch zwischen

²³ Auf die besondere Rolle von Aufgaben v. a. für den Mathematikunterricht sei an dieser Stelle kurz verwiesen auf Blum, Drüke-Noe, Hartung und Köller (2012), Blum, Vogel, Drüke-Noe und Roppelt (2015), Bromme, Seeger und Steinbring (1990), Bruder (2003), Christiansen und Walther (1986), Biermann, Wiegand und Blum (2003), Girmes (2003), Leisen (2006) sowie Neubrand, Jordan, Krauss, Blum und Löwen (2011).

Lehrkräften und Lernenden) zusammen (Blum, Drücke-Noe, Leiß & Wiegand, 2005; Blum, Drücke-Noe, Leiß, Wiegand & Jordan, 2005).

Generell kann es jedoch *a priori* lediglich um die bestmögliche Gestaltung eines Angebots gehen, das von Schülerinnen und Schülern genutzt werden soll, denn: „Thus, what a teacher does in her classroom depends greatly on what her students do, and vice versa“ (O'Day, 2002, S. 297). Diesen Sachverhalt stellt das vielzitierte und häufig Anwendung findende Angebots-Nutzungs-Modell dar (Lipowsky, 2006, 2009).

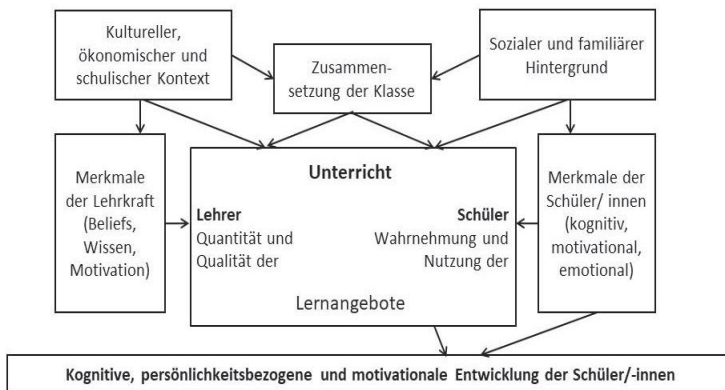


Abbildung 12: Angebots-Nutzungs-Modell von Unterricht (Lipowsky, 2009, S. 81)

Sowohl auf Seiten des Angebots als auch auf Seiten der Nutzung ist ein komplexes Hintergrundgefüge beeinflussend. An dieser Stelle sei nur kurz auf die bereits bekannten Ausführungen zur *professionellen Kompetenz* von Lehrkräften verwiesen (Seite des Angebots, links in Abbildung 11). Des Weiteren finden sich a. a. O. weitere Literaturverweise, die die peripheren Aspekte des Angebots-Nutzungs-Modells näher beschreiben. Eine ebenfalls detailreiche Übersicht über mögliche Einflussgrößen auf den Lernerfolg von Schülerinnen und Schülern bzw. die Wirksamkeit von Unterricht findet sich bei Blum et al. (2006) sowie Ditton (2000). Im Weiteren wird sich auf die Angebotsqualität beschränkt.

1.5.4.2 Kriterien von Unterrichtsqualität

In den vergangenen Jahrzehnten wurde in verschiedenen Forschungstraditionen die Erklärung für Unterrichtserfolg untersucht (Helmke, 2009). Diese Paradigmen sind in der jüngeren Vergangenheit im verbindenden Angebots-Nutzungs-Modell aufgegangen (Abbildung 12). Für die Angebotsseite wurden wiederum vielfach Kriterien

angeführt, die einen *guten* Unterricht ausmachen. Dabei ist grundlegend wichtig, dass es nicht „den“ guten Unterricht gibt und dieser sich auch nicht zwingend dadurch auszeichnet, dass *alle* Kriterien in vollem Umfang erfüllt sind (Bauch et al., 2011; Blum et al., 2006; Blum & Leiß, 2007b; Helmke, 2009). Die Vielfalt an verschiedenen Kriterien lässt sich dabei vor allem durch unterschiedlich hohen Auflösungsgrad erklären. In Modellen mit weniger Kriterien werden diese gröber gefasst und subsumieren mehrere Kriterien aus kleinschrittigeren Modellen.

Prominente Vertreter der Unterrichtsqualitätsforschung sind dabei aus der allgemeinen Pädagogik und Erziehungswissenschaft Helmke (2009) und Meyer (2016) sowie – für hessische Schulen direkt relevant – der Hessische Referenzrahmen Schulqualität (HRS) (Diel, 2011; Hessisches Kultusministerium & IQ, 2011). Mit dem Fokus auf Mathematik lassen sich hier die Projekte SINUS (Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts; Prenzel, 2001; Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung [BLK], 1997), DISUM (Didaktische Interventionsformen für einen selbstständigkeitsgesteuerten Unterricht am Beispiel Mathematik; Blum & Leiß, 2003, 2005, 2006, 2007a, 2007b; Schukajlow, 2006), COACTIV (Kunter et al., 2011) und die Pythagoras-Studie (Unterrichtsqualität, Lernverhalten und mathematisches Verständnis; Klieme, Lipowsky, Rakoczy & Ratzka, 2006; Lipowsky et al., 2009; Rakoczy, 2006) anführen.

Im Folgenden wird eine Synthese verschiedener Modelle der Unterrichtsqualität präsentiert, die in ähnlicher Form im später beschriebenen Projekt VELM-8 (Verbesserung der Effektivität der Lernstandserhebungen Mathematik Klasse 8; Teil 2) genutzt wurde. Die näher betrachteten Unterrichtsqualitätskriterien sind:

1. Effektive Klassenführung
2. Schülerorientiertes, unterstützendes Klassenklima
3. Kognitive Aktivierung der Lernenden
4. Metakognitive Aktivierung der Lernenden
5. Fachlich gehaltvolle Unterrichtsgestaltung

1.5.4.2.1 Effektive Klassenführung

In fast allen Operationalisierungen von Unterrichtsqualität taucht eine effektive Klassenführung auf (Blum & Leiß, 2006, 2007a, 2007b; Helmke, 2009; Klieme et al., 2001, 2001; Klieme et al., 2006; Kunter & T. Voss, 2011; Lipowsky et al., 2009; Lipowsky,

2009; Ramm, 2014; Seidel, 2009; van Ackeren & Klemm, 2011). Als Grundlagenwerk und Ursprung der Forschung zur *Klassenführung* bzw. *classroom management* gelten die Arbeiten von Kounin (2006; Reprint des Originals von 1976). Ausgehend vom *Wellen-Effekt*²⁴ (Seidel, 2009) zeigt Kounin die folgenden Dimensionen der *Klassenführung* als wirksam – im Sinne positiver Korrelationen zwischen diesen Merkmalen und dem Ausbleiben von Fehlverhalten im Unterricht – auf (vgl. auch Brophy & Good, 1984):

- Allgegenwärtigkeit und Überlappung
- Reibungslosigkeit und Schwung
- Gruppenmobilisierung und Rechenschaftsprinzip
- Valenz und intellektuelle Herausforderung
- Abwechslung und Herausforderung bei der Stillarbeit

Hinter diesen Merkmalen/Dimensionen einer „guten“ Klassenführung steckt ein Verhalten, das den Schülerinnen und Schülern deutlich macht, dass die Lehrkraft *allgegenwärtig* über die Situation im Klassenraum informiert ist und dabei *überlappend* sowohl dem Unterricht weiter folgen als auch simultan einzelne Schülerinnen und Schüler in den Fokus nehmen kann. Dabei sollte es an Gelenkstellen nicht zu Brüchen oder Leerlauf kommen, sodass die Lerngruppe den Fokus auf den Lerninhalt nicht verliert, und die Klasse als Gruppe sollte aktiviert werden, jedoch ohne dabei die oder den Einzelnen (mit den individuellen Bedürfnissen; vgl. schülerorientiertes, unterstützendes Unterrichtsklima) aus dem Blick zu verlieren. Die beiden letztgenannten Dimensionen beschreiben *abwechselnde* und *herausfordernde* Aufgaben- und Problemstellungen (vgl. auch Kognitive Aktivierung; Brophy & Good, 1984; Helmke, 2009; Kiel, Frey & Weiß, 2013; Schönbachler, 2008; Seidel, 2009).

Alles in allem soll eine Lernumgebung geschaffen werden, in der es gelingt, die Lernzeit bei gleichzeitiger Prävention von Störungen sowie höchstmöglicher Motivation, Konzentration und Selbstständigkeit der Schülerinnen und Schüler zu maximieren (van Ackeren & Klemm, 2011; Helmke, 2009; Kiel et al., 2013; Schönbachler, 2008; Seidel, 2009). Dabei ist die Klassenführung einerseits geeignet, die Komplexität und

²⁴ Der *Wellen-Effekt* beschreibt das Phänomen, dass bei einer geringen Unterrichtsstörung durch die Intervention der Lehrkraft letztendlich eine größere Störung bzw. Behinderung im Unterrichtsfluss entsteht als durch die Störung an sich.

Unsicherheit im Unterricht zu reduzieren (Brophy & Good, 1984; Hiebert & Wearne, 1993; Kiel et al., 2013) und so einer erhöhten Lehrerbelastung entgegenzuwirken (Schaarschmidt, 2005, 2007; Schönbachler, 2008). Andererseits bedeutet eine maximierte Lernzeit (*time on task*) eine erhöhte Lernleistung bei Schülerinnen und Schülern (Hiebert & Wearne, 1993; Klieme et al., 2006; Lipowsky et al., 2009; Lipowsky, 2009).

Besondere Anforderungen an die Lehrkräfte stellt das Kriterium der *Effektiven Klassenführung* nicht nur – wie u. U. intuitiv zu vermuten – in der eigentlichen Unterrichtssituation, sondern auch in deren Vorbereitung: Im Rahmen der Planung von Unterricht(-seinheiten) gilt es auf einzuhaltende Regeln zu achten bzw. diese vorab zu berücksichtigen, Übergänge im Unterricht möglichst fließend und ohne Leerlauf zu gestalten, im direkten Vorfeld des Unterrichts den Klassenraum infrastrukturell vorzubereiten und letztlich auch die inhaltliche Planung des Unterrichts zu leisten (Brophy & Good, 1984; Evertson, Emmer & Worsham, 2006; Haag, 2012; Helmke, 2009; Kiel et al., 2013; Schönbachler, 2008).

Unter *Klassenführung* ist somit sowohl der Umgang mit als auch die Prävention von Störungen zu verstehen. Im schulischen Alltag gibt es eine Vielzahl verschiedener Störungen, die immer dann vorliegen, „wenn das Lehren und Lernen stockt, aufhört, pervertiert, unerträglich oder inhuman wird“ (Winkel, 2011, S. 29). Im Umgang mit Störungen ist es nach dem *low-profile-Ansatz* (Borich, 2007; Borich & Martin, 2008; Shavelson & Stern, 1981) wichtig, entsprechende *anticipation* (Vorbeugung; z. B. regelmäßiges Herumgehen in der Klasse, störungsvermeidende Methodenwahl, Vorbereitung des Klassenraumes) zu betreiben und geringfügige Störungen, die den Unterrichtsfluss nur in kleinem Maße beeinträchtigen, mit sparsamen Reaktionen zu beantworten (*deflection*), um keinen *Wellen-Effekt* auszulösen. Dies fällt v. a. jungen Lehrkräften schwer (Thiel, Richter & Ophardt, 2012). Wesentliche Aspekte der Vorbeugung sind dabei klare Regeln, Routinen und Rituale, die „gleich zu Beginn des Schuljahres in der Klasse etabliert werden“ (Seidel, 2009, S. 143; Evertson et al., 2006; Helmke, 2009; Schönbachler, 2008).

In vielen empirischen Untersuchungen sowie Meta-Analysen zeigt sich die Wirksamkeit von Klassenführung, wobei sie in den verschiedenen Studien oftmals unterschiedlich bzw. nur in Teilaspekten operationalisiert wird: In der Meta-Analyse von Wang, Haertel und Walberg (1993) liegen „classroom practices“ auf dem zweiten

Platz in der Effektrangfolge direkt hinter den Charakteristika von Schülerinnen und Schülern. Etwas detaillierter aufgeschlüsselt erlangt das „classroom management“ sogar eine noch höhere Wirksamkeit und liegt in der Wirkrangfolge an der Spitze. In der bekannten Meta-Analyse von Hattie (2009) zeigt sich die Quintessenz einer *effektiven Klassenführung*, nämlich die „aktive Lernzeit“, als Merkmal mit einem mittleren Effekt ($d = 0.38$; vgl. Hattie et al., 2013). Seidel und Shavelson (2007) mit Bezug auf Scheerens und Bosker (1997) unterstreichen diese Wirkung in der eigenen Untersuchung bezüglich positiver Einflüsse auf den Lernprozess und die Motivation. Seidel und Shavelson (2007) finden in ihrer Meta-Analyse ebenfalls quer durch verschiedene Studien signifikante Einflüsse von *Strukturierung* (als Merkmal von *effektiver Klassenführung* zur Reduktion von Komplexität und der Wahrscheinlichkeit von Störungen) sowohl auf den Lernprozess als auch insbesondere auf motivational-affektiven Merkmale. Vor allem für den deutschsprachigen Raum gibt es einige mathematikspezifische Untersuchungen zum Einfluss der Unterrichtsqualität auf Lernleistung, Motivation und das Interesse: Lipowsky et al. (2009) berichten einen mehrbenen-analytischen Effekt von $\beta = .09$ ($p < .05$) des *classroom managements* auf Posttestleistung unter Kontrolle der Prätestleistung, der Unterrichtszeit zwischen Prä- und Posttest sowie Merkmalen auf Individualebene (beispielsweise dem mathematischen Interesse). Wenig überraschend ist nach vorherigen Ausführungen der korrelative Zusammenhang zwischen intrinsischer Motivation (hier operationalisiert als *basic needs* der Selbstbestimmungstheorie; Deci, 1972; Deci & Ryan, 1993; Ryan & Deci, 2000) sowie dem Interesse (A. Krapp & Weidenmann, 2005; K.-P. Wild & A. Krapp, 2005) und Strategien der Klassenführung (operationalisiert durch Regelklarheit und Monitoring im Sinne Kounins *Allgegenwärtigkeit*; Kunter, Baumert & Köller, 2007). Den direkten Einfluss auf die Motivation (operationalisiert durch die *basic needs*) zeigt Rakoczy (2006) für *Strukturiertheit*²⁵ ($\beta = .13$, $p < .05$ für Autonomieunterstützung; $\beta = .14$, $p < .05$ für Kompetenzerleben; $\beta = .09$, $p < .05$ für *soziale Eingebundenheit*; vgl. auch Klieme et al., 2006; Rakoczy et al., 2007; Rakoczy, 2008): Ein störungsarmer und disziplinierter Unterricht führt demnach zu einer erhöhten Motivation und einer erhöht wahrgenommenen Unterstützung. Als wichtige Ergänzung zu diesen auf den

²⁵ Den positiven Einfluss von *Strukturierung* in geöffneten Lernumgebungen im naturwissenschaftlichen Unterricht weisen Hardy, Jöns, Möller und Stern (2006) nach.

deutschsprachigen Raum bezogenen Ergebnissen lässt sich der große Effekt von $d = 0.75$ der Klarheit der Lehrperson auf die Leistung bei Hattie (2009) anführen.

1.5.4.2.2 Schülerorientiertes²⁶, unterstützendes Unterrichtsklima

Bereits oben wurde angedeutet, dass eine *effektive Klassenführung* den Schülerinnen und Schülern ein Gefühl von Unterstützung gibt und dabei die *basic needs* bedient, also motivationsförderlich ist (Kunter et al., 2007; Rakoczy, 2006; Seidel & Shavelson, 2007). Einige Studien belegen meta-analytisch einen mittleren Effekt (Hattie, Biggs & Purdie, 1996) zwischen *unterstützendem Klassenklima* und Leistung (vgl. auch Lipowsky, 2009), doch wird von diesen die Qualität der Beziehungen vernachlässigt. Lipowsky (2009) unterstreicht daher die theoretisch erwartbaren und bereits deutlich gemachten indirekten Einflüsse auf Leistung über eine erhöhte Motivation und gesteigertes Interesse (vgl. auch Lipowsky et al., 2009; Ames & Archer, 1988; Baumert & Lehmann, 1997; Furrer & Skinner, 2003; Goodenow, 1993; Gruehn, 2000; Ryan, Stiller & Lynch, 1994; Turner et al., 1998; Wentzel, 1997). Dieser nur mittelbare Einfluss verdeutlicht sich auch in der geringeren Bedeutsamkeit des „classroom climate“ in der Meta-Analyse von Wang et al. (1993). Hattie (2009) zeigt in der viel zitierten Meta-Analyse hingegen einen großen Effekt von $d = 0.72$ für die Lehrer-Schüler-Beziehung auf (vgl. Hattie et al., 2013).

Die Wirkungen von *schülerorientiertem und unterstützendem Klassenklima* liegen eher im motivationalen Bereich (Rakoczy, 2006, 2008): Dies belegen Mehrebenenanalysen aus dem deutsch-schweizerischen Projekt „Pythagoras“ (Lipowsky et al., 2009). Zwar fehlt dem direkten Einfluss des „supportive climate“ auf die Posttestleistung die Signifikanz, doch zeigt sich bei Kontrolle u. a. von Vortest, Unterrichtszeit und Interesse eine signifikante Interaktion des *Unterrichtsklimas* auf Klassenebene mit dem Interesse auf der Individualebene: $\beta = .07$ ($p < .01$), also profitieren Schülerinnen und Schüler bei geringerem Interesse umso stärker von einem *unterstützenden Klassenklima* bei der Zielvariable Leistung. In der Meta-Analyse von Seidel und Shavelson (2007) zeigen sich kleine Effekte für Feedback als Teil eines *unterstützenden Klimas* ($d = 0.11$; vgl. Abschnitt 1.3) auf Motivation sowie von kooperativem Lernen auf Motivation ($d = 0.12$) und auf den Lernprozess ($d = 0.11$).

²⁶ Selbstverständlich sind hierbei alle Geschlechter gemeint. Aufgrund der besseren Lesbarkeit wird jedoch bei diesem Begriff auf eine genderkonforme Formulierung verzichtet.

Eine genaue und übergreifende Definition von *unterstützendem Klassenklima* lässt sich jedoch nicht finden, allerdings meinen viele Autoren eine konstruktive, fehlertolerante und anerkennende Atmosphäre, in der vorrangig die individuelle Bezugsnorm herangezogen (geringer Effekt $d = 0.23$ für Individualisierung bei Hattie, 2009), kein Leistungsdruck aufgebaut wird und Schülerinnen und Schüler in ihren eigenen Arbeitsweisen unterstützt werden; insbesondere, wenn diese Arbeitsweisen möglicherweise nicht den „direkten“ Weg zur Lösung eines Problems bedeuten (Baumert & Lehmann, 1997, S. 51; Blum & Leiß, 2005, 2006, 2007a, 2007b; Blum, 2015). Dabei sollten Lehrkräfte nach Möglichkeit selbstständigkeitsfördernde Interventionen liefern, da diese Fehler im Sinne kognitiver Konflikte Schülerinnen und Schüler kognitiv aktivieren (vgl. auch Metakognitive Aktivierung der Lernenden; Blum & Leiß, 2007a, 2007b; Neubrand, 2012). Optimistisch stimmen die Ergebnisse von Reeve (2004), nach denen autonomieunterstützende Klassenführung von Lehrkräften erlernt werden kann. In diesem Kontext spielen u. a. Feedback, die Lehrkraft-SchülerInnen-Beziehung sowie Fragen von Lehrkräften wichtige Rollen.

Insbesondere aufgrund der unterrichtsleitenden Funktion der Fragen von Lehrkräften und dem daraus resultierenden Anspruch des Unterrichts kommt dieser Thematik eine besondere Wichtigkeit zu. Theoretisch lassen sich Fragen nach der sog. *Bloomschen Taxonomie* in sechs hierarchisch gegliederte Anspruchsniveaus der intendierten kognitiven Lernziele unterteilen (Bloom, 1976), wobei jedoch weniger anspruchsvolle Fragen im Schulalltag dominieren (Niegemann & Stadler, 2001):

1. **Wissen** („Definiere...“)
2. **Verstehen** („Erkläre mit eigenen Worten...“)
3. **Anwenden** („Wende die Regel ... an, um die Lösung zu finden.“)
4. **Analyse** („Was beeinflusste die Schriftstellerei von...?“)
5. **Synthese** („Wie könnten wir Geld einwerben für...?“)
6. **Bewertung** („Was bevorzugst du und warum?“)

Diese verschiedenen Anspruchsniveaus sind vergleichbar mit den bereits bekannten Anforderungsbereichen der Bildungsstandards (vgl. Abschnitt 1.1). Der Anspruch von Lehrerfragen leitet somit direkt vom *unterstützenden und schülerorientierten Klassenklima* über zur *kognitiven Aktivierung*.

1.5.4.2.3 Kognitive Aktivierung

Während *effektive Klassenführung* und *unterstützendes Klassenklima* lerngruppen-einheitliche Merkmale von Unterricht darstellen, handelt es sich bei der *kognitiven Aktivierung* eher um ein individuelles „internes“ Merkmal, welches erst seit vergleichsweise kurzer Zeit Beachtung findet (Baumert et al., 2004; Klieme et al., 2006; Lipowsky, 2009). Schülerinnen und Schüler sollen geistig aktiv sein (Blum, 2015; Schoenfeld, 1992) und möglichst selbstständig arbeiten (Blum, 2006), was in der Unterrichtsbeobachtung beispielsweise daran festgemacht wird, ob komplexere Fragen/Probleme gestellt werden, Begründungen und Erklärungen eingefordert werden und wie kleinschrittig die Lehrperson Erklärungen und Anweisungen gibt (Klieme et al., 2006; Lipowsky, Rakoczy, Buff & Klieme, 2005; Lipowsky, Drollinger-Vetter, Hartig & Klieme, 2006; Lipowsky et al., 2009)²⁷. In COACTIV wurde beispielsweise kontrastierend das Potenzial eingesetzter Aufgaben zur *kognitiven Aktivierung* durch Analysen untersucht (Baumert et al., 2008; Jordan et al., 2006; Kunter et al., 2011; Neubrand, Jordan, Krauss, Blum & Löwen, 2011).

Der *kognitiven Aktivierung* liegt aus theoretischer Sicht u. a. der (Sozial-)Konstruktivismus zu Grunde (Lipowsky, 2009; Ufer, Heinze & Lipowsky, 2015; E. Wild & Möller, 2009, 2015). Schülerinnen und Schüler werden bei kognitiven Konflikten mit Informationen konfrontiert, die sich mit den bisherigen Vorstellungen nur bedingt in Einklang bringen lassen. Dieses etwaige Ungleichgewicht muss entweder durch Eingliederung der Informationen in vorhandene Schemata (Assimilation) oder durch Anpassung der vorhandenen Schemata (Akkommodation) ausgeglichen werden (Äquilibration; Piaget, 1976). Ebenfalls Bezüge finden sich zum Sozialkonstruktivismus von Vygotskiĭ et al. (1978), nach dem Lernen und kognitive Aktivität sich vor allem im Austausch mit (kompetenteren) Anderen vollzieht (vgl. auch Reusser, 2001). Ausgangspunkt für einen kognitiven Konflikt können im Unterricht beispielsweise „Irrwege“ bei Aufgabebearbeitungen sein, die von der Lehrkraft in der selbstständigen Arbeit der Schülerinnen und Schüler zunächst zugelassen werden, um diese im Nachhinein aufzugreifen, einen Vergleich von verschiedenen Wegen/Lösungen anzustreben oder

²⁷ Wenngleich hier keinesfalls ausschließlich komplexe und schwere Fragen mit hoher *kognitiver Aktivierung* gleichzusetzen sind. Vielmehr muss das Frageniveau dem Wissens- und Kompetenzstand der Schülerinnen und Schüler entsprechen (Brophy & Thomas, 1984; Katzenbach, 2011a; Katzenbach, 2011b). Hierbei wird der Begriff der *Aptitude-Treatment-Interaction* (ATI) angesprochen. Hattie (2009) berichtet eine insgesamt jedoch nur gering ausfallende Effektstärke ($d = 0.19$) für den Zuschnitt von Methoden auf Schülermerkmale auf die Leistung.

als Diagnosemöglichkeit bei der Suche nach Schwierigkeiten und Fehlvorstellungen zu nutzen (Blum & Leiß, 2003, 2005, 2006, 2007a, 2007b; Blum, 2015; Neubrand, 2012; Schukajlow & Krug, 2013; Sjuts, 2008, 2012; Stein & Lane, 1996). Insbesondere durch den Sozialkonstruktivismus sind die Ergebnisse von Hattie et al. (1996) und Kramarski und Mevarech (2003) zu erklären, die bei kooperativ arbeitenden Schülerinnen und Schülern eine erhöhte Flexibilität in den Erklärungen und den Leistungen finden (vgl. auch Blum, 2015; Lipowsky et al., 2009; Lipowsky, 2009). Ähnlich positive Effekte berichtet auch Hattie (2009) für kooperativ, kognitiv aktivierende Unterrichtsmethoden wie das Peer-Tutoring ($d = 0.55$), wobei der Lerngewinn sowohl für die Unterrichtenden als auch die Unterrichteten gleichermaßen gilt. Den direkten Einfluss von *kognitiver Aktivierung* auf die Lernleistung belegen viele Studien in verschiedenen Designs und Auswertungsmethoden (Hiebert & Wearne, 1993; Klieme et al., 2006; Shayer & Adhami, 2007; Stein & Lane, 1996; Taylor, Pearson, Peterson & Rodriguez, 2003; Wang et al., 1993): Beispielsweise zeigen Baumert und Lehmann (1997) sowie Klieme et al. (2001) positive Zusammenhänge zwischen der *kognitiven Aktivierung* und der Leistung in TIMSS. Lipowsky et al. (2009) zeigen unter Beachtung der Mehrebenenstruktur mit einem Regressionskoeffizient von $\beta = .10$ ($p < .05$) den positiven Einfluss der *kognitiven Aktivierung* auf die Lernleistung als Posttestleistung bei Kontrolle von Vortestleistung (sowohl auf Klassen- als auch Individualebene), der Beschulungszeit, Interesse etc. Rakoczy et al. (2007) bestätigen diese Befunde mit einem mittleren Effekt ($\beta = .33$, $p < .05$) der kognitiven Aktivierung auf die Motivation und einem kleinen Effekt ($\beta = .13$, $p < .05$) auf die Leistung.

Kognitive Aktivierung steht darüber hinaus in direktem Zusammenhang mit den zuvor genannten Qualitätskriterien, was die eng ineinander übergehenden Ausführungen bereits nahelegen: Ein gutes *classroom management* wird von Baumert und Lehmann (1997, S. 54) als Voraussetzung für *kognitive Aktivierung* beschrieben (auf Basis eines Zusammenhangs beider Merkmale). Analog unterstützen Rakoczy et al. (2007) die Behauptung mit positiven Regressionskoeffizienten von $\beta = .14$ ($p < .05$) für die Strukturierung der Lernumgebung und $\beta = .20$ ($p < .05$) für die *Strukturierung* der inhaltlichen Präsentation auf die *kognitive Aktivierung* auf Klassenebene unter Kontrolle weiterer Klassenebenen- und Individualebenenmerkmale. Im Projekt PALMA (Projekt zur Analyse der Leistungsentwicklung in Mathematik; Pekrun et al., 2006; Vom Hofe, Pekrun, Kleine & Götz, 2006) wurden u. a. ebenfalls die Zusammenhänge von Unterrichtsqualitätskriterien untersucht und ein negativer Zusammenhang von

Teil 1 – Theorie

Schulentwicklung

$r = -.22$ ($p < .01$) für *ineffektive Klassenführung* und *kognitive Aktivierung* gefunden, also wiederum: Bei schlechterer *Klassenführung* fällt die *kognitive Aktivierung* der Lerngruppe geringer aus. Turner et al. (1998) beschreiben darüber hinaus einen positiven Zusammenhang mit dem Klassenklima, sodass sich bei einem *unterstützend wahrgenommenen Klassenklima* mehr *kognitive Aktivierung* ergibt (Lipowsky, 2009). Im Rahmen von PALMA wurde das Unterrichtsqualitätsmerkmal der *Individualisierung* (individuelle Bezugsnormorientierung, Unterstützung nach Misserfolg etc.) betrachtet und ein Zusammenhang von $r = .78$ ($p < .01$) zur *kognitiven Aktivierung* gefunden (Pekrun et al., 2006, S. 44).

In Abbildung 13 zeigen sich die drei erstgenannten Qualitätskriterien in einem visuellen Zwischenfazit aufgegliedert nach der vermuteten bzw. weitestgehend bestätigten Wirkung sowie den zugrundeliegenden Nutzungsaspekten.

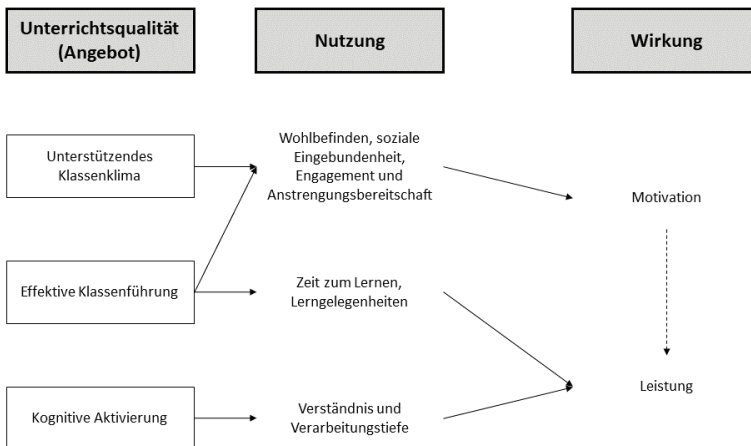


Abbildung 13: Basisdimensionen guten Unterrichts und deren vermutete Wirkung (Klieme et al., 2006, S. 131)

Klieme et al. (2006, S. 130) beschreiben im Zusammenhang mit den sog. Basisdimensionen mit dem Blick auf die *kognitive Aktivierung* eine zentrale Einschränkung: „Was als ‚kognitiv aktivierendes‘ Unterrichtsmerkmal gelten kann, ist [...] abhängig von fachdidaktischen und denkpsychologischen Erkenntnissen zu bestimmen.“ Damit lassen sich die bisherigen Kriterien (vor allem das letztgenannte) nicht ohne den fachlichen Hintergrund betrachten (Schlesinger & Jentsch, 2016). Bevor nun näher auf den

fachlichen Anspruch eingegangen wird, soll eine kurze Erweiterung der *kognitiven Aktivierung*, die *metakognitive Aktivierung*, kurz erläutert werden.

1.5.4.2.4 Metakognitive Aktivierung der Lernenden

Vielfach wird unter der *kognitiven Aktivierung* auch die *metakognitive Aktivierung* subsumiert: Insbesondere für die Fortbildungsgestaltung ist es jedoch lohnend, diese explizit zu betonen, da Metakognition und Selbstregulation wichtige und wirkungsvolle Aspekte von Unterricht darstellen. Meta-analytisch wird die Bedeutung von metacognitive instruction 1996 von Hattie et al. bestätigt ($d = 0.71$; bei 20 Studien aus der Domäne Sprache) und später nochmals abgesichert: meta-kognitiven Strategien ($d = 0.69$) und Lerntechniken ($d = 0.59$; Hattie, 2009). Bei Wang et al. (1993) liegt die *metakognitive Aktivierung* als zweitwirksamstes Kriterium direkt hinter dem *classroom management*. In der Wirkung scheint der Einsatz von Lernstrategien (als Mediator der Gehemmtheit) positiv auf Lernerfolg zu wirken (Artelt, 1999). Positive Zusammenhänge der Selbstregulation mit Interesse ($r = .73$, $p < .001$) und Selbstwirksamkeitserwartung ($r = .44$, $p < .001$) sowie Leistung ($r = .36$, $p < .001$) berichten Pintrich und Groot (1990). Cardelle-Elawar (1995) zeigt varianzanalytisch einen positiven Einfluss einer auf Metakognition ausgerichteten Intervention auf die Mathematikleistung ($F(1,17) = 85.90$, $p < .001$).

Aus theoretischer Sicht bekommt die *metakognitive Aktivierung* durch den Paradigmenwechsel zur Outputorientierung und der damit verbundenen Neuausrichtung von Unterricht vor dem Hintergrund des Angebots-Nutzungs-Modells (Abbildung 12) eine wesentliche Bedeutung, da Schülerinnen und Schüler „zunehmend eigenständige Lernerinnen und Lerner“ sein müssen, um „in wachsendem Maße Selbstverantwortung für ihr Lernen zu übernehmen und ihre Lernstrategien weiterzuentwickeln“ (Oelkers & Reusser, 2008, S. 19). In diesem Sinne sollten metakognitive Aktivitäten systematisch und langfristig angelegt in den Fachunterricht integriert werden (Ufer et al., 2015). Die *metakognitive Aktivierung* soll dazu dienen, dass Schülerinnen und Schüler ihre eigenen Lernprozesse planen, überwachen und evaluieren können. In diesem Sinne liegen die Modelle des selbstregulierten Lernens zugrunde. Hier soll nun lediglich kurz auf das Prozessmodell eingegangen werden (vgl. S. Krapp, 2006; Landmann et al., 2009; Schmitz, Landmann & Perels, 2007; zum Schichtmodell der Selbstregulation vgl. Boekaerts & Minnaert, 1999): In diesem Modell werden eine *präaktionale*, eine *aktionale* und eine *postaktionale* Phase unterschieden, in der die

Lernhandlung von den Schülerinnen und Schülern mit Rücksicht auf Ziele, Ressourcen und Anforderungen geplant, im Verlauf reflektiert und rückblickend mit vorliegendem Ergebnis evaluiert wird (vgl. auch Sembill, Wuttke, Seifried, Egloffstein & Rausch). Ähnlich den Ausführungen zu Evaluation (vgl. Kapitel 1.4) sollten sich auch beim selbstregulierten Lernen aus der *postaktionalen* Phase Erkenntnisse für die nächste *präaktionale* Phase ableiten bzw. diese bei Nichterreichen des Ziels direkt eingeleitet werden. Dementsprechend handelt es sich bei diesem Modell zwar im Namen um ein „Phasenmodell“ zugrunde liegt aber vielmehr ein Kreislauf.

Ebenfalls als Kreislauf sind der Modellierungskreislauf (vgl. Abbildung 3) und der Problemlösekreislauf (Polya, 1967) angelegt. Aus den Untersuchungen der TIMSS-Videostudie ist bekannt, dass „typische“ Mathematikstunden im deutschsprachigen Raum gegenüber Unterrichtsskripten aus Japan nur wenig auf Reflexion und Metakognition, sondern vielmehr auf Einüben von Routinen und Verfahren ausgerichtet sind (Baumert & Lehmann, 1997; Blum & Neubrand, 1998; Blum, 2001). In DISUM (Blum & Leiß, 2003, 2005, 2006, 2007a, 2007b) wurde daher als letzte Phase im Unterricht explizit eine Reflexionsphase eingefordert. Dies ergibt vor dem Hintergrund der TIMSS-Ergebnisse Sinn, da metakognitiv geschulte Schülerinnen und Schüler zu höheren Leistungen kommen und eine größere Flexibilität bei Erklärungen aufweisen (Baumert & Lehmann, 1997; Blum, 2015; Kramarski & Mevarech, 2003). Auch hier gilt wiederum, dass von Seiten der Lehrkräfte diese Heuristiken Schülerinnen und Schülern als Strategien vermittelt werden müssen (vgl. Lösungsplan in DISUM). Durch diese *Strukturierung* von geöffneten Lernumgebungen wird in der Selbstständigkeit ein größerer Lerngewinn erreicht als in direkter Instruktion ($d = 0.30$, $p < .05$; Alfieri, Brooks, Aldrich & Tenenbaum, 2011), was vor dem Hintergrund der *cognitive-load-theory* sehr gut nachvollziehbar ist (Hardy, Jonen, Möller & Stern, 2006; Paas, Renkl & Sweller, 2003; Sweller, van Merriënboer & Paas, 1998; E. Wild & Möller, 2015).

1.5.4.2.5 Fachlich gehaltvolle Unterrichtsgestaltung

Alle bisher genannten Kriterien zielen darauf ab, den Lernerfolg von Schülerinnen und Schüler zu verbessern. Baumert und Kunter (2011) greifen den Gedanken des Angebots-Nutzungs-Modells (Abbildung 12) auf und erweitern dies an entscheidender Stelle, nämlich dem Übergang von Angebot und Nutzung zur Wirkung. An dieser Stelle sind „verständnisvolle Lernprozesse“ notwendig, um aus einer veranlassten

kognitiven Aktivierung tatsächlichen Lerngewinn auf Seiten der Schülerinnen und Schüler zu generieren (Leuders, 2012).

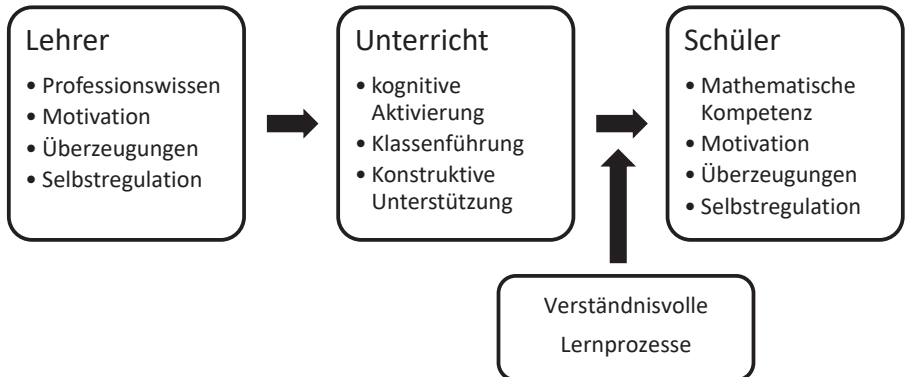


Abbildung 14: Wirkung professioneller Kompetenz (nach Löwen, Baumert, Kunter, Krauss & Brunner, 2011, S. 70)

Die Bedeutung eines fachlich anspruchsvollen Unterrichts scheint zunächst intuitiv verständlich und lässt sich auch empirisch belegen: Höhere Ansprüche seitens der Lehrkraft sind vielfach als Indikator für mehr Lernerfolg genannt worden (zsf. Hattie, 2009 mit $d = 0.43$). In Verbindung zur *metakognitiven Aktivierung* zeigen Hattie et al. (1996), dass mit dem Ziel des einfachen Behaltens reine Wiederholungsstrategien (also im Terminus der Lernstrategien *kognitive Strategien*; Landmann et al., 2009; K.-P. Wild & Schiefele, 1994), für anspruchsvollere Unterrichtsinhalte hingegen (vgl. Bloomsche Taxonomie) entsprechend komplexere Strategien notwendig sind.

Lambert (2011) führt aus fachdidaktischer Sicht den „wissensbasierter Konstruktivismus“ an, der ergänzend zum Konstruktivismus darauf fußt, dass räumlich-zeitliche und soziale Lernkontexte den Erfahrungsraum vorgeben (Reinmann-Rothmeier & Mandl, 2005). Entsprechend weisen die Ergebnisse von Hattie et al. (1996) darauf hin, dass Transfer, insbesondere Ferntransfer nicht erwartbar ist (Blum, 2015), also durch Nutzung eines Rechenverfahrens nicht gelernt wird, dieses zu bewerten o. ä. Dies legt nahe, dass zum Erwerb von Wissen und Fähigkeiten, Lernmöglichkeiten geboten werden müssen, die inhaltlich in etwa dem entsprechen, was gelernt werden soll („Es gibt kein Stricken ohne Wolle“), da insbesondere die Kompetenzdefinition

der Bildungsstandards von konkreten Anforderungssituationen spricht (Bruder, 2012).

Um einen verständnisorientierten Unterricht mit fachlichem Gehalt zu erreichen, gilt es eine *neue Aufgabenkultur* zu implementieren (Blum & Leiß, 2006, 2007a, 2007b; Bruder, 2006), die eher verstehens- statt verfahrensorientierte Aufgaben nutzt (Blum & Neubrand, 1998; Leuders, 2012, 2015; Wynands, 2012). Dies wird erreicht, indem beispielsweise

- inner- und außermathematische Kontexte genutzt und entsprechende Vernetzungen gefordert werden,
- Begründungen erwartet werden (vgl. Bloomsche Taxonomie),
- Fragestellungen umgekehrt werden,
- (Gegen-)Beispiele angegeben werden sollen,
- Fragestellungen dynamisiert werden („Was passiert, wenn...“).

Entsprechende Aufgaben sollten dazu ausreichend offen und nicht zu kleinschrittig gestaltet sein, sodass Schülerinnen und Schüler sofern benötigt Unterstützung finden (etwa durch gestufte Hilfen), herausfordernde Probleme gestellt bekommen sowie Kommunikation und Kooperation gefördert werden (Leuders, 2012; Biermann, Wiegand & Blum, 2003; Bruder, 2003, 2006, 2012; Büchter & Leuders, 2007; Girmes, 2003; Hammer & Ufer, 2015; Leuders, 2011, 2015).

Neben der Betrachtung des reinen Unterrichts und dessen Qualität geht insbesondere der Hessische Referenzrahmen Schulqualität (Diel, 2011; Hessisches Kultusministerium & IQ, 2011) eine Stufe höher und analysiert darüber hinaus auch die Schulqualität. Dabei bildet selbstverständlich der Unterricht den Kern der Schule (A. Hosenfeld, 2010; Schildkamp & Kuiper, 2010), doch spielen weitere Aspekte wie etwa die Führungsqualität der Schulleitung, Schulkultur bzw. Evaluationskultur der Schule, Schulprogrammarbeit, Verhalten von Schülerinnen und Schülern untereinander sowie gegenüber den Lehrkräften etc. ebenfalls eine Rolle (Bischof et al., 2013; Blum et al., 2006; Fend, 1986, 2008, 2008; Petersen, 2013; van Ackeren & Klemm, 2011). An den angegebenen Orten kann hierzu vertiefend recherchiert werden.

Mit diesen Ausführungen ist nun der Evaluationskreislauf geschlossen und die möglichen Entwicklungsrichtungen, die aus den Informationen der Evaluation resultieren, wurden mit der Personal-, Organisations- und Unterrichtsentwicklung aufgezeigt.

Alles zusammen mündet in eine tragfähige Schulentwicklung. Dies gelingt nach Ramm (2014, S. 46) nur, „wenn Rückmeldungen über den Stand eingeholt, darüber reflektiert und Steuerungsmaßnahmen ergriffen werden“. Im kommenden Kapitel werden die Vergleichsarbeiten (VerA) bzw. Lernstandserhebungen als Operationalisierung eines bildungsstandardsbezogenen Testverfahren näher vorgestellt, die auf den Erhalt von Steuerungswissen für Unterricht und Schule zielen, Schulentwicklung durch Entwicklung von Fachunterricht anstoßen und Schülerleistungen diagnostizieren wollen (Nachtigall & Jantowski, 2007, S. 402).

1.6 Vergleichsarbeiten (VerA) und Lernstandserhebungen

Eine überwältigende Mehrheit von 83% der deutschen Bevölkerung spricht sich demgegenüber dafür aus, mit Hilfe von einheitlichen Vergleichstests für alle SchülerInnen zu überprüfen, ob die SchülerInnen in jeder Schule die in den Bildungsstandards deutschlandweit festgelegten Lernziele erreichen [...]. Unter den Lehrkräften ist die Zustimmung zu einheitlichen Vergleichstests für alle SchülerInnen zwar signifikant niedriger als in der Gesamtbevölkerung, jedoch spricht sich auch unter den LehrerInnen eine klare Mehrheit von 66% dafür aus. (Wössmann et al., 2016, 24f.)

Die Einführung der Bildungsstandards in Deutschland (Abschnitt 1.1) brachten Veränderungen für das deutsche Schulsystem sowie die Einzelschule mit sich. Auf diesen abschlussbezogenen Bildungsstandards bauen die Vergleichsarbeiten (VerA) bzw. Lernstandserhebungen (LSE) in der dritten und achten Jahrgangsstufe auf²⁸ und bilden damit eine Operationalisierung der *Output*erwartungen, die in den Bildungsstandards formuliert sind (Maag Merki, 2010). Die ersten Überlegungen zur regelmäßigen Teilnahme an Leistungsvergleichen sowie zu deren Ergänzung durch länderbezogene Qualitätssicherungsmaßnahmen gehen in der jüngeren Bildungsdebatte auf die Konstanzer Beschlüsse von 1997 zurück (Hessisches Kultusministerium, 2009; KMK, 1997; LSA, 2015). Einige Jahre später wurden im Zuge der konkreten Vorbereitung von Bildungsstandards für die Primarstufe sowie den Hauptschulabschluss und den mittleren Bildungsabschluss auf den Plenarsitzungen in Eisenach und Würzburg von der Kultusministerkonferenz die Einführung von „Orientierungs- und Vergleichsarbeiten“ zur Überprüfung der Einhaltung von Bildungsstandards beschlossen (Helmke & I. Hosenfeld, 2003; KMK, 2002a, 2002c; Leutner et al., 2008). Ausgehend von diesen Beschlüssen gestalteten ab Anfang der 2000er Jahre einzelne Bundesländer, v. a. Rheinland-Pfalz mit der Universität Koblenz-Landau, Vorläuferversionen der mittlerweile deutschlandweit einheitlichen Vergleichsarbeiten mit dem Ziel, „Qualitätsentwicklung und -sicherung [...] auf allen Ebenen des Schulsystems“ – hier konkret auf der Ebene von Einzelschulen und Unterricht – aufzubauen (KMK, 2001, S. 1). Mit der

²⁸ Dieses Instrument wird in einzelnen Bundesländern z. T. unter anderem Namen geführt, so beispielsweise *Kompetenztest*, *KERMIT* (Kompetenzen ERMITteln) etc.

Gründung des Instituts für Qualitätsentwicklung im Bildungswesen (IQB) stand ein nationales Institut zur Koordinierung von Vergleichsarbeiten und deren Entwicklung zur Verfügung (Hessisches Kultusministerium, 2009; KMK & IQB, 2006; KMK, 2015; Leutner et al., 2008).

Seit dem Schuljahr 2008/2009 werden unter der administrativen Leitung des IQB deutschlandweit Vergleichsarbeiten in den Jahrgangsstufen 3 und 8 in den Fächern Mathematik und Deutsch sowie in Jahrgang 8 in der ersten Fremdsprache (Englisch/Französisch) durchgeführt. Dieses zentral administriert und dezentral durchgeführte Verfahren lässt sich als bundeslandinternes Verfahren zur Qualitätsentwicklung einstufen. Dabei sind Vergleichsarbeiten eng an den Bildungsstandards orientiert und werden „eingesetzt (werden) mit dem Ziel, die Leistungen der Schüler an einer klassen- und schulübergreifenden sozialen und/oder kriterialen Bezugsnorm zu messen“ (Helmke & I. Hosenfeld, 2003, S. 10). Die Vergleichsarbeiten gehen mit umfassenden Auswertungen u. a. auch mit überregionalem, jedoch nicht bundeslandübergreifendem Fokus einher (Korngiebel, 2009).²⁹

Vergleichsarbeiten sind zunächst einmal nichts anderes als zentral entwickelte Leistungstests, zu denen verschiedene Rückmeldungen bereitgestellt werden. Zu Grunde liegen bei der Entwicklung die Bildungsstandards: Nach den bisherigen Ausführungen zu Bildungsstandards wird deutlich, dass Vergleichsarbeiten nicht den Anspruch eines inhaltlich allumfassenden Tests führen; vielmehr sollen zentrale und grundlegende Kompetenzen auf dem Weg zum angestrebten Abschluss überprüft werden (vgl. Abbildung 2; Dobbelstein & Peek, 2008; Hessisches Kultusministerium, 2009; LSA, 2015). Einschränkend muss jedoch angemerkt werden, dass mit den Vergleichsarbeiten lediglich eine querschnittliche Messung der Leistung von Schülerinnen und Schülern vorhanden ist, die keine kausalen Schlüsse bzgl. der Leistung zulässt, sondern diese zunächst nur beschreibt (Kuhn, 2014, S. 423). Diese Problematik lässt sich jedoch im Falle von VerA umgehen, sodass auch kausale Schlüsse sowie individual-diagnostische Urteile aus den Vergleichsarbeiten abzuleiten sind, wenn neben der allgemeinen

²⁹ Mit den Vergleichsarbeiten werden von kultusministerieller Seite in den einzelnen Bundesländern verschiedene Ziele verbunden, woraus unterschiedliche Umsetzungen in den einzelnen Bundesländern resultieren: Im Speziellen soll sich als Grundlage für die später beschriebene Fortbildungsreihe VELM-8 (Teil 2) nun auf die für Hessen gültigen Informationen zu Lernstandserhebungen beschränkt werden. Ausführungen zu den Umsetzungen und Zielen anderer Bundesländer finden sich beispielsweise für Thüringen bei Nachtigall (2014) oder für Nordrhein-Westfalen in Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (2011, 2012).

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

Rückmeldung einzelne individuelle Aufgabenbearbeitungen diagnostisch analysiert werden und mit vorherigen diagnostischen Informationen zu dieser Schülerin bzw. diesem Schüler verbunden werden (vgl. auch Bartnitzky, 2005). In diesem Sinne wird die Vergleichsarbeit dann zur Kompetenzdiagnostik genutzt (vgl. Tabelle 4).

1.6.1 Ziele der Vergleichsarbeiten

Von vielen Seiten werden an die Vergleichsarbeiten unterschiedliche Funktionen und Ziele herangetragen: Dabei nehmen die Lehrkräfte diese Funktionen jedoch in unterschiedlichem Maße wahr bzw. schreiben den Vergleichsarbeiten keine eindeutige Funktion zu (Richter et al., 2014). Die Möglichkeit, die Vergleichsarbeiten zu nutzen, unterscheidet sich jedoch durch föderal bedingte Unterschiede. Insbesondere kann das Instrument nur bedingt den Spagat zwischen individueller Förderung und Systemmonitoring leisten: Die Vergleichsarbeiten stehen wegen der zentralen Planung in der Einordnung als *interne* und *externe* Evaluation in Kombination (durch zentrale Planung und dezentrale Durchführung) quasi an der Schnittstelle, sodass Ziele aus beiden Evaluationsformen zu finden sind (vgl. Abschnitt 1.4.1; Baumert, 2001b). Die meisten der mit diesem Verfahren verbundenen Ziele sind trotz föderal bedingter Unterschiede gesamtdeutsch gültig und lassen sich im Wesentlichen auf drei Ziele zusammenführen:

1. Ergänzung des Systemmonitorings
2. Orientierung
3. Evaluation zur Schul- und Unterrichtsentwicklung

Tabelle 4: Zielsetzung Kompetenzdiagnostik, Vergleichsarbeiten und Bildungsmonitoring (verändert nach Leutner et al., 2008, S. 151)

	Kompetenzdiagnostik	Vergleichsarbeiten	Bildungsmonitoring (der KMK)
Zielsetzung	Vorbereitung von Entscheidungen im <u>Einzelfall</u> (auch für Selektion und/oder Platzierung)	Vorbereitung pädagogischer, didaktischer und/oder curricularer Entscheidungen auf <u>Schul- und Unterrichtsebene</u> (Selbstevaluation)	Vorbereitung politischer Entscheidungen auf <u>Schulsystemebene</u> (Fremdevaluation)
Fokus	Inhaltliche Tiefe in einem Teilgebiet des Faches: Kompetenzschätzung des Individuums	Inhaltliche Breite (jahrgangsübergreifend); relative inhaltliche Tiefe: Vollerhebung aller Personen Stichprobe von Items Kompetenzverteilung in Klassen, Schulen (keine Individualdiagnostik)	Relative fachliche Tiefe und zugleich fachliche Breite: Multi-Matrix-Sampling; Stichprobe von Items und Stichprobe von Personen Kompetenzverteilung in Bundesländern, Staaten (keine Individualdiagnostik, keine Aussagen auf Schul- und Klassenebene)

1.6.1.1 Systemmonitoring

Auf Ebene des gesamten Schulsystems respektive dem Schulsystem der einzelnen Bundesländer sollen die Vergleichsarbeiten ergänzend zu anderen Verfahren Informationen zur Leistungsfähigkeit des Bildungssystems liefern (Burkard & G. Orth, 2005; Oelkers & Reusser, 2008), Planungs- und Steuerungswissen erzeugen, um so die Schulaufsicht zu unterstützen (Richter et al., 2014) und Ergebnisse der Schulinspektionen erweitern/ergänzen (Wacker & Kramer, 2012) sowie Unterstützungs- und Entwicklungsbedarf (wiederum auf Bundesländerebene) identifizieren (Kühle & Peek, 2007; Maier, 2010; EMSE, 2006). Diese Funktion wird jedoch von einigen Autoren als eher nachrangig beschrieben: Ebenso wenig wie zur Grundlagenforschung eignen sich Vergleichsarbeiten zum *Systemmonitoring*, da es sich nicht um repräsentative Erhebungen handelt, also die Ergebnisse keine Verlässlichkeit in der Verallgemeinerung bieten (I. Hosenfeld et al., 2006; Leutner et al., 2008). Es handelt sich jedoch auch nicht um Vollerhebungen (bezogen auf alle Schülerinnen und Schüler eines Bundeslandes), sodass es zu systematischen Verzerrungen aufgrund der Stichprobe kommen kann, was letztlich im *Systemmonitoring* zu falschen Schlüssen führen würde. Tabelle 4 zeigt im oberen Teil der Zeile „Fokus“ die Unterschiede zwischen System-/Bildungsmonitoring, Kompetenzdiagnostik und Vergleichsarbeiten. Vergleichsarbeiten sind nicht darauf ausgelegt, „Ergebnisqualität des Fachunterrichts insgesamt zu erfassen“ (Dobbelstein & Peek, 2008, S. 19) und werden nicht von externen Testleitern in den Schulen durchgeführt, sodass die Objektivität des Verfahrens nicht sichergestellt ist (Leutner et al., 2008). Insbesondere der Unterschied in der Auswahl von Testinhalten ist relevant: Während im *Systemmonitoring* die gänzliche fachliche Breite und Tiefe abgedeckt werden sollte, was üblicherweise durch Multi-Matrix-Sampling geschieht, sind die Vergleichsarbeiten auf eine bildungsstandardskonforme, abschlussbezogene fachliche Breite ausgelegt. Entsprechend dem Rat der Expertengruppe um Eckhard Klieme nach „deutlicher Trennung zwischen [...] Tests für Evaluation, Bildungsmonitoring und [...] individuelle Förderung“ bei der Entwicklung von Bildungsstandards (Klieme et al., 2003, 2009, S. 49) fordert Maritzen (2014, S. 408) „eine systematische Fokussierung der Instrumente und Verfahren auf diesen Zweck“, nämlich die „Grundfunktion von VerA 3 und VerA 8 als Instrument der Schul- und Unterrichtsentwicklung“ anstelle der systematischen Überfrachtung des Instruments (Kuper & Diemer, 2012; Maier, Metz, Bohl, Kleinknecht & Schymala, 2012).

1.6.1.2 Orientierung

Bevor das hier schon genannte Hauptziel „Schul- und Unterrichtsentwicklung“ näher thematisiert wird, soll zunächst verdeutlicht werden, dass die Vergleichsarbeiten – quasi als Zwischenschritt (über Tabelle 4 hinausgehend) – in ihrer Anlage für Lehrkräfte eine gute *Orientierung* bieten (Bonsen et al., 2006; Burkard & G. Orth, 2005; EMSE, 2006; Groß-Ophoff, Koch, Hosenfeld & Helmke, 2006; Korngiebel, 2009; Kühle & Peek, 2007; Strähling, 2005; Stryck, 2005). Seit ihrer Einführung sind die Bildungsstandards verpflichtende Grundlage für die Ausbildung von Schülerinnen und Schülern sowie die universitäre Ausbildung von Lehrkräften (vgl. auch Standards für die Lehrerbildung; KMK, 2004c). Schulbücher, Unterrichtsmaterialien und Unterrichtsabläufe werden von Lehrkräften jedoch nicht in jedem Schuljahr gewechselt und im Rahmen von Weiterbildungen sind ggf. auch andere Themen relevant, sodass die Implementierung von Bildungsstandards in den Schulalltag einige Zeit braucht (Herzog, 2006; Holtappels, 2014b). Aus den Bildungsstandards resultiert im Fachunterricht allerdings u. a. eine neue Aufgabenkultur (Baptist, 1998, 2001; Baptist & Raab, 2007; Bruder, 2006; Hessisches Kultusministerium, 2009; LSA, 2015). Hierzu bieten Vergleichsarbeiten mit bildungsstandardsbezogenen Aufgaben eine gute *Orientierung* und illustrieren Bildungsstandards in einer unterrichtsnahen Form (vgl. auch Aufgabensammlungen in Blum et al., 2012; Blum et al., 2015; Katzenbach et al., 2009a, 2009b). Implizit wird diese Funktion der Vergleichsarbeiten ebenfalls daran deutlich, dass im Sinne der Ergänzung des *Systemmonitorings* die Erreichung der Bildungsstandards geprüft und eine gewisse Standardsicherung erreicht werden sollen (Bremerich-Vos, Groß-Ophoff, Helmke, Hosenfeld & Wagner, 2005; Burkard & G. Orth, 2005; Lorenz, 2005; Peek & Dobbstein, 2006a; Peek et al., 2006). Durch diesen *washback*-Effekt ist zu erwarten, dass die Grundlage von Prüfungen bzw. *Systemmonitoring*-Untersuchungen sich im Unterricht widerspiegelt und in diesem Sinne eine (unbeabsichtigte) Normendurchsetzung erfolgt (vgl. Funktionen von *externer* Evaluation in Abschnitt 1.4.1; Anderson & Wall, 1993; Wall, 2000).

1.6.1.3 Evaluation zur Schul- und Unterrichtsentwicklung

Neben der *Orientierung* werden mehrere Funktionen bzw. Ziele mit den Vergleichsarbeiten verbunden, die im Folgenden unter dem Punkt *Evaluation* zur SE und UE zusammengefasst werden sollen: Vergleichsarbeiten sollen eine

1. Standortbestimmung der Fachleistungen der Schülerinnen und Schüler erlauben und davon ausgehend
2. zur Entwicklung der diagnostischen und professionellen Kompetenz von Lehrkräften beitragen sowie
3. Schul- und Unterrichtsentwicklung anregen (Strähling, 2005).

Separiert man die erstgenannte Funktion der Standortbestimmung von den anderen beiden *Entwicklungsfunktionen*, so erkennt man die Beziehung zum Evaluationskreislauf: Aufbauend auf einer Messung und deren Interpretation werden entsprechende Maßnahmen ergriffen.

1.6.1.3.1 Standortbestimmung der Fachleistung

Ziel der Lernstandserhebungen ist eine „frühzeitige, solide Standortbestimmung“ (Hessisches Kultusministerium, 2009, S. 10). Damit bekommen die Vergleichsarbeiten in Klasse 3 und 8 als Lernstandserfassung (Büchter & Leuders, 2005a) auf dem Weg zu den Abschlussprüfungen die Funktion einer „Warnlampe“ bzw. eines „Screenings“ (Brügelmann, 2005b; Burkard & G. Orth, 2005; Leutner et al., 2008). Diese Standortbestimmung erfolgt dabei jedoch nicht (nur) ipsativ, sondern an verschiedenen Normen (v. a. kriterial, sozial) orientiert, sodass eine objektive und an Vergleichbarkeit orientierte Messung und Interpretation möglich ist (Groß-Ophoff et al., 2006; Helmke et al., 2003; A. Hosenfeld, 2010; Kühle & Peek, 2007; Leuders, 2006; Peek & Dobbstein, 2006b; Posch, 2009; Reimers et al., 2009; Strähling, 2005), aus der jedoch insbesondere keine Selektion folgen soll (Kuper & Diemer, 2012). Nach der Testdurchführung und deren Bewertung erhalten die Lehrkräfte Rückmeldungen zu den Ergebnissen der Vergleichsarbeiten auf Klassenebene (der detailliertere Ablauf folgt in Abschnitt 1.6.2). Für Hessen gliedern sich die Rückmeldeformate in online unmittelbar verfügbare Rückmeldungen in Form einer tabellarischen Übersicht und eines Sofortberichts sowie je einen etwas später ebenfalls online verfügbaren Ergebnis- und Schulberichts (Hessisches Kultusministerium, 2009; LSA, 2015). Die unmittelbar verfügbaren Berichte dienen dazu, Vergleiche mit Parallelklassen anzustreben, die Erwartungen mit den tatsächlichen Leistungen der Schülerinnen und Schüler in Relation zu setzen oder etwa im Sinne eines Prä-Post-Test-Designs vorherige Leistungsmessungen mit den Ergebnissen der VerA zu vergleichen (Helmke & I. Hosenfeld, 2005). Büchter und Leuders (2005b) führen an, dass hierbei v. a. die quantitativen Daten als Vergleichsmöglichkeit auf Klassenebene herangezogen werden.

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

Individualdiagnostische Schlüsse lassen sich eher aus qualitativen Daten, etwa der Analyse von individuellen Aufgabenbearbeitungen ableiten, um eine ausreichende fachliche Tiefe gewährleisten zu können (Kompetenzdiagnostik im Sinne von Tabelle 4).

Auf den folgenden Seiten sollen nun die vier genannten Rückmeldungen der Lernstandserhebungen mit den zur Verfügung stehenden Vergleichsinformationen auf Klassenebene näher betrachtet werden³⁰. Da sich die Rückmeldungen auf Klassenebene befinden, eignen sie sich – anders als intuitiv bei einem Leistungstest für *Schülerinnen und Schüler* zu vermuten – nicht zur Beurteilung dieser, sondern v. a. als Grundlage für die Reflexion und Evaluation von Unterrichtsprozessen zum einen von der einzelnen Lehrkraft und zum anderen in der Fachgruppe, womit letztlich die *professionelle Kompetenz* der Lehrkräfte im Fokus steht (Helmke et al., 2003; Leuders, 2006; Maier et al., 2012; Petersen, 2013). Zielgruppe von Vergleichsarbeiten bzw. deren Rückmeldungen sind aus diesem Blickwinkel also sowohl die jeweiligen Fachlehrkräfte als auch Fachkollegien (Bonsen et al., 2006; Dobbstein & Peek, 2008; Peek & Dobbstein, 2006a).

Vergleichsmöglichkeiten/Rückmeldungen auf Klassen-/Schulebene³¹

Für jede Schülerin und jeden Schüler bietet die *tabellarische Übersicht* Aufschluss, ob einzelne Items sprich Teilaufgaben korrekt (Score 1) oder falsch (Score 0) bearbeitet wurden. Zusätzlich gibt es neben einem Testgesamtscore Lösungsanteile für die einzelnen Dimensionen der Bildungsstandards: Aufgeschlüsselt nach Leitideen, Kompetenzen und Anforderungsbereichen sowie Kompetenzstufen ist für jede Schülerin und jeden Schüler angegeben, wie hoch der Anteil der richtig gelösten Items der in diese Dimension respektive in diese Kompetenzstufe fallenden Items ist. Der *Sofortbericht* bietet auf Klassenebene eine Zusammenfassung dieser Informationen der Art, dass der Klassenmittelwert bzw. Mittelwert der Lerngruppe im Gesamten, in den einzelnen Dimensionen der Bildungsstandards sowie den Kompetenzstufen in Form von Balkendiagrammen visualisiert wird (Abbildung 15). Zusätzlich wird die Information

³⁰ Dabei werden die individuellen Rückmeldungen an die Schülerinnen und Schüler nicht näher berücksichtigt. Zum Zeitpunkt der Durchführung des Projekts VELM-8 und auch der als Datenmaterial genutzten Online-Fragebögen erfolgten keine individualisierten Rückmeldungen.

³¹ In diesen Ausführungen wird sich weitestgehend an der Umsetzung in Hessen orientiert. Dabei bieten andere Bundesländer (sehr) ähnliche Vergleichsmöglichkeiten an.

angeboten, zu welchem Anteil die Lerngruppe die einzelnen Items korrekt gelöst hat (Abbildung 16).

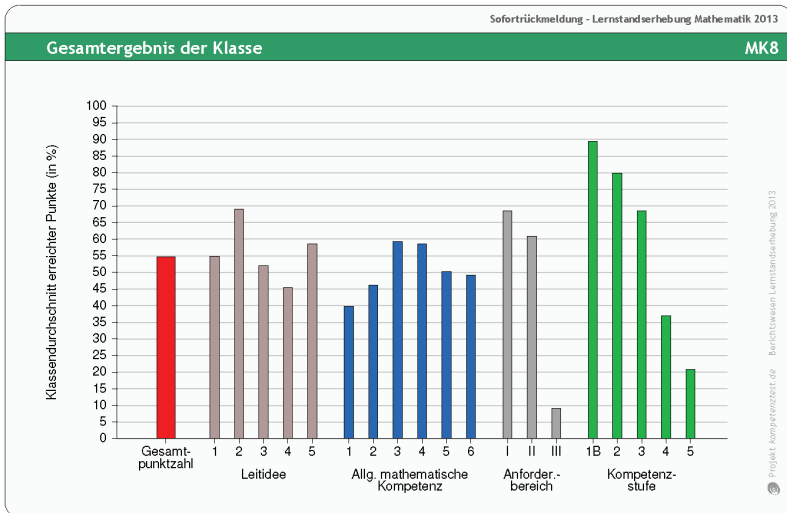


Abbildung 15: Gesamtergebnis der Klasse (Sofortbericht)

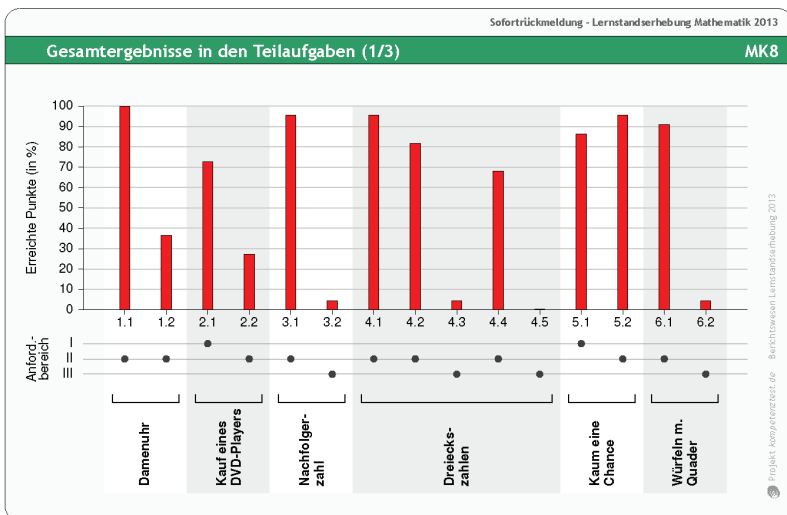


Abbildung 16: Gesamtergebnis in den Teilaufgaben (Sofortbericht)

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

Mit der Orientierung an den Dimensionen der Bildungsstandards und insbesondere durch den Bezug zum Kompetenzstufenmodell wird ein curriculumorientiertes Kriterium formuliert, welches einen „erweiterten Blick“ auf die Leistungen der Lerngruppe sowie der einzelnen Schülerinnen und Schüler erlaubt, die über die alltäglichen Unterrichtsszenarien und auch die Testung durch Klassenarbeiten hinausgeht (Dobbelstein & Peek, 2008; Drüke-Noe, 2014; EMSE, 2006; Helmke et al., 2003; Peek & Dobbelstein, 2006b).

Einige Wochen nach der Verfügbarkeit der tabellarischen Übersicht und des Sofortberichts werden den teilnehmenden Lehrkräften mit dem *Ergebnisbericht* und dem *Schulbericht* weitere Perspektiven auf die Ergebnisse zur Verfügung gestellt. Der Ergebnisbericht erweitert die Informationen des *Sofortberichts* um eine *soziale* Bezugsnorm, indem „adjustierte Vergleichswerte“ präsentiert werden (vgl. Abbildung 17; EMSE, 2006; Nachtigall & Kröhne, 2006; Peek & Dobbelstein, 2006b). Unter „adjustierten Vergleichswerten“ ist für Hessen der sog. *korrigierte Landesmittelwert* zu verstehen: Dabei werden aus der Klassenleistung diejenigen Faktoren regressionsanalytisch „herausgerechnet“, auf die die Schule, die Lehrkräfte sowie der Unterricht keinen Einfluss haben, welche jedoch nach aktuellem Forschungsstand Einfluss auf die Leistung nehmen (etwa Geschlechterverhältnis, Sprache zu Hause oder sozioökonomischer Status; Hessisches Kultusministerium, 2009; LSA, 2015). Alternativ schlagen Nachtigall und Kröhne (2006) die Rückmeldung von unkorrigierten Populationsmittelwerten oder von realen „Vergleichsklassen“ vor. Ersteres ist jedoch vor dem Hintergrund der nicht repräsentativen Stichprobenuntersuchung nahezu unmöglich wegen der Anfälligkeit der Stichprobe gegenüber Verzerrungen. Für die Nutzung von „realen“ Vergleichsklassen müsste die Anonymität der Untersuchung zumindest stückweise aufgehoben werden, worin ein Risiko des Akzeptanzverlustes und erhöhter Kontrollwahrnehmung bestünde. Neben der Ergänzung einer sozialen Bezugsnorm zu den Ergebnissen trägt die Vermittlung von solchen adjustierten Vergleichen positiv zur Akzeptanz der Vergleichsarbeiten durch die Lehrkräfte bei, da keine „Äpfel mit Birnen“ verglichen werden und Schulen die Leistungen nicht aufgrund von Voraussetzungen der Schülerschaft über- oder unterschätzen (Stryck, 2005, S. 5; Bonsen et al., 2006; Dobbelstein & Peek, 2008; Klieme et al., 2003, 2009; Peek & Dobbelstein, 2006a, 2006b).

Im Schulbericht werden die Leistungen der Schülerinnen und Schüler im zeitlichen Verlauf angegeben. Dabei ist aus psychometrisch-forschungsmethodischer Sicht kritisch anzumerken, dass es sich hierbei nicht um eine Panel-Untersuchung handelt und daher in den einzelnen Jahren verschiedene Schülerinnen und Schüler getestet wurden. Demnach lassen sich Veränderungen auf keinen Fall kausal interpretieren. Neben dem Gesamtergebnis aller getesteten Fächer werden die Ergebnisse für Mathematik nach den Leitideen aufgeschlüsselt (Abbildung 18). Diese Rückmeldungen geben den Lehrkräften und der Schule als Ganzes die Möglichkeit, auch auf der Ebene von Jahrgängen, Stärken und Schwächen (im sehr groben Raster der Leitideen) zu erkennen (EMSE, 2006) und beschreibend die Veränderungen über die Zeit zu dokumentieren (um etwa Veränderungen im Schulcurriculum zu evaluieren). Dabei sind aufgrund der höheren Aggregationsebene der Daten die Ergebnisse des Schulberichtes (der sich auf die Ergebnisse aller teilnehmenden Klassen stützt) verlässlicher als Ergebnisse von einzelnen Schülerinnen oder Schülern (Brügelmann, 2005b; A. Hosenfeld, 2010; Isaac et al., 2006; Korngiebel, 2009).

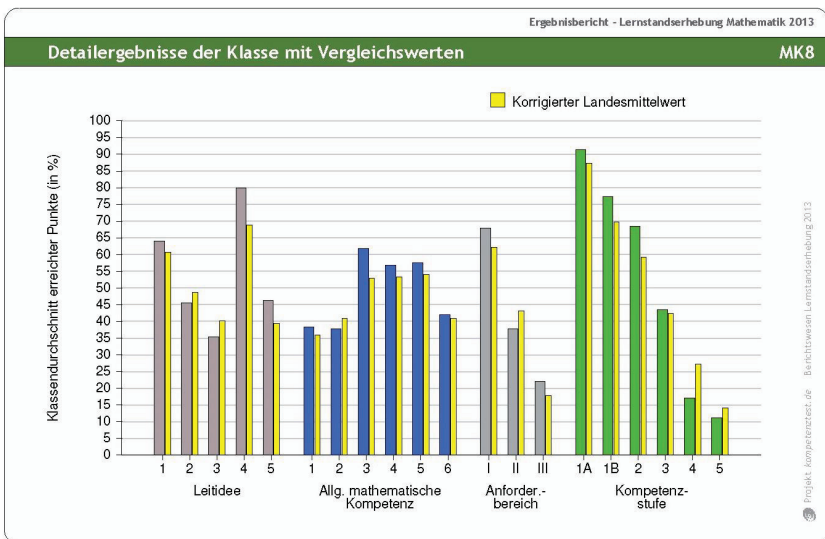


Abbildung 17: Detailergebnisse der Klasse mit Vergleichswerten

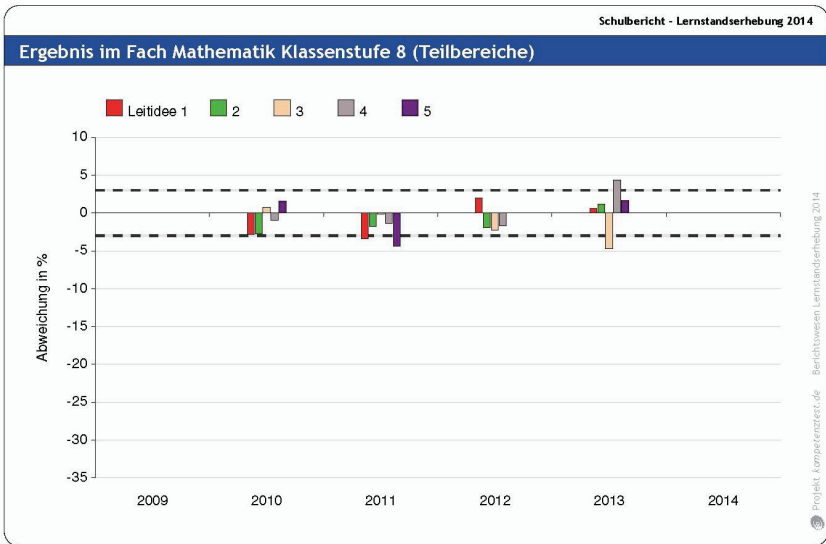


Abbildung 18: Ergebnis im Fach Mathematik Klassenstufe 8 (nach Leitideen getrennt)

Vergleichsmöglichkeiten auf Individualebene

Die bisher genannten Rückmeldungen bewegen sich auf der Ebene von Lerngruppen oder sind sogar lerngruppenübergreifend. Zu einem erheblichen Teil müssen Bildungsprozesse jedoch für jede Schülerin und jeden Schüler differenziert betrachtet werden. Entsprechend dieses Ansatzes der Aggregation von Daten ist die Verlässlichkeit von Individualdiagnosen basierend auf einem einmaligen Leistungstest als eher gering einzuschätzen, worin begründet liegt, dass die Lernstandserhebungen nicht zu Selektionszwecken genutzt werden sollen (Hessisches Kultusministerium, 2011; vgl. auch Tabelle 4). Einige Autoren gehen sogar weiter und sprechen davon, dass das Instrument individualdiagnostische Schlüsse nicht zulasse oder damit überfordert sei (Brügelmann, 2005b; Korngiebel, 2009; Leuders, 2006; Leutner et al., 2008; Selter, 2005b). Die Gründe für diese Beschränkungen liegen zum einen in der Kürze der zur Verfügung stehenden Testzeit bei gleichzeitigem Anspruch, den kumulativen Kompetenzaufbau seit Beginn des Bildungsabschnittes zu messen (also im Falle der Lernstandserhebungen in Klasse 8 seit Beginn der fünften Jahrgangsstufe), zum anderen in der fehlenden Differenzierung in der Auswertung der Tests: Es wird lediglich zwischen korrekt oder inkorrekt unterschieden und die Ursache für ein „falsch“ können dabei von einem simplen Zahlendreher bis zu tiefgreifenden Fehlkonzepten reichen.

Allerdings enthalten viele Teilaufgaben von VerA beispielsweise den Zusatz „Begründe deine Aussage.“ Dieser und ähnliche Zusätze werden mit Blick auf die Auswertungsobjektivität bei der Kodierung vielfach vernachlässigt, liefern jedoch den auswertenden Lehrkräften wichtige Diagnosehinweise (Büchter & Leuders, 2005a, S. 329). Da die Lernstandserhebungen von den hessischen Lehrkräften vor Ort basierend auf einem deutschlandweit einheitlichen Kodier-/Bewertungsmanual selbstständig ausgewertet werden, haben die Lehrkräfte jede einzelne Bearbeitung ihrer Schülerinnen und Schüler vor Augen. Durch diese Diagnosemöglichkeit liefern die Lernstandserhebungen also ein Instrument, welches das Verständnis und Denken von Schülerinnen und Schülern differenziert sichtbar macht und zwar „to both teachers and themselves“ (Pellegrino et al., 2001, S. 4; Leuders, 2006). Deutlicher formuliert es Isaac et al. (2006, S. 109): „Lehrkräfte, die VERA nicht sinnvoll nutzen, nehmen sich die Chance, wichtige zusätzliche Informationen, z. B. zu Fähigkeiten ihrer Schüler zu erhalten.“

Ergänzend zur Betrachtung der individuellen Bearbeitung einzelner Teilaufgaben (im Sinne einer Kompetenzdiagnostik aus Tabelle 4) steht den Lehrkräften auch die tabellarische Übersicht zur Verfügung. Gemeinsam mit weiteren diagnostischen Informationen, die Lehrkräfte aus dem alltäglichen Unterricht oder aus anderen Leistungsmessungen (etwa Klassenarbeiten) über einzelne Schülerinnen und Schüler haben, können diese Erkenntnisse aus der Bearbeitung der Lernstandserhebung ein konkreteres diagnostisches Urteil formen: Vergleichsarbeiten haben also insbesondere keinen Anspruch, alleiniges Diagnoseinstrument zu sein, sondern müssen als Ergänzung zu anderen Verfahren verstanden werden, sodass als Quintessenz aller individualdiagnostisch relevanten Informationen konkrete Fördermaßnahmen abgeleitet werden können (Bremerich-Vos et al., 2005; Helmke & I. Hosenfeld, 2003; Kuper & Hartung, 2007; Selter, 2005b; Siebenborn, 2005). Damit sind VerA eher als integrativer Teil der Arbeit von Lehrkräften zu verstehen denn als zusätzliches externes Mittel der dezentralen Leistungsmessung (Posch, 2009, S. 125).

1.6.1.3.2 Konsequenzen der Standortbestimmung

Aufbauend auf dieser *Lernstandserhebung* zeichnen sich drei wesentliche Bereiche ab, in denen die Informationen genutzt werden und „innovative Impulse [...] setzen“ sollen (Büchter & Leuders, 2005b, S. 15):

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

1. Schulentwicklung
2. Unterrichtsentwicklung
3. Individuelle Förderung

Analog zu den Ausführungen zur SE ist die Erweiterung der *professionellen Kompetenz* der Lehrkräfte ein Teil der Schulentwicklung (Koch et al., 2006; Posch, 2009), der zwar für die UE und die individuelle Förderung wesentlich ist, als Ziel der Lernstandserhebungen an dieser Stelle jedoch von nachrangiger Bedeutung ist. Mit Hilfe der Rückmeldungen aus den Vergleichsarbeiten können Lehrkräfte, wie dargelegt, ihre eigenen diagnostischen Urteile erweitern, kontrollieren und reflektieren (pädagogisch-psychologisches Wissen von Lehrkräften im Kompetenzmodell von COACTIV; Helmke et al., 2003; Kuper & Diemer, 2012; Switalla, 2005). Des Weiteren stellen die Aufgaben Illustrationen der Bildungsstandards zur Weiterentwicklung des fachdidaktischen Wissens dar (vgl. Kompetenzmodell von COACTIV). Als weitere Lerngelegenheit informeller Art lassen sich die *didaktischen Kommentare* verstehen, die neben Aufgabenkommentaren auch Unterrichtsvorschläge und Fehleranalysen bereithalten (fachdidaktisches Wissen).

Schulentwicklung

Wie die einzelnen Entwicklungsperspektiven inhaltlich gefüllt werden, hängt in starkem Maße von den jeweiligen Ausgangssituationen ab und lässt sich hier nicht konkret darstellen. Vielmehr sollen ausgewählte potentielle Entwicklungslinien genannt werden: Im Rahmen einer Schulentwicklung geben VerA mit dem Adressatenkreis „Lehrkräfte und Fachkollegien“ bereits eine Entwicklungslinie zu mehr Kooperation und Kommunikation im (Fach-/Jahrgangs-)Kollegium vor (vgl. OE in Abschnitt 1.5.3). Darüber hinaus können die Vergleichsarbeiten (entweder durch die Ergebnisse oder auch durch Aufgaben bzw. *didaktische Kommentare*) auf kollegialer Ebene didaktische oder methodische Diskussionen mit Blick auf die Gestaltung von Unterrichtskonzepten und -reihen anregen (vgl. UE in Abschnitt 1.5.4; hierzu werden von diversen Autoren Anregungen gemacht und Vorschläge unterbreitet: Bonsen et al., 2006; Helmke & I. Hosenfeld, 2003; Leutner et al., 2008; Maier et al., 2012; Maritzen, 2014; Nachtigall & Jantowski, 2007; Peek & Dobbstein, 2006a; Peek et al., 2006; A. Voss & Blatt, 2010). Insbesondere sollte auf dieser Ebene auf etwaige systematische Stärken und Schwächen lerngruppenübergreifend oder in der zeitlichen Entwicklung geachtet werden, um ggf. schul-/ fachprogrammatrisch eingreifen zu können.

Unterrichtsentwicklung

Ähnlich wie mögliche didaktische Diskussionen auf kollegialer Ebene ergibt sich die Möglichkeit für die einzelne Lehrkraft, die didaktischen Anregungen zur Reflexion des eigenen Unterrichts zu nutzen. Die die Bildungsstandards illustrierenden Aufgaben können beispielsweise den Ansatzpunkt für eine Veränderung der bestehenden Aufgabekultur im Klassenraum bilden (Helmke & I. Hosenfeld, 2003; Koch et al., 2006). Sollten sich in den beschriebenen Rückmeldungen Besonderheiten zeigen (etwa absolut oder relativ zum korr. Landesmittelwert besonders hohe oder niedrige Lösungshäufigkeiten einzelner Aufgaben oder Leitideen), sollte dies Ausgangspunkt korrespondierender Maßnahmen im Unterricht sein (wiederum werden von verschiedenen Autoren Anregungen und Hinweise zur unterrichtsentwickelnden Nutzung gegeben: Bensen et al., 2006; Büchter & Leuders, 2005a; Burkard & G. Orth, 2005; EMSE, 2006; Helmke et al., 2003; Koch et al., 2006; Koch, 2011; Korngiebel, 2009; Leutner et al., 2008; Maier et al., 2012; Maritzen, 2014; Peek & Dobbelsstein, 2006a; Peek et al., 2006; Posch, 2009; Richter et al., 2014; Stryck, 2005; A. Voss & Blatt, 2010).

Individuelle Förderung

Ausgehend von den beschriebenen diagnostischen Schlüssen auf individueller Ebene, ließen sich entsprechende Fördermaßnahmen anschließen (Maier et al., 2012; Peek et al., 2006). Sind beispielsweise Fehlvorstellungen zu erkennen, so gilt es im Anschluss an die VerA diese gezielt zu beheben. Ansatzpunkte können hierbei die *didaktischen Kommentare* liefern.

1.6.2 Ablauf VerA

Bei internationalen und nationalen Vergleichsuntersuchungen wird immer wieder bemängelt, dass diese für eine direkte schulische Nutzung keine Möglichkeit böten, da u. a. die Rückmeldungen erst mit deutlichem zeitlichen Abstand zur Erhebung und in sehr abstrakter Form bereitgestellt werden (I. Hosenfeld et al., 2006). Die Vergleichsarbeiten hingegen sind in ihrer Konzeption und den unmittelbar verfügbaren Rückmeldungen direkt auf eine unterrichtsentwickelnde und individuell-fördernde Nutzung ausgerichtet. Besonders und speziell für die Vergleichsarbeiten sind dabei die jährlich eigens für die im Testheft enthaltenen Aufgaben entwickelten *didaktischen Kommentare* (vgl. unten). Im Folgenden soll der exemplarische Durchlauf durch

das „VerA-Jahr“ den Ablauf zeigen (in Anlehnung an Helmke & I. Hosenfeld, 2003; I. Hosenfeld et al., 2008)³².

1. Aufgabenentwicklung
2. Pilotierung der Aufgaben & Entwicklung von Testheften
3. Vorbereitung der Durchführung der VerA an den Schulen
4. Durchführung der VerA und Auswertung durch die Lehrkräfte
5. Nutzung der Rückmeldungen aus den VerA zur UE und individuellen Förderung

1.6.2.1.1 Testentwicklung

Die Aufgabenentwicklung wird vom IQB zentral koordiniert und unter fachdidaktischer Unterstützung und Beratung durchgeführt. Die Aufgaben werden dabei von erfahrenen, aktiven Lehrkräften aus ganz Deutschland entwickelt und bedienen somit den Anspruch von Siebenborn (2005, S. 46), die fordert: „Wer Testaufgaben entwickelt, braucht eine Vorstellung davon, wie die fachspezifischen Anforderungen angemessen zu repräsentieren sind. Es geht vor allem auch darum, schwierigkeitsbestimmende Merkmale möglichst treffsicher einzuschätzen.“ In einer ersten unabhängigen fachdidaktischen Bewertung der Items wird darauf geachtet, dass der Bezug der Items zu den Bildungsstandards bzw. Curricula vorhanden ist und die Items authentisch, differenzierend und kognitiv aktivierend sind (Spoden et al., 2014, S. 82). Vor dem Hintergrund der individual-diagnostischen Schlüsse werden Items hier u. a. mit Blick auf eine für Diagnose ausreichende Produktion der Schülerinnen und Schüler betrachtet (Leuders, 2006).

Bei der Aufgabenentwicklung ist grundsätzlich das Dilemma vorhanden, fachlich möglichst anspruchsvolle und vielschichtige Aufgaben konstruieren zu wollen, die bei differenzierter Bewertung eine Vielzahl an diagnostischen Informationen liefern. Andererseits sollen sie jedoch bei der späteren Auswertung den Ansprüchen der Psychometrie genügen, nach denen eine Kodierung ohne Schulung und umfangreiche Kodiermanuale dennoch ausreichend objektiv in richtig oder falsch erfolgen können soll (Spoden et al., 2014). Insofern werden an Aufgaben für die Vergleichsarbeiten Ansprüche von verschiedenen Seiten gestellt: Zum einen von Seiten der

³² Bei den Ausführungen stütze ich mich neben den angegebenen Quellen auf meine eigenen Erfahrungen im Rahmen meiner Mitarbeiterstelle bei Prof. Dr. Werner Blum an der Universität Kassel zwischen 2011 und 2014. In dieser Zeit waren Teile der Arbeitsgruppe von Prof. Dr. Werner Blum mit der Aufgabenentwicklung und -überarbeitung im Auftrag des IQB betraut.

Fachdidaktik (Aufgaben sollen Mathematik als Ganzes abbilden; vgl. Grunderfahrungen nach H. Winter, 1995), zum zweiten von der Psychometrie (deren Ansprüche v. a. in der Pilotierung geprüft werden) und letztlich sollen die Aufgaben anschlussfähig an den Unterricht sein, um aus der Erfahrungswelt der Lehrkräfte zu kommen (Büchter & Leuders, 2005a). Dieser Prozess der Aufgabenerstellung, Begutachtung, Präpilotierung, Überarbeitung und Finalisierung dauert etwa ein Viertel Jahr für jede Aufgabe. Parallel zur Entwicklung und Überarbeitung werden Kodieranweisungen erstellt, die eine objektive Bewertung jeder Teilaufgabe mit *richtig* (Code 1) oder *falsch* (Code 0) anhand von Kriterien und abgrenzenden Beispielen ermöglichen sollen.

Nachdem alle Teilaufgaben eines Jahrgangs fertiggestellt wurden und samt Kodieranweisungen dem IQB vorliegen, werden hieraus leitideenausgewogene Testhefte erstellt. Bei der Testhefterstellung wird darauf geachtet, dass nach Möglichkeit auch die prozessbezogenen Kompetenzen ausgeglichen gefordert sind. Ziel der folgenden Pilotierung ist es, die Teilaufgaben an einer ausreichend großen Stichprobe auf ihre Testgüte und tatsächliche empirische Schwierigkeit zu prüfen: Letztlich sollen die Testitems dem Anspruch der probabilistischen Testtheorie genügen, um eine entsprechende Skalierung und Anknüpfung an die Kompetenzstufen zu ermöglichen. Die Kodierung der Items durch geschultes Personal anhand der angefertigten Kodiervorgaben ermöglicht parallel eine finale Anpassung/Überarbeitung der Kodiervorgaben. Nach Abschluss der Kodierung und statistischer Auswertung erfolgt die Zusammenstellung der finalen Testhefte in Rücksprache mit den Kultusministerien der Bundesländer bzw. Landesinstituten. Eine weitere wichtige Funktion der Pilotierung ist die Generierung von exemplarischen Bearbeitungen von Schülerinnen und Schülern. Ziel ist das Auffinden *typischer Schülerfehler*, um diese Informationen bei der Konzeption der *didaktischen Kommentare* zu nutzen (Helmke & I. Hosenfeld, 2003).

1.6.2.1.2 Didaktische Kommentare

Die *didaktischen Kommentare* stellen den Versuch eines Brückenschlags zwischen VerA-Entwickelnden und Schule dar und sind als Unterstützungs- und Informationsangebot an die teilnehmenden Lehrkräfte zu verstehen (I. Hosenfeld et al., 2006; Koch et al., 2006; Peek & Dobbelstein, 2006a, 2006b; Selter, 2005a). Die *didaktischen Kommentare* umfassen standardmäßig für jede Teilaufgabe folgende Informationen:

- Itemlösung
- Itemklassifikation (Leitidee, Kompetenz, Anforderungsbereich, Kompetenzstufe)
- Beschreibung des Lösungsprozesses
- Beschreibung potenzieller, häufiger Fehler von Schülerinnen und Schülern
- Anregungen zur unterrichtlichen Weiterarbeit

Während die Itemlösung und die Itemklassifikation bereits aus der Itementwicklung feststehen bzw. im Rahmen der psychometrischen Prüfung der Aufgabengüte (Kompetenzstufe) ermittelt werden, werden die detaillierte Beschreibung des Lösungsprozesses, die Beschreibung und Erläuterung typischer Fehlermuster und die Anregungen zur unterrichtlichen Weiterarbeit von fachdidaktischem Personal mit eigener Unterrichtserfahrung/-tätigkeit im Anschluss an die Pilotierung verfasst. Für diese Arbeit werden die in der Pilotierung gewonnenen Bearbeitungen der Schülerinnen und Schüler mit einbezogen und bei Bedarf als Illustration in die *didaktischen Kommentare* genutzt.

In der Gestaltung und mit diesen Ausführungen zu jeder einzelnen Teilaufgabe stellen die *didaktischen Kommentare* ein informelles Fortbildungsangebot für praktizierende Lehrkräfte dar: Es wird auf Basis dieses Materials die Möglichkeit geboten, beispielsweise Fehlvorstellungen bewusst vor Augen zu führen, Schwierigkeit generierende Merkmale von Aufgaben zu reflektieren oder auch die unterrichtliche Begegnung von Fehlvorstellungen zu betrachten (I. Hosenfeld et al., 2006). Darüber hinaus werden die diagnostischen Kompetenzen von Lehrkräften u. a. im Wissen über Fehler von Schülerinnen und Schülern gefordert und gefördert. Mit den Anregungen für eine unterrichtliche Weiterarbeit mit den Teilaufgaben der Vergleichsarbeiten stehen den Lehrkräften mit den *didaktischen Kommentaren* bildungsstandardkonforme Unterrichts Anregungen zur Verfügung, die eine Entlastung in der Unterrichtsvorbereitung darstellen können. Distales Ziel der *didaktischen Kommentare* ist damit eine Steigerung der direkten Nutzungsmöglichkeiten der VerA für die Lehrkraft, um so eine gesteigerte Akzeptanz auf Seiten der Lehrkräfte zu erreichen (Korngiebel, 2009).

1.6.2.1.3 Testdurchführung und Auswertung

Die Vorbereitung von Seiten des IQB und der unterstützenden Fachdidaktik ist mit Fertigstellung der finalen Testhefte sowie der zugehörigen *didaktischen Kommentare* abgeschlossen. Die Landesinstitute sind im Folgenden dafür zuständig, dass die

teilnehmenden Schulen bzw. Klassen ausreichend Material zugesandt bekommen und die Lehrkräfte hierüber entsprechend informiert sind. Es wird für Schulen dabei empfohlen, eine Lernstandskoordinatorin/einen Lernstandskoordinator zu bestimmen, die/der als Ansprechpartner/in die Kommunikation mit dem jeweiligen Landesinstitut übernimmt (Bonsen et al., 2006; Burmeister et al., 2005).

Für die einzelne Lehrkraft ist eine bewusste Vorbereitung weder selbst noch für die Schülerinnen und Schüler weder notwendig (Hessisches Kultusministerium, 2009) noch sinnvoll, wenn nicht sogar kontraproduktiv, vor dem Hintergrund des Ziels einer Standortbestimmung für die Lehrkraft selbst, um Entwicklungsmöglichkeiten aufzudecken. Die Lehrkraft sollte selbstverständlich für einen störungsfreien Ablauf und ausreichend Information sorgen, sodass auch den Schülerinnen und Schülern der Zweck der Lernstandserhebungen – Unterrichtsentwicklung und Förderung; insbesondere jedoch keine individuelle Bewertung oder Selektion – bewusst ist. Ebenfalls kann es sinnvoll sein, Schülerinnen und Schüler anhand von (Teil-)Aufgaben vergangener Jahre mit möglichen Aufgabenformaten vor Testdurchführung vertraut zu machen, um das Aufgabenformat als Fehlerquelle auszuschließen (I. Hosenfeld et al., 2006).

Für die eigentliche Durchführung wird den teilnehmenden Lehrkräften Testmaterial zur Verfügung gestellt, welches die Schülerinnen und Schüler in der zur Verfügung stehenden Zeit von 80 Minuten bearbeiten. Im Anschluss an die Datenerhebung erfolgt die Kodierung bzw. Bewertung der Testhefte durch die betreffende Lehrkraft selbstständig (um bereits bei diesem Schritt einen ersten diagnostischen Blick auf die individuelle Bearbeitung werfen zu können). Die *richtig-/falsch*-Bewertungen jeder individuellen Bearbeitung einer Teilaufgabe werden im Anschluss in ein Online-Portal eingegeben, in dem die Lehrkräfte die bereits beschriebenen Rückmeldungen finden (Leutner et al., 2008, 154ff.).

An dieser Stelle befinden sich die Lernstandserhebungen an einer sehr wichtigen Gelenkstelle: Die bisher zentral entwickelten, zentral pilotierten und zentral finalisierten Instrumente gelangen in die Breite und werden dezentral durchgeführt, bewertet und die Ergebnisse zur Weiterverarbeitung online eingegeben. Entsprechend stellen sich hier besondere Anforderungen an die Objektivität dieses Instrumentes: Hier zeigen sich einige Verfälschungstendenzen durch die Lehrkräfte. Spoden et al. (2014) verweist auf einen Tagungsbeitrag von Hosenfeld (2007) in der Unterscheidung von

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

kognitiven (eher unsystematisch) und *motivationalen* (systematisch) Fehlern der Lehrkräfte. Zweifelsfrei bieten sich gewisse Spielräume, innerhalb derer die Durchführung und die Beurteilung der Itemlösung vorgenommen werden (Brügelmann, 2005b; Strähling, 2005), jedoch gilt es, *motivationale*, also systematische Beurteilungsfehler der Lehrkräfte zu verhindern (vgl. Kapitel 1.6.4).

Nach Abschluss der Eingabe der *richtig-/falsch*-Beurteilung durch die Lehrkraft gibt es unmittelbar die Möglichkeit auf den Sofortbericht und die tabellarische Übersicht zuzugreifen. Bereits oben wurde darauf hingewiesen, welche Informationen in diesen Materialien enthalten sind. Vielfach bieten die Landesinstitute bei der Nutzung der Rückmeldeformate exemplarische Bearbeitungsvorlagen oder Beispielfragen zur Auswertung an (Hessisches Kultusministerium, 2009; LSA, 2015). Von dieser Fülle an Informationen ist jedoch zweifelsfrei produktiv und ohne eine Überforderung nur ein von den Lehrkräften selbst ausgewählter Bereich detaillierter zu betrachten (Kuper & Diemer, 2012). Damit beginnt für die Lehrkräfte die eigentliche Arbeit erst nach der Lernstandserhebung in Form der resultierenden Unterrichts- und Schulentwicklung bzw. der individuellen Förderung (Baumert, 2001b). Die Testhefte sollten nach einiger Zeit an die Schülerinnen und Schüler wieder ausgeteilt werden, sodass auch diese die Möglichkeit der Reflexion der eigenen Leistungen erhalten.

1.6.3 Exkurs: Abgrenzung VerA

Bereits in den bisherigen Ausführungen wurden an einzelnen Stellen implizit Unterschiede zwischen den Lernstandserhebungen bzw. Vergleichsarbeiten und anderen Verfahren der Leistungsmessung deutlich. Hier sollen kurz die wesentlichen Unterschiede im Vergleich zu

1. Klassenarbeiten sowie Parallelarbeiten,
2. Zentralen Abschlussprüfungen und
3. (Inter-)Nationalen Vergleichsuntersuchungen (large-scale)

verdeutlicht und zusammengefasst werden (Tabelle 5).

Klassenarbeiten und *Parallelarbeiten* sind vorrangig innerhalb der Schule von Lehrkräften entwickelte Verfahren zur möglichst gerechten Notenvergabe, wobei sozial als Vergleichsperspektive bei Klassenarbeiten lediglich die Leistungen anderer Schülerinnen und Schüler der eigenen Klasse bzw. bei Parallelarbeiten anderer Klassen sowie kriterial die Erwartungen der Lehrkräfte zur Verfügung stehen

(„Erwartungshorizont“; vgl. Kompetenzdiagnostik in Tabelle 4). Die Durchführung der Arbeiten erfolgt wie bei den Lernstandserhebungen dezentral und nicht durch geschultes Personal. Der Fokus der Untersuchungen liegt inhaltlich begrenzt auf der Bewertung einer zurückliegenden Unterrichtseinheit und nimmt weniger den kumulativen Kompetenzaufbau mit dem Abschluss als Fernziel in den Blick (Bonsen et al., 2006; Lorenz, 2005; Stryck, 2005; van Ackeren & Klemm, 2011). Ebenfalls wird Klassenarbeiten die Möglichkeit zugesprochen, dass Lehrkräfte diese Leistungsmessungen als Rückmeldungen zum eigenen Unterricht verstehen und entsprechende Schlüsse für die Unterrichtsgestaltung ziehen (Büchter & Leuders, 2008; Schwippert, 2005).

Eine solche auf zukünftige Unterrichtsgestaltung in der jeweiligen Lerngruppe gerichtete Perspektive bieten *zentrale Abschlussprüfungen* verständlicherweise weniger (Baumert, 2001b). Nichtsdestotrotz können Lehrkräfte auch diese Informationen zur Reflexion über den Unterrichtsprozess nutzen und die gezogenen Schlüsse für eine erneute Unterrichtsreihe anwenden. Vorrangiges Ziel zentraler Abschlussprüfungen ist jedoch die Vergabe von Abschlüssen zum Ende eines Bildungsabschnittes und damit die Zertifizierung und Selektion ähnlich den *Klassenarbeiten* und *Parallelarbeiten* (van Ackeren & Klemm, 2011). Unterschiede zu den vorherigen Verfahren zeigen sich jedoch in der zentralen, zumindest zu einem gewissen Grad an Gütekriterien wissenschaftlicher Messungen orientierten Testentwicklung.

Alle drei beschriebenen Verfahren (*Klassen-* und *Parallelarbeiten* sowie *zentrale Abschlussprüfungen*) dienen der Selektion und Zertifizierung und grenzen sich damit eindeutig von den Vergleichsarbeiten und (*inter-*)*nationalen Vergleichsuntersuchungen* ab (Büchter & Leuders, 2005a; Hessisches Kultusministerium, 2009; LSA, 2015; A. Voss & Blatt, 2010). Bei (*inter-*)*nationalen Studien* begründet sich dies vor allem dadurch, dass es sich um stichprobenbasierte Erhebungen handelt (vgl. Bildungsmonitoring in Tabelle 4; KMK, 2015). So konzipiert können diese Verfahren keine direkt nutzbaren Informationen auf Einzelklassen- oder Lehrkraftebene liefern, sondern sind auf der Systemebene verortet (van Ackeren & Klemm, 2011). In punkto testtheoretischer Grundlagen und empirischen Standards steigen die Ansprüche bei den genannten Verfahren von *Klassen-* und *Parallelarbeiten* über *zentrale Abschlussprüfungen* zu *Lernstandserhebungen/Vergleichsarbeiten* und schließlich zu (*inter-*)*nationalen Vergleichsuntersuchungen* an (Peek & Döbelstein, 2006a).

Tabelle 5: Vergleich vers. Leistungsmessungen in Abgrenzung der Lernstandserhebungen (zusammenfassend nach Bensen et al., 2006; Helmke & I. Hosenfeld, 2003; Hessisches Kultusministerium, 2009; I. Hosenfeld et al., 2006; KMK & IQB, 2006; KMK, 2015; Koch, 2011; LSA, 2015; Peek et al., 2006; Schwippert, 2005; A. Voss & Blatt, 2010)

	LSE	Klassenarbeit/Parallelarbeit	Abschluss	(inter-)nationale Vergleichsun- tersuchung
Stichprobe	Vollerhebung	Vollerhebung	Vollerhebung	(Zufalls-)Stichprobe
Bewertung/Selektion	nein	ja	ja	nein
Externe Testkonzeption	ja	nein	ja	ja
Externe Testdurchführung	nein	nein	nein	ja
Inhaltlicher Fokus	grundlegende Kompetenz kumulativer Kompetenzaufbau (orientiert an Bildungsstan- dards; deutschlandweit)	(kurzlich) vergangene Unter- richtsinhalte	breit angelegt über unterrichtli- che Inhalte (orientiert am Kerncurriculum; bundeslandab- hängig)	grundlegende Kompetenz
Fokus der Untersuchung	Klassen und Schülerinnen und Schüler	Schülerinnen und Schüler (Unterricht)	Schülerinnen und Schüler (Unterricht)	Schulsysteme Population von Schülerinnen und Schülern
Zielgruppe der Untersuchung	Lehrkräfte und Fachkollegien	Lehrkräfte	Schülerinnen und Schüler	Bildungspolitik
Vergleichsmöglichkeit	korrigiertes Landesmittel absolute Häufigkeiten Erwartungen Parallelklassen klassenintern	absolute Häufigkeiten Erwartungen klassenintern	absolute Häufigkeiten Erwartungen Parallelklasse klassenintern	nationale Mittelwerte
Ranking der Ergebnisse	nein	nein	nein	ja
Datenhoheit	Schule	Lehrkraft	Fachkollegium	öffentlich
Ziel	Schul-/Unterrichtsentwicklung individuelle Förderung	gerechte Leistungsbewertung Diskussion innerhalb der Fach- gruppe	Notenbildung und Abschluss- vergabe	Systemmonitoring Vergleich von Schulsystemen

1.6.4 Bedingungen für die Nutzung von Vergleichsarbeiten

Die Vergleichsarbeiten bieten Lehrkräften diverse Rückmeldungen und zusätzliche Vergleichsinformationen sowie Unterrichts Anregungen. Damit sichergestellt werden kann, dass die Lernstandserhebungen auch in der intendierten Art genutzt werden und dabei den mit ihnen verbundenen Zielen und an sie gestellten Ansprüchen gerecht werden, gilt es, einige Voraussetzungen und Bedingungen zu beachten. Verschiedene Autoren beschäftigen sich teils vom theoretischen, teils vom empirischen Standpunkt aus mit dieser Frage und beleuchten auch benachbarte Fragen zu Bedingungen und Voraussetzungen zur Nutzung anderer Leistungsmessungen.

1.6.4.1 Relevante Tests/Close-to-the-job

Ziel der Vergleichsarbeiten ist es u. a., dass sich Entwicklungen auf verschiedenen Ebenen ableiten lassen und die Arbeiten nach ihrer Durchführung nicht „in der Schublade“ verschwinden. Damit ist als erste Voraussetzung – fast schon selbstverständlich – notwendig, dass Vergleichsarbeiten als relevante Tests erfahren werden und dabei *close-to-the-job* sind (vgl. *Erwartungs-Wert-Modell*; Koch, 2011; Kohler & Schrader, 2004; Nachtigall & Kröhne, 2006). Popham (2003, S. 51) wird dabei noch deutlicher: „First, educators should disregard data from any test that isn't instructionally useful. Second, they should push for the installation of instructionally useful test so that the data that those assessments yield will lead to better-taught student.“ Im Sinne von Rolff (2001, S. 337) stellen Vergleichsarbeiten für Schulen und Lehrkräfte aufgrund der Erhebungen *aller* Schülerinnen und Schüler der jeweiligen Lerngruppe Verfahren mit *unmittelbarem Nutzen* dar, die zusätzlich in Ergänzung zu Roh- auch adjustierte Daten zurückmelden (I. Hosenfeld et al., 2008). Maier (2008b) beispielsweise mutmaßt, dass durch das Vorhandensein adjustierter Landesmittelwerte und generell differenzierterer Rückmeldung die Vergleichsarbeiten von Lehrkräften aus Thüringen gegenüber Lehrkräften aus Baden-Württemberg deutlich bessere Akzeptanz- und Nutzungseinschätzungen erhalten. Um Lehrkräften die Relevanz der Vergleichsarbeiten für ihre tägliche Arbeit aufzuzeigen, sollten nach Koch (2011) beispielsweise durch Fortbildungen (vgl. Teil 2) die Transparenz gesteigert und die intendierten Ziele des Instruments verdeutlicht werden. Dabei sollte auch insbesondere auf die Datenhoheit der Schulen hingewiesen werden (vgl. Abschnitt 1.6.7.2). Die Bedeutung der wahrgenommenen Relevanz von Vergleichsarbeiten unterstreicht Ramm (2014, S. 21) mit einem positiven Zusammenhang zwischen der Bedeutsamkeit der Beurteilung und der Veränderungsbereitschaft in Bezug auf eine Verbesserung des

Unterrichts und führt umgedreht fehlende Kenntnisse über Qualitätsmanagement als einen von mehreren Hinderungsgründen an. Zusätzlich beeinflusst wird dieser Punkt durch den positiven Effekt, den die Bezugsfähigkeit von Rückmeldungen auf bereits bestehende/vorhandene Daten bzw. Rückmeldungen innerhalb der Schulen hat (also insbesondere wie anschlussfähig bzw. widersprüchlich die LSE-Rückmeldungen zu vorhandenen diagnostischen Informationen sind; Kühle & Peek, 2007; Kuper & Hartung, 2007).

1.6.4.2 Akzeptanz

Womöglich noch grundlegender als die Relevanz der Tests bzw. der Rückmeldungen für die Lehrkräfte ist die Verfügbarkeit der Tests bzw. der Rückmeldungen. Dies ist die Grundlage für jeglichen Umgang der Lehrkräfte mit den Tests bzw. den Rückmeldungen. Lehrkräfte müssen problemlosen Zugang zu den Tests, den Rückmeldungen und v. a. zu Informationen haben (KMK, 2010; Peek et al., 2006). Dies ist Voraussetzung dafür, dass der Evaluationsablauf nach Helmke (2004) mit der Rezeption überhaupt beginnen kann.

Vergleichsarbeiten werden dezentral von den Lehrkräften vor Ort durchgeführt und anschließend anhand eines Kodiermanuals von ihnen ausgewertet. Dieses Vorgehen stellt enorme Anforderungen an den Versuch, Objektivität zu gewährleisten, und stellt damit das Instrument testtheoretisch grundsätzlich in Frage: Voraussetzung für eine möglichst verzerrungsfreie Durchführung und Nutzung der Vergleichsarbeiten ist daher, dass die beteiligten Lehrkräfte dieses Verfahren *akzeptieren*. In der Literatur finden sich verschiedene Berichte, die in der Regel von indifferenten Haltungen bis hin zu ansatzweise vorhandener Akzeptanz sprechen: Mittelwerte liegen meist nahe dem semantischen Median³³ etwa bei Groß-Ophoff und Koch (2007b, S. 9) mit Mittelwerten über mehrere Jahre hinweg von $2.3 \leq M \leq 2.7$ bzgl. der Vergleichsarbeiten in Rheinland-Pfalz bei einem semantischen Median von 2.5 (vgl. auch Bensen et al., 2006; Groß-Ophoff et al., 2006; Groß-Ophoff & Koch, 2007a; A. Hosenfeld, 2010; Klieme et al., 2003, 2009; Koch, 2011; Kuper & Diemer, 2012; Maier, 2008b, 2009a, 2009b; Peek et al., 2006; Ramm, 2014; Wacker & Kramer, 2012). Hieraus sind vor

³³ Mit dem Terminus *semantischer Median* wird der mittlere Wert der Likert-Skala bezeichnet, der der Stelle der Likert-Skala entspricht, an der die Bedeutung wechselt, also beispielsweise der Wert 3 (als indifferente Haltung) bei einer 5-stufigen Skala von 1: „vollkommene Ablehnung“ bis 5: „vollkommene Zustimmung“. Die Bedeutung des *semantischen Medians* entspricht dem *theoretischen Mittelwert*.

allen die Ergebnisse von Maier (2009b) besorgniserregend, da von einer deutlich abnehmenden Akzeptanz berichtet wird. Diesem Befund widersprechen allerdings die Ergebnisse von A. Hosenfeld (2010, 45f.), die berichtet, dass „nach anfänglichen Implementationsproblemen die Akzeptanz der Beteiligten von Jahr zu Jahr steigt und die Ergebnisse in vielfältiger Weise zur Unterrichtsentwicklung genutzt werden.“

Die Akzeptanz der Lehrkräfte kann dabei durch verschiedene Stellschrauben positiv (oder auch negativ) beeinflusst werden: Insbesondere im anglo-amerikanischen Sprachraum sind sog. *high-stakes*-Testungen weit verbreitet. Bei diesen Tests werden aus den Resultaten Entscheidungen abgeleitet, welche die Lehrkräfte direkt betreffen, also beispielsweise Personalentscheidungen oder leistungsangepasste Bezahlung. Diese Tests bergen dementsprechend für Lehrkräfte einen deutlichen Kontrollcharakter und erheblichen Druck. Dieser Charakter liegt bei Lernstandserhebungen nicht vor, da die Datenhoheit bei der Schule liegt. Nach Maier (2010) und Maag Merki (2010) führt eine Beteiligung von Lehrkräften an der Entwicklung und möglichst geringe Sanktionen zu einer erhöhten Akzeptanz sowie zu einem höher eingeschätzten Potenzial für eine produktive Nutzung. Ergänzend steigern die Partizipation auch die wahrgenommene Validität der Ergebnisse (Koch, 2011).

Damit Lehrkräfte Vergleichsarbeiten nicht als *high-stakes*-Test oder externe Kontrolle wahrnehmen, bedarf es ausreichender *Informationen*, da die Mehrheit der Lehrkräfte (80%) prinzipiell bereit ist, die Ergebnisse schulischer Arbeit prüfen zu lassen (Rolff, 2001). Lehrkräfte müssen die Ziele und Funktionen von Vergleichsarbeiten kennen und diese „als für ihre schulische Arbeit sinnvoll und nützlich erachten“ (relevante Tests; Peek et al., 2006, 228ff.). Bensen et al. (2006) führen darüber hinaus an, dass Vergleichsarbeiten vor allem dann als sinnvoll erachtet und akzeptiert werden, wenn sie in der intendierten Art und Weise eingesetzt werden, was letztlich die Kenntnis über die intendierte Nutzung seitens der Lehrkräfte voraussetzt. Diese Informationen können Lehrkräften beispielsweise über Fortbildungen, Informationsmaterial oder bereits in der Ausbildung vermittelt werden (Oelkers & Reusser, 2008) und scheinen notwendig zu sein, da über die Hälfte der via Interview befragten Lehrkräfte in einer Studie von Maier (2009a, S. 140) angibt, dass „ein positives Gesamtergebnis ausreichend für die Außendarstellung“ sei; mit der „Außendarstellung“ wird dabei ein im Sinne der Vergleichsarbeiten zur Schulentwicklung deutlich

nachrangiger Aspekt in den Fokus gestellt. Ein weiterer Aspekt zur Steigerung der Akzeptanz liegt in der Nutzung von adjustierten Vergleichswerten.

1.6.4.3 Verständlichkeit

Die Akzeptanz eines Verfahrens wird des Weiteren durch dessen *Verständlichkeit* bzw. die der Rückmeldungen beeinflusst. Dieses Merkmal ist dabei sicherlich noch essenzieller als eine eher abstrakte Akzeptanz, um Vergleichsarbeiten im intendierten Maße einsetzen zu können. Prinzipiell werden zustimmende Mittelwerte für die Verständlichkeit der Rückmeldungen angegeben (u. a. Dederling, 2011; Groß-Ophoff et al., 2006; Groß-Ophoff & Koch, 2007a; Kohler & Schrader, 2004; Nachtigall & Kröhne, 2006). Durch die *didaktischen Kommentare* sowie die erläuterten Rückmeldungen und Informationen wird versucht, die Verständlichkeit für die Lehrkräfte zu steigern (I. Hosenfeld, 2005; Koch et al., 2006). So soll sichergestellt werden, dass empirisch wenig bewanderte Lehrkräfte nicht mit dem Angebot an Informationen überfordert sind und es zu „hochkomplexen Rezeptionsproblemen“ (etwa Fehlinterpretationen) kommt (Rolff, 2001, S. 343; Bensen et al., 2006; Dobbstein & Peek, 2008; I. Hosenfeld et al., 2006; Peek & Dobbstein, 2006a; Peek et al., 2006). Mathematiklehrkräfte berichten intuitiv erwartbar von weniger Verständnisschwierigkeiten als Lehrkräfte anderer Fächer (Kühle & Peek, 2007). Um diesen Verständnisproblemen zu begegnen, werden von vielen Autoren neben entsprechenden Erläuterungen und Informationsmaterial Unterstützungsangebote wie etwa Fortbildungen angeführt (Groß-Ophoff & Koch, 2007a; I. Hosenfeld et al., 2006), da sich Interpretationsfähigkeit der Lehrkräfte und Datenqualität u. U. unterscheiden. Generell gilt: „Je besser die Datenqualität, desto wahrscheinlicher entsteht daraus datengesteuerte Schulentwicklung. [...] Letztlich entscheiden die Nutzer selbst über die Nutzung“ (Rolff, 2002, S. 93).

Ein ausschlaggebender Faktor der Verständlichkeit ist die *professionelle Kompetenz* der Lehrkräfte: Es setzt diagnostische Kompetenz voraus, mit den Schülerproduktionen diagnostisch und unterrichtsleitend umzugehen. Es erfordert statistisches Grundverständnis, um mit Diagrammen verständlich zu arbeiten und bedarf einer funktionierenden innerschulischen Steuerung, um Ergebnisse der Vergleichsarbeiten etwa in zukünftigen Fortbildungsplanungen zu berücksichtigen (Peek et al., 2006). In diesem Zusammenhang müssen Vergleichsarbeiten die Balance zwischen Aufwand

(Auswertung, Interpretation etc.) und Nutzen (z. B. Optimierung von Unterrichtsabläufen) wahren.

Empirisch zeigt sich in vielen Befragungen eine vergleichsweise gute Verständlichkeit: Groß-Ophoff und Koch (2007b) berichten einen Mittelwert von 3.20 im Mittel über drei Jahre bei der rheinland-pfälzischen Vorläuferversion der Vergleichsarbeiten; Koch et al. (2006, 193ff.) von $M = 3.07$ jeweils auf einer vierstufigen Skala. Ähnlich positive Werte finden I. Hosenfeld et al. (2008, S. 55): $M = 3.20$ ($SD = 0.58$) respektive $M = 3.10$ ($SD = 0.68$) für die Verständlichkeit der Fähigkeitsniveau-Rückmeldungen zwei Wochen bzw. einige Zeit nach der Dateneingabe. Die Verständlichkeit auf aggregierter Ebene liegt ebenso hoch: $M = 3.20$ ($SD = 0.61$) für die Aggregation auf Klassen-/Schulebene und $M = 3.00$ ($SD = 0.79$) für den fairen Vergleich. Trotz oder gerade wegen dieser hoch ausfallenden Werte zeigt sich bei Kühle und Peek (2007, S. 436) kein direkter Einfluss der Verständlichkeit auf die Bereitschaft zur nutzungsorientierten Rezeption ($\beta = .01$, $p = .216$; evtl. handelt es sich hier um Deckeneffekte oder soziale Erwünschtheit), während die Akzeptanz mit $\beta = .35$ ($p < .001$) einen signifikanten Einfluss hat, ebenso wie die Nützlichkeit für den pädagogischen Alltag $\beta = .50$ ($p < .001$). Allerdings zeigt sich in diesen Daten eine hohe signifikante Korrelation von $r = .50$ ($p < .01$) zwischen der Nützlichkeit und der Verständlichkeit. Bonsel et al. (2006, S. 147) weisen darauf hin, dass der Einfluss der Verständlichkeit auf die Bedeutung der Rückmeldungen für die Unterrichtsentwicklung durch die Nützlichkeit mediiert wird (vgl. Abbildung 19). Ebenfalls in dieser Analyse wird deutlich, dass auch die Intensität der Auseinandersetzung einen positiven Einfluss auf die Verständlichkeit und die Nützlichkeit hat (A. Hosenfeld, 2010, S. 46). Insbesondere für die Intensität der Auseinandersetzung mit den Ergebnissen berichten jedoch Groß-Ophoff und Koch (2007b) im zeitlichen Verlauf abnehmende Werte ($M_{2004} = 3.10$; $M_{2005} = 2.80$; $M_{2006} = 2.60$). Aufgrund der fehlenden Angaben zur Standardabweichung und der unterschiedlichen Stichproben zu den Messzeitpunkten lassen diese Ergebnisse aber nur sehr ungenaue Trendaussagen zu. Interpretieren lassen sich die zurückgehenden Intensitätswerte zwar einerseits so, dass eine längere Auseinandersetzung nicht mehr nötig ist, andererseits u. U. jedoch auch nicht mehr durchgeführt wird. Hierzu wären Forschungen im Panel-Design sehr aufschlussreich, inwiefern wiederholte Teilnahme und Nutzung der Vergleichsarbeiten selbige beeinflussen. Fraglich erscheint vor diesem Hintergrund die Richtung der Beziehung zwischen Intensität und Verständlichkeit bzw. Nützlichkeit. Die in Abbildung 19 angenommene Wirkrichtung geht

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

von Auseinandersetzung als Prädiktor für Verständlichkeit bzw. Nützlichkeit aus, jedoch zeigt Koch (2011, S. 148) die umgedrehte Wirkrichtung mit durchweg positiven Regressionskoeffizienten der Verständlichkeit auf die Intensität ($\beta_{2004} = .47, p < .01$; $\beta_{2005} = .29, p < .01$; $\beta_{2006} = .16, p < .01$; $\beta_{2007} = .38, p < .01$; $\beta_{2008} = .33, p < .01$) sowie auf die Nützlichkeit ($\beta_{2004} = .51, p < .01$; $\beta_{2005} = .19, p < .01$; $\beta_{2006} = .35, p < .01$; $\beta_{2007} = .28, p < .01$; $\beta_{2008} = .29, p < .01$).

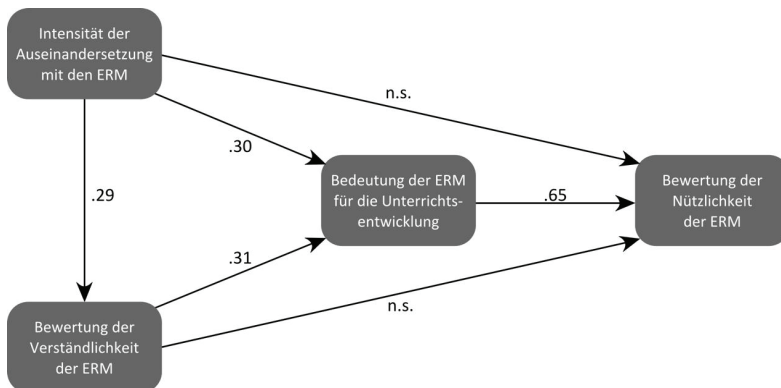


Abbildung 19: Zusammenhänge zwischen auf die Vergleichsarbeiten bezogenen Einschätzungen der Fachlehrkräfte (Bonsen et al., 2006, S. 147)

Anmerkung: Die berichteten Koeffizienten werden signifikant ($p < .05$).

$N = 2162$; $\chi^2 = 1738.1515$, $df = 98$; CFI = .951; TLI = .939; RMSEA = .088

1.6.4.4 Nützlichkeit

In Abbildung 19 wird deutlich, dass allein das Verständnis nicht ausreichend für die Einschätzung der Bedeutung der Vergleichsarbeiten bzw. deren Rückmeldungen ist, sondern die Intensität ebenfalls einen positiven Effekt hat. Die Bedeutung zeigt seinerseits einen großen Einfluss auf die Nützlichkeit, ohne dass dessen Prädiktoren Intensität und Verständlichkeit einen direkten Effekt auf die Nützlichkeitsbewertung hätten (Groß-Ophoff et al., 2006; Kohler & Schrader, 2004; Kuhn, 2014; Nachtigall & Kröhne, 2006; Peek & Dobbelsstein, 2006a). Exemplarisch zeigt sich bei Nachtigall und Jantowski (2007, 406f.) mit 80% Zustimmung für Nutzbarkeit und Verständlichkeit von Lehrkräften ein positives Bild.

Einen entscheidenden Ausschlag für die Nützlichkeit gibt die Balance zwischen Aufwand und Ertrag, bei der vor allem der zeitliche Aufwand, der mit der Korrektur und Online-Dateneingabe verbunden ist, eine Hälfte der imaginären Waage ausmacht.

Dieser liegt laut Bensen et al. (2006, 142ff.) für Mathematik-Lehrkräfte etwas niedriger als für andere Fachlehrkräften³⁴. Demnach bleibt zu vermuten, dass vor allem Deutsch- und Englischlehrkräfte der Aussage „Der mit der Lernstandserhebung verbundene Arbeitsaufwand steht in keinem Verhältnis zum Nutzen.“ zugestimmt haben und für einen entsprechend zustimmenden Mittelwert bei den Lehrkräften gegenüber der Schulleitung und Koordinatoren sorgen³⁵. In einer Untersuchung der Vergleichsarbeiten in Thüringen kommt Nachtigall (2014, S. 87) ebenfalls zu dem Ergebnis, dass die Nützlichkeit von Mathematiklehrkräften gegenüber Deutsch- und Englischlehrkräften positiver eingeschätzt wird, wobei die Nützlichkeit in die Facetten „Lernstandsdiagnostik“, „Unterrichtsentwicklung“ und „Schulentwicklung“ untergliedert wurde (vgl. Fachmittelwerte Tabelle 6; geringere Werte repräsentieren höhere Zustimmung). Dabei korrelieren die drei Facetten der Nützlichkeit hoch mit jeweils etwa $r = .70$ (leider ohne Angabe des Signifikanzniveaus). Ebenfalls positiv hängen Lehrplanadäquatheit und die Facetten der Nützlichkeit zusammen (etwa jeweils $r = .50$; wiederum ohne Angabe eines Signifikanzniveaus). Diese positive Korrelation ist anschlussfähig an Befunde von Dobbelsstein und Peek (2008, S. 28), nach denen die Nützlichkeit höher eingeschätzt wird, wenn Rückmeldungen mit vorhandenen Daten verknüpft werden können; insofern sind Rückmeldungen der Vergleichsarbeiten anknüpfungsfähig an den an Bildungsstandards orientierten Unterricht, wenn sie lehrplanadäquat sind (vgl. Bartnitzky, 2005).

Tabelle 6: Einschätzung der Kompetenztests im Hinblick auf Nützlichkeit (Mittelwertangaben; Nachtigall, 2014, S. 89)³⁶

	Lehrplan- adäquatheit	Lernstands- diagnostik	Unterrichts- entwicklung	Schul-ent- wicklung
Deutsch	2.62	3.02	3.08	3.32
Mathematik	2.49	2.66	2.8	3.04
Englisch	2.72	2.89	3.07	3.31

Die optimistische Einschätzung seitens der Schulleitungen sowie Koordinatorinnen und Koordinatoren aus der Befragung von Bensen et al. (2006) bestätigt sich auch in den Ergebnissen von Nachtigall und Jantowski (2007, 404ff.): Der Großteil der

³⁴ $M_{\text{Mathematik}} = 3.15$, $SD = 1.57$; $M_{\text{Deutsch}} = 5.45$, $SD = 2.16$; $M_{\text{Englisch}} = 5.11$, $SD = 2.17$

Die Auswertungskategorien sind 10-Minuten-Schritte von 1 = „unter 10 Minuten“ bis 8 = „71-80 Minuten“ pro Testheft.

³⁵ $M_{\text{Lehrkräfte}} = 1.70$, $SD = 0.86$; $M_{\text{Schulleitungen}} = 3.58$, $SD = 0.86$; $M_{\text{Koordinatoren}} = 3.64$, $SD = 0.76$ bei vierstufigem Antwortformat: 1 = „trifft zu“ bis 4 = „trifft nicht zu“

³⁶ Leider fehlen Angaben zur Standardabweichung auf der Skala von 1 = „volle Zustimmung“ bis 5 = „gar keine Zustimmung“.

befragten Schulleitungen sieht in den Vergleichsarbeiten wichtige Impulse für die gezielte Förderung und Differenzierung (70.4%), spricht ihnen partiell wichtige Impulse für die Schulentwicklung zu (85.0%) und schätzt die Ergebnismeldungen als hilfreich für Schul- und Unterrichtsentwicklung ein (90.6%).

1.6.4.5 Kooperation

Die Vergleichsarbeiten richten sich in ihrer Zielsetzung ganz bewusst sowohl an Lehrkräfte als auch an Fachkollegen. Eine kollegiale Nutzung und Thematisierung in Fachkonferenzen oder die Besprechung im Fachteam setzt jedoch bereits eine gewisse Kommunikations- und Kooperationskultur voraus (Bonsen et al., 2006; Dobbelsstein & Peek, 2008; Helmke et al., 2003; Holtappels, 2014b; Kuper & Hartung, 2007; Maier et al., 2012; Peek & Dobbelsstein, 2006a; Peek et al., 2006). Beispielsweise zeigt A. Hosenfeld (2010, S. 144) einen positiven Zusammenhang zwischen der Reflexionsbereitschaft und der kollegialen Ergebnisdiskussion ($r = .30$, $p = .02$). Vor allem bestehende Strukturen scheinen positiv von den Vergleichsarbeiten beeinflusst zu werden, wodurch die Bedeutung der Kooperation als Voraussetzung noch deutlicher wird (Koch, 2011).

Eine solche Kooperation (vgl. auch Abschnitt 1.5.2.2) sollte dabei von der Schulleitung sowie von Schulaufsicht angeregt und unterstützt werden (Burmeister et al., 2005; Holtappels, 2014a, 2014b; I. Hosenfeld, 2005), kann jedoch auch in größerem Zusammenhang erkannt werden, im Sinne von schulischen Netzwerken und folglich schulübergreifender sowie außerschulischer Kooperation (Projektgruppe "Innovationen im Bildungswesen", 2005).

1.6.4.6 Unterstützungssysteme

Lehrkräfte sind Experten für die Vermittlung von Wissen und die Gestaltung von Unterricht. Lehrkräfte sind jedoch in der Regel keine Experten darin, statistische Daten, Diagramme und Rückmeldungen zu interpretieren, entsprechende Schlüsse daraus zu ziehen oder die Grenzen von Untersuchungen korrekt einzuschätzen, wie man es beispielsweise von empirisch Bildungsforschenden erwarten würde. Aus diesem Grund wird von diversen Autoren ein ausreichendes Unterstützungsangebot für Lehrkräfte gefordert. Diese Unterstützung kann dabei institutionell vonseiten der Schulleitung durch entsprechende Möglichkeiten zur Datennutzung (etwa Bereitstellung von entsprechenden Zeitfenstern) kommen, von der Schulaufsicht durch Beratung, „best-practice-Beispiele“ oder ressourcielle Ausstattung (Ramm, 2014, S. 28) oder

durch Fortbildungsinstitutionen (etwa Universitäten) erfolgen, denn „die Ursache für fehlerhafte statistische Schlussfolgerungen liegt normalerweise nicht an dem zugrunde liegenden Rechenverfahren, sondern an mangelnden Kenntnissen bei der Anwendung und der Interpretation von Statistiken – und zwar sowohl auf Seiten des Forschenden als auch auf Seiten des Konsumenten“ (vgl. auch bei I. Hosenfeld et al., 2006; KMK, 2010; Ditton & Arnoldt, 2004; EMSE, 2008; Koch, 2011; Projektgruppe "Innovationen im Bildungswesen", 2005; Rolff, 2001). Diese Unterstützung wird seitens der Autoren jedoch nicht nur aus theoretischer Perspektive gefordert, sondern auch als in großen Studien erkanntes entscheidendes Merkmal für einen effektiven Umgang mit den Vergleichsarbeiten angegeben (Maag Merki, 2010; Maier, 2009a; Schildkamp & Visscher, 2009).

1.6.5 Wirkung von Lernstandserhebungen

Im Kapitel zur Evaluation wurde das theoretische Ablaufmodell nach Helmke (2004; vgl. Abbildung 10) vorgestellt. Nachdem nun Bedingungen für eine Nutzung bereits geschildert wurden, soll im Folgenden kurz dargelegt werden, welche Auswirkungen sich aus den Vergleichsarbeiten ergeben können. Es wird sich hierbei grob am genannten Modell orientiert. Leider ergeben sich aufgrund der föderalen Unterschiede in der Datenhoheit und der konkreten Umsetzung auch für die Nutzung Unterschiede, da beispielsweise die Testhefte gänzlich vom zuständigen Landesinstitut korrigiert werden. Dies hat selbstverständlich einen großen Einfluss auf die konkrete Nutzung – insbesondere auf die vorgeschlagene individualdiagnostische Nutzung, da keine Möglichkeit eines individualdiagnostischen Blickes während der Korrektur besteht. Entsprechend gibt es keine systematischen Wirkungsanalysen über Bundesländer hinweg, sondern als Orientierung dienen im Weiteren eher internationale Vergleichsuntersuchungen bzw. deren Rezeption und Wirkung (Maier, 2009a, S. 131). Effekte, wie sie v. a. von *high-stakes*-Testungen ausgehen – etwa eine Einschränkung des Unterrichtsinhaltes (*narrowing-the-curriculum*) oder gezielte Vorbereitung auf Tests (*teaching-to-the-test*) – werden dabei vorrangig aus der englischen Literatur berichtet (u. a. Wall, 2000; vgl. auch Abschnitt 1.6.7).

1.6.5.1 Rezeption und Reflexion

Nach den vorliegenden Informationen ist davon auszugehen, dass Lehrkräfte die vorhandenen Ergebnismeldungen wahrnehmen und diese rezipieren. Ohne eine Wahrnehmung und Auseinandersetzung mit den Rückmeldungen ist eine Reflexion

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

der Ergebnisse unmöglich (zsf. Dederling, 2011; auch bei Kühle & Peek, 2007; Nachtigall & Jantowski, 2007; Richter et al., 2014).

Für die Reflexion der Ergebnisse wird vielfach und von der großen Mehrheit der Lehrkräfte (70.1%; Groß-Ophoff & Koch, 2007a, S. 7) der Vergleich mit anderen Lerngruppen als zentral angegeben – sowohl im schulinternen Vergleich als auch bundeslandweit über den korrigierten Landesmittelwert (Stryck, 2005). Dabei bemerken Kuper und Diemer (2012, S. 238), dass neben der „analytisch differenzierenden Verwendung“, also der Suche nach Ursachen zur Maßnahmenableitung, „auch eine kompetitive Lesart der Ergebnismeldungen und der Vergleiche zwischen den einzelnen Klassen zu beobachten“ ist. Weiter wird berichtet, dass vor allem bei „schlechten“ Ergebnissen (absolut geringe Lösungsquoten, Zurückbleiben hinter den Erwartungen etc.) Handlungsbedarf gesehen wird, wobei die Ursachen selten in der eigenen *professionellen Kompetenz* gesucht werden, sondern eher extern attribuiert werden (Groß-Ophoff & Koch, 2007b; A. Hosenfeld, 2010; Maier et al., 2012; O'Day, 2002). Dies kann vor dem Hintergrund eines hohen Ich-Involvements bei Lehrkräften als Schutzmechanismus verstanden werden (vgl. auch *endemische Unsicherheit*; vgl. Abschnitt 1.4.3.1; A. Hosenfeld, 2010; Kuper & Diemer, 2012). Entsprechende Befunde berichtet Maier (2009a, S. 140): Danach hat lediglich die Hälfte der befragten Lehrkräfte die Frage nach einer kritischen Reflexion des eigenen Unterrichts im Anschluss an die Vergleichsarbeiten mit „Ja“ beantwortet. Dieser Wert fällt etwas pessimistischer aus als Berichte von Koch et al. (2006, 193f.): Die Diskussion erfolgt in 63% der betrachteten Schulen im Kollegium, jedoch lediglich in knapp 40% in der Fachkonferenz. Entsprechend fallen die Konsequenzen/Aktionen (vgl. nächster Abschnitt) auf der Fachkonferenz-Ebene noch geringer aus.

1.6.5.2 Aktion

Auch wenn vergleichsweise wenige Befunde zur Rezeption und Reflexion vorliegen, so zeigen jedoch die Ergebnisse zum Bereich *Aktion*, dass eine Auseinandersetzung mit den Ergebnissen erfolgen muss. Leider gibt es für die Vergleichsarbeiten keine Untersuchungen der Wirkungen auf die Schülerleistungen (Maier et al., 2012; Ramm, 2014) und die meisten Auswertungen bewegen sich auf der Ebene von Selbstauskünften der Lehrkräfte. Einschränkend muss mit Verweis auf Kuper und Hartung (2007, S. 223) angemerkt werden, dass wie bereits dargelegt, Vergleichsarbeiten keine

unmittelbaren und direkt ablesbaren Ableitungen ermöglichen, sondern zunächst eine Interpretationsleistung von Lehrkräften erfordern.

Die Vergleichsarbeiten scheinen nur in geringem Maße tatsächlichen Innovationscharakter zu besitzen: Durch die Vergleichsarbeiten werden in vielen Fällen vorhandene Maßnahmen intensiviert oder Strukturen bestätigt, jedoch beispielsweise „innovative Formen der Zusammenarbeit [werden] eher selten eingeführt“ (A. Hosenfeld, 2010, 45f.; *symbolic use*). Darüber hinaus scheinen die gezogenen Konsequenzen vielfach auf den konkreten Unterricht der einzelnen Lehrkraft abzielen und weniger die Arbeit der Fachkonferenz zu betreffen. Diese Befunde belegen Kuper und Hartung (2007, S. 225) qualitativ mit Interviews. Maier (2009a, S. 140) berichtet ebenfalls aus qualitativen Interviews in Baden-Württemberg, dass eine „systematische, schulinterne Kommunikation“ von keiner Lehrkraft genannt wurde, jedoch gut die Hälfte der befragten Lehrkräfteangaben, im informellen Austausch über die Testergebnisse mit Kolleginnen und Kollegen geredet zu haben (19 von 36 befragten Lehrpersonen). Hierbei bleibt zu hoffen, dass sich gegebenenfalls im Laufe der Zeit die systematische Auswertung der Vergleichsarbeiten etabliert, wie es Reimers et al. (2009, S. 7) für die Unterrichtsgestaltung berichten, deren Einschätzung sich von 2007 bis 2009 „deutlich verbessert“ hat. Ebenfalls Anlass zur Hoffnung geben Werte aus Thüringen, nach denen in 82.4% der Fälle eine Besprechung in der Schulkonferenz erfolgt (Nachtigall & Jantowski, 2007).

Nach Koch et al. (2006) lassen sich u. a. die Bereiche der Unterrichtsinhalte, Unterrichtsgestaltung und Individualförderung unterscheiden, die sich jedoch fast ausschließlich auf den Unterricht in der einzelnen Lerngruppe beziehen: Unter dem erstgenannten Bereich werden u. a. Wiederholungen (von knapp 70% der Lehrkräfte realisiert, von knapp 20% erwogen), Üben von Aufgabenformaten (61.5% bzw. 16.1%), eine Neuaufnahme von Inhalten (32.2% bzw. 23.3%) und das Vorziehen von Inhalten (21.0% bzw. 27.5%) zusammengefasst. In Thüringen werden von 40% der Lehrkräfte Veränderungen an Unterrichtsinhalten berichtet (Nachtigall & Jantowski, 2007; vgl. auch Richter et al., 2014; Groß-Ophoff et al., 2006; Koch et al., 2006; Kühle & Peek, 2007; Schneewind & Kuper, 2009). In Bezug auf eine Individualförderung wird mehrheitlich von der Nutzung der Vergleichsarbeiten in Elterngesprächen berichtet (46.8% realisiert und 14% erwogen; Koch et al., 2006, S. 195), was jedoch von Maier (2009a, S. 134) nur bedingt bestätigt werden kann: Lediglich 5 von 36 befragten Lehrkräften

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

gaben an, Gespräche mit Eltern geführt zu haben.³⁷ Zusätzlicher Unterricht (17.7% bzw. 14.6%) oder auch der Einbezug von Förderlehrkräften (20.9% bzw. 14.6%) – also insbesondere Maßnahmen, die über die einzelne Lerngruppe hinausgehen – werden nur in vergleichsweise wenigen Fällen realisiert bzw. erwogen. In der Unterrichtsgestaltung erkennen in Thüringen 35% der Lehrkräfte Einflüsse und 30% eine Wirkung auf die Binnendifferenzierung (Nachtigall & Jantowski, 2007). Die entsprechenden Angaben in der Untersuchung von Koch et al. (2006) fallen hingegen etwas höher aus: „Überdenken der Methoden“ von 51.9% realisiert und 21.5% erwogen und jeweils etwa ein Drittel der Lehrkräfte realisiert eine leistungshomogene (33.8%) oder -heterogene Differenzierung (36.8%) oder nutzt leistungsstarke Schülerinnen und Schüler als Tutoren (43.2%). In Bezug auf Lernkontrollen oder das Unterrichtsmaterial lässt sich feststellen, dass die Vergleichsarbeiten ihrer Orientierungsfunktion gerecht werden, da gut ein Drittel der Befragten in der Folge bildungsstandardorientierte Testaufgaben einsetzt (35.7%) und entsprechende Unterrichtsmaterialien (41.5%) sowie Übungsaufgaben entwickelt (35.4%). Generell lässt sich jedoch feststellen, dass die durch die Vergleichsarbeiten angeregten Veränderungen in den verschiedenen Untersuchungen eher als *instrumental/symbolic use* nach Visscher und Coe (2003; vgl. Abschnitt 1.3.3) zu beschreiben sind, also eher vorhandene Maßnahmen intensiviert oder unmittelbar umsetzbare Projekte abgeleitet werden. Dabei werden vor allem Maßnahmen getroffen, die von einzelnen Lehrkräften unmittelbar zu leisten sind und keine Abstimmung oder Kooperation mit anderen Lehrkräften erfordern sowie keine finanziellen Ressourcen benötigen (etwa Lehrbuchwechsel, erwogen von 17.5%, realisiert von 5.9%, oder zusätzlicher Unterricht; Koch et al., 2006, S. 195; vgl. *Einzelkämpfertum* in Abschnitt 1.1.3). Ein großer Teil der Lehrkräfte scheint jedoch keinerlei Nutzen aus den Vergleichsarbeiten zu ziehen (vgl. Reflexionswerte): So geben 20 von 36 befragten Lehrkräften keine Veränderungen bzw. eine Bestätigung des eigenen Unterrichts an (Maier, 2009a, S. 134). Für vergleichbare Testverfahren aus Österreich werden etwa für 60% der Lehrkräfte keine Veränderungen berichtet, wobei andererseits mehr als die Hälfte der Lehrkräfte zumindest berichtet, nach den Rückmeldungen einen besseren Informationsstand über Stärken und Schwächen ihrer Schülerinnen und Schüler zu haben (Grabensberger et al., 2008). Hierbei spielen

³⁷ Wobei nicht auszuschließen ist, dass in diesem Falle die Vergleichsarbeiten als Auslöser des Elterngespräches eingeschätzt werden, während im vorangegangenen genannten Fall regelmäßige Elterngespräche durch die Rückmeldungen der Vergleichsarbeiten bereichert werden.

jedoch wiederum die bereits genannten Bedingungen und Voraussetzungen eine wesentliche Rolle, da „Lehrkräfte nach eigenen Angaben häufiger Veränderungen im Unterricht vornehmen, wenn VERA aus ihrer Sicht zur Unterrichtsentwicklung dient und diagnostische Informationen bereitstellt“ ($\beta = -.33$, $p < .05$ unter Kontrolle von Geschlecht, Berufserfahrung, der Sichtweise der Vergleichsarbeiten als externe Kontrolle sowie Schülermerkmalen; Richter et al., 2014, 234f.). Analog zeigt sich eine höhere Kompetenzorientierung ($\beta = .52$, $p < .05$) und mehr Differenzierung ($\beta = .38$, $p < .05$) bei unterrichtsentwickelnder Einschätzung der Vergleichsarbeiten.

1.6.5.3 Evaluation von Folgemaßnahmen

Nach aktuellem Stand gibt es leider noch keine Befunde darüber, wie und ob die ergriffenen Maßnahmen nach Vergleichsarbeiten evaluiert werden. Dies ist zwar einerseits schade und lässt eine Forschungslücke offen, ist jedoch andererseits im schulischen Alltag nachvollziehbar. Eine kausal-analytische Untersuchung, die belastbare Ergebnisse hervorbringen könnte, würde einen enormen Eingriff in den Schulalltag und die schulischen Entscheidungs- sowie Entwicklungsprozesse bedeuten, da letztlich in der Zeit der Evaluation keine oder nur kaum andere Veränderungen vorgenommen werden dürften.

1.6.6 Typen von Lehrkräften

Einfache Erklärungsmuster, die allgemein gültige Aussagen für alle Lehrkräfte und alle Bundesländer propagieren, sind nicht geeignet, die föderale und individuelle Vielfalt abzubilden. Vor dem Hintergrund der uneinheitlichen Ergebnisse im vorherigen Abschnitt scheinen auch die bisher berichteten Mittelwerte ganzer Stichproben m. E. nicht detailliert genug zu sein. Um der Heterogenität Rechnung zu tragen, haben sich in Bezug auf die Vergleichsarbeiten Groß-Ophoff, I. Hosenfeld und Koch (2007) sowie etwas später A. Hosenfeld (2010) an eine Typisierung von Lehrkräften gewagt (ohne konkreten Bezug zu Vergleichsarbeiten gibt es weitere typenbildende Untersuchungen bzgl. Rückmeldungsnutzung für Lehrkräfte, beispielsweise Ditton & Arnoldt, 2004).

In der erstgenannten Untersuchung wurden Fragebögen von 1510 Lehrkräften aus sieben Bundesländern (1199 verschiedene Schulen; Rücklaufquote 34.6%) ausgewertet. Hierzu wurde mithilfe von Mixed-Rasch-Modellen der beste Modellfit für eine Vier-Klassen-Lösung gefunden, in der sich die Stichprobe zu 38.9% in Typ 1, 35.2% in Typ 2 und 25.0% in Typ 3 aufgliedert (die Zuordnungswahrscheinlichkeiten liegen

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

jeweils bei mehr als .89). Die restlichen 0.9% werden von den Autoren als „Unskalierbare“ von weiteren Analysen ausgeschlossen. Die Ergebnisse lassen sich qualitativ folgendermaßen zusammenfassen (vgl. Tabelle 7):

Tabelle 7: Qualitativer Überblick über die Ergebnisse von Groß-Ophoff et al. (2007) im Vergleich der einzelnen Cluster zueinander

	Intensität	Verständlichkeit	Nützlichkeit
	der Ergebnismrückmeldungen		
Typ 1	+	+	+
Typ 2	-	-	+
Typ 3	+	-	-

Die Unterschiede zwischen diesen Typen zeigen sich dabei nicht nur bezüglich der Nutzung, sondern insbesondere auch in Bezug auf die Realisierung von Maßnahmen: $M_{Typ\ 1} = 7.4$, $M_{Typ\ 2} = 7.0$, $M_{Typ\ 3} = 6.3$, $F(2) = 8.55$, $p \leq .05$. Dies wird von den Autoren noch detaillierter aufgeschlüsselt: Im Wesentlichen zeigt sich jedoch Typ 1 als in vielfältiger Weise sehr aktiv bei der Nutzung der Vergleichsarbeiten. Typ 2 setzt im Anschluss an die Vergleichsarbeiten ähnlich häufig wie Typ 1 zu Wiederholungen und Vertiefungen an (72.1% gegenüber 75.3%) und übt sogar ausgeprägter bestimmte Aufgabentypen (66.5% gegenüber 61.6%), zeigt aber beispielsweise weniger Engagement bei der Eigenentwicklung von an die Vergleichsarbeiten angelehnte Materialien oder Übungsaufgaben (etwa 36% zu 46%). Gemäß dem globalen Mittelwert oben fallen die Prozentangaben zu Realisierungen bei Typ 3 am geringsten aus: Lediglich etwa 60% der Lehrkräfte wiederholen im Anschluss an die Vergleichsarbeiten und sogar nur etwa ein Drittel entwickelt oder überarbeitet Unterrichtsmaterialien und/oder Übungsaufgaben. Vor diesem Hintergrund sind die Ergebnisse der vergangenen beiden Abschnitte zur Wirkung und zu Bedingungen differenzierter zu betrachten, da die beschriebenen Zusammenhänge beispielsweise zwischen Intensität der Auseinandersetzung und der Verständlichkeit, Nützlichkeit oder der Bedeutung (vgl. Abbildung 19) durch Typen beeinflusst werden.

A. Hosenfeld (2010) nutzt in ihrer Untersuchung latente Klassenanalysen (LCA) zur Typenfindung und wählt eine 3-Klassen-Lösung, wobei die zugrundeliegenden Skalen für die LCA dichotomisiert wurden. Die Stichprobe umfasst leider nur 43 Lehrkräfte. In der Untersuchung ging es um etwaige Unterschiede in der Wahrnehmung von schriftlicher und videobasierter Rückmeldung sowie den Einstellungen hierzu. Typ 1 (59%) charakterisiert sich dabei durch eine hilfreiche Einschätzung beider Rückmeldevarianten und korrespondierend hoher Reflexion, wobei die schriftlichen

Rückmeldungen positiver bewertet werden als das Video. Typ 2 (24%) zeigt in Anlehnung an das Evaluationsmodell von Helmke (2004; Abbildung 10) eine ebenfalls hohe Rezeption der schriftlichen Rückmeldungen (vergleichbar mit Typ 1) und eine mittlere Rezeption des Videos, jedoch deutlich abfallende Werte für die Reflexion und Aktion aus beiden Rückmeldevarianten. Typ 3 (18%) bevorzugt eindeutig eine video-basierte Rückmeldung. Für eine qualitative Übersicht der drei Typen wird sich in der folgenden Tabelle 8 auf die Darstellung der schriftlichen Rückmeldungen beschränkt.³⁸

Tabelle 8: Qualitative Unterscheidung zwischen Typen in A. Hosenfeld (2010, S. 161) bzgl. schriftlicher Rückmeldungen

	Rezeption	Reflexion	Aktion
Typ 1	+	+	+
Typ 2	+	0	-
Typ 3	-	-	-

Die Typen unterscheiden sich des Weiteren beispielsweise signifikant in der selbst angegebenen Kenntnis des eigenen Unterrichts ($F(2,37) = 5.34, p = .009, \eta^2 = .224^{39}$). Keine signifikanten Unterschiede werden für die Berufszufriedenheit, die Reflexionsbereitschaft und die Selbstwirksamkeitserwartung berichtet. Die Zufriedenheit mit der schriftlichen Rückmeldung zeigt dabei ein zwar nicht signifikantes, aber deutliches Ausreißen des Typs 3 ($M_{Typ\ 1} = 2.85, M_{Typ\ 2} = 2.9, M_{Typ\ 3} = 2.45$; bei vierstufigem Antwortformat⁴⁰). Ebenfalls nicht signifikant unterscheiden sich die Typen bei der Verständlichkeit der schriftlichen Rückmeldung. Die Mittelwerte aller Typen liegen über dem semantischen Median von 2.5, wie auch bei der Nützlichkeit der schriftlichen Rückmeldungen (wiederum ohne statistisch signifikanten Unterschied). Die tendenziell negativere Haltung von Typ 3 wird zusammenfassend im signifikanten Unterschied zwischen den Typen im Punkt Einstellung zur Untersuchung deutlich ($M_{Typ\ 1} = 3.19, M_{Typ\ 2} = 2.88, M_{Typ\ 3} = 2.57, F(2,36) = 3.24, p = .05, \eta^2 = .152$). Kritisch anzumerken bleibt für diese Untersuchung jedoch die geringe Stichprobengröße und

³⁸ Selbstverständlich ist eine gänzliche Vernachlässigung der Ergebnisse zum Video nicht vollkommen korrekt, da es durch Weglassen dieser Variablen zweifelsfrei zu anderen Ergebnissen in der LCA gekommen wäre. Da jedoch im weiteren Verlauf dieser Arbeit lediglich schriftliche Rückmeldungen von Lehrkräften genutzt werden, erscheint diese Einschränkung als Vereinfachung sinnvoll. Umfassende Informationen zum Vergleich von Video und schriftlicher Rückmeldungen können der Originalquelle entnommen werden.

³⁹ Die Mittelwerte sind leider nicht konkret angegeben. Aus dem Diagramm der Originalquelle lassen sich die Mittelwerte etwa ablesen (leider ohne Standardabweichung): $M_{Typ\ 1} = 3.0, M_{Typ\ 2} = 3.5, M_{Typ\ 3} = 3.25$ (bei vierstufigem Antwortformat).

⁴⁰ Die Mittelwerte konnten wegen mangelnder Angabe lediglich aus einem Diagramm abgelesen werden.

die resultierenden kleinen Klassen von nur acht (Typ 2) bzw. sieben Lehrkräften (Typ 3), wobei sich die Typen jedoch im Geschlechterverhältnis, Alter und Berufserfahrung nicht unterscheiden.

1.6.7 Kritik an VerA

Mit Vergleichsarbeiten stehen Lehrkräften ein Instrument zur Verfügung, welches vielfältige Einsatzmöglichkeiten bietet. Die geschilderten Befunde zur Wirkung zeigen jedoch auf, dass die Vergleichsarbeiten nicht in vollem Maße in der Praxis ankommen und genutzt werden. Dies mag zu einem Teil an zeitlich und regional beschränkten Problemen wie etwa lokaler Überbelastung (beispielsweise durch die Auswertungszeiten; Bonsen et al., 2006) liegen, allerdings gibt es einige grundsätzliche Schwierigkeiten und Kritikpunkte, die jedoch „nicht in Frage [stellen, S.V.], dass die Ergebnisse als wichtige Indikatoren für die angepeilten Leistungen dienen können“ (Brügelmann, 2005b, S. 8).

Häufig wird gegenüber Leistungstests angebracht, eine geringe Auswirkung auf die tägliche Praxis zu haben. Allerdings wird hierbei vernachlässigt, dass (vergleichende) Leistungstests und insbesondere die Vergleichsarbeiten „nur *einen* ‚Mosaikstein‘ ausmachen“ (Siebenborn, 2005, S. 47), also die Diagnose und korrespondierende Maßnahmen weiterhin durch die Lehrkräfte erfolgen müssen (F. Winter, 2005, S. 76) und einfache Erklärungsmuster oder Ursachengefüge basierend lediglich auf einer Messung „unterkomplex“ bleiben (Kuper & Hartung, 2007, S. 226).

1.6.7.1 Top-Down-Maßnahme/Verordnung

In den vorherigen Ausführungen wurde bereits deutlich, dass die mit den Vergleichsarbeiten verbundenen Hoffnungen z. T. das Instruments überfordern und damit Funktionen für Lehrkräfte u. U. undeutlich werden. Dies zeigt sich beispielsweise auch in einem Mittelwert nahe dem semantischen Median bei Richter et al. (2014, S. 238) für die Funktionen *Unterrichtsentwicklung*⁴¹. Der Vorwurf, dass die behördliche Seite lediglich Defizite ausmache, ohne daraus Handlungen abzuleiten (Burchard, 2014), ist zumindest für die Vergleichsarbeiten in Hessen nicht haltbar, da die Datenhoheit bei den Schulen liegt und damit keine Informationen über Defizite zu Behörden gelangen und folglich die Handlungspflicht ebenfalls bei den Schulen bzw. Lehrkräften liegt. Dennoch scheinen in diesem Sinne „Vergleichsarbeiten eine paradoxe

⁴¹ $M = 2.42$, $SD = 0.64$; $\mu_{\text{semantisch}} = 2.5$ bei vierstufigem Antwortformat

Intervention. Der Staat zwingt die Schulen und Lehrkräfte zur Selbstevaluation“ (Maier et al., 2012, S. 200). Insbesondere diese *top-down*-Verordnung birgt Problem-potential (O'Day, 2002). Pointierter formulieren Böttcher und Rürup (2010, S. 61): „Als Problem aber dürfte sich erweisen, dass die einschlägigen Reformmaßnahmen lediglich an die Schulen und ihr Personal adressiert werden. Während die Schulen eine Vielzahl von Rechenschaftsbelegen beibringen müssen, scheint der Wille der Politik, die eigenen Aktivitäten – vergleichend – zu evaluieren, nicht sonderlich ausgeprägt“ (vgl. auch Campbell, 1979). Zur Problematik von *top-down*-Entwicklungen wird hier auf die entsprechenden Ausführungen in den vorangegangenen Kapiteln verwiesen (vgl. Abschnitt 1.1.1.4 und 1.1.3).

1.6.7.2 Angst von Konsequenzen

Mit der steigenden Gefahr der Kontrolle oder – allgemeiner gesagt – von Konsequenzen nimmt die Gefahr der Verfälschung ebenfalls dramatisch zu (A. Hosenfeld, 2010; Leutner et al., 2008; Richter et al., 2014). Hierzu wird vielfach Bezug auf Campbell (1979, S. 85) genommen: „I come to the following pessimistic laws (at least for the U.S. scene): *The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social process it is intended to monitor.*“ Dieser Zusammenhang, der ursprünglich bei bezahlten Nachhilfelehrkräften nachgewiesen wurde, wird vielfach als *campbell's law* zitiert. Die Veröffentlichung von Testergebnissen führt danach zunächst zu Sorgen um negative Folgen und letztlich, um dies zu umgehen, zu Verfälschungstendenzen (Wall, 2000). Ausgehend vom Kodiermanual (welches eingehalten werden soll; Strähling, 2005) haben Lehrkräfte Freiheiten und damit gleichzeitig Schwierigkeiten mit der eindeutigen, Lösungsschritte ignorierenden richtig-/falsch-Bewertung (Brügelmann, 2005b, S. 8). Dies kann bei „vielerorts geschilderten und geschürten Ängsten und Verunsicherungen“ (Isaac et al., 2006, S. 109) zu systematischen, sog. *motivationalen* Beurteilungsfehlern führen (Spoden et al., 2014). Somit ist die Auswertungsobjektivität kritisch zu betrachten und auch bei der Durchführungsobjektivität gibt es wegen der dezentralen Durchführung viele Möglichkeiten der Verfälschung der Tests (etwa Hilfestellungen, zusätzliche Zeit etc.; Strähling, 2005).

Die Ursachen für Ängste und Kontrollvermutungen liegen dabei zum einen darin, dass „Evaluation und datenbasierte Schulentwicklung nicht zur Arbeitskultur von

Lehrkräften gehören“ (Maier, 2010, S. 65; Posch, 2009) bzw. erst vor kurzem im Zuge der Outputorientierung zur Arbeitskultur ergänzt wurden, und zum anderen in externen Attributionstendenzen der Lehrkräfte bedingt durch das *Einzelkämpfertum* (Kuper & Hartung, 2007, S. 224). Vergleichsarbeiten sind an sich jedoch nicht als kontrollierendes Instrument intendiert, sondern versuchen als *low-stakes*-Verfahren Lehrkräften Informationen zu liefern, zu denen sie mangels methodischer Kompetenz nicht gelangen würden (Ditton & Arnoldt, 2004). Sie sollten daher ohne die Gefahr von Konsequenzen einen höheren Nutzen erlangen als *high-stakes*-Tests (Brügelmann, 2005a; Isaac et al., 2006; Maag Merki, 2010; Maier, 2009a; Petersen, 2013), sofern den Lehrkräften diese Unterschiede respektive Besonderheiten der Vergleichsarbeiten bekannt sind.

1.6.7.3 Fehlnutzung der Ergebnisse

Wird davon ausgegangen, dass die Vergleichsarbeiten trotz Gefahren in Durchführung und Auswertung erfolgreich und so objektiv wie möglich durchgeführt wurden, erhalten Lehrkräfte umgehende Rückmeldung sowie (für Hessen) die um den korrigierten Landesmittelwert ergänzte Auswertung einige Wochen später. Als nächste Gefahr der Nutzung der Vergleichsarbeiten können diese Ergebnisse in nicht intendierter Art und Weise oder überhaupt nicht genutzt werden (I. Hosenfeld et al., 2006, S. 13; Rolff, 2001).

Dabei wird ein Großteil der fehlerhaften Interpretation und Nutzung nicht zu verhindern sein, weil sie nicht bewusst erfolgt, sondern etwa durch mangelnde Datenkompetenz begründet ist (Brügelmann, 2005b; *misuse*; vgl. Abschnitt 1.3.3). Ein bewusster Missbrauch der Ergebnisse oder allgemeiner von Daten durch die Lehrkräfte (z.B. bewusstes Entfernen von Teildatensätzen, Kürzen von Darstellungen oder Auswertungen) ist jedoch ebenfalls nicht auszuschließen (*abuse*; Lorenz, 2005; Stamm, 2003). Beispielweise liefern Vergleichsarbeiten zwar zuverlässige Aussagen auf Ebene ganzer Klassen, doch sind diese Aussagen notwendigerweise ausschnittshaft (Tillmann, 1999; Burkard & G. Orth, 2005; Nachtigall & Kröhne, 2006). Diese Daten eignen sich darüber hinaus ebenfalls nur eingeschränkt und v. a. weder statistisch abgesichert noch pädagogisch intendiert für die Erstellung von Rankings (Büchter & Leuders, 2005a; Isaac et al., 2006; Rolff, 2001; Schrader, 2010; Strähling, 2005; Uhl, 2006; A. Voss & Blatt, 2010). Die Gründe hierfür liegen u. a. im methodischen Bereich, da es sich lediglich um querschnittliche Messungen ohne Anspruch auf Repräsentativität

handelt (Arnold, 1999; Bühner, 2011), die zugrundeliegenden Rohwerte vielfach einflussreiche Faktoren außer Acht lassen (Helmke & I. Hosenfeld, 2005, S. 3) und darüber hinaus widersprechen Rankings aus pädagogischer Sicht den Zielen der Vergleichsarbeiten (Rolff, 2001, S. 345).

1.6.7.4 Teaching to the test/Narrowing the curriculum

Ein genereller Einwand gegenüber zentral entwickelten Tests sind ungewollte Einflüsse auf den Unterricht. Diese lassen sich als gezielte Vorbereitung auf Testinhalte (*teaching-to-the-test*) oder als eine bewusste Einschränkung der Unterrichtsinhalte auf die Testinhalte bzw. testbare Inhalte verstehen (*narrowing-the-curriculum*), wobei in der Literatur zum Teil unter dem ersten Begriff beides subsumiert wird (Brügelmann, 2005a; I. Hosenfeld et al., 2008; A. Hosenfeld, 2010; Jerald, 2006; Landwehr, 2011; Lorenz, 2005; Popham, 2001; Wall, 2000).

Wie bereits oben ausgeführt, werden jedoch die Vergleichsarbeiten unter direktem Bezug auf die verbindlichen Bildungsstandards entwickelt und sind entsprechend lehrplanvalide (ohne über die Passung zum tatsächlichen Unterricht eine Aussage machen zu können; Bremerich-Vos et al., 2005; Brügelmann, 2005b; Redaktion neues Deutschland, 2012; Selter, 2005b; Strähling, 2005; Switalla, 2005; Uhl, 2006). In diesem Sinne müsste ein *teaching-to-the-test*, also eine gezielte Testvorbereitung, zu einem an den Bildungsstandards orientierten und damit auf einen kumulativen Kompetenzaufbau zielenden Unterricht führen. Analog ist eine inhaltseinschränkende Wirkung ebenfalls kaum zu erwarten, sondern eher davon auszugehen, dass den Vergleichsarbeiten die erwünschte *Orientierungsfunktion* zukommt (vgl. 1.6.1; Helmke & I. Hosenfeld, 2003, S. 43). Mehr noch: Lehrkräften sowie Schülerinnen und Schüler sollte verdeutlicht werden, dass die Funktion der Vergleichsarbeiten die Evaluation des Lernstandes auf dem Weg zum Abschluss des Bildungsabschnittes ist, und die Vergleichsarbeiten konsequent ein Instrument darstellen, das bewusst ohne Kontrolle und ohne negative Konsequenzen auskommt. Vielmehr haben Vergleichsarbeiten das Ziel, Grundlage für individuelle Maßnahmen zu der Erreichung des Bildungsabschlusses jedes Einzelnen zu sein. So „ist eine gezielte Vorbereitung auf die Vergleichsarbeiten im Sinne von Übung der Aufgaben weder sinnvoll noch erforderlich“ (Hessisches Kultusministerium, 2009, S. 21). Andererseits bleibt die Frage offen, inwiefern Vergleichsarbeiten „unterrichtsvalide“ sind. Vor dem Hintergrund der Aufgabenentwicklung durch erfahrene aktive Lehrkräfte ist davon auszugehen, dass

Teil 1 – Theorie

Vergleichsarbeiten/Lernstandserhebungen

Vergleichsarbeiten sowohl lehrplan- als auch unterrichtsvalide sind, letztgenannter Punkt jedoch in der Verantwortung jeder einzelnen Lehrkraft liegt und entsprechend nicht dem Instrument angelastet werden kann: Im schlimmsten Fall ist der Unterricht selbst dann nicht lehrplanvalide.

Empirisch wurden die Auswirkungen der Vergleichsarbeiten von Wacker und Kramer (2012) untersucht. Von ihnen wurden für Baden-Württemberg die Orientierungsfunktion, die Übungserfordernisse und drei Facetten der Curriculumseinschränkung in Bezug auf Kompetenzbereiche, auf Aufgabenformate und allgemeine Fachvernachlässigung untersucht.⁴² Dabei zeigen sich bei Betrachtung der Querschnittsstichprobe sowie der Längsschnittstichprobe zu zwei Messzeitpunkten (T_1 : 2005, T_2 : 2009) folgende Ergebnisse (alle durchgeführten t -Tests sind signifikant auf .001-Niveau; vgl. Tabelle 9): Im Laufe der Zeit kommt es zu einer signifikanten Abnahme der wahrgenommenen Einschränkungen durch die Vergleichsarbeiten wie auch der Orientierungsfunktion. Die signifikant verringert wahrgenommenen Übungserfordernisse können als Beleg dafür dienen, dass es nicht (mehr) zu einem ausgeprägten *teaching-to-the-test* kommt, durch das Unterrichtszeit verloren geht (Wacker & Kramer, 2012, S. 690). Zusammenfassend lässt sich also durch diese Untersuchung sagen, dass negative Auswirkungen auf den Unterricht, wie beispielsweise *teaching-to-the-test* oder *narrowing-the-curriculum* bei den Vergleichsarbeiten als *low-stakes*-Tests kaum vorkommen, sondern eher bei vor allem im anglo-amerikanischen Sprachraum verbreiteten *high-stakes*-Testungen (Jerald, 2006; Popham, 2001; Wacker & Kramer, 2012; Wall, 2000).

⁴² Erstgenannter Aspekt wurde mit drei Items gemessen (Cronbachs $\alpha_1 = .77$, $\alpha_2 = .80$). Die Übungserfordernisse sowie die drei Perspektiven der Curriculumseinschränkung wurden mit Einzelitems erhoben:

Übungserfordernisse: „Weil sich die Vergleichsarbeiten auf Themen und Kompetenzen von zwei Schuljahren beziehen, muss ich mit meinen Klassen sehr viel mehr üben und wiederholen als bisher.“

NTC-Kompetenzbereiche: „Ich vermittele hauptsächlich jene Kompetenzbereiche, die in den Vergleichsarbeiten abgeprüft werden.“

NTC-Aufgabenformate: „Die Vergleichsarbeiten führen mich dazu, meinen Unterricht sehr stark an der Art ihrer Aufgabenstellungen zu orientieren.“

NTC-Fachvernachlässigung: „Die Vergleichsarbeiten führen zur Vernachlässigung derjenigen Fächer, die nicht abgeprüft werden.“ (Wacker & Kramer, 2012, S. 696)

Tabelle 9: t-Test für unabhängige Stichproben im Querschnitt sowie Wald-Test für abhängige Stichproben im Längsschnitt: T_1 vs. T_2 (Wacker & Kramer, 2012, 697ff.)

	Querschnitt						Längsschnitt						Cohens <i>d</i>			
	T1			T2			Cohens <i>d</i>	<i>t</i>	T1			T2				
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		<i>N</i>		
Orientierungsfunktion	652	3.71	1.06	435	2.42	1.17	18.43	1.16	3.74	0.99	2.44	1.13	1135	108.31	1.05	
NTC-Kompetenzbereiche	632	3.36	1.32	426	2.30	1.24	13.02	0.83	3.42	1.28	2.27	1.21	132	43.53	0.84	
NTC-Aufgabenformate	623	3.42	1.28	428	2.03	1.15	18.39	1.14	3.49	1.25	1.95	1.11	134	171.21	1.11	
NTC-Fachvernachlässigung	592	3.16	1.50	395	2.23	1.44	9.75	0.63	3.04	1.55	2.04	1.23	128	39.28	0.66	
Übungsfordernisse	619	3.81	1.23	408	2.55	1.742	41.55	0.95	3.77	1.22	2.46	1.32	109.43	109.43	0.93	

1.7 Fazit und Forschungsfragen

Die Vergleichsarbeiten sind ein Instrument, das in konkreter Form die im Vorfeld näher beschriebenen Aspekte vereint: Dem bundeslandinternen Verfahren liegen die Bildungsstandards zu Grunde und als Evaluationsverfahren sollen Informationen zur Schulentwicklung bereitgestellt werden. Im ersten Teil wurden die Bildungsstandards in ihrer Entstehung und ihrem Anspruch näher vorgestellt. Dabei fast schon selbstverständlich ist mittlerweile das dreidimensionale Kompetenzstrukturmodell mit dem Dimensionen *Leitideen*, *prozessbezogene Kompetenzen* und *Anforderungsbereiche* als Repräsentationen der Bereiche Inhalt, Tätigkeit und Anspruch. Mit der Einführung von Bildungsstandards wurde sichergestellt, dass anstelle von *Inputvorgaben* verbindliche *Outputerwartungen* als *Kompetenzanforderungen* (Weinert, 2001) formuliert sind. Damit dienen die Bildungsstandards sowohl der Orientierung, was Schülerinnen und Schüler zum Ende eines Bildungsabschnittes können sollen, als auch der Evaluation der Zielerreichung, um etwaig steuernd eingreifen zu können. Die Möglichkeiten der Bildungsstandards entfalten sich jedoch auch auf übergreifender Ebene für das deutsche Bildungssystem: Durch die verbindliche Einführung nationaler Bildungsstandards ist eine nationale Referenz angelegt worden, an der sich verschiedene Verfahren orientieren und so miteinander in Bezug gesetzt werden können (*Verzahnung* nach Rolff, 2001). Dieser gegenseitige Bezug wurde 2006 (Neuaufgabe von 2015) in der Gesamtstrategie zum Bildungsmonitoring erstmals auch schriftlich fixiert (KMK & IQB, 2006; KMK, 2015).

Das steuernde Eingreifen entspricht dem Verständnis von Evaluation: anhand von Daten Maßnahmen ableiten, um die Zielerreichung optimieren zu können. Im schulischen Kontext ist die Zielgröße das bestmögliche Lernen und Verstehen der Schülerinnen und Schüler, was vielfach durch Leistungsmaße operationalisiert ist (u. a. in (inter-)nationalen Vergleichsuntersuchungen oder in Abschlusstests). Wichtig ist, dass die erhobenen Daten in verständnisvoller Art und Weise genutzt werden und dabei die methodischen Einschränkungen berücksichtigt werden; so kann beispielsweise eine einmalige Messung keine kausalen Schlüsse nahelegen oder aus einer nicht repräsentativen Untersuchung können keine Maßnahmen für das gesamte System abgeleitet werden. Ein wesentlicher Schritt der Evaluation ist die Rückmeldung der Informationen (*Rezeption & Reflektion* in Abbildung 10 nach Helmke, 2004). Die Evaluierenden müssen berücksichtigen, dass wirksame Rückmeldung Informationen für die Nutzenden bereitstellen muss, die sich auf verschiedenen Ebenen bewegt und

nach Möglichkeit die drei Feedback-Fragen nach Hattie und Timperley (2007) beantwortet: (1) Where am I going? (What are the goals?) – *Feed Up*; (2) How am I going? (What progress is being made toward the goal?) – *Feed back*; (3) Where to next? (What activities need to be undertaken to make better progress?) – *Feed forward*.

Ziel des Feedbacks im Rahmen der Vergleichsarbeiten soll es sein, den Lehrkräften Lerngelegenheiten zu bieten, aus denen Schulentwicklung als Summe gezielter Unterrichts-, Personal- oder Organisationsentwicklung resultiert. Aus den vorhandenen Rückmeldungen bzw. gestützt auf Daten werden nach dem Wesen der Evaluation Maßnahmen abgeleitet, die in der Lage erscheinen, die ausgemachten Defizite bzw. Probleme auf dem Weg zur Zielerreichung zu beheben: Es können im Sinne der Organisationsentwicklung also Maßnahmen resultieren, die strukturelle Veränderungen der einzelnen Schule mit sich bringen wie etwa die Einrichtung von Jahrgangsteams, Teamteaching o. ä. Den Fokus auf die Lehrkraft legend und damit auch auf die Personalentwicklung wurde von COACTIV ein umfassendes Modell *professioneller Kompetenz* von Lehrkräften vorgestellt, das neben dem Wissen von Lehrkräften in verschiedenen Bereichen (u. a. Fachwissen und fachdidaktisches Wissen) deren Motivation, Beliefs und Selbstregulation in den Blick nimmt. Eine persönliche Weiterbildung in diesen Bereichen wäre dementsprechend die Variante der Personalentwicklung. Der letzte Bereich der Schulentwicklung betrifft die Entwicklung von Unterricht: Durch die Anpassung von Methoden und Inhalten wird der Kernbereich der Lehrkräfte mit aktuellen Forschungsergebnissen zur Unterrichtsqualität fortwährend optimiert. Ziel ist dabei immer der bestmögliche Lerngewinn bei den Schülerinnen und Schülern. Einen an aktuellen Forschungsergebnissen ausgerichteten Unterricht zu planen und durchzuführen, bedarf jedoch eben damit der persönlichen Weiterentwicklung/-qualifikation, für die im organisationalen Rahmen entsprechende Möglichkeiten geschaffen werden müssen. Eine isolierte Entwicklung in nur einem der drei Bereiche Unterricht, Personal oder Organisation ist also weder denkbar noch wünschenswert, sondern von jedem Aspekt kann eine Entwicklung beginnen und auf die anderen übergehen.

Die Vergleichsarbeiten stellen ein Verfahren dar, das auf diesen Ausführungen aufbaut und letztlich die Schulentwicklung veranlassen will und dabei die Bildungsstandards als Grundlage hat. Damit erfüllen die Vergleichsarbeiten nicht nur die Ansprüche der *Evaluation* als durchführbare, faire und genaue Verfahren (DeGEval, 2002),

sondern genügen darüber hinaus den vielen Ansprüchen, die eine erfolgreiche Nutzung bedingen. Beispielsweise liefern die Vergleichsarbeiten mit dem Bezug zu den Bildungsstandards ein Testinstrument, das die gleiche Basis hat wie der Unterricht und insofern direkt anschlussfähig an diesen ist. Ebenso handelt es sich um ein Verfahren, welches auf hochkomplexe und zeitaufwendige Analysen verzichtet, sodass die Lehrkräfte verständliche und unmittelbare/zeitnahe Rückmeldungen erhalten, aus denen Maßnahmen abgeleitet werden können. Grundsätzlich liegt mit den Vergleichsarbeiten also ein vermeintlich sehr potentes Instrument vor, um Schulentwicklung anzustoßen. Allerdings zeigen Untersuchungen z. T. nur indifferente Haltungen bzgl. der *Akzeptanz* des Verfahrens oder zeigen nur bedingt die Entwicklung von Maßnahmen. Für eine genauere Betrachtung sind die folgenden Forschungsfragen als interessant und untersuchungsleitend anzusehen:

1. Vorhandene Untersuchungen stützen sich vielfach auf die VerA im Jahrgang 3 und/oder Vorläuferversionen der national einheitlichen Arbeiten. Insbesondere für die föderal bedingt leicht variierenden Formen der Vergleichsarbeiten, fehlen Forschungsergebnisse für Hessen. Die erste Forschungsfrage zielt darauf, diese Lücken mit der vorliegenden Untersuchung für Hessen und den Jahrgang 8 zu schließen.
2. Zentral ist dabei die Frage nach dem Ist-Stand der Vergleichsarbeiten in Hessen (Lernstandserhebungen) in Bezug auf *Akzeptanz*, *Nützlichkeit*, *Nutzung* etc. Im Rahmen von Personalentwicklung soll es gelingen die bei Lehrkräften verfügbaren Kompetenzen zu verändern. Trotz aller zusätzlichen Informationen können die Vergleichsarbeiten zu einer zusätzlichen Belastung werden und die Nutzung für die Lehrkräfte ohne entsprechende Unterstützung unklar bleiben bzw. im Alltagsgeschäft Unterricht untergehen. Daher stellt sich die Forschungsfrage, inwieweit eine externe Unterstützung Lehrkräften dabei helfen kann, die Vergleichsarbeiten in der intendierten Art zu nutzen und daraus weniger Belastung zu fahren.
3. Die dritte Forschungsfrage ist die Frage nach Heterogenität. Bereits im Theorieeil wurden vielfach zunächst vergleichsweise einfache und direkte Ursache-Wirkungs-Gefüge beschrieben. Allerdings geben bestimmte Ergebnisse den Hinweis (Groß-Ophoff et al., 2007; A. Hosenfeld, 2010), dass sich verschiedene Typen von Lehrkräften ausmachen lassen, die die Vergleichsarbeiten in unterschiedlicher Weise nutzen. Daraus leitet sich die Frage ab, ob sich für

Hessen ebenfalls unterschiedliche Typen von Lehrkräften klassifizieren lassen, die sich in den Einstellungen zu den Vergleichsarbeiten sowie dem Nutzen der Vergleichsarbeiten unterscheiden.

2 Projektbeschreibung: Verbesserung der Effektivität der Lernstandserhebungen Mathematik Klasse 8 (VELM-8)

2.1 Projektskizze

Das Projekt VELM-8 (Verbesserung der Effektivität der Lernstandserhebungen Mathematik Klasse 8) wurde im Schuljahr 2012/2013 unter der Leitung von Prof. Dr. Werner Blum⁴³ durchgeführt. Die Fortbildungsreihe richtete sich an interessierte Lehrkräfte aus dem regionalen Umfeld der Universität Kassel. Grundsätzliches Ziel war es, exemplarisch zu zeigen, wie Lehrkräfte in einer schuljahrbegleitende Fortbildungsreihe Informationen über die Lernstandserhebungen und deren Entwicklung erhalten sowie bei deren Nutzung bestmöglich unterstützt werden können. VELM-8 ist daher eher als Unterrichtsentwicklungsprojekt zu verstehen denn als ausgewiesenes Forschungsprojekt. Darüber hinaus war geplant, aus den Erkenntnissen aus VELM-8 eine entsprechende Fortbildungsreihe für Unterrichtsentwicklungsberaterinnen und -berater in ganz Hessen zu erstellen (ergänzende Informationen zum Projekt vgl. auch Vogel, Achmetli, Krawitz & Blum, 2013; Vogel, Blum, Achmetli & Krawitz, 2013; Vogel, Achmetli, Krawitz & Blum, 2014; 2016).

Nachfolgend sollen Zielsetzung des Projektes und anschließend die einzelnen Fortbildungen im Detail betrachtet werden.

2.2 Ziele der Fortbildungsreihe

Das übergeordnete Ziel war es, exemplarisch zu zeigen, wie die Lernstandserhebungen bestmöglich genutzt werden können. Dabei ging es neben dem Aufzeigen von Handlungsoptionen und Präsentieren von vorhandenem Material insbesondere um die Beeinflussung der Einstellungen der Lehrkräfte gegenüber dem Instrument *Lernstandserhebungen*. Die inhaltlichen Schwerpunkte der Fortbildungen lassen sich aus den behandelten Themen des Theorieteils bereits gut ableiten. Mit Hilfe von Inputvorträgen und zugehörigen Arbeits- und Reflexionsphasen sollten den Lehrkräften Lerngelegenheiten zur Erweiterung des persönlichen Fachwissen sowie

⁴³ In der beteiligten Arbeitsgruppe haben am Projekt mitgearbeitet: Kay Achmetli, Michael Besser, Christina Drücke-Noe, Janina Krawitz, André Krug und Sebastian Vogel.

fachdidaktischen und pädagogisch-psychologischen Wissens geboten werden, die sich thematisch auf folgende Themen bezogen:

- Bildungsstandards als Grundlage der Lernstandserhebungen und von Mathematikunterricht
- Leistungsmessung und Rückmeldung/Feedback
- Unterrichtsqualität in verschiedenen Operationalisierungen
- Konzeption, Funktionen und Rückmeldungen der Lernstandserhebungen

Neben der Erweiterung des Wissens lag dabei auch der Fokus darauf, die kooperativ erzeugten Arbeitsergebnisse in jeder Veranstaltung kritisch zu reflektieren, um so eine fehlertolerante Evaluationskultur (in kleinem Stile) zu schaffen und gleichzeitig die Vorteile einer kooperativen Arbeit hervorzuheben. Insbesondere die abschließende Reflexion jeder Veranstaltung diene im Sinne eines *pädagogischen Doppeldeckers* der meta-kognitiven Aktivierung bei gleichzeitiger Verknüpfung der Arbeitsergebnisse mit den Inputinhalten (Wahl, 2013). Durch diese angestrebten Veränderungen in der *professionellen Kompetenz* der Lehrkräfte wurde erwartet, dass sich die Unterrichtspraxis (u. a. durch Thematisierung von Unterrichtsqualität) positiv verändert und resultierend nach Projektende eine erhöhte Leistung der Schülerinnen und Schüler zu verzeichnen ist.

Im Konkreten lassen sich die einzelnen Fortbildungsinhalte⁴⁴ wie folgt darstellen.

2.3 Fortbildungsbeschreibung

2.3.1 Informationsveranstaltung – 30.08.2012

In einer ersten Auftaktveranstaltung ging es vor dem eigentlichen Projektbeginn darum, interessierten Lehrkräften einen Überblick über die Projektdurchführung und die Projektziele zu geben. In einer anschließenden Arbeitsphase wurden Lernstandserhebungsaufgaben analysiert und diskutiert. Dieses Zusammenspiel aus Input- und Arbeitsphase mit abrundender Diskussion sollte den interessierten Lehrkräften als Vorlage für die Gestaltung der vier Fortbildungen dienen. Im Anschluss an diese

⁴⁴ Die Dauer jeder Veranstaltung lag bei 3 Stunden, wobei die Fortbildung 3 wegen der Arbeit am konkreten Material auf 4 Stunden verlängert wurde.

Veranstaltung konnten sich Fachkollegien um die Teilnahme bewerben (vgl. auch Stichprobenbeschreibung; Abschnitt 3.3).

2.3.2 Fortbildung 1 – 11.09.2012

In der ersten inhaltlichen Fortbildung standen die Bildungsstandards im Fokus. Diese stellen die Grundlage der Lernstandserhebungen dar. Um im Sinne effektiver Fortbildungen an das Wissen der Lehrkräfte anzuknüpfen, wurde als Grundlage für die weiteren Sitzungen exemplarisch anhand der Kompetenz „Probleme mathematisch lösen“ die Idee der Bildungsstandards vorgestellt. Hierzu ging es zunächst um die Theorie des Problemlösens mit einem Vergleich verschiedener Definitionen (u. a. Bruder & Collet, 2011; KMK, 2004a; Leiß & Blum, 2012). Insbesondere wurde der Problemlöseprozess als Bewältigen eines zu Beginn nicht klar erkennbaren Lösungsprozesses betont (vgl. Polya, 1967), welcher sowohl für die Analyse von Problemlösesituationen als auch heuristisch für deren Bewältigung genutzt werden kann.

Um dieses theoretische Wissen in einen schul- und unterrichtspraktischen Kontext zu stellen, wurde die Verknüpfung zu diagnostischen Situationen angestrebt: Anhand von exemplarischen, bildungsstandardorientierten Aufgaben (u. a. aus Blum et al., 2012) erläuterte das Fortbildungs-Team die heuristischen Hilfsmittel Tabelle, Gleichung und informative Figur zur Unterstützung des Lösungsprozesses. Als heuristische Strategien wurden Vorwärts- und Rückwärtsarbeiten, das Invarianz-, Analogie-, Extremal-, Zerlegungs- und Ergänzungsprinzip illustriert. Diese Informationen zur Erweiterung der professionellen Kompetenz der Lehrkräfte erfolgte in Form von zwei Inputs mit kurzen Übungs- und Aktivierungsphasen. Im dritten Teil dieser Fortbildung ging es konkret um die Kompetenz „Probleme mathematisch lösen“ in den Lernstandserhebungen Mathematik Klasse 8. Hierzu wurden zwei exemplarische Aufgaben aus den Lernstandserhebungen von den Lehrkräften in einer längeren Übungsphase bearbeitet, analysiert und mit Blick auf den erwartbaren Einsatz heuristischer Hilfsmittel und Strategien seitens der Schülerinnen und Schüler in Partnerarbeit besprochen.

Um in den weiteren Fortbildungen an den Unterrichtssituationen der Lehrkräfte ansetzen zu können und so eine inhaltliche Relevanz zu schaffen, hatten die Lehrkräfte als Aufgabe, bis zur zweiten Fortbildung weitere Problemlöse-Aufgaben aus den Lernstandserhebungen in ihren Lerngruppen des Jahrgangs 8 einzusetzen, sodass

nachfolgend je Lehrkraft ein eigener Klassensatz an Bearbeitungen von Schülerinnen und Schülern vorlag.

2.3.3 Fortbildung 2 – 10.01.2013

In der zweiten Fortbildung standen mit den Themen Feedback/Rückmeldung und Unterrichtsqualität zwei inhaltliche Aspekte auf der Agenda. Zunächst losgelöst von den Lernstandserhebungen wurden in einer ersten Input-Phase den Lehrkräften essentielle Grundlagen von Leistungsmessung, -bewertung und -rückmeldung vermittelt (u. a. beziehend auf Hattie & Timperley, 2007). Zu Feedback und Rückmeldung von Leistungswerten wurde sich für den Mathematikunterricht u. a. auf das Projekt *Co²CA* (**C**onditions and **C**onsequences of **C**lassroom **A**ssessment; Better, Blum & Klimczak, 2013; Bürgermeister et al., 2011; Harks, Rakoczy, Hattie, Better & Klieme, 2013) bezogen. Vor allem dienten die Diagnosebögen des Projekts als Heuristik. Diese Diagnosebögen nutzen bei der Rückmeldung an Schülerinnen und Schüler den Dreischritt, der eng an die drei „Feedback-Fragen“ (Hattie & Timperley, 2007) angelehnt ist:

1. Mit folgenden Inhalten kannst du bereits gut umgehen.
2. Bei folgenden Inhalten kannst du dich noch verbessern, wenn du meine Tipps beachtest.
3. Tipps, wie du dich verbessern kannst.

In einer anschließenden Aktivierungsphase sollten die Lehrkräfte diesen Dreischritt bei den von ihnen selbst erzeugten Klassensätzen nachvollziehen und exemplarisch die Rückmeldung für einzelne Schülerinnen und Schüler notieren. Im Anschluss tauschten sich die Lehrkräfte gemeinsam mit den Referierenden über diese Rückmeldungen aus.

Während der erste Teil dieser Fortbildungsveranstaltung eher den Fokus auf eine individuelle Rückmeldung und tendenziell eher den Blick auf eine Nutzung von Lernstandserhebungen bzw. allgemein von Aufgaben zur individuellen Förderung legte, wurde im zweiten Teil gezielt die Nutzung der Lernstandserhebung bzw. einzelner Aufgaben daraus für die Unterrichtsentwicklung angesprochen. Zu Beginn dieser Phase wurde mit einem Input-Teil sichergestellt, dass alle Lehrkräfte über einen ähnlichen Wissensstand zu Unterrichtsqualitätskriterien verfügen. Entsprechend den Ausführungen vorne (Abschnitt 1.5.4) orientierten sich die Ausführungen der

Fortbildenden an den mathematikspezifischen Projekten SINUS, DISUM, COACTIV, TEDS-M, aber auch dem Hessischen Referenzrahmen Schulqualität (BLK, 1997; Blum & Leiß, 2003, 2007a; Diel, 2011; Hessisches Kultusministerium & IQ, 2011; Kunter et al., 2011; Prenzel, 2001). Vorgestellt und exemplarisch mit Unterrichtssequenzen erläutert wurden folgende Unterrichtsqualitätskriterien:

1. Effektive und schülerorientierte Unterrichtsführung⁴⁵
2. Kognitive Aktivierung der Lernenden
3. Metakognitive Aktivierung der Lernenden
4. Fachlich gehaltvolle Unterrichtsgestaltung

In einer anschließenden Präsentation zeigte eine erfahrene Lehrkraft, wie der Einsatz einer Lernstands-erhebungsaufgabe und die aus der Auswertung gezogenen diagnostischen Schlüsse für die Unterrichtsgestaltung genutzt werden können. Diese Stunde wurde im Plenum vor dem Hintergrund der Unterrichtsqualitätskriterien reflektiert. Den Abschluss dieser Fortbildungsveranstaltung stellte eine Praxisphase in Kleingruppen dar, in der die Lehrkräfte gemeinsam aus eigenen Bearbeitungen der Schülerinnen und Schüler diagnostische Informationen ableiten und darauf bezugnehmende Stunden planen, diese vorstellen und im Plenum kritisch reflektieren sollten.

Die zweite Fortbildung lag vor der eigentlichen Durchführung der Lernstandserhebungen. So konnte die Nutzung der Lernstandserhebungen exemplarisch als „Trockenübung“ erfolgen. Zwischen der zweiten und der dritten Fortbildung wurden die Lernstandserhebungen tatsächlich durchgeführt. Alle Lehrkräfte, die am Projekt teilnahmen, haben die Lernstandserhebungen in ihren Lerngruppen durchgeführt, sodass nachfolgend am konkreten Material (sowohl Aufgaben als auch deren Bearbeitungen und der didaktischen Handreichungen zu den Aufgaben) gearbeitet werden konnte.

2.3.4 Fortbildung 3 – 13.03.2013

Als informativen Input startete diese Fortbildung mit einem Exkurs zum Kompetenzstufenmodell, den verschiedenen Rückmeldeformaten und deren

⁴⁵ Da es sich bei den Unterrichtsqualitätskriterien „Effektive Klassenführung“ und „Schülerorientierte Unterrichtsführung“ (vgl. Abschnitt 1.5.4.2) eher um allgemeine und nicht fachspezifische Kriterien handelt, wurden diese als didaktische Reduzierung für die Lehrkräfte zu einem Punkt zusammengefasst.

Interpretationsmöglichkeiten. In einer ersten Arbeitsphase wurden nun die *didaktischen Kommentare* näher analysiert. Viele der in den ersten beiden Fortbildungen selbst erarbeiteten Schritte (Aufgabenanalysen, Identifizierung möglicher Schülerfehler, Anregungen zur unterrichtlichen Weiterarbeit etc.) wurden den Lehrkräften hierin aufbereitet zur Verfügung gestellt.

Im Vorfeld der Fortbildung konnten die Lehrkräfte Aufgaben mitteilen, zu denen sie gerne im weiteren Verlauf arbeiten würden. Zu drei dieser Aufgaben (je eine pro Bildungsgang/Testheft) wurden in der Hauptarbeitsphase in Kleingruppen aufbauend auf den Ergebnissen der Lernstandserhebungen und den konkreten Bearbeitungen der einzelnen Schülerinnen und Schüler Unterrichtsstunden erstellt. Hierbei sollten wiederum die Unterrichtsqualitätskriterien der zweiten Fortbildung genutzt werden und die Stunden nach Planung kritisch reflektiert werden⁴⁶.

2.3.5 Fortbildung 4 – 07.05.2013

Die abschließende vierte Fortbildung erweiterte den Blick auf die Lernstandserhebung um die kollegiale Ebene. Bisher standen Auswertungen im Fokus, die v. a. für einzelne Lehrkräfte mit dem Blick auf die eigene Lerngruppe und/oder einzelne Schülerinnen und Schüler Relevanz haben. Die Adressaten von Lernstandserhebungen sind jedoch neben den Einzellehrkräften auch Fachkollegien, sodass auf einer hierarchisch höheren schulischen Ebene eine Nutzung ebenfalls möglich ist. Hierzu wurde auf *best-practice*-Beispiele zurückgegriffen, die im Vorfeld mit der Projektleitung abgestimmt wurden. Nach diesen Beispielen hatten die Lehrkräfte Gelegenheit, anhand von vorläufigen korr. Landesmittelwerten (diese wurden der Projektleitung vom Land Hessen zur Verfügung gestellt, da sie zu diesem Zeitpunkt offiziell noch nicht verfügbar waren) im Team der teilnehmenden Fachkolleginnen und -kollegen die Ergebnisse zu analysieren und dabei auf die Unterstützung des Fortbildungsteams zurückzugreifen.

⁴⁶ Die entwickelten Unterrichtsstunden wurden mit dem Einverständnis der Lehrkräfte sowie der Schülerinnen und Schüler videografiert und im Rahmen von Examensarbeiten z. T. ausgewertet; diese Auswertungen wurden jedoch nicht zusammenhängend oder systematisch koordiniert bzw. aggregiert.

2.3.6 Projektausblick

Nachdem VELM-8 nach Einschätzung von Projektleitung und Lehrkräften erfolgreich durchgeführt⁴⁷ und von den teilnehmenden Lehrkräften als gute Erfahrung mit interessanten Fortbildungsinhalten bewertet wurde, schloss sich die Entwicklung einer dreiteiligen Fortbildungsreihe in Kooperation mit dem Hessischen Kultusministerium und der Lehrkräfteakademie für das folgende Schuljahr (2013/2014) an, die (zu großen Teilen fachspezifisch⁴⁸) die Unterrichtsentwicklungsberaterinnen und -berater (UEB) zum selben Thema fortbilden sollte. Ziel war es, den UEBs das notwendige Wissen zu vermitteln, über Lernstandserhebungen dezentral in den ihnen zugeordneten Schulamtsbezirken Fortbildungen anzubieten. So sollte eine von VELM-8 aufgrund des Modellcharakters nicht realisierbare Breitenwirkung erreicht werden.

⁴⁷ Persönliche Rückmeldungen der teilnehmenden Lehrkräfte ohne statistisch belastbare Ergebnisse (vgl. hierfür Teil 4).

⁴⁸ Das Hessische Kultusministerium und die Lehrkräfteakademie übernahm die Fortbildungen für die UEBs der Fächer Deutsch und Englisch/Französisch.

3 Methode

3.1 Hypothesen

Mit der Darlegung der Fortbildungsreihe VELM-8 bereits im Vorfeld der nun nachfolgend formulierten Hypothesen und dem Methodenteil wird die typische Struktur empirischer Arbeiten bewusst durchbrochen. Hierdurch soll verdeutlicht werden, dass die inhaltliche Gestaltung der Fortbildungsreihe weniger Mittel zum Zweck als vielmehr Hauptanliegen des Projekts war. Entsprechend resultieren zunächst vor allem projektevaluationsbezogene Hypothesen in dem Sinne, dass die Wirkung der Fortbildungen auf die Einstellungen der Lehrkräfte geprüft werden soll (Forschungsfrage 2). Dabei bewegen sich die Hypothesen quasi zwischen der ersten und zweiten Ebene von Fortbildungserfolg; es werden nicht die unmittelbaren Einschätzungen/Wahrnehmungen der Fortbildungsveranstaltungen an sich (Ebene 1) evaluiert, jedoch ebenfalls „noch“ nicht veränderte Kognitionen und Wissen (Ebene 2; vgl. 1.5.2.1). Die formulierten Projektziele beziehen sich hingegen vorrangig auf veränderte Kognitionen und verändertes Unterrichtshandeln (Ebene 3). Im Rahmen des Projektes wurde sich bewusst – trotz dieser Diskrepanz zwischen Zielen und Hypothesen/Evaluationsmethoden – für dieses Vorgehen entschieden, um den Charakter von VELM-8 als *Fortbildungsprojekt* zu unterstreichen. Die Erhebung von Einstellungen ist um ein Vielfaches einfacher zugänglich als die Erhebung von Wissen oder gar unterrichtlichem Handeln. Mit dem vorliegenden Projekt VELM-8 liegt daher zweifelsfrei kein empirisch und forschungsmethodisch einwandfreies Unterfangen vor. In Anbetracht der vergleichsweise geringen Fortbildungsdauer (max. 4stündige Fortbildungsveranstaltungen) und der extrem langen Laufzeit (ein knappes Schuljahr) mit entsprechend resultierenden langen Phasen zwischen den einzelnen Sitzungen, ohne eine dezierte Begleitung durch die Fortbildenden, erscheinen vermutlich einige der Hypothesen vor dem Hintergrund des berichteten Forschungsstands eher optimistisch. Allerdings sind Veränderungen im Zusammenhang mit der Fortbildungsreihe VELM-8 insofern erwartbar und nicht zwingend unrealistisch, da die Konzeption der Fortbildung an den Merkmalen erfolgreicher Fortbildungen orientiert war. Einschränkend an dieser Stelle sei bereits angeführt, dass die Daten weitestgehend als Selbstauskünfte der Lehrkräfte im Fragebogen zu verstehen sind und insbesondere keine Unterrichtsbeobachtungen einschließen (vgl. Ebene 3 des Fortbildungserfolges).

3.1.1 VELM-8

3.1.1.1 Projektevaluation

3.1.1.1.1 Einstellungen

Im Speziellen als direkte Wirkungen aus der Fortbildungsreihe VELM-8 werden Effekte in verschiedenen Dimensionen erwartet: Wie theoretisch dargelegt, stellt die *Akzeptanz gegenüber den Lernstandserhebungen* wegen der dezentralen Durchführung und Auswertung eine grundlegende Bedingung dar. Die Akzeptanz lässt sich dabei durch die Informiertheit der Lehrkräfte positiv beeinflussen: Im Rahmen der Fortbildungssitzungen wurden den Lehrkräften der Projektgruppe umfassende Informationen zu den Lernstandserhebungen angeboten, z. B. die Bildungsstandards als Grundlage der Aufgaben, der Ablauf der Lernstandserhebungen oder die verfügbaren Rückmeldungen. So ist davon auszugehen, dass sich die Einschätzung der Akzeptanz im Rahmen von VELM-8 in der Projektgruppe positiver entwickelt als in der Kontrollgruppe (Hypothese 1). Neben der unmittelbaren Akzeptanz spielt die Einschätzung der *Lehrplanvalidität der Lernstandserhebungen* ebenfalls eine entscheidende Rolle, damit Lehrkräfte den direkten Bezug zwischen Unterricht bzw. Curriculum/Lehrplan und Lernstandserhebungen erkennen. Wesentlicher Bestandteil der Fortbildungen, explizit in der ersten und implizit in den weiteren, waren der Bezug der Lernstandserhebungen zu den Bildungsstandards sowie die Umsetzung von Bildungsstandards durch kompetenzorientierten Unterricht im Alltag. Damit ist auch hier zu erwarten, dass Lehrkräfte nach der Teilnahme an der Fortbildungsreihe eine positive Entwicklung in der Einschätzung der Lehrplanvalidität zeigen (Hypothese 2). Ein weiterer wesentlicher Einflussfaktor für die generelle Akzeptanz ist die wahrgenommene *Belastung durch die Lernstandserhebungen*. Vielfach wird den Lernstandserhebungen vorgeworfen, lediglich „Zusatz“ zu sein und Lehrkräfte zu belasten. Aus der Forschung von Schaarschmidt (2005, 2007) lässt sich die belastungsreduzierende Wirkung von Fortbildungen ableiten. Ebenfalls weniger Belastung wird durch die Lehrkräfte wahrgenommen, wenn mehr Wissen über die Nutzung eines Instrumentes vorhanden ist (Heymann, 2005). Innerhalb der Fortbildungen wurden den Lehrkräften theoretisch sowie exemplarisch vielfältige Nutzungsmöglichkeiten aufgezeigt, die verfügbaren Rückmeldungen und didaktischen Materialien als die Unterrichtsplanung erleichternde Elemente vorgestellt sowie durch Gruppenarbeit wie auch durch Rückmeldung der Fortbildenden bei der Interpretation von Schülerbearbeitungen als Grundlage von weiterer Unterrichtsplanung unterstützt. Folglich sollte die (zusätzliche)

Belastung durch die Lernstandserhebung von Projektgruppen-Lehrkräften geringer eingeschätzt werden als von „untrainierten“ Lehrkräften (Hypothese 3).

Neben diesen grundsätzlichen Einschätzungen der Lernstandserhebungen sollten sich jedoch nach Betrachtung der Fortbildungsgestaltung und -schwerpunkte ebenfalls Veränderungen in Bezug auf eine individuelle sowie gruppenspezifische Förderung ergeben. Im Projekt wurde wiederholt betont, dass die Lernstandserhebungen sowohl *individualdiagnostisch* (Hypothese 4) genutzt werden können als auch Informationen über die gesamte Lerngruppe mit Blick auf eine angepasste *Unterrichtsplanning* (Hypothese 5) liefern können. Entsprechend sollten die Aufgaben sowie die Rückmeldungen von Lehrkräften dazu genutzt werden, Schülerinnen und Schüler individuell bestmöglich zu fördern (Fortbildungen 1 und 2) und gleichzeitig Schlüsse zuzulassen, welche Anpassungen im nachfolgenden Unterricht sinnvoll erscheinen (Fortbildungen 2 und 3). Diese beiden Nutzungsmöglichkeiten – Nutzung eher auf Individualebene und zusätzlich auf Ebene der gesamten Lerngruppe bzw. Klasse – lassen sich als intendierte Nutzungsarten beschreiben und sollten sich durch die Fortbildungsteilnahme positiv entwickeln (Visscher & Coe, 2003).

Ebenfalls denkbar wäre eine Nutzung der Lernstandserhebungen zu *selektionsdiagnostischen* Zwecken, etwa Benotung oder Versetzung. Diese Nutzungsvariante ist (in Hessen) jedoch nicht intendiert. Daher ist zu erwarten, dass diese Nutzungsvariante bei hessischen Lehrkräften zwar auftritt; durch die Projektfortbildungen, in denen eine solche Nutzung nicht angeregt, sondern eher abgelehnt wurde, sollte sich die Einschätzung der Lehrkräfte allerdings dahingehend verändern, dass Lernstandserhebungen im Projektverlauf weniger für selektionsdiagnostische Entscheidungen herangezogen werden (Hypothese 6).

Teil 3 – Methode

Hypothesen

Zusammenfassend lassen sich also zunächst sechs Hypothesen aufstellen, die direkt die Lernstandserhebungen betreffen:

Tabelle 10: Übersicht der Hypothesen zu Einstellungen und Nutzen der Lernstandserhebungen

Hypothese 1:	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, eine höhere <i>Akzeptanz gegenüber den Lernstandserhebungen</i> zu haben als Lehrkräfte der Kontrollgruppe.
Hypothese 2:	Im Verlauf der Projektteilnahme schätzen Lehrkräfte der Projektgruppe die wahrgenommene <i>Lehrplanvalidität der Lernstandserhebungen</i> positiver ein als Lehrkräfte der Kontrollgruppe.
Hypothese 3:	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe eine geringere wahrgenommene <i>Belastung durch die Lernstandserhebungen</i> an als Lehrkräfte der Kontrollgruppe.
Hypothese 4:	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des <i>förderdiagnostischen Nutzens der Lernstandserhebungen</i> für die Lehrkräfte der Projektgruppe positiver als für die Kontrollgruppe.
Hypothese 5:	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des <i>unterrichtsplanerischen Nutzens der Lernstandserhebungen</i> für die Lehrkräfte der Projektgruppe positiver als für die Kontrollgruppe.
Hypothese 6:	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des <i>selektionsdiagnostischen Nutzens der Lernstandserhebungen</i> für die Lehrkräfte der Projektgruppe negativer als für die Kontrollgruppe.

3.1.1.1.2 Unterrichtsqualität

Im Rahmen der theoretischen Ausführungen wurde deutlich gemacht, dass Unterrichtsqualität einen wesentlichen Einfluss auf die Lernleistung von Schülerinnen und Schülern hat. Diese enorme Relevanz eines ungestörten (Stichwort: Klassenführung), kooperativen (Stichwort: Unterstützendes Klassenklima), aktivierenden (Stichwort: kognitive Aktivierung), Reflexionen anregenden (Stichwort: Metakognitive Aktivierung) und inhaltlich anspruchsvollen (Stichwort: Fachlicher Gehalt) Unterrichts wurde als wesentlicher Baustein in die Fortbildungen unter Berufung auf die im Theorieteil vorgestellten Qualitätskriterien und ausgewählte Studienergebnisse wiederholt eingebaut. Insbesondere der Aspekt der kognitiven Aktivierung spielte in den Fortbildungen eine wesentliche Rolle. Wie dargelegt lässt sich kognitive Aktivierung auf verschiedene Arten operationalisieren: Ausgehend von der exemplarischen Einführung der Bildungsstandards am Beispiel der prozessbezogenen Kompetenz „Probleme mathematisch lösen“ (K2) sowie dem aufbauend auf den Schülerlösungen ermittelten Schwerpunkt „Mathematisch argumentieren“ (prozessbezogenen Kompetenz K1) ist zu erwarten, dass Projektgruppenlehrkräfte nach Fortbildungsteilnahme angeben, in Ihrem Unterricht mehr *Begründungen und Erklärungen* einzufordern als Lehrkräfte der Kontrollgruppe (Hypothese 7). Wichtig erscheint es im Sinne der kognitiven Aktivierung sowie möglicher Überforderungen (durch erhöhte Aufgabenschwierigkeit vgl. *cognitive-load-theory*), jede Schülerin und jeden Schüler auf dem

aktuellen Leistungsstand anzusprechen und damit notwendigerweise in der Aufgabenschwierigkeit Variationen zu ermöglichen (vgl. z. B. die Blütenaufgaben bei Bruder, 2003). Die im Laufe der Fortbildungen wiederholt hervorgehobenen *didaktischen Kommentare* bieten solche Variationsmöglichkeiten für die eingesetzten Aufgaben an. Entsprechend wurde in der Fortbildung die Bedeutung von Variation sowie die variationsanregenden *didaktischen Kommentare* betont. Hieraus folgend sollte die selbsteingeschätzte Häufigkeit des Einsatzes von Aufgaben, die in der Schwierigkeit variieren, bei Projektgruppenlehrkräften höher ausfallen als bei Kolleginnen und Kollegen aus der Kontrollgruppe (Hypothese 8).

Insbesondere im Fach Mathematik ist die besondere Bedeutung von Aufgaben vielfach berichtet worden (vgl. u. a. Bromme, Seeger & Steinbring, 1990; Christiansen & Walther, 1986). Deutlich wird dies auch durch die den Bildungsstandards beiliegenden, illustrierenden Aufgaben sowie den illustrierenden Büchern (Blum et al., 2012; Blum et al., 2015; Walther et al., 2012). In den Lernstandserhebungen werden die Kompetenzanforderungen der Bildungsstandards ebenfalls durch Aufgaben operationalisiert. Genau diese Aufgaben sollten im weiteren Verlauf des Schuljahres wiederum eine Nutzung im Unterricht erfahren bzw. Grundlage für Weiterentwicklungen im Unterricht sein. Der Zusammenhang zwischen Aufgaben und kognitiver Aktivierung im Mathematikunterricht wurden innerhalb der Fortbildungen mehrfach angesprochen und in den Arbeitsphasen wiederum kritisch reflektiert. Daher sollten Lehrkräfte die Bedeutung von Aufgaben im bildungsstandardbezogenen Unterricht zunehmend erkennen. Daraus sollte resultieren, dass Projektgruppenlehrkräfte im Projektverlauf angeben, häufiger *Aufgaben kognitiv aktivierend* einzusetzen als die Lehrkräfte der Kontrollgruppe (Hypothese 9). Neben dem reinen Einsatz ist der konkrete Umgang mit Aufgaben im Mathematikunterricht ebenfalls relevant. Unter anderem von Leuders (2009, 2012, 2015) wird in diesem Zusammenhang von „Intelligentem Üben“ gesprochen, welches nicht eine sture Reproduktion und wiederholte Anwendung von Rechenverfahren darstellt, sondern beispielsweise eine kognitiv anspruchsvolle Suche nach Mustern in verschiedenen Aufgaben. Ein solches „intelligentes“ Üben als Merkmal guten Mathematikunterrichts an der Schnittstelle der Qualitätskriterien *kognitive Aktivierung* und *fachlicher Gehalt* wurde in den Fortbildungen (insbesondere bei der Planung von Unterrichtsstunden und deren Reflexion) vielfach betont. Folglich wäre ein vermehrter Einsatz „intelligenten“, *kognitiv aktivierenden Übens* im Unterricht in der Projektgruppe gegenüber der Kontrollgruppe erwartbar

Teil 3 – Methode

Hypothesen

(Hypothese 10). Umgedreht formuliert und als negative Indikatoren dieses Merkmals wird erwartet, dass ein wiederholendes, *repetitives Üben* nach Projektverlauf durch die Lehrkräfte der Projektgruppe seltener angegeben wird als von den Lehrkräften der Kontrollgruppe (Hypothese 11). Ebenfalls als negativer Indikator kognitiver Aktivierung ist ein enggeführter Unterricht zu sehen, in dem Schülerinnen und Schüler keine Impulse einbringen können, sondern eher rezeptiv den fehlerfreien Ausführungen der Lehrkraft folgen, ihre Beiträge nicht als Ausgangspunkt weiterer Überlegungen genutzt und die Schülerinnen und Schüler kleinschrittig angeleitet bzw. kontrolliert werden: Wesentlicher Inhalt der Fortbildungsreihe VELM-8 war es, zu verdeutlichen, dass die Schülerbearbeitungen der Testaufgaben diagnostische Situationen beinhalten, die es seitens der Lehrkräfte zu nutzen gilt. Dementsprechend ist der Umgang mit Schülerbeiträgen wesentlicher Bestandteil der Fortbildungen gewesen, zumal u. a. bei den Stundenplanungen und zugehörigen Reflexionen der Rückbezug zu den korrespondierenden Schülerbeiträgen kritisch geprüft wurde. Letztlich sind die Lernstandserhebungen also Schülerbeiträge, die als Ausgangspunkt von individueller Förderung und weiterer Unterrichtsplanung genutzt werden sollen. Entsprechend dürfte die Häufigkeits-Einschätzung des *Umgangs mit Schülerbeiträgen* durch die Projektgruppenlehrkräfte höher ausfallen als bei den Kontrollgruppenlehrkräften (Hypothese 12). In engem Zusammenhang mit der Nutzung von Schülerbeiträgen und der *kognitiven Aktivierung* von Schülerinnen und Schülern steht das Verständnis von Unterricht. Den (Sozio-)Konstruktivismus zugrundelegend und Schülerbeiträge als Bezugspunkt im Unterricht ansehend, scheint eine rezeptive Rolle der Lernenden nur sehr eingeschränkt erstrebenswert. Die wesentliche Rolle der Schülerbeiträge wurde in den Fortbildungen wiederholt thematisiert und setzt eine Aktivität der Schülerinnen und Schüler voraus: Im Projektverlauf sollten die Lehrkräfte der Projektgruppe gegenüber der Kontrollgruppe infolgedessen angeben, ein geringeres *rezeptives Unterrichtsverständnis* (Hypothese 13) zu haben und weniger kleinschrittige, „*fürsorgliche*“ *Anleitung* (Hypothese 14) durchführen, da enggeführter Unterricht mit der lehrerinduzierten Fehlerfreiheit in diesem Sinne nur wenig zur kognitiven Aktivierung beiträgt. Vielmehr sollten die Lehrkräfte Beiträge von Schülerinnen und Schüler aufgreifen und diese in ihrer kognitiven Selbstständigkeit unterstützen. Die Lehrkräfte wurden im Rahmen der Fortbildung über die Relevanz eines solchen Vorgehens insbesondere auch aus lerntheoretischer Perspektive hingewiesen, sodass eine erhöhte

Unterstützung kognitiver Selbstständigkeit (Hypothese 15) in der Selbstauskunft durch Projektgruppenlehrkräfte nach Projektteilnahme erwartbar ist.

Mit den Lernstandserhebungen steht den Lehrkräften ein standardisiertes Diagnoseinstrument zur Verfügung. Im Rahmen der Fortbildungen wurde als theoretische Grundlage vor der Auseinandersetzung mit den Lernstandserhebungen „Feedback/Rückmeldung“ und „Diagnose“ als dessen Voraussetzung thematisiert (vgl. Fortbildung 2 im Teil 2). Die Möglichkeit, mit Hilfe der Lernstandserhebungen diagnostische Informationen zu gewinnen, ist folglich den teilnehmenden Lehrkräften ebenso offenbart worden wie die Notwendigkeit regelmäßiger Diagnosesituationen im alltäglichen Unterricht zu schaffen, um den Unterricht an die Schülerinnen und Schüler und ihr Leistungsniveau anpassen zu können (vgl. *Zone of proximal development*, Vygotskii et al., 1978). Entsprechend ist zu vermuten, dass die Häufigkeit der Nutzung von *leistungsbezogenen Diagnosesituationen im Unterricht* von fortgebildeten Lehrkräfte nach Projektablauf in der Selbstauskunft höher ausfällt als bei nicht fortgebildeten Lehrkräften (Hypothese 16).

Die bisherigen Hypothesen lassen sich weitestgehend als Subfacetten der *kognitiven Aktivierung* verstehen. Neben der *kognitiven Aktivierung* spielt ein kooperatives und schülerorientiertes Unterrichtsklima eine wichtige Rolle. Dabei liegen einige der nachfolgend beschriebenen Aspekte quasi an der Schnittstelle von Unterrichtsklima und kognitiver Aktivierung. Leitendes Prinzip der Fortbildung war es, den Lehrkräften neben theoretischem Input ausreichend Zeit zu geben, im angeleiteten Rahmen in Kleingruppen an konkreten Aufgaben zu arbeiten. Dieses Prinzip der kognitiven Selbstständigkeit und des Soziokonstruktivismus sollten Lehrkräfte auch in ihrem Unterricht umsetzen, da durch die soziale Interaktion ein erhöhter Lerngewinn zu erwarten ist. Neben der kognitiven Seite dieses Vorgehens spielt die Gruppenzugehörigkeit nach der Selbstbestimmungstheorie auch für die Motivation eine wesentliche Rolle. Durch die theoretische Erläuterung (Fortbildung 2) sowie die praktische Umsetzung im Rahmen der Fortbildung ist ein häufigerer Einsatz/häufigere Nutzung sozialer Wissenskonstruktion durch die Lehrkräfte der Projektgruppe im Vergleich zu Kontrollgruppenlehrkräften nach Projektteilnahme erwartbar. Um den Begriff der sozialen Interaktion deutlicher zu operationalisieren, werden hier eine *gemeinsame Wissenskonstruktion* (Hypothese 17) und die *Nutzung von Lernpartnerschaften* analysiert (Hypothese 18).

Teil 3 – Methode

Hypothesen

Ein wesentlicher Teilaspekt eines unterstützenden Klassenklimas ist eine respektvolle Atmosphäre zwischen Lehrkraft und Schülerinnen sowie Schülern. In diesem Sinne ist analog zur Diagnose im Leistungsbereich eine „soziale Diagnose“ für Lehrkräfte notwendig, um Schülerinnen und Schüler bei Problemen unterstützen zu können. Damit jedoch Schülerinnen und Schüler mit Lehrkräften vertrauensvoll umgehen, ist umgekehrt auch ein persönliches Vertrauen zwischen Lehrkräften und Schülerinnen/Schülern notwendig. Basierend auf der Thematisierung der erläuterten Kriterien von Unterrichtsqualität und der Betonung von leistungsbezogener Diagnose sowie kooperativem, konstruktivem Umgang mit Fehlern wäre es zu erwarten, dass sich mit dem Fortbildungsbesuch bei den Projektgruppenlehrkräften positive Veränderungen für das wahrgenommene *persönliche Vertrauen* (Hypothese 19; höhere Werte) sowie die *Diagnose im sozialen Bereich* (Hypothese 20; häufigere Nutzung) ergeben als bei der Kontrollgruppe.

Letztlich sollten Lehrkräfte Schülerinnen und Schülern deutlich aufzeigen, dass trotz aller Standardisierung durch Bildungsstandards, Abschlussprüfungen, Vergleichsarbeiten innerhalb der Schulen und Lernstandserhebungen die individuelle Bezugsnorm für Lehrkräfte wesentlich ist. Mit der individuellen Bezugsnorm ist u. a. Lob verknüpft, das bei Schülerinnen und Schülern zu positiven Emotionen führen und damit zu einem schülerorientierten Unterrichtsklima beitragen kann. Die Differenzierung von Feedback und Lob wurde in der Fortbildung ebenfalls vorgenommen, sodass Projektgruppenlehrkräfte diese Facette von unterstützendem Klassenklima häufiger einsetzen sollten als Lehrkräfte der Kontrollgruppe (Hypothese 21). Grundsätzlich zeigt sich jedoch in Bezug auf Lob die Problematik von ambivalenten Wirkungen beispielsweise in Abhängigkeit vom Selbstwertgefühl (Brummelman et al., 2014) oder durch das Ablenken von Aufmerksamkeit von Tätigkeiten auf die Ebene der Person (Ditton & Arnoldt, 2004).

Mit der Zusammenfassung der Hypothesen in Tabelle 11 vor dem Hintergrund der Fortbildungsinhalte aus der Projektbeschreibung (Teil 2) wird bereits deutlich, dass Unterrichtsqualitätsaspekte zwar in Fortbildung 2 explizit angesprochen wurden und auch implizit in den nachfolgenden Fortbildungen v. a. in den Stundenreflexionen relevant waren, jedoch der Fokus der Fortbildung eher auf der Nutzung der Lernstandserhebungen lag, bei der das Stichwort Unterrichtsqualität für die Lehrkräfte im Hinterkopf präsent sein sollte. Daher sind m. E. die formulierten Hypothesen – und damit

die Hoffnungen durch die Fortbildungsreihe Veränderungen anzustoßen – sehr weitreichend. Jedoch wurde deutlich darlegt, dass diese Aspekte von Unterrichtsqualität, wie in den Hypothesen 7 bis 21 formuliert, Teil der Fortbildungsreihe waren und daher entsprechende Hypothesen aufgestellt werden.

3.1.1.1.3 Leistung

Distal liegt das Ziel von Unterrichts- und Schulentwicklung in der Optimierung der Leistung von Schülerinnen und Schülern. Diesem Ziel ist auch VELM-8 verpflichtet und soll nicht ausschließlich Einstellungen und Wahrnehmungen von Lehrkräften verändern. Theoretisch belegt ist der positive Einfluss von *kognitiver Aktivierung* und allgemein höherer Unterrichtsqualität sowie detaillierter Diagnose und präziser Rückmeldung/Feedback auf die Leistung der Schülerinnen und Schüler. Daher wäre der Nachweis einer überzufälligen positiven Abweichung der Schülerinnen und Schüler der Projektgruppenlehrkräfte in der schuljahresbegleitenden *Leistungsentwicklung* gegenüber den Schülerinnen und Schülern der Kontrollgruppenlehrkräfte denkbar (Hypothese 22), wobei für die Dauer von Yoon et al. (2007) angegeben wird, dass mehr als 14 Stunden einen deutlichen Einfluss haben, die von VELM-8 jedoch nicht erreicht werden. Allerdings relativieren Timperley et al. (2007) diese Einschätzung mit durch den Hinweis, dass insbesondere der Inhalt einer Fortbildung gewichtiger sei als rein die mit der Fortbildung verbrachte Zeit.

Teil 3 – Methode

Hypothesen

Tabelle 11: Übersicht der Hypothesen zum Bereich Unterrichtsqualität und Leistung der Schülerinnen und Schüler

Hypothese 7:	Im Verlauf der Projektteilnahme entwickelt sich die selbst wahrgenommene Häufigkeit des <i>Insistierens auf Begründungen und Erklärungen</i> für Projektgruppenlehrkräfte positiver als für Kontrollgruppenlehrkräfte.
Hypothese 8:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger Aufgaben mit <i>Variation in der Aufgabenschwierigkeit</i> einzusetzen als Lehrkräfte der Kontrollgruppe.
Hypothese 9:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>kognitiv anspruchsvolle Aufgaben</i> zu nutzen als Lehrkräfte der Kontrollgruppe.
Hypothese 10:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>kognitiv herausforderndes Üben</i> einzusetzen als Lehrkräfte der Kontrollgruppe.
Hypothese 11:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, ihre Schülerinnen und Schüler seltener <i>repetitiv Üben</i> zu lassen als Lehrkräfte der Kontrollgruppe.
Hypothese 12:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>Schülerbeiträge (Ideen/Fehler)</i> als Ausgangspunkt für den weiteren Unterricht zu nutzen als Lehrkräfte der Kontrollgruppe.
Hypothese 13:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, seltener <i>fehlervermeidende Erarbeitungsphasen</i> durch die Lehrkraft selbst zu gestalten als Lehrkräfte der Kontrollgruppe.
Hypothese 14:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, weniger bzw. seltener <i>kleinschrittige, fürsorgliche Anleitung und Kontrolle</i> zu geben als Lehrkräfte der Kontrollgruppe.
Hypothese 15:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, Schülerinnen und Schüler stärker in der <i>kognitiven Selbstständigkeit zu unterstützen</i> als Lehrkräfte der Kontrollgruppe.
Hypothese 16:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>leistungsbezogene Diagnose</i> durchzuführen als Lehrkräfte der Kontrollgruppe.
Hypothese 17:	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, größeren Wert auf eine <i>gemeinsame Wissenskonstruktion</i> zu legen als Lehrkräfte der Kontrollgruppe.
Hypothese 18:	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, <i>Lernpartnerschaften</i> mehr zu unterstützen als Lehrkräfte der Kontrollgruppe.
Hypothese 19:	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, größeren Wert auf <i>persönliches Vertrauen</i> zu legen als Lehrkräfte der Kontrollgruppe.
Hypothese 20:	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, häufiger <i>Diagnosen im sozialen Bereich</i> durchzuführen als Lehrkräfte der Kontrollgruppe.
Hypothese 21:	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, häufiger <i>nach individueller Bezugsnorm zu loben</i> als Lehrkräfte der Kontrollgruppe.
Hypothese 22:	Die Schülerinnen und Schüler der Projektgruppenlehrkräfte zeigen einen größeren <i>Leistungszuwachs</i> im Verlauf des Projektes gegenüber den Schülerinnen und Schülern der Kontrollgruppenlehrkräfte.

3.1.1.2 Typenbildung

In der Theorie wurden wiederholt Determinanten der Nutzung von Lernstandserhebungen bzw. deren Rückmeldungen genannt, wobei deutlich wird, dass simple monokausale Erklärungsmuster zu kurz greifen und nur auf den ersten Blick weiterhelfen. Zweifelsfrei lässt sich (noch) keine allumspannende theoretische Erklärung für die Nutzung von Lernstandserhebungen und deren Rückmeldungen aufbauen. Dabei meint „allumspannend“ zum einen die verschiedenen Ausgestaltungen des zentral entwickelten Instruments VerA in den einzelnen Bundesländern und zum anderen nimmt dies Bezug auf die enorme Heterogenität der Lehrkräfte, die mit der

Durchführung und Nutzung des Instruments betraut sind. Bereits mehrfach angesprochen wurde das Forschungsdesiderat fehlender Studien aus Hessen. Ein weiteres Manko des aktuellen Forschungsstands ist der überwiegende Bezug auf die Vergleichsarbeiten in der Grundschule. Diese beiden Lücken versucht diese Arbeit zu schließen und dabei der Heterogenität der Lehrkräfte ansatzweise gerecht zu werden. In diesem explorativen Sinne werden Gruppen von Lehrkräften gesucht, die die Lernstandserhebungen in ähnlicher Weise einschätzen und nutzen (Forschungsfrage 3). Im Anschluss an dieses explorative Vorgehen sollen die Ergebnisse mit denen anderer Studien verglichen werden (Ditton & Arnoldt, 2004; Groß-Ophoff et al., 2007; A. Hosenfeld, 2010).

3.1.2 Hypothesen zur Situation in Hessen

Über diese projektevaluationsbezogenen Hypothesen zu VELM-8 hinaus lassen sich weitere Annahmen über die Wahrnehmung der Lernstandserhebungen durch Lehrkräfte machen (die dabei nicht zwingend im Zusammenhang zu VELM-8 stehen, sondern eher einen Ist-Stand in Hessen beschreiben sollen). Mit den Bildungsstandards als Grundlage der Lernstandserhebungen sowie des Unterrichts ist zu vermuten, dass die *Lehrplanvalidität* der Lernstandserhebungen von den Lehrkräften recht hoch eingeschätzt wird und sie dabei insbesondere eher zustimmen als ablehnen (Mittelwert größer als semantischer Median; Hypothese 23a; vgl. auch m. E. Hypothese 2, die jedoch die Entwicklung im Projektverlauf in den Blick nimmt, während hier der reine Ist-Stand betrachtet wird). Vor dem Hintergrund der Ergebnisse anderer Bundesländer sowie der Vielfalt bildungsreformorientierter Projekte in Mathematik liegt des Weiteren die Vermutung nahe, dass Mathematiklehrkräfte die Validität der Lernstandserhebungen höher einschätzen als Fachkolleginnen und -kollegen der Sprachen (Hypothese 23b; Maier, 2008b; Maier et al., 2012).

Einige weitere Hypothesen betreffen nachfolgend den Bereich von Fachunterschieden. Die Aufgaben der Lernstandserhebungen unterliegen den testtheoretischen Auflagen, insbesondere der Objektivität (vgl. Kapitel 1.2). Daraus resultiert, dass die Kodierung möglichst einheitlich erfolgen soll, weshalb sich – wie bereits dargelegt – v. a. vergleichsweise kurze Items im finalen Testinstrument finden. Dies gilt dabei insbesondere für Mathematik gegenüber den Fächern Deutsch und der ersten Fremdsprache. Entsprechend sollte die Auswertungszeit für Mathematiklehrkräfte geringer ausfallen als für Kolleginnen und Kollegen der Fächer Deutsch und Englisch

Teil 3 – Methode

Hypothesen

(Hypothese 24a). Diese Tendenz zeigte sich auch in anderen Untersuchungen (vgl. u. a. Bensen et al., 2006). Aus einer kürzeren Auswertungszeit lässt sich eine größere Zustimmung zur *Angemessenheit der Auswertungszeit* ableiten (Hypothese 24b).

Sowohl die Auswertungszeit, deren Angemessenheit als auch die Lehrplanvalidität lassen sich zumindest theoretisch als Voraussetzungen für eine weitere Beschäftigung mit den Lernstandserhebungen einordnen und wirken so bezugnehmend auf Helmkes Modell (2004; Abbildung 10) auf die Rezeption von Rückmeldungen. Intuitiv nachvollziehbar ist ein Zusammenhang zwischen der Balance zwischen Aufwand und Ertrag und der eingeschätzten Nützlichkeit⁴⁹ als Bedingung für die Rezeption, Reflexion und vor allem die Aktion im genannten Modell. Entsprechend den vorherigen Ausführungen und den postulierten Zusammenhängen sollten die Fachunterschiede zugunsten der Mathematik auch für die Nützlichkeit zu finden sein (Hypothese 25b). Aus dem Modell von Helmke lässt sich außerdem eine prädiktive Wirkung der Auswertungszeit auf die Nützlichkeit ableiten (Hypothese 25a). Diese Vermutung stützt Nachtigall (2014) mit Befunden aus Thüringen.

In diesem Zusammenhang stellt die wahrgenommene Nützlichkeit also die abhängige Variable dar. Die wahrgenommene Nützlichkeit ist jedoch letztlich ein *a priori* Merkmal, welches wiederum als unabhängige Variable prädiktiv positiv auf die *Durchsicht* der verfügbaren Rückmeldungen (Hypothese 26a) auswirken sollte. Die *Durchsicht* selbst sollte dem Modell folgend wiederum ein Prädiktor der *Besprechung der Rückmeldung* (Hypothese 26b) sein.⁵⁰ Über die prädiktiven Zusammenhänge hinaus, lassen Befunde aus Österreich eine intensivere Durchsicht (Hypothese 26c) und eine intensivere Besprechung mit allen verschiedenen Austauschgruppen (Hypothese 26d) durch Mathematik-Lehrkräfte als bei Kolleginnen und Kollegen mit sprachlichem Hintergrund vermuten (Grabensberger et al., 2008).

Unter erneutem Rückgriff auf das Modell von Helmke wurden nun also bereits Voraussetzungen, die *Durchsicht (Rezeption)*, die *Besprechung (Reflexion)* in der

⁴⁹ In Abgrenzung zur tatsächlichen Nutzung ist die Einschätzung der Nützlichkeit als eine *a-priori*-Einschätzung zu verstehen. Dieser Aspekt sollte eine Einschätzung über die Lernstandserhebungen zulassen, der unabhängig von etwaigen Alltagsproblemen (wie etwa mangelnde Zeit) zu sehen ist. Die tatsächliche Nutzung wiederum ist von diesen etwaigen Alltagsproblemen abhängig und direkt beeinflusst. Insofern unterscheiden sich *Nützlichkeit* und *Nutzung* in der Operationalisierung voneinander.

⁵⁰ Diese Wirkungskette legt bereits nahe, ein komplexeres Modell ebenfalls zu Rate zu ziehen. Hierauf wird im Rahmen dieser Arbeit jedoch in Anbetracht der Datenqualität verzichtet.

Untersuchung berücksichtigt. Im Folgenden postuliert das Modell die Nutzung (*Aktion*), also das Ableiten von Entwicklungsmaßnahmen. Dem Modell folgend ist der prädiktive Zusammenhang der *Nützlichkeit* (Hypothese 27a) und der *Besprechung* (Hypothese 27b) auf die *Nutzung* zu prüfen. Naheliegend ist die konsequente Weiterverfolgung der oben erwarteten Fachunterschiede auch auf diesen Aspekt, sodass von höheren Nutzungswerten bei Mathematiklehrkräften ausgegangen werden kann (Hypothese 27c). Fraglich bleibt jedoch, ob die Einschätzung der Nützlichkeit und der tatsächlichen Nutzung bei Selbstauskünften der Lehrkräfte empirisch zu trennen ist.

Die folgenden drei Hypothesen nehmen Bezug zu den Hypothese 3 bis 5, sollen jedoch einen Ist-Stand beschreiben und keine projektbezogene Entwicklung evaluieren. Als Voraussetzung für eine Nutzung der Lernstandserhebungen wurde bereits die Balance aus Aufwand und Ertrag angesprochen, wozu aus der Literatur *Belastungsindikatoren*/Indikatoren für Aufwand abgeleitet werden können. In klarem Zusammenhang steht die bereits aufgeführte Auswertungszeit mit der Klassengröße, da mehr Schülerinnen und Schüler sicher mehr Korrekturaufwand bedeuten. Ebenfalls als Prädiktor für eine geringere Nutzung stellt nach Korngiebel (2009) das *Alter* dar. Entsprechend lässt sich vermuten, dass die *Klassengröße* als Prädiktor für erhöhte Belastung durch die Lernstandserhebungen auffällt (Hypothese 28; Maier, 2008b; Maier et al., 2012) und das *Alter* ebenso mit einer erhöhten Belastung durch die Lernstandserhebungen einhergeht (Hypothese 29a). Korngiebel (2009, S. 96) geht in ihren Ausführungen jedoch über die Belastung einen Schritt weiter und spricht davon, dass „...Vorbehalte, Skepsis und Ignoranz eher bei den älteren Lehrergenerationen anzutreffen sind“. Darauf aufbauend ist ein negativer Einfluss des *Alters* auf den *förderdiagnostischen* und *unterrichtsplanerischen Nutzen* der Lernstandserhebungen zu vermuten (Hypothesen 29b und c).

Teil 3 – Methode

Hypothesen

Tabelle 12: Übersicht der allgemeinen Hypothesen zur Lage in Hessen

Hypothese 23a:	Die Lehrplanvalidität wird durch die Lehrkräfte eher zustimmend als ablehnend eingeschätzt.
Hypothese 23b:	Mathematiklehrkräfte schätzen die Lehrplanvalidität positiver ein als Kolleginnen und Kollegen der Fächer Deutsch oder Englisch.
Hypothese 24a:	Die Auswertungszeit fällt in Mathematik geringer aus als in Deutsch oder Englisch.
Hypothese 24b:	Die Angemessenheit der Korrektur- und Eingabezeit wird in Mathematik positiver eingeschätzt als in Deutsch und Englisch.
Hypothese 25a:	Die Auswertungszeit sagt die Nützlichkeit voraus.
Hypothese 25b:	Die Nützlichkeit wird von Mathematiklehrkräften positiver eingeschätzt als von Deutsch- oder Englischlehrkräften.
Hypothese 26a:	Die Einschätzung der Nützlichkeit bestimmt die Durchsicht der verfügbaren Rückmeldungen.
Hypothese 26b:	Die Nützlichkeit bestimmt die Intensität der Besprechung der Ergebnisse maßgeblich.
Hypothese 26c:	Die Durchsicht der verfügbaren Rückmeldungen wird von Mathematiklehrkräften höher angegeben als von Kolleginnen und Kollegen der sprachlichen Fächer.
Hypothese 26d:	Die Intensität der Besprechung wird von Mathematiklehrkräften höher berichtet gegenüber Kolleginnen und Kollegen mit sprachlichem Hintergrund.
Hypothese 27a:	Die Einschätzungen der Nützlichkeit sagen die Angaben zur Nutzung voraus. Die Intensität hat eine prädiktive Wirkung auf die Nutzung der Lernstandserhebungen.
Hypothese 27b:	Die Intensität hat eine prädiktive Wirkung auf die Nutzung der Lernstandserhebungen.
Hypothese 27c:	Die Nutzung wird von Mathematiklehrkräften positiver eingeschätzt als von Deutsch- oder Englischlehrkräften.
Hypothese 28:	Die Klassengröße steht in einem positiven prädiktiven Zusammenhang mit dem Belastungserleben der Lehrkräfte.
Hypothese 29a:	Das Alter bzw. die Berufserfahrung zeigt einen positiven Einfluss auf die eingeschätzte Belastung durch die Lernstandserhebungen.
Hypothese 29b:	Das Alter bzw. die Berufserfahrung zeigt einen negativen Einfluss auf den eingeschätzten förderdiagnostischen Nutzen der Lernstandserhebungen.
Hypothesen 29c:	Das Alter bzw. die Berufserfahrung zeigt einen negativen Einfluss auf den eingeschätzten unterrichtsplanerischen Nutzen der Lernstandserhebungen.

Über die konkreten Hypothesen hinaus besteht ein grundlegendes Forschungsinteresse, Informationen über die Nutzung und Wahrnehmung der Lernstandserhebungen für Hessen zu erlangen und damit die bisherigen Daten anderer Bundesländer zu ergänzen und für die hessische Situation ggf. Unterschiede oder Ähnlichkeiten aufzuzeigen. Vor dem Hintergrund der hohen Priorisierung der *didaktischen Kommentare* in der Fortbildungsreihe VELM-8 ist von besonderem Interesse, wie die Bekanntheit und Zufriedenheit mit den *didaktischen Kommentaren* ausfallen.

Naheliegender wäre es, den Ist-Stand in Hessen zunächst zu untersuchen, bevor mit VELM-8 ein in diese Gesamtsituation eingebettetes Projekt evaluiert wird. Von dieser intuitiven Reihenfolge wird zugunsten einer bewussten Schwerpunktsetzung abgewichen: Der Kern dieser Arbeit stellt die Evaluation von VELM-8 und die damit verbundene Typisierung der Lehrkräfte dar. Die Untersuchung des Ist-Standes in Hessen soll diese Informationen eher erweitern und einbetten als deren Grundlage sein.

3.2 Forschungsdesign

Zur Klärung der gestellten Hypothesen werden zwei verschiedene Datengrundlagen genutzt: Zum einen die jährliche Online-Befragung der an den Lernstandserhebungen teilnehmenden Lehrkräfte, die im Online-Portal (<http://www.lernstandsportal.de>) durchgeführt wird, und zum anderen die Daten der am Projekt VELM-8 Teilnehmenden.

In VELM-8 wurde ein Prä-Post-Test-Design mit Kontroll- und Projektgruppe gewählt, um die Wirkungen der Fortbildungen auf die Lehrkräfte und auf die Leistung der Schülerinnen und Schüler zu messen. Nach der Informationsveranstaltung im August 2012 konnten sich interessierte Lehrkräfte für die Teilnahme am Projekt bewerben (erste Erhebungen erfolgten bereits auf der Auftaktveranstaltung). Bei der Zuweisung in Kontroll- und Projektgruppe wurde nicht randomisiert, sondern auf die Wünsche der Lehrkräfte Bezug genommen. Voraussetzungen für die Teilnahme in der Projektgruppe war die Teilnahme an den Lernstandserhebungen 2013 und die Teilnahme von mindestens zwei Lehrkräften derselben Schule, um im Verlauf der Fortbildung insbesondere die Funktion von Lernstandserhebungen in der kollegialen Nutzung ansprechen zu können (vgl. Fortbildung 4 – 07.05.2013). Nach Aufteilung der Bewerbungen in Kontroll- und Projektgruppe wurde eine kurze Eingangsbefragung der Lehrkräfte zu den Lernstandserhebungen und zur Unterrichtsqualität sowie der Leistungstest für Schülerinnen und Schüler durchgeführt. Dabei wurde darauf geachtet, dass die Tests durch geschulte Testleiter mit Hilfe eines einheitlichen Manuals administriert wurden, was jedoch logistisch nicht in allen Lerngruppen realisiert werden konnte. Die Auswertung der Fragebögen und Leistungstests erfolgte hingegen durchgängig durch geschulte Kodierende. Die Lehrkräfte der Projektgruppe erhielten im Laufe des Schuljahres vier Fortbildungen (vgl. Teil 2), wobei je zwei vor und nach der tatsächlichen Durchführung der Lernstandserhebungen Ende Februar 2013 erfolgten. Nach Ende der vierten Projektgruppenfortbildung erfolgte der Posttest (wiederum in Form zweier Fragebögen für Lehrkräfte zu Lernstandserhebungen und Unterrichtsqualität sowie eines Leistungstests für Schülerinnen und Schüler). Die Durchführung erfolgte analog zum Prätest nach Möglichkeit mit geschultem Personal. Im Anschluss wurde eine Kompaktfortbildung für die Lehrkräfte der Kontrollgruppe angeboten, in der die Inhalte der Projektgruppenfortbildungen zusammenfassend dargestellt wurden. Insgesamt lässt sich der Projektverlauf im Schuljahr 2012/13 inkl. der Prä- und Post-Tests folgendermaßen darstellen (vgl. Abbildung 20).

Teil 3 – Methode

Hypothese



Abbildung 20: Projekttablauf VELM-8

3.3 Stichprobe

3.3.1 VELM-8

Im Projekt VELM-8 wurden interessierte Lehrkräfte aus Nordhessen per Fragebogen befragt und die zugehörigen Lerngruppen in Bezug auf die bildungsstandardsorientierte Leistung gemessen. Wie dargelegt, wurde bei der Einteilung in Projekt- und Kontrollgruppe des Projekts keine Randomisierung vorgenommen und auch *a priori* nicht die Äquivalenz der Gruppen in Bezug auf einzelne Merkmale wie beispielsweise Alter oder Geschlechterverhältnis kontrolliert. Entsprechend handelt es sich bei der vorliegenden Stichprobe um eine *Gelegenheitsstichprobe*, die keinerlei Anspruch auf Repräsentativität erhebt.

Insgesamt interessierten sich nach der Auftaktveranstaltung 34 Lehrkräfte von 12 Schulen im Umkreis von etwa 150 Kilometer um Kassel für die Projektteilnahme. Im Mittel interessierten sich drei Lehrkräfte einer Schule für die Teilnahme ($min = 1$; $max = 7$; $SD = 1.47$). Für die Projektevaluation liegt eine Panel-Stichprobe vor, die Prä- und Posttest der Schülerleistung inkl. der Lehrkräftebefragung absolvierten (Stichprobe 1)⁵¹. Die Panel-Stichprobe umfasst 27 Lehrkräfte, davon 16 in der Projektgruppe und 11 in der Kontrollgruppe. Die Einschätzungen zu den Lernstandserhebungen *a priori* liegen aus Auftaktveranstaltung und Prätest für insgesamt 67 Lehrkräfte vor (Stichprobe 2). In Stichprobe 2 sind also alle interessierten Lehrkräfte aus dem nordhessischen Raum enthalten, während Stichprobe 1 eine Teilmenge hiervon ist und den Lehrkräften entspricht, für die bei Projektteilnahme zu Prä- und Posttest Daten vorliegen. Im Folgenden sollen diese zwei Stichproben näher beschrieben werden:

⁵¹ Ergänzend zur Panel-Stichprobe wäre die Möglichkeit gegeben, die Hypothesen an einer minimal größeren Stichprobe von 35 Lehrpersonen zu testen. Da in dieser Stichprobe jedoch für Lehrkräfte z. T. Einstellungen und/oder Leistungskennwerte nur zu einem Messzeitpunkt vorliegen und insbesondere die Ergebnisse in der Interpretation nicht von denen der Panel-Stichprobe abweichen, werden in dieser Arbeit lediglich Panel-Ergebnisse berichtet.

Aufgrund der vor der Erhebung erwarteten geringen Stichprobengröße und der regionalen Nähe zwischen Lehrkräften und Forschenden wurde darauf verzichtet, umfangreiche Hintergrundvariablen zu erheben. Daher fallen die Stichprobenbeschreibungen eher knapp aus. Die zwei genannten Stichproben unterscheiden sich dabei nicht wesentlich in den betrachteten Merkmalen. Neben der deskriptiven Darstellung für die einzelnen Stichproben finden sich in der kommenden Tabelle die Ergebnisse von χ^2 - und Mann-Whitney-U-Tests zur Überprüfung etwaiger Unterschiede zwischen der Verteilung des jeweiligen Merkmals in Projekt- und Kontrollgruppe für die Stichprobe 1 (Tabelle 14).

Beide Stichproben weisen einen leicht höheren Frauenanteil auf. Bei der statistischen Prüfung zeigen sich mit Hilfe des χ^2 -Anpassungstests allerdings keine Signifikanzen ($p > .20$; Bortz, 2005; Bortz & Döring, 2006; Luhmann, 2011). Die Altersverteilung stellt sich mit einer Spitze bei knapp 60 Jahren und einer weiteren zu Beginn der beruflichen Tätigkeit (etwa 30 Jahre) bimodal dar. In der Verteilung der Altersstufen unterscheiden sich die Gruppen der Evaluationsstichprobe nicht signifikant. Als grobe Heuristik für einen Mittelwert des Alters kann die Unterrichtserfahrung dienen. Wenn davon ausgegangen wird, dass der Beginn der beruflichen Tätigkeit bei Mitte zwanzig liegt, entspricht die mittlere Berufserfahrung von knapp 15 Jahren (Stichprobe 1) einem Altersdurchschnitt von etwa 40 Jahren. Wiederum zeigen sich keine Unterschiede zwischen den Gruppen in den Evaluationsstichprobe.

Ebenfalls keine Gruppenunterschiede lassen sich bezüglich der bisherigen Vorerfahrungen mit den Lernstandserhebungen ausmachen. Allerdings fallen die Werte für dieses Merkmal erstaunlich niedrig aus. Dabei ist zu bedenken, dass die Lernstandserhebungen für die befragten Lehrkräfte ausschließlich in Jahrgang 8 und erst seit dem Schuljahr 2008/2009 durchgeführt werden. Die rechtliche Vorgabe der Teilnahme an den Lernstandserhebungen in mindestens einem Fach pro Schuljahr führt vermutlich zu innerschulischen Absprachen über ein Abwechseln der einzelnen Fächer, sodass die geringen Vorerfahrungen auf den zweiten Blick verständlich werden.

In Bezug auf die Ausbildung zeigt sich in den Stichproben eine überwiegende Mehrheit mit einem mathematischen Fachstudium. Ebenso haben fast alle Lehrkräfte in den Stichproben eine Ausbildung im gymnasialen oder Haupt-/Realschul-Bereich durchlaufen. Dabei teilen sich die Lehrkräfte in beiden Stichproben etwa jeweils zur Hälfte auf das gymnasiale und das Haupt- und Realschul-Lehramt auf. Wegen der

Teil 3 – Methode

Stichprobe

kooperativen und integrierten Gesamtschulen zeigt sich diese Aufteilung bei den Schularten jedoch nicht, spiegelt sich allerdings in der Verteilung der Lernstandserhebungstesthefte wieder: Die beiden Testhefte A (Hauptschule-Niveau) und B (Realschulniveau) weisen gemeinsam etwa dieselbe Häufigkeit auf wie Testheft C (gymnasiales Niveau). Die hohe Anzahl an fehlenden Werten (missings) für Stichprobe 2 erklärt sich dadurch, dass diese Information erst nach der Auftaktveranstaltung erfragt wurde (zur Organisation der Leistungstests für Schülerinnen und Schüler), sodass Lehrkräfte, die lediglich die Auftaktveranstaltung besucht haben, diese Information nicht geben konnten.

Auf Seiten der Schülerinnen und Schüler – selbstverständlich alle aus dem achten Jahrgang – ergeben sich gegliedert nach den Testheften bzw. Bildungsgängen folgende Stichprobengrößen zu beiden Messzeitpunkten:

Tabelle 13: Stichprobengrößen für Leistungstests der Schülerinnen und Schüler zu beiden Messzeitpunkten

$N_{\text{Hauptschule MZP1}}$	=	75	$N_{\text{Hauptschule MZP2}}$	=	67
$N_{\text{Realschule MZP1}}$	=	286	$N_{\text{Realschule MZP2}}$	=	284
$N_{\text{Gymnasium MZP1}}$	=	384	$N_{\text{Gymnasium MZP2}}$	=	338

Zwar wurden die Leistungstests von Schülerinnen und Schülern ausgefüllt, jedoch zielt das Forschungsinteresse nicht auf den Vergleich von Schülerinnen oder Schülern, sondern aggregiert auf den von Lerngruppen. Daher wurden die Prätestleistungen der Schülerinnen und Schüler nicht mit ihren Posttestleistungen verlinkt. Es liegt somit keine Panel-Stichprobe für die Leistung (Ebene 4) vor, sondern lediglich ein Längsschnitt im Trenddesign (Stein, 2019).

Tabelle 14: Stichprobenbeschreibungen

		Stichprobe 1		Stichprobe 2
			Gruppenunterschiede (Projekt-/ Kontroll- gruppe)	
Geschlecht	weiblich	58%	$\chi^2(1) = .007, p = .93$	58%
	männlich	42%		42%
	missings	1		5
Alter	> 65	0	$\chi^2(6) = 2.234, p = .90$	0
	61-65	0		0
	56-60	9		16
	51-55	1		1
	46-50	2		5
	41-45	3		8
	36-40	2		5
	31-35	4		12
	26-30	6		15
	< 26	0		1
	missings	0		4
Unterrichtserfahrung	Minimum	1.00	$W = 96.500, p = .40$	0.00
	1. Quartil	5.50		4.00
	Median	11.00		7.00
	Mittelwert	14.89		12.49
	3. Quartil	23.50		19.00
	Maximum	37.00		37.00
	SD	12.11		10.81
	missings	0		4
bisherige LSE-Teil- nahmen	0	16	$\chi^2(3) = 3.555, p = .31$	39
	1	8		15
	2	2		6
	3	1		3
	4	0		0
Ausbildung	2. Staatsexamen (Gymnasial-LA)	11	$\chi^2(2) = 1.994, p = .37$	28
	2. Staatsexamen (HR-LA)	15		28
	2. Staatsexamen (GS-LA)	0		0
	1. Staatsexamen	1		5
	Sonstiges	0		1
	missings	0		5
Fachstudium Mathematik	vorhanden	96%	$\chi^2(1) = 4.134 \cdot 10^{-32}, p = 1.00$	94%
	missings	0		4
Schulform	Gymnasium	4	$\chi^2(3) = 5.086, p = .17$	21
	Realschule	0		2
	Hauptschule	0		0
	KGS	16		28
	IGS	6		11
	sonstige	1		1
	missings	0		4
LSE-Testheft	A (Hauptschule)	4	$\chi^2(2) = 1.505, p = .47$	5
	B (Realschule)	7		8
	C (Gymnasium)	15		16
	missings	1		38

3.3.2 Online-Erhebung

Von der Hessischen Lehrkräfteakademie (früher Amt für Lehrerbildung), welches die Lernstandserhebungen in Hessen organisatorisch begleitet, werden jährlich nach Abschluss der Dateneingabe Einschätzungen der teilnehmenden Lehrkräfte erfragt. Dies geschieht über einen Online-Fragebogen, der von der Lehrkräfteakademie selbst erstellt wurde.⁵² Die Stichprobe der Online-Erhebung steht dabei in keinerlei Zusammenhang zu den beiden berichteten Stichprobe aus VELM-8.

Da es sich um eine online bereitgestellte Befragung handelt, liegt auch hier lediglich eine *Gelegenheitsstichprobe* vor. Die Stichprobe umfasst Aussagen von $N_{Gesamt} = 4401$ Personen, wobei hier jedoch nicht zwischen den verschiedenen Fächern (Deutsch, Englisch/Französisch⁵³ und Mathematik) sowie der Jahrgangsstufe (Jahrgangsstufe 3, 6 oder 8) differenziert wird. Auf die Frage nach persönlicher Teilnahme antworten lediglich 620 Lehrkräfte mit „ja“. Entsprechend reduziert sich die Stichprobe deutlich, was nochmals drastischer wird, wenn nach den Lernstandserhebungen 2013 für die einzelnen Fächer gefragt wird:

Tabelle 15: Stichprobengröße differenziert nach Fächern bei Berücksichtigung der Teilnahme an den LSE 2013

	<i>N</i> Teilnahme 2012
Mathematik	35
Deutsch	81
Englisch	95
Mathematik & Deutsch	1
Mathematik & Englisch	1
Mathematik & Deutsch & Englisch	3
Deutsch & Englisch	4

Leider wurden in diesem Fragebogen außer der Schulform keine Hintergrundinformationen erhoben. Bei der Schulform wird zwischen Grund-, Förder-, Regelschulen und Gymnasien unterschieden: Es stammen lediglich knapp 1% der Teilnehmenden von Förderschulen. Der Großteil von 38% unterrichtet an Grundschulen, knapp 11% an Gymnasien und etwa 28% an Regelschulen. Von insgesamt 23% der Teilnehmenden liegen keine Aussagen zur Schulform vor.

⁵² An späteren Versionen des Online-Fragebogens war ich beratend ebenfalls beteiligt, jedoch nicht an der dieser Arbeit zugrundeliegenden Version.

⁵³ Für Französisch Jahrgang 8 liegen keinerlei Daten vor, sodass Französisch in den weiteren Teilen der Arbeit nicht weiter aufgeführt wird.

Aufgrund der verschiedenen Stichproben, die für die Untersuchung unterschiedlicher Hypothesen herangezogen werden, verdeutlicht Tabelle 16 die Zuordnung der Stichproben zu den Hypothesen.

Tabelle 16: Übersicht Stichproben-Hypothesen-Zuordnung

Hypothese	Stichprobe	Inhaltsbereich	Hypothesenkapitel
1 – 6	Stichprobe 1	Projektevaluation VELM-8 Einstellungen & Nutzen Lernstandserhebungen	Kapitel 3.1.1.1
7 – 21	Stichprobe 1	Projektevaluation VELM-8 Unterrichtsqualität	Kapitel 3.1.1.1
22	SuS-Stichprobe	Projektevaluation VELM-8 SuS-Leistung	Kapitel 3.1.1.1
23-27	Online-Stichprobe	Ist-Stand Lernstandserhebungen in Hessen	Kapitel 3.1.2
28 – 29	Stichprobe 2	Ist-Stand Lernstandserhebungen in Hessen	Kapitel 3.1.2

3.4 Instrumente

3.4.1 VELM-8

Um die gestellten Hypothesen beantworten zu können, wurden neben Fragen zu persönlichen Hintergrundmerkmalen bei den teilnehmenden Lehrkräften Skalen zu Einstellungen gegenüber den Lernstandserhebungen nach Maier (2008a) sowie ausgewählte Skalen zur Unterrichtsqualität nach COACTIV (Baumert et al., 2008) eingesetzt. Dabei liegen die Daten zur Unterrichtsqualität lediglich für die Stichprobe 1 vor, während die Skalen zu Einstellungen gegenüber den Lernstandserhebungen für alle beide Stichproben (1 und 2) erhoben wurden.

3.4.1.1 Einstellungen gegenüber den Lernstandserhebungen

Für die Skalen zu Einstellungen gegenüber den Lernstandserhebungen ergeben sich über beide Stichproben und beiden Messzeitpunkte durchweg akzeptable und fast überwiegend gute Reliabilitäten (Cronbachs $\alpha > .68$; Bortz & Döring, 2006; Streiner, 2003). Hierbei fällt vor allem die Skala *Lehrplanvalidität der Lernstandserhebungen* auf, deren Reliabilität zum Prätest für die Stichprobe 1 geringfügig unter $.70$ liegt (jedoch über dem Grenzwert von $\alpha = .60$). Für die zugehörigen Items liegen jedoch die Trennschärfen durchweg im akzeptablen Bereich von $r_{it} > .30$ (Bortz & Döring, 2006; Bühner, 2011), sodass diese geringfügigen Abweichungen akzeptiert werden. Genauere Aufschlüsselungen u. a. Trennschärfen etc. sind im Anhang zu entnehmen (vgl. Anhang 7.2).

Tabelle 17: Reliabilitäten (Cronbachs α) der Skalen zu den Lernstandserhebungen bzgl. der verschiedenen Stichproben

	Stichprobe 1		Stichprobe 2
	Prätest	Posttest	
Allgemein Akzeptanz (H1)	.89	.74	.81
Lernstandserhebungen als Belastung (H2)	.72	.75	.82
Lehrplanvalidität der Lernstandserhebungen (H3)	.69	.78	.70
Förderdiagnostischer Nutzen (H4)	.84	.83	.80
Unterrichtsplanerischer Nutzen (H5)	.89	.86	.84
Selektionsdiagnostischer Nutzen (H6)	.78	.73	.77

Tabelle 18: Einstellungen zu den Lernstandserhebungen: Skalennittelwerte, Standardabweichung getrennt nach Stichproben⁵⁴

		Stichprobe 1				Stichprobe 2			
		Mittelwerte nach Maier (2008a)	Prätest	t-Test zu Maier (2008a)	Posttest	t-Test zu Maier (2008a)	Prätest	t-Test zu Maier (2008a)	
Allgemeine Akzeptanz (H1)	M	2.75	3.17	t(26) = 2.47, p < .05	3.40	t(25) = 4.40, p < .01	3.03	t(58) = 2.89, p < .01	
	SD		0.87		0.76		0.74		
LSE als Belastung (H2)	M	2.71	2.14	t(26) = -4.30, p < .001	2.01	t(25) = -5.52, p < .001	2.44	t(59) = -2.43, p < .05	
	SD		0.69		0.65		0.86		
Lehrplanvalidität der LSE (H3)	M	2.78	3.79	V = 250, p < .001 ⁵⁵	3.89	V = 339, p < .001 ⁵¹	3.64	t(46) = 10.66, p < .001	
	SD		0.49		0.60		0.55		
Förderdiagnostischer Nutzen (H4)	M	2.68	3.36	t(25) = 5.17, p < .001	3.58	t(25) = 7.04, p < .001	3.32	t(56) = 7.70, p < .001	
	SD		0.67		0.65		0.63		
Unterrichtsplanerischer Nutzen (H5)	M	2.71	3.22	t(25) = 3.57, p < .01	3.40	t(25) = 4.73, p < .001	3.19	t(56) = 5.57, p < .001	
	SD		0.73		0.74		0.66		
Selektionsdiagnostischer Nutzen (H6)	M	2.82	2.49	t(25) = -2.39, p < .05	2.75	t(25) = -0.51, p = .61	2.51	t(58) = -3.13, p < .01	
	SD		0.71		0.73		0.77		

⁵⁴ Antwortformat: 1: „stimmt gar nicht“ bis 5: „stimmt genau“

⁵⁵ Aufgrund der nicht anzunehmenden Normalverteilung dieser Skala wird hier auf den Wilcoxon-Rang-Test zurückgegriffen.

Tabelle 19: Übersicht (marginal) signifikant von der Normalverteilung abweichender Skalen (nach Kolmogorow-Smirnow-Lilliefors-Test)

Stichprobe	Messzeitpunkt	Skala	D	p
Stichprobe 1	Prätest	Lehrplanvalidität der LSE	0.21	.013
	Posttest	Lehrplanvalidität der LSE	0.18	.030
Stichprobe 2	Prätest	LSE als Belastung	0.12	.046
		Lehrplanvalidität der LSE	0.17	.001
		Unterrichtsplanerischer Nutzen	0.11	.062
		Selektionsdiagnostische Nutzung	0.13	.017

Aus Tabelle 18 sind die Mittelwerte und Standardabweichungen der einzelnen Skalen getrennt für die Stichproben zu entnehmen. Für fast alle Mittelwerte der Stichproben zeigen sich signifikante Abweichungen gegenüber der Stichprobe von Maier (2008a)⁵⁶. Hierin deuten sich wohlwollendere Einschätzungen der Lernstandserhebungen durch die Lehrkräfte in der vorliegenden Untersuchung an. Einschränkend anzumerken ist, dass als Grundlage für die berechneten *t*-Tests die Annahme der Normalverteilung nicht für alle Skalen zu allen Messzeitpunkten basierend auf dem Kolmogorow-Smirnow-Lilliefors-Test bestätigt werden kann (vgl. Tabelle 19); in der Tabelle sind auch diejenigen Skalen-/ Messzeitpunkte aufgeführt, die lediglich einen marginal signifikanten Wert⁵⁷ aufweisen. Auch in diesen Analysen zeigt sich die Skala *Lehrplanvalidität der LSE* wiederum als auffällig, sodass die Ergebnisse zu dieser Skala mit Vorsicht zu betrachten sind. Zur Absicherung der diese Skala betreffenden Ergebnisse werden neben den parametrischen Verfahren non-parametrische Verfahren eingesetzt (vgl. hinten in Analysemethoden).

3.4.1.2 Unterrichtsqualität

Bei den erhobenen Skalen zur Unterrichtsqualität handelt es sich um Skalen aus COACTIV (Baumert et al., 2008). Diese Skalen weisen ein vierstufiges Antwortformat auf (entweder 1: „trifft nicht zu“ bis 4: „trifft völlig zu“ oder 1: „selten oder nie“ bis 4: „regelmäßig“). Die exakten Itemstatistiken sowie Korrelationen zwischen Items der einzelnen Skalen finden sich im Skalenhandbuch im Anhang (vgl. Anhang 7.2).

⁵⁶ Bei der Stichprobe von Maier (2008a, S. 104) handelt es sich um eine Gelegenheitsstichprobe, von der versandte Fragebögen retourniert wurden. Die Stichprobe weist ein leicht verschobenes Geschlechterverhältnis von 63.1% Lehrerinnen auf und stammt zu je 36.8% vom Gymnasium und von der Hauptschule. Die restlichen 26.4% der Lehrkräfte stammen von der Realschule.

⁵⁷ Marginale Signifikanz meint in Anlehnung an Pospeschill (2006, S. 158) einen *p*-Wert zwischen .05 und .10 ($.05 \leq p < .10$; vgl. auch Bortz, 2005; Hosenfeld, 2010).

Tabelle 20: Reliabilitäten der Skalen zur Unterrichtsqualität im Vergleich zu Baument et al. (2008) für Stichprobe 1

	VELM-8		Posttest				COACTIV				Mittelwertsunterschiede	
	Prätest	Items	Prätest		Posttest		2003		2004		Prätest zu 2003	Posttest zu 2004
			α	M (SD)	α	M (SD)	α	M (SD)	α	M (SD)		
Insistieren auf Erklärungen und Begründungen (H7)		4	.74	3.11 (.62)	.66	2.95 (.51)	.69	2.80 (.58)	.73	2.95 (.57)	$t(24) = 2.48$, $p = .02$	$t(23) = -0.02$, $p = .98$
Differenzierung durch variierende Aufgabenschwierigkeit (H8)		6	.69	2.71 (.47)	.76	2.74 (.51)	.83	2.34 (.58)	.85	2.25 (.62)	$t(22) = 3.74$, $p = .001$	$t(23) = 4.64$, $p = .001$
Kognitiv herausfordernde Aufgaben (H9)		8	.85	2.79 (.50)	.86	2.68 (.47)	.82	2.51 (.50)	.82	2.55 (.49)	$t(24) = 2.76$, $p = .01$	$t(20) = 1.25$, $p = .23$
Kognitiv herausforderndes Üben (H10)		8	.78	2.83 (.37)	.80	2.95 (.41)	.82	2.55 (.44)	.79	2.64 (.39)	$t(23) = 3.72$, $p = .001$	$t(21) = 3.56$, $p = .002$
Einschleifendes repetitives Üben (H11)		5	.85	2.38 (.65)	.85	2.49 (.55)	.71	2.49 (.46)	.71	2.55 (.43)	$t(23) = -0.86$, $p = .40$	$t(22) = -0.55$, $p = .59$
Kognitiv herausfordernder Umgang mit Schülerbeiträgen (Ideen/Fehler) (H12)		4	.53	2.86 (.43)	.36	2.92 (.30)	.75	2.91 (.49)	.81	3.01 (.48)	$t(25) = -0.65$, $p = .52$	$t(23) = -1.52$, $p = .14$
Fehlervmeidende Erarbeitung durch den Lehrer selbst (H13)		5	.84	2.53 (.57)	.76	2.50 (.55)	.80	2.71 (.55)	.84	2.61 (.57)	$t(25) = -1.61$, $p = .12$	$t(22) = -0.93$, $p = .36$
Fürsorgliche Anleitung und Kontrolle (H14)		6	.58	2.90 (.39)	.62	2.95 (.35)	.69	2.81 (.46)	.70	2.76 (.44)	$t(25) = 1.24$, $p = .23$	$t(23) = -2.62$, $p = .02$
Unterstützung kognitiver Selbstständigkeit (H15)		4	.59	3.64 (.33)	.70	3.49 (.38)	.69	3.16 (.48)	.74	3.24 (.45)	$t(25) = 7.42$, $p < .001$	$t(23) = 3.22$, $p = .004$
Diagnose im Leistungsbereich (H16)		5	.70	3.08 (.45)	.15	2.82 (.29)	.73	3.09 (.36)	.78	3.07 (.38)	$t(25) = -0.15$, $p = .88$	$t(20) = -4.30$, $p < .001$
Aktive soziale Konstruktion von Verständnis (H17)		8	.89	2.28 (.61)	.88	2.30 (.58)	.83	1.97 (.51)	.82	1.93 (.50)	$t(23) = 2.44$, $p = .02$	$t(23) = 3.13$, $p = .005$
Tutorielle Unterstützung in Lernpartnerschaften und Gruppen (H18)		6	.75	2.90 (.54)	.77	2.87 (.55)	.78	2.38 (.58)	.81	2.49 (.63)	$t(24) = 4.78$, $p < .001$	$t(23) = 3.40$, $p = .002$
Persönliches Vertrauen (H19)		4	.91	3.63 (.47)	.22	3.01 (.35)	.85	3.50 (.45)	.84	3.49 (.46)	$t(25) = 1.46$, $p = .16$	$t(23) = -6.42$, $p < .001$
Diagnostische Kompetenz im sozialen Bereich (H20)		4	.81	3.09 (.50)	.62	3.08 (.51)	.84	3.03 (.51)	.88	2.94 (.55)	$t(25) = 0.58$, $p = .57$	$t(23) = 1.37$, $p = .19$
Lob nach individueller Bezugsnorm (H21)		3	.76	3.56 (.44)	.33	2.90 (.48)	.80	3.53 (.45)	nicht erhoben	nicht erhoben	$t(24) = 0.34$, $p = .74$	nicht erhoben

Teil 3 – Methode Instrumente

Bei einigen Skalen zeigen sich Reliabilitäten, die für einzelne Messzeitpunkte unter .60 liegen. Dazu zählen:

- Unterstützung kognitiver Selbstständigkeit
- Fürsorgliche Anleitung und Kontrolle

Für die beiden Skalen liegt eine geringe Abweichung vom Grenzwert $\alpha = .60$ lediglich zum Prätest vor. Entsprechend weisen die Instrumente zum jeweils anderen Messzeitpunkt eine ausreichende Reliabilität auf. Daher habe ich mich für den unveränderten Einsatz dieser Skalen entschieden.

Einige weitere Skalen werden bei der schlichten Beurteilung durch die Reliabilität ebenfalls auffällig und sollen im Folgenden etwas detaillierter betrachtet werden:

3.4.1.2.1 Kognitiv herausfordernder Umgang mit Schülerbeiträgen (Ideen/Fehler)

Tabelle 21: Itemformulierungen Skala „Kognitiv herausfordernder Umgang mit Schülerbeiträgen (Ideen/Fehler)“

mside1	Ich gehe von den Ideen der Schüler/innen aus und spiele mit ihnen die Konsequenzen durch, bis die Schüler/innen erkennen, ob ihre Gedanken zum Ziel führen oder Ungereimtheiten offenkundig werden.
mside2	Wenn ein/eine Schüler/in bei der Erarbeitung eines Sachverhalts einen Fehler macht, nehme ich die Vorschläge zunächst ohne Korrektur an und verfolge gemeinsam mit den Schüler/innen die Konsequenzen, bis der Fehler offenkundig wird.
mside3	Ich lasse die Schüler/innen auch einmal bewusst in die Irre laufen, bis sie sehen, dass etwas nicht stimmen kann.
mside4	Ich gehe von Vorstellungen der Schüler/innen aus und spiele mit ihnen die Konsequenzen durch, bis die Schüler/innen verblüffende Ungereimtheiten erkennen.

Aus der inhaltlichen Betrachtung der Items wird bereits deutlich, dass mside1 und mside4 eher von den Vorstellungen und Ideen der Schülerinnen und Schüler sprechen, während mside2 und mside3 eher auf den Umgang mit Fehlern abzielen.

Tabelle 22: Itemanalyse „Kognitiv herausfordernder Umgang mit Schülerbeiträgen“

	Prätest			Posttest		
	r_{it}	M	α wenn Item entfernt	r_{it}	M	α wenn Item entfernt
mside1	.61	2.69		.14	.56	2.92
mside2	.23	3.08		.52	.31	3.00
mside3	.14	3.08		.59	-.15	3.00
mside4	.34	2.58		.44	.19	2.75

In dieser Skala fällt das Item mside3 zu beiden Messzeitpunkten durch geringe, z. T. negative Trennschärfen auf. Zur Kontrolle wurde diese Skala faktorenanalytisch untersucht. Dabei wurde zunächst die Anzahl der empfohlenen Faktoren geprüft:

Hierbei ergeben sich für beide Stichproben und Messzeitpunkte ähnliche Scree-Plots mit lediglich einem Eigenwert größer 1, jedoch ergeben sich nach Parallelanalyse zwei Faktoren (Bühner, 2011; Luhmann, 2011; Revelle, 2017b, 2017a). Wiederholt man diesen Schritt ohne mside3, so kommen beide Verfahren zur Ein-Faktoren-Lösung. Daraus leite ich ab, dass diese Skala ohne Item 3 genutzt wird mit einer Reliabilität von $\alpha_{\text{Prätest}} = .59$ und $\alpha_{\text{Posttest}} = .62$.

3.4.1.2.2 Persönliches Vertrauen

Tabelle 23: Itemformulierungen Skala „Persönliches Vertrauen“

mverp1	Ich kümmere mich um meine Schüler/innen, wenn sie Probleme haben.
mverp2	Ich baue Vertrauen zu meinen Schüler/innen auf.
mverp3	Ich zeige Verständnis für meine Schüler/innen.
mverp4	Ich nehme mir Zeit, wenn meine Schüler/innen etwas mit mir bereden möchten.

In der genaueren Betrachtung wird das Item mverp1 zum zweiten Messzeitpunkt auffällig. Es weist negative Korrelationen zu den anderen Items der Skalen *persönliches Vertrauen* im geringen und mittleren Bereich auf ($-0.05 \leq r_{ij} \leq -0.48$). Entsprechend fällt die Trennschärfe ebenfalls negativ aus (vgl. Tabelle 24). Woran diese Abweichung liegt, lässt sich an dieser Stelle ohne genaueren Blick in die Daten nicht erkennen. Da durch das Entfernen von mverp1 weder zum ersten noch zum zweiten Messzeitpunkt negative Folgen entstehen, reduziere ich die Skala in den Auswertungen auf drei Items: mverp2, mverp3 und mverp4. Damit ergeben sich entsprechend Tabelle 24 akzeptable Reliabilitäten von $\alpha_{\text{Prätest}} = .87$ bzw. $\alpha_{\text{Posttest}} = .65$.

Tabelle 24: Itemanalyse Skala „Persönliches Vertrauen“

	Prätest			Posttest		
	<i>r_{it}</i>	<i>M</i>	$\alpha_{\text{wenn Item entfernt}}$	<i>r_{it}</i>	<i>M</i>	$\alpha_{\text{wenn Item entfernt}}$
mverp1	.84	3.58	.87	-.28	2.29	.65
mverp2	.93	3.69	.85	.47	2.67	-.50
mverp3	.64	3.65	.94	.41	3.42	-.23
mverp4	.82	3.62	.88	.06	3.67	.23

3.4.1.2.3 Lob nach individueller Bezugsnorm

Tabelle 25: Itemformulierungen Skala „Lob nach individueller Bezugsnorm“

mlob1	Ich lobe auch die schlechten Schüler/innen in dieser Klasse, wenn ich merke, dass sie sich verbessern.
mlob2	Wenn sich ein/eine Schüler/in besonders angestrengt hat, lobe ich ihn/sie meistens, auch wenn andere Schüler/innen noch besser sind als er/sie.
mlob3	Wenn ein/eine Schüler/in seine/ihre Mathematikleistungen verbessert, lobe ich ihn/sie, auch dann, wenn er/sie im Vergleich zur Klasse unter dem Durchschnitt liegt.

Teil 3 – Methode Instrumente

Mit dieser Skala aus COACTIV wird erhoben, ob Lehrkräfte das Lob von Schülerinnen und Schüler an der individuellen Leistung(-sentwicklung) messen, also beispielsweise Schülerinnen und Schüler auch dann loben, wenn trotz einer Leistungssteigerung die individuelle Leistung dennoch unter dem Klassendurchschnitt liegt. Wie bereits in den Ausführungen zur entsprechenden Hypothese dargelegt, ist jedoch der Aspekt von Lob im Unterricht ambivalent zu sehen. Aufgrund der nun vorliegenden schlechten Reliabilitäten, die eine zuverlässige Messung mit dieser Skala nicht sicher gewährleisten, wird auf diese Skala in der Auswertung verzichtet.

3.4.1.2.4 Diagnose im Leistungsbereich

Tabelle 26: Itemformulierungen Skala „Diagnose im Leistungsbereich“

mdial1	Es fällt mir leicht festzustellen, ob ein/eine Schüler/in eine Aufgabe verstanden hat.
mdial2	Ich merke sehr schnell, wenn jemand etwas nicht verstanden hat.
mdial3	Ich weiß, bei welchen Aufgaben die einzelnen Schüler/innen Schwierigkeiten haben.
mdial4	Ich kenne die Stärken und Schwächen der einzelnen Schüler/innen.
mdial5	Ich merke sofort, wenn ein/eine Schülerin im Unterricht nicht mitkommt.

Tabelle 27: Itemanalyse Skala „Diagnose im Leistungsbereich“

	Prätest			Posttest		
	<i>r_{it}</i>	<i>M</i>	α wenn Item entfernt	<i>r_{it}</i>	<i>M</i>	α wenn Item entfernt
mdial1_1	.61	3.08	.57	-.29	1.96	.33
mdial2_1	.68	3.23	.58	.17	2.96	-.43
mdial3_1	.40	2.85	.67	.05	3.00	-.21
mdial4_1	.43	3.12	.65	-.13	3.13	.06
mdial5_1	.22	3.12	.73	.22	3.04	-.35

Tabelle 28: Korrelationen zwischen Einzelitems der Skala „Diagnose im Leistungsbereich“

	mdial1_1	mdial2_1	mdial3_1	mdial4_1	mdial5_1	mdial1_2	mdial2_2	mdial3_2	mdial4_2	mdial5_2
mdial1_1	1.00	.75	.48	.34	.16	.11	.33	.00	.15	-.13
mdial2_1	.75	1.00	.53	.41	.18	.11	.27	-.11	.04	-.18
mdial3_1	.48	.53	1.00	.28	.08	.17	.23	-.01	.39	-.11
mdial4_1	.34	.41	.28	1.00	.29	-.33	.31	-.15	.23	-.25
mdial5_1	.16	.18	.08	.29	1.00	-.03	.20	-.15	.21	.36
mdial1_2	.11	.11	.17	-.33	-.03	1.00	0.00	.40	.01	-.01
mdial2_2	.33	.27	.23	.31	.20	.00	1.00	-.36	.46	.10
mdial3_2	.00	-.11	-.01	-.15	-.15	.40	-.36	1.00	.05	-.35
mdial4_2	.15	.04	.39	.23	.21	.01	.46	.05	1.00	.02
mdial5_2	-.13	-.18	-.11	-.25	.36	-.01	.10	-.35	.02	1.00

Auch bei dieser Skala weisen einige Items deutliche Differenzen in der Trennschärfe zwischen den Messzeitpunkten auf. Dabei scheinen einzelne Items im Posttest

anders verstanden bzw. anders beantwortet worden zu sein als im Prätest, denn die Trennschärfen von mdial1 und mdial4 liegen beim Prätest im positiven und beim Posttest im negativen Bereich. Ähnliches zeigt sich bei der Betrachtung der Korrelationen aller Items untereinander, die für den Prätest im Positiven liegen, im Posttest jedoch z. T. negativ auffallen (vgl. Tabelle 28).

Die Prüfung der Skala mit Hilfe der Faktorenanalyse fällt für den Prätest unproblematisch aus, was sich auch in den Reliabilitäten widerspiegelt, und es wird eine Ein-Faktoren-Lösung vorgeschlagen. Für die Posttests zeigt sich jedoch ein anderes Bild, so dass 2 Faktoren aufgrund des Paralleltests empfohlen werden, von denen jedoch lediglich ein Eigenwert größer 1 ist. Ohne nähere Analysen, welche den Rahmen der Arbeit übersteigen würden, lässt sich hier nur vermuten, dass die Items der Skala zu den einzelnen Messzeitpunkten unterschiedlich interpretiert wurden. Da so keine zuverlässige Messung erwartbar ist, verzichte ich in den Auswertungen auf diese Skala.

3.4.1.3 Leistung

Bei den Leistungstests der Schülerinnen und Schüler wurden Lernstandserhebungen der vergangenen Jahre eingesetzt. Konkret wurde auf die Lernstandserhebung des Schuljahres 2009/2010 zurückgegriffen (Roppelt, Katzenbach & Pant, 2010). Da es sich bei den Lernstandserhebungen 2010 bereits um deutschlandweit einheitlich entwickelte Leistungstests sowie nach psychometrischen Maßstäben geprüfte Testhefte handelt, kann davon ausgegangen werden, dass die Gütekriterien für diese Tests erfüllt sind. Es ergeben sich für die einzelnen Testheftversionen (Heft A, Heft B und Heft C) die folgenden Reliabilitäten zu den beiden Messzeitpunkten:

Tabelle 29: Reliabilität und deskriptive Statistiken der Leistungsdaten der Schülerinnen und Schüler

	Prätest				Posttest			
	α	N	M	SD	α	N	M	SD
Testheft A (Hauptschule)	.62	75	.21	0.09	.74	67	.21	0.11
Testheft B (Realschule)	.73	286	.26	0.12	.80	283	.35	0.14
Testheft C (Gymnasium)	.79	384	.44	0.14	.84	338	.53	0.16

Im Rahmen der Testauswertung wurden Teile der Testhefte doppelkodiert, wobei ich als Referenz bei jedem Testheft kodiert habe. Hierbei wurden etwa jeweils mehr als 20% der Fälle doppelkodiert (TH A: 33%; TH B: 38%; TH C: 25%). Eine tabellarische Aufstellung der prozentualen Übereinstimmungen sowie der Kappa-Werte (J. Cohen, 1960; Fleiss & Cohen, 1973) findet sich im Anhang (vgl. 7.3.1). Die Doppelkodierung ergab größtenteils sehr zufriedenstellende Werte in der prozentualen Betrachtung. Die z. T. nicht zufriedenstellenden K-Werte sind vermutlich darauf zurückzuführen,

Teil 3 – Methode Instrumente

dass die Rater verschieden „streng“ waren, in dem Sinne, dass in einfach zu kodierenden Aufgaben verschiedene Ansprüche angelegt wurden: Konkret also bei einer hohen erwarteten Übereinstimmung zwischen den Ratenden eine überraschend geringe tatsächliche Übereinstimmung auftrat und folglich der Quotient aus $K = \frac{p_o - p_e}{1 - p_e}$ mit p_o als den tatsächlichen und p_e den erwarteten Übereinstimmungen gering ausfällt. Nichtsdestotrotz lag die große Mehrheit (vgl. Medianwerte und Wert für 1. Quartil in Tabelle 30) der K-Werte im Bereich moderater ($K > .4$) bis nahezu perfekter ($K > .8$) Übereinstimmung (Landis & Koch, 1977, S. 165; vgl. auch z.T. strengere Interpretationen bei Bakeman & Gottman, 2010; Fleiss & Cohen, 1973; Margraf & Fehm, 1996; Wirtz & Caspar, 2002). In der folgenden Tabelle sind jeweils die minimalen Werte der prozentualen Übereinstimmung sowie von Cohens K angegeben. Wird ergänzend der Odds Ratio betrachtet, der nach Wirtz und Caspar (2002) weniger anfällig für Unterschiede in den Randsummen, also prinzipielle „Strenge“ einzelner Rater, ist, so zeigen sich erhebliche Übereinstimmungswerte selbst für Testheft A, welches die schlechtesten prozentualen Übereinstimmungen und Kappa-Werte aufweist (TH A: 112.29, also Übereinstimmungen über 100mal häufiger ist als eine Nicht-Übereinstimmung; TH B: 1496.63; TH C: 802.33).

Tabelle 30: Prozentuale Übereinstimmung und Cohens Kappa der Doppelkodierung im Leistungstest

	Prozentuale Übereinstimmung			Cohens Kappa						
	Minimum	M	SD	Minimum	1. Quartil	Median	M	SD	Anzahl Items $K < .4$	Gesamtzahl Items
Testheft A	77.3%	92.7%	7.27	-.05	.53	.65	.64	0.30	3	33
Testheft B	90.7%	97.45%	2.58	-.01	.86	.93	.87	0.23	3	35
Testheft C	88.1%	96.46%	3.61	-.02	.79	.93	.82	0.27	4	36

Im Nachgang der Doppelkodierung wurden die kritischen Fälle, also alle nicht übereinstimmenden Ratings für Items, bei denen die Übereinstimmung nach Cohens Kappa kleiner .4 war, gemeinsam gesichtet, diskutiert und die Kodierenden einigten sich auf ein Konsensurteil.

Im Rahmen der Testauswertung wird für die Leistungsdaten der Schülerinnen und Schüler auf die Parameter zurückgegriffen, die sich aus einer Raschskalierung ergeben. Damit wird hier „das derzeit wohl am häufigsten verwendete probabilistische Testmodell“ angewendet (Bortz & Döring, 2006, S. 208; vgl. auch Hecht, Roppelt &

Siegle, 2013). Die Voraussetzungen einer Raschskalierung, eine Eindimensionalität und lokale stochastische Unabhängigkeit wurden geprüft: Die Eindimensionalität des Tests wurde getrennt für die drei Testhefte geprüft, um fehlende Werte zu vermeiden. Hierbei wurde für alle drei Testhefte messzeitpunktübergreifend im Wesentlichen ein Hauptfaktor gefunden, sodass diese Voraussetzung erfüllt ist (Bühner, 2011; Luhmann, 2011; Revelle, 2017b, 2017a).

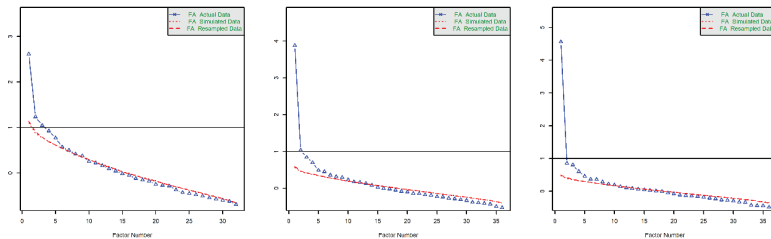


Abbildung 21: Ergebnisse der explorativen Faktorenanalyse zur Prüfung der Eindimensionalität der Testhefte (links: Testheft A; Mitte: Testheft B; rechts: Testheft C)

Tabelle 31: Auffällige Q_3 -Werte (vgl. ausführliche Darstellung im Anhang 7.2.2)

	A4.1	A4.2	C9.1	C9.2	C11.1	C11.2
A4.1		.46				
A4.2	.46					
C9.1				.59		
C9.2			.59			
C11.1						.41
C11.2					.41	

Für die Prüfung der lokalen stochastischen Unabhängigkeit wird die Q_3 -Statistik nach Yen (1993) genutzt (implementiert im Paket *TAM* für die Software *R*; Kiefer, Robitzsch & Wu, 2013, 2016). Hierbei erweisen sich lediglich drei Aufgabenpaare als auffällig mit residualen Korrelationen von $|Q_3| < 0.4$.

Bei diesen Aufgabenpaaren handelt es sich um die Aufgaben „Schultaschen“ aus Testheft 1 (A4.1 und A4.2) sowie aus dem gymnasialen Testheft die Aufgaben „Durchschnittslinie“ (C9.1 und C9.2) und „Fehler in der Gleichung“ (C11.1 und C11.2). Da diese Items jedoch jeweils den gleichen Aufgabenstamm und somit auch einen gemeinsamen Stimulus haben, werden diese Residualkorrelationen nicht als grundlegender Verstoß gegen die Annahme der lokalen stochastischen Unabhängigkeit des gesamten Testinstrumentes angesehen. Eine detailliertere Betrachtung soll kurz analysieren, worin die Abhängigkeiten liegen könnten. Für Aufgabe A4 zeigt sich eine Q_3 -

Teil 3 – Methode

Instrumente

Statistik von .46. In der Aufgabenstellung geht es um das Gewicht von Schultaschen (Abbildung 22). Die grundlegende Information für die korrekte Bearbeitung der Aufgaben liegt im Verständnis des Stimulus mit der Definition von „gut gepackt“. Teilaufgabe 1 erfordert die direkte Nutzung des „gut gepackt“-Prozentsatzes von 12%. Teilaufgabe 2 stellt eine Umkehraufgabe hierzu dar, benötigt jedoch ebenfalls die korrekte Interpretation der 12% des Körpergewichts. Insofern ist eine erhöhte Abhängigkeit dieser Items unmittelbar nachvollziehbar.

Oft tragen Schülerinnen und Schüler viel zu schwere Schultaschen. Deshalb sollte das Gewicht der Schultasche regelmäßig geprüft werden.

Eine Schultasche wird als „gut gepackt“ bezeichnet, wenn ihr Gewicht ungefähr 12 % des Körpergewichts eines Kindes nicht übersteigt.

Teilaufgabe 4.1: Schultaschen

Gib an, welches Gewicht eine „gut gepackte“ Schultasche eines 50 kg wiegenden Kindes höchstens haben sollte.

_____ kg

Teilaufgabe 4.2: Schultaschen

Eine Schultasche wiegt gefüllt 7,5 kg. Das Mädchen, das diese Schultasche trägt, wiegt 53 kg.

Ist diese Schultasche „gut gepackt“?

Kreuze an.

☐ Ja

☐ Nein

Begründe.

Abbildung 22: Aufgabe „Schultaschen“ aus Testheft 1 (Hauptschule)

Ähnlich der Aufgabe aus dem Testheft A liegt für die Aufgabe „Durchschnittslinie“ (Abbildung 23) ebenfalls eine grundlegende Information im Stimulus vor. Im Stimulus wird die durchschnittliche Höhe der gegebenen Säulen anhand einer gestrichelten Linie visualisiert. Dies muss von den Schülerinnen und Schülern verstanden sein, d. h. es muss die Verbindung hergestellt werden können, dass insbesondere die gestrichelte Linie nicht einfach nur die Höhe des mittleren Balkens darstellt. Wird dies nicht korrekt geleistet, ist es weder möglich, im direkten Vorwärtsarbeiten in Teilaufgabe 1 die Durchschnittslinie einzuzeichnen, noch in der Rückwärtsaufgabe in Teilaufgabe 2 die fehlende Säule einzuzeichnen. Potenzielle Missverständnisse könnten dadurch entstehen, dass im Stimulus lediglich drei Säulen eingezeichnet sind, der Durchschnitt

genau der mittleren Säule entspricht und die Säulen aufsteigend nach Höhe sortiert sind.

Bei der letzten Aufgabe „Fehler in der Gleichung“ handelt es sich wiederum um einen Stimulus, der beide Teilaufgaben betrifft (Abbildung 24). In diesem Falle muss anstelle eines Textes oder einer Graphik eine Gleichung nachvollzogen werden. Der Fehler in der Gleichungsumformung muss von den Schülerinnen und Schülern entdeckt werden. Finden sie diesen nicht, ist eine Bearbeitung der Teilaufgaben nicht möglich bzw. in Teilaufgabe 2 wird lediglich notiert, was im Stimulus steht.

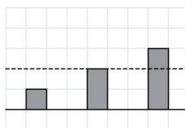
Da in der Entwicklung der Testhefte jedoch bereits ausgiebig auf die statistische Eignung der einzelnen Items geachtet wird, bleiben diese Items im Test enthalten, um nicht die Verteilung von prozessbezogenen Kompetenzen, Leitideen und Anforderungsbereichen zu verändern. Der Einsatz von Lernstandserhebungen früherer Jahre wurde schließlich bewusst gewählt, um mit dem Leistungstest möglichst nahe an den Lernstandserhebungen als Grundlage für VELM-8 zu sein.

Teil 3 – Methode

Instrumente

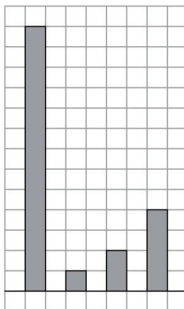
Gegeben ist ein Säulendiagramm.

Die gestrichelte Linie zeigt die durchschnittliche Höhe der Säulen.



Teilaufgabe 9.1: Durchschnittslinie

Zeichne in das folgende Säulendiagramm die durchschnittliche Höhe der Säulen mit einer waagerechten Linie ein.



Teilaufgabe 9.2: Durchschnittslinie

Zeichne in das folgende Säulendiagramm eine vierte Säule so ein, dass die gestrichelte Linie die durchschnittliche Höhe aller vier Säulen zeigt.

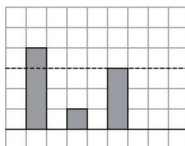


Abbildung 23: Aufgabe „Durchschnittslinie“ aus Testheft 3 (Gymnasium)

Folgende Bearbeitung einer Gleichung ist gegeben:

$$4x + 2 = 3x - 5 \quad | - 2$$

$$4x = 3x - 5 \quad | - 3x$$

$$x = - 5$$

Teilaufgabe 11.1: Fehler in der Gleichung

Markiere den Fehler in der obigen Rechnung.

Teilaufgabe 11.2: Fehler in der Gleichung

Löse diese Gleichung richtig.

$$4x + 2 = 3x - 5$$

Abbildung 24: Aufgabe „Fehler in der Gleichung“ aus Testheft 3 (Gymnasium)

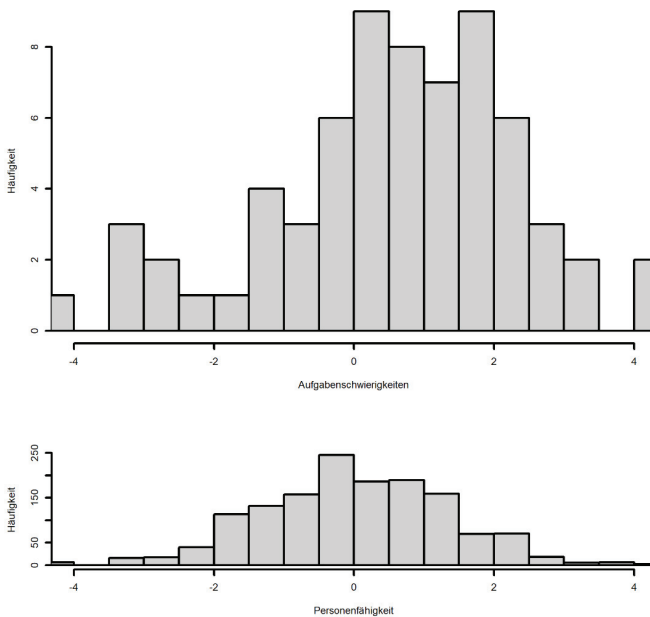


Abbildung 25: Verteilung der Aufgaben- und Personenparameter des Leistungstests bei gemeinsamer Skalierung der Messzeitpunkte und Testhefte

Die Skalierung der Aufgaben- und Personenparameter wird in einer gemeinsamen Skalierung mit dem Paket *TAM* (Robitzsch, 2017) im Programm *R* – und somit analog zur Skalierung in Pilotierungen der Vergleichsarbeiten (Penk, Pöhlmann, Roppelt, Pitsch & Pant, 2013) – vorgenommen. Die Skalierung weist eine gute EAP-Reliabilität von .88 auf. Dabei wird nach den gängigen Infit-Kriterien ($0.8 \leq \text{Infit} \leq 1.2$; Adams & Wu, 2002; OECD, 2014b; Wright & Linacre, 1994) keines der Items negativ auffällig. Der Infit aller Items liegt zwischen 0.87 und 1.19 und damit im akzeptablen Bereich (vgl. Anhang 7.2.3). Die Aufgaben weisen ein Schwierigkeitsspektrum von -4.15 bis 5.99 bei einem Mittelwert von 0.69 auf. Die Personenfähigkeiten liegen zwischen -5.57 und 6.15 bei einem Mittelwert von 0.01. Auch aus Abbildung 25 wird deutlich, dass die Verteilung der Aufgabenschwierigkeiten und der Personenfähigkeiten gut zueinander passt und der eingesetzte Test sich somit als reliables Messinstrument erweist.

3.4.2 Online-Fragebogen

Im Rahmen des Online-Fragebogens wurden diverse einzelne Items formuliert. Einige Items lassen sich dabei zu Skalen zusammenfassen, welche im Weiteren geprüft werden sollen (Reliabilitätsbetrachtungen und Faktoranalysen). Die Items des Fragebogens sind entweder dichotom (ja/nein bzw. vorhanden/nicht vorhanden), als Schulnoten (Ausprägungen 1: „sehr gut“ bis 6: „ungenügend“) oder als vier- oder fünf-stufige Likert-Skala (1: „stimme voll und ganz zu“ bis 4 bzw. 5: „stimme gar nicht zu“) gestellt. Die genaue Formulierung aller einzelnen Items ist dem Skalenhandbuch zu entnehmen (vgl. Anhang 7.1.4). Ebenfalls im Anhang finden sich weitere Informationen zu den beiden vorgestellten Skalen *Nützlichkeit* und *Schulentwicklung/Nutzung*. Alle Items stellen Selbstauskünfte der Lehrpersonen dar.

3.4.2.1 Nützlichkeit

Einige Items des Online-Fragebogens zielen auf die *Nützlichkeit*. Die Items sind im Wortlaut mit den üblichen Statistiken in Tabelle 32 abgedruckt. Die einzelnen Items weisen dabei ein fünfstufiges Antwortformat von 1: „++“ bis 5: „--“ und haben den folgenden Stimulus: „Inwieweit stimmen Sie den folgenden Aussagen zu?“. Die Vermutung des inhaltlichen Zusammenhangs der Items wird statistisch durch die hohe Reliabilität ($\alpha = .93$) sowie die hoch ausfallenden Trennschärfen und die eindeutig zu einem Faktor hinweisende Faktorenanalyse (vgl. Scree-Plot in Abbildung 26) unterstützt.

Tabelle 32: Itemanalyse Nützlichkeit

		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
prof	Die Schüler/-innen profitieren von der Durchführung der LSE.	.76	3.17	.92
paed	Von den LSE sind wichtige Impulse für die pädagogische Diskussion in meiner Schule ausgegangen.	.73	3.36	.92
qual	Die LSE leisten einen Beitrag zur langfristigen Qualitätsentwicklung an meiner Schule.	.84	3.22	.91
vergl	Die landesweite Durchführung der LSE nehme ich als Chance wahr, über den eigenen 'Tellerrand' zu blicken.	.77	2.79	.92
gst	Die LSE nehme ich als Baustein in einem Gesamtkonzept zur Unterrichts- und Schulentwicklung wahr.	.84	2.96	.91
sinn	Ich halte die LSE insgesamt für eine sinnvolle Maßnahme.	.82	3.11	.91

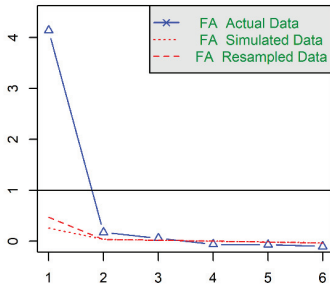


Abbildung 26: Scree-Plot Nützlichkeit

Über diese Skala hinaus wurden drei konkrete Einzelitems zur Nützlichkeit erhoben, welche ebenfalls ein fünfstufiges Antwortformat aufweisen (1: „++“ bis 5: „--“). Die Items haben einen gemeinsamen Stimulus, sind jedoch inhaltlich nicht sinnvoll zu einer Skala zusammenzufassen.

Tabelle 33: Items Nützlichkeit der Lernstandserhebung

Die Lernstandserhebungen sind ein nützliches Instrument...	<i>M</i>	<i>SD</i>
...zur Lernstandsdiagnose.	2.63	1.12
...für die Unterrichtsentwicklung.	2.90	1.08
...für die Schulentwicklung.	3.14	1.05

3.4.2.2 Schulentwicklung/Nutzung

Im Fragebogen der Lehrkräfteakademie wird ebenfalls die *Nutzung* der Lernstandserhebungen zur Schulentwicklung abgefragt. Dabei werden folgende Items unter dem gemeinsamen Stimulus „Inwiefern haben Sie Maßnahmen/Schritte aus den Lernstandserhebungen abgeleitet?“ mit einem dreistufigen Antwortformat (1: „systematisch/vollständig, 2: „punktuell/teilweise“, 3: „gar nicht“) gestellt:

Teil 3 – Methode
Instrumente

Tabelle 34: Itemanalyse Schulentwicklung/Nutzung

		<i>f</i> / <i>n</i>	<i>M</i>	α wenn Item entfernt
lerns	Veränderung der Lernschwerpunkte für den aktuellen Unterricht	.60	2.14	.87
lehrin	Änderungen der Anordnung der Unterrichtsthemen im schuleigenen Stoffverteilungsplan/Fachcurriculum	.52	2.20	.88
verlk	Veränderung der Aufgabenformen in Leistungskontrollen	.58	2.13	.87
komp	Einsatz kompetenzorientierter Unterrichtsmethoden	.70	1.76	.86
met- did	methodisch-didaktische Überlegungen zum eigenen Unterricht	.72	1.88	.86
diff	Veränderung der Differenzierungsstruktur in meiner Klasse	.68	1.90	.87
neu	Neuanschaffung moderner Lehrwerke, die kompetenzorientiertes Lernen fördern	.60	1.91	.87
fortb	Besuch von Fortbildungsveranstaltungen	.65	1.96	.87
koor	stärkere Koordination der Arbeit in den Jahrgangsstufen	.64	1.94	.87

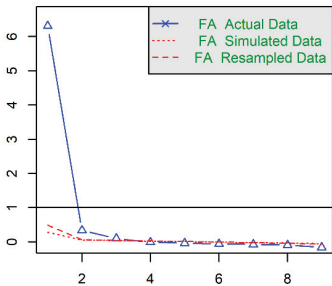


Abbildung 27: Scree-Plot Nutzung

Aus der Itemanalyse (Tabelle 34) und dem Scree-Plot der exploratorischen Faktorenanalyse (Abbildung 27) wird ersichtlich, dass die Items statistisch nachvollziehbar zusammenfassbar sind und mit einer Reliabilität von $\alpha = .95$ einen guten Wert erreichen.

3.5 Analysemethoden

3.5.1 VELM-8

3.5.1.1 Projektevaluation

Im Rahmen der Auswertung werden verschiedene methodische Zugänge gewählt. Durch den wiederholten Einsatz der identischen Instrumente zu Prä- und Posttest sowie die vorliegende Panel-Stichprobe bietet sich eine messwiederholte Varianzanalyse (mANOVA) zur Auswertung an – vor allem vor dem Hintergrund des recht geringen Stichprobenumfanges der Evaluationsstichprobe (Eid et al., 2011; Girden, 1992; Keselman et al., 1998; Pospeschill, 2006). Mit dieser Auswertungsmethode werden die Hypothesen 1 bis 22 untersucht. Hierzu wird das Paket *ez* in *R* genutzt (Lawrence, 2016; Luhmann, 2011). Bortz und Döring (2006, S. 560) weisen darauf hin, dass bei der messwiederholten Varianzanalyse die Voraussetzung des gleichen Ausgangsniveaus zu beachten ist. Entsprechend wird die Aussagekraft bei zu großen Prätestunterschieden zwischen den zu betrachtenden Gruppen in Frage gestellt. Alternativ wird eine Kovarianzanalyse (ANCOVA) vorgeschlagen (Baguley, 2012; Wildt & Ahtola, 1978⁵⁸). Entsprechend wird vor der Bearbeitung der Hypothesen 1 bis 22 jeweils die Voraussetzung von Prätestunterschieden zwischen Kontroll- und Projektgruppe mithilfe von *t*-Tests für unabhängige Stichproben geprüft (Bortz & Döring, 2006; Bühner & Ziegler, 2012). Aufgrund der geringen Stichprobengröße wird als Entscheidungsgrundlage neben der Signifikanz ebenfalls die Effektstärke *Cohens d* für Mittelwertunterschiede herangezogen (J. Cohen, 1988)⁵⁹. Sofern ein signifikanter *t*-Test oder eine Effektstärke für den Mittelwertunterschied im Prätest von $|d| > .2$ vorliegt, werden die zugehörigen Hypothesen mittels einer ANCOVA zusätzlich abgesichert. Neben der Suche nach signifikanten Ergebnissen werden zusätzlich Effektstärken berichtet, wie beispielsweise η^2 (mANOVA) bzw. β (ANCOVA) (Bakeman & Gottman, 2010; Bortz & Döring, 2006; Cortina & Nouri, 2000; Eid et al., 2011).

⁵⁸ Die Verfahren mANOVA und ANCOVA sind in der Interpretation jedoch keinesfalls deckungsgleich: Während die messwiederholte Varianzanalyse (mANOVA) in diesem Falle untersucht, ob es zwischen den beiden Gruppen eine unterschiedliche Entwicklung zwischen Prä- und Posttest gab, betrachtet die Kovarianzanalyse unter Berücksichtigung der Prätest-Werte Gruppenunterschiede im Posttest. Insofern ist die messwiederholte Varianzanalyse prinzipiell das für diese Untersuchung passendere Analyseverfahren, sofern die Voraussetzungen erfüllt sind.

⁵⁹ Für die Berechnung der Effektstärke wird das Paket *compute.es* genutzt (AC Del Re, 2015).

Teil 3 – Methode

Analysemethoden

Bevor die Auswertungen durch *t*-Tests oder mANOVA bzw. ANCOVA durchgeführt werden, werden die Voraussetzungen 1) Normalverteilung und 2) Varianzhomogenität geprüft. Dabei wird auf den Kolmogorow-Smirnow-Lilliefors-Test aus dem Paket *norstest* (Gross & Ligges, 2015) und den Levene-Test aus dem Paket *car* (Fox et al., 2016) zurückgegriffen (Luhmann, 2011). Sofern in den Auswertungen nicht anders vermerkt, wurden diese Voraussetzungen geprüft und zeigen keine Auffälligkeiten, sodass Normalverteilung und Varianzhomogenität (auf üblichem 5%-Niveau) anzunehmen sind. Für die ANCOVA wird die jeweilige Prätest-Einstellung zentriert und vergleichend ein Modell mit und ohne Interaktionsterm (Zeit x Gruppe) berechnet. Beim Vergleich der Modelle wird das komplexere Modell mit Interaktionsterm lediglich dann berichtet, wenn der Verzicht auf den Interaktionsterm zu signifikanten Einbußen in der aufgeklärten Varianz führt (dies ist in Fußnoten vermerkt; Crawley, 2013).

In Abschnitt 3.4 wurde bereits angedeutet, dass für einige der eingesetzten Skalen die Annahme der Normalverteilung nicht bestätigt werden kann. Daher werden für diese Skalen die Ergebnisse der üblichen parametrischen Verfahren (mANOVA und ANCOVA) durch non-parametrische Verfahren abgesichert (konkret eine rangtransformierte Variante der messwiederholten Varianzanalyse und Kovarianzanalyse: als Abkürzungen im Weiteren rmANOVA und rANCOVA). Diese Verfahren sind voraussetzungsärmer und weniger restriktiv in den Annahmen (Bortz, Lienert & Boehnke, 2008; Gibbons, 1993; Lienert, 1978, 1986; Olejnik & Algina, 1985; Stemmler, 1998; Toothaker & Newman, 1994; Wobbrock, Findlater, Gergle, Higgins & Kay, 2016). Die konkret eingesetzte rangtransformierte Varianzanalyse wird von diversen Autoren dem verallgemeinerten Vorzeichentest nach Friedman vorgezogen, welche wiederum lediglich die Analyse verbundener Stichproben jedoch keine Gruppenvergleiche ermöglicht (Bortz et al., 2008; Conover & Iman, 1981; Pospeschill, 2006; Thom, 2012; D. Zimmerman & Zumbo, 1993).

Tabelle 35 zeigt eine Übersicht der eingesetzten Auswertungsverfahren für die Hypothesen 1 bis 22, aufbauend auf den nachfolgend berichteten Prätestunterschieden bzw. Verletzungen der Voraussetzungen. Für die Skala *Diagnose im sozialen Bereich* fällt die Reliabilität leider problematisch aus, sodass hier ebenfalls auf non-parametrische Verfahren zurückgegriffen wird, obwohl die Voraussetzungen der

parametrischen erfüllt sind. Genauere Ausführungen sowie die Einführung von Hypothesen „a“ und „b“ wird auf den folgenden Seiten erläutert.

Für die einstellungsbezogenen Skalen zeigt sich vor allem die Skala *Lehrplanvalidität* in den Voraussetzungen der parametrischen Verfahren kritisch: Die Normalverteilungsannahme ist zu beiden Messzeitpunkten verletzt (Kolmogorow-Smirnow-Lilliefors-Test: $D_{\text{Prätest}} = 0.21, p < .05$; $D_{\text{Posttest}} = 0.18, p < .05$), die Annahme von Varianzhomogenität zum Posttest (Levene-Test auf Varianzhomogenität: $F_{\text{Posttest}}(1,23) = 4.30, p < .05$). Entsprechend wird für diese Skala auf non-parametrische Verfahren zurückgegriffen.

Tabelle 35: Übersicht der Auswertungsverfahren nach Hypothesen unter Prüfung von Prätestunterschieden und Voraussetzungen der parametrischen Verfahren

H	Formulierung	mANOVA	ANCOVA	rmANOVA	rANCOVA
1a	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, eine höhere Akzeptanz gegenüber den Lernstandserhebungen zu haben als Lehrkräfte der Kontrollgruppe.				
1b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einstellung eine höhere Akzeptanz der Lernstandserhebungen.				
2	Im Verlauf der Projektteilnahme schätzen Lehrkräfte der Projektgruppe die wahrgenommene Lehrplanvalidität der Lernstandserhebungen positiver ein als Lehrkräfte der Kontrollgruppe.				
3a	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe eine geringere wahrgenommene Belastung durch die Lernstandserhebungen an als Lehrkräfte der Kontrollgruppe.				
3b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Belastung eine geringere Belastung durch die Lernstandserhebungen.				
4a	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des förderdiagnostischen Nutzens der Lernstandserhebungen für die Lehrkräfte der Projektgruppe positiver als für die Kontrollgruppe.				
4b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einschätzung eine höhere Einschätzung des förderdiagnostischen Nutzens.				
5a	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des unterrichtsplanerischen Nutzens der Lernstandserhebungen für die Lehrkräfte der Projektgruppe positiver als für die Kontrollgruppe.				
5b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einschätzung eine höhere Einschätzung des unterrichtsplanerischen Nutzens.				
6	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des selektionsdiagnostischen Nutzens der Lernstandserhebungen für die Lehrkräfte der Projektgruppe negativer als für die Kontrollgruppe.				
7	Im Verlauf der Projektteilnahme entwickelt sich die selbst wahrgenommene Häufigkeit des Insistierens auf Begründungen und Erklärungen für Projektgruppenlehrkräfte positiver als für Kontrollgruppenlehrkräfte.				
8	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger Aufgaben mit Variation in der Aufgabenschwierigkeit einzusetzen als Lehrkräfte der Kontrollgruppe.				
9	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger kognitiv anspruchsvolle Aufgaben zu nutzen als Lehrkräfte der Kontrollgruppe.				
10	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger kognitiv herausforderndes Üben einzusetzen als Lehrkräfte der Kontrollgruppe.				
11a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, Ihre Schülerinnen und Schüler seltener repetitiv Üben zu lassen als Lehrkräfte der Kontrollgruppe.				
11b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, weniger repetitives Üben zu nutzen.				

H	Formulierung	mANOVA	ANCOVA	rmANOVA	rANCOVA
12a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger Schülerbeiträge (Ideen/Fehler) als Ausgangspunkt für den weiteren Unterricht zu nutzen als Lehrkräfte der Kontrollgruppe.				
12b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, häufiger Schülerbeiträge (Ideen/Fehler) als Ausgangspunkt für den weiteren Unterricht zu nutzen.				
13a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, seltener fehlervermeidende Erarbeitungphasen durch sich selbst zu gestalten als Lehrkräfte der Kontrollgruppe.				
13b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, seltener fehlervermeidende Erarbeitungsphasen durch die Lehrkraft selbst zu gestalten.				
14a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, weniger bzw. seltener Kleinschrittige, fürsorgliche Anleitung und Kontrolle zu geben als Lehrkräfte der Kontrollgruppe.				
14b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, seltener Kleinschrittige, fürsorgliche Anleitung und Kontrolle zu geben.				
15a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, Schülerinnen und Schüler stärker in der kognitiven Selbstständigkeit zu unterstützen als Lehrkräfte der Kontrollgruppe.				
15b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, Schülerinnen und Schüler stärker in deren kognitiven Selbstständigkeit zu unterstützen.				
16	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger leistungsbezogene Diagnose durchzuführen als Lehrkräfte der Kontrollgruppe.				
17	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, größeren Wert auf eine gemeinsame Wissenskonstruktion zu legen als Lehrkräfte der Kontrollgruppe.				
18	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, Lernpartnerschaften mehr zu unterstützen als Lehrkräfte der Kontrollgruppe.				
19a	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, größeren Wert auf persönliches Vertrauen zu legen als Lehrkräfte der Kontrollgruppe.				
19b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, größeren Wert auf persönliches Vertrauen zu legen.				
20	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, häufiger Diagnosen im sozialen Bereich durchzuführen als Lehrkräfte der Kontrollgruppe.				
21	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, häufiger nach individueller Bezugsnorm zu loben als Lehrkräfte der Kontrollgruppe.				
22a	Die Schülerinnen und Schüler der Projektgruppenlehrkräfte zeigen einen größeren Leistungszuwachs im Verlauf des Projektes gegenüber den Schülerinnen und Schülern der Kontrollgruppenlehrkräfte.				
22b	Nach Abschluss des Projektes zeigen die Schülerinnen und Schüler der Projektgruppenlehrkräfte gegenüber denen der Kontrollgruppenlehrkräfte nach Kontrolle der Prätestleistung eine höhere Leistung.				

Teil 3 – Methode

Analysemethoden

Aus Tabelle 36 lässt sich ablesen, dass keiner der durchgeführten *t*-Tests einen signifikanten Mittelwertunterschied anzeigt. Betrachtet man jedoch die Effektstärke *Cohens d*, zeigen sich vor allem bei den beiden Skalen *Akzeptanz* und *förderdiagnostischer Nutzen* Unterschiede mit mittleren Effekten von $d = -0.39$ bzw. $d = -0.42$. Bei den Skalen *unterrichtsplanerischer Nutzen* sowie *Belastung* liegen mit $d = 0.26$ bzw. $d = -0.22$ kleine Effekte vor. Damit scheint es sich um zwar möglicherweise zufällige, jedoch im Ausmaß m. E. bedeutsame Unterschiede zu handeln. Aus diesem Grund werden die Ergebnisse der messwiederholten Varianzanalyse für diese vier genannten Skalen durch eine Kovarianzanalyse abgesichert. Für die Skala *Lehrplanvalidität* ergibt der Mann-Whitney-Test keinen signifikanten Prätestunterschied ($Z = -1.07$, $p = .29$).

Tabelle 36: Mittelwerte, *t*-Test und Cohens *d* für Einstellungsskalen (Stichprobe 1) im Prätest

Skala	<i>M</i> _{Projektgruppe}	<i>M</i> _{Kontrollgruppe}	<i>t</i>	<i>df</i>	<i>p</i>	Cohens <i>d</i>	Konfidenzintervall Cohens <i>d</i>	
Akzeptanz (H1)	3.26	2.93	-0.96	24	.35	-0.39	-1.23	0.45
Belastung (H3)	2.23	2.08	-0.55	24	.59	-0.22	-1.05	0.61
Förderung (H4)	3.48	3.18	-1.04	23	.31	-0.42	-1.26	0.42
Unterrichtsplanung (H5)	3.12	3.31	0.71	23	.49	0.26	-0.57	1.10
Selektion (H6)	2.49	2.43	-0.22	23	.83	-0.09	-0.92	0.74

Bezogen auf die technischen Unterschiede zwischen den Verfahren *MANOVA* und *ANCOVA* sollten an dieser Stelle die Hypothesen zu den auffälligen Skalen korrekterweise leicht angepasst werden. Die ursprünglichen Formulierungen aus Abschnitt 3.1.1.1 zusammengefasst in Tabelle 10 können unter Hypothese 1a, 2a, 4a und 5a weitergeführt werden. Die angepassten Hypothesen lauten:

- Hypothese 1b: Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einstellung eine höhere *Akzeptanz der Lernstandserhebungen*.
- Hypothese 3b: Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Belastung eine geringere *Belastung durch die Lernstandserhebungen*.
- Hypothese 4b: Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einschätzung eine höhere *Einschätzung des förderdiagnostischen Nutzens*.

- Hypothese 5b: Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einschätzung eine höhere Einschätzung des *unterrichtsplanerischen Nutzens*.

Für die Skalen zur Unterrichtsqualität ergibt sich bei der Prüfung der Voraussetzung parametrischer Verfahren folgendes Bild für die Normalverteilungs- und Varianzhomogenitätsannahme (Tabelle 37), wobei lediglich Skalen aufgeführt sind, die ein signifikantes Testergebnis aufzeigen, also mindestens eine der Voraussetzungen verletzen. Anstelle der parametrischen Verfahren wird bei jenen Skalen non-parametrisch vorgegangen, die zu beiden Messzeitpunkten die Normalverteilungsannahme verletzen oder mindestens zu einem Zeitpunkt Varianzheterogenität vorliegt (H12, H13, H14, H15, H16, H19).

Tabelle 37: Überprüfung der Voraussetzungen für t-Test und messwiederholte Varianzanalyse

Skala	Normalverteilung		Varianzhomogenität	
Insistieren auf Begründungen (H7)	$D_{\text{Prätest}} = .18$ $p = .002$			
Differenzierung durch variierende Aufgabenschwierigkeit (H8)		$D_{\text{Posttest}} = .178$ $p < .05$		
Repetitives Üben (H11)		$D_{\text{Posttest}} = .174$ $p = .05$		
Kognitiv herausfordernder Umgang mit Schülerbeiträgen (H12)	$D_{\text{Prätest}} = .02$ $p = .005$	$D_{\text{Posttest}} = .280$ $p < .001$		
Fehlervermeidende Erarbeitung durch den Lehrer selbst (H13)	$D_{\text{Prätest}} = .21$ $p = .002$		$F_{\text{Prätest}}(1,24) = 6.05$ $p = .02$	
Fürsorgliche Anleitung und Kontrolle(H14)			$F_{\text{Prätest}}(1,24) = 2.97$ $p < .10$	$F_{\text{Posttest}}(1,21) = 4.65$ $p = .04$
Unterstützung kognitiver Selbstständigkeit (H15)	$D_{\text{Prätest}} = .240$ $p < .001$		$F_{\text{Prätest}}(1,24) = 2.98$ $p < .10$	
Gemeinsame Wissenskonstruktion (H17)				$F_{\text{Posttest}}(1,21) = 3.44$ $p = .08$
Tutorielle Unterstützung von Lernpartnerschaften (H18)		$D_{\text{Posttest}} = .159$ $p = .08$		
Persönliches Vertrauen (H19)	$D_{\text{Prätest}} = .353$ $p < .001$			$F_{\text{Posttest}}(1,21) = 5.94$ $p = .02$
Diagnose im sozialen Bereich (H20)	$D_{\text{Prätest}} = .223$ $p = .002$			

Mit Blick auf Prätestunterschiede lässt sich zunächst ableiten, dass anstelle der t-Tests auf Prätestunterschiede für die Skalen mit signifikantem Kolmogorow-Smirnow-Lilliefors-Test, also nicht anzunehmender Normalverteilung, der Mann-

Teil 3 – Methode Analysemethoden

Whitney-Test⁶⁰ durchgeführt wird. Sofern die Normalverteilung anzunehmen ist, jedoch Varianzheterogenität vorliegt (etwa für die Skala *Fürsorgliche Anleitung und Kontrolle*), wird auf den Welch-Test zurückgegriffen.

Tabelle 38: Ergebnisse der Suche nach Prätestunterschieden für Unterrichtsqualität mittels t-Test und Welch-Test

Skala	<i>M</i> _{Projektgruppe}	<i>M</i> _{Kontrollgruppe}	<i>t</i>	<i>df</i>	<i>p</i>	Cohens <i>d</i>	Konfidenzintervall Cohens <i>d</i>	
Kognitiv aktivierende Aufgaben (H9)	2.77	2.74	-0.13	24	.89	-0.05	-0.89	0.78
Kognitiv aktivierendes Üben (H10)	2.83	2.84	0.06	24	.96	0.02	-0.81	0.85
Repetitives Üben (H11)	2.11	2.78	2.85	24	.01	1.15	0.26	2.05
Fürsorgliche Anleitung ⁶¹ (H14)	2.81	3.05	1.75	24	.09	0.71	-0.15	1.56
Gemeinsame Wissenskonstruktion (H17)	2.22	2.33	0.44	24	.67	0.18	-0.66	1.01
Lernpartnerschaften (H18)	2.96	2.88	-0.36	24	.72	-0.15	-0.98	0.69

Tabelle 39: Ergebnisse der Suche nach Vortestunterschieden mittels Mann-Whitney-Test

Skala	<i>M</i> _{Projektgruppe}	<i>M</i> _{Kontrollgruppe}	<i>Z</i>	<i>p</i>	Cohens <i>d</i>	Konfidenzintervall Cohens <i>d</i>	
Begründungen (H7)	3.05	3.15	0.40	.69	0.16	-0.67	1.00
Variation Schwierigkeit (H8)	2.73	2.70	-0.19	.85	-0.06	-0.89	0.78
Schülerbeiträge (H12)	2.85	2.67	-1.28	.20	-0.38	-1.18	0.45
Erarbeitung LPS (H13)	2.31	2.88	2.46	.01	1.13	0.23	2.02
Kognitive Selbstständigkeit (H15)	3.77	3.45	-2.18	.03	-1.05	-1.94	-0.17
Persönliches Vertrauen (H19)	3.71	3.57	-0.72	.47	-0.31	-1.14	0.53
Soziale Diagnose (H20)	3.09	3.08	0.28	.78	-0.04	-0.87	0.80

Bei der Suche nach Prätestunterschieden wird auch für die Skalen zur Unterrichtsqualität neben der reinen Signifikanzprüfung das Kriterium Cohens $|d| > .2$ herangezogen. Aus den Ergebnissen dieser Tests (Tabelle 38 und Tabelle 39) lässt sich erkennen, dass mehrere Skalen mit Hilfe einer ACNOVA bzw. rANCOVA ergänzend geprüft werden sollten. Analog zu den Skalen der Lernstandserhebungseinstellungen sollen hierfür die Hypothesenformulierungen mit Blick auf die technischen Unterschiede zwischen den Verfahren angepasst werden. Die ursprünglichen Formulierungen aus Abschnitt 3.1.1.1 zusammengefasst in Tabelle 11 können mit dem Zusatz „a“ weitergeführt werden. Die angepassten Hypothesen lauten wie folgt:

⁶⁰ Durchgeführt mit dem Paket *coin* (Hothorn, Hornik, van de Wiel, Mark, Winell & Zeileis, 2017; Hothorn, Hornik, van de Wiel, Mark & Zeileis, 2008).

⁶¹ Welch-Test

- Hypothese 11b: Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, weniger *repetitives Üben* zu nutzen.
- Hypothese 12b: Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, häufiger *Schülerbeiträge (Ideen/Fehler)* als Ausgangspunkt für den weiteren Unterricht zu nutzen.
- Hypothese 13b: Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, seltener *fehlervermeidende Erarbeitungsphasen* durch die Lehrkraft selbst zu gestalten.
- Hypothese 14b: Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, seltener *kleinschrittige, fürsorgliche Anleitung und Kontrolle* zu geben.
- Hypothese 15b: Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, Schülerinnen und Schüler stärker in deren *kognitiven Selbstständigkeit zu unterstützen*.
- Hypothese 19b: Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, größeren Wert auf *persönliches Vertrauen* zu legen.

Die Hypothese 22 postuliert eine bessere Leistung der Schülerinnen und Schüler, deren Lehrkräften an VELM-8 teilgenommen haben. Zur Überprüfung der Prätestunterschiede wurde ein Welch-Test durchgeführt, da die Varianzhomogenität nicht gewährleistet ist ($F(1,29) = 5.073$, $p = .03$ im Levene-Test). Für die Untersuchung der Prätestunterschiede liegen Daten von 31 Lehrkräften vor, darunter 20 Projekt- und entsprechend 11 Kontrollgruppenlehrkräfte. Aus Tabelle 40 lässt sich ablesen, dass sich die Mittelwerte zum Prätest signifikant und in großem Ausmaß voneinander unterscheiden. Entsprechend werden die Ergebnisse der messwiederholten Varianzanalyse durch eine Kovarianzanalyse abgesichert. Die aktualisierte Hypothese 22b lautet folglich: Nach Abschluss des Projektes zeigen die Schülerinnen und Schüler der Projektgruppenlehrkräfte gegenüber denen der Kontrollgruppenlehrkräfte nach Kontrolle der Prätestleistung eine höhere Leistung.

Tabelle 40: Prätestunterschiede der Schülerleistung Welch-Test

Skala	$M_{\text{Projektgruppe}}$	$M_{\text{Kontrollgruppe}}$	t	p	Cohens d	Konfidenzinter- vall Cohens d	
Leistung (H22)	-0.06	-0.70	-2.31	.03	-0.87	-1.67	-0.07

3.5.1.2 Typenbildung

Der Typenbildung liegt zunächst die Vorstellung von Punktwolken in einem n -dimensionalen Raum zugrunde. Dieser Raum hat als theoretische Basis im linear-algebraischen Sinn die zugrundeliegenden Skalen zu den Einstellungen der Lehrkräfte gegenüber den Lernstandserhebungen. Diese Einstellungen sind als theoretische Annahme zunächst voneinander unabhängig und damit orthogonal. Das *Ward*-Verfahren zielt dabei darauf ab, Cluster so zu bilden, dass die Streuung zwischen den Clusterzentren maximiert wird (u. a. Backhaus, Erichson, Plinke & Weiber, 2016). Dieses Vorgehen wird von Bacher, Pöge und Wenzig (2010) anderen Methoden insbesondere bei kleinen Stichproben vorgezogen. Des Weiteren gibt es die Möglichkeit Verschmelzungen nach dem Kriterium *average* und *complete* durchzuführen, bei denen der durchschnittliche bzw. summierte Abstand aller bisherigen Datenpunkte des Clusters berücksichtigt wird. Hierbei können verschiedene Metriken als Abstandsmaß eingesetzt werden (Hatzinger, Hornik, Nagel & Maier, 2014; Kaufman & Rousseeuw, 1990; Maechler et al., 2017). Als Maß für den „Abstand“ der Cluster wird die Manhattan-Metrik gewählt, nach der der „Abstand“ zweier Punkte die Summe der Komponentendifferenzen darstellt: $d(a, b) = \sum_i |a_i - b_i|$ (vgl. u. a. Bosch, 2006).

Neben der Entscheidung für die Clusteranzahl mit Hilfe der grafischen Darstellung der Verschmelzungshöhen im Dendrogramm (sog. *Elbow*-Kriterium ähnlich dem Scree-Plot der Faktorenanalysen; Backhaus et al., 2016) liefern einige Indizes Empfehlungen. Von Milligan und Cooper (1985; Milligan, 1981) wird vor allem die Nutzung der folgenden Indizes empfohlen, welche jedoch eine Abhängigkeit von den zugrundeliegenden Daten zeigen und insofern nur bedingt objektiv einzuschätzen sind: 1) *C-Index*, 2) *Gamma*, 3) *PTbiserial Index* und mit 4) *G Plus*. Backhaus et al. (2016) empfehlen zusätzlich unter Bezug auf Milligan (1980) das Kriterium *Calinski/Harabasz* (Calinski & Harabasz, 1974; Milligan & Cooper, 1985). Die Berechnung dieser Indizes wird mit dem Paket *NbClust* vorgenommen (Charrad, Ghazzali, Boiteau & Niknafs, 2014; Charrad, Ghazzali, Boiteau & Niknafs, 2015). Da die Power-Analyse von Baker und Hubert (1975) für den Index *Gamma* sich jedoch auf *single*- und *complete-linking*-Clusterverfahren bezieht, wird auf diesen Index verzichtet. Ebenfalls einen Hinweis

auf die Güte der Clusterlösung liefert der Dunn-Index zwischen 0 und maximal 1 (Brock, Pihur, Datta & Datta, 2008; Dunn, 1974; Hatzinger et al., 2014).

Als elaborierte Methoden neben der Clusteranalyse (Aldenderfer & Blashfield, 1984; Kaufman & Rousseeuw, 1990) würden sich auch die bereits in Abschnitt 1.6 angesprochenen *mixed-Rasch-Modelle* (Abschnitt 1.6.5) oder auch latente Klassenanalysen (LCA; Clogg, 1995; Hagenaars & McCutcheon, 2002; Kleine, 2005; McCutcheon, 1987; Vermunt & Magidson, 2002) anbieten. Aufgrund der geringen Stichprobengröße und der hohen Anzahl zu schätzender Parameter ist eine Analyse mit diesen Methoden jedoch nicht sinnvoll. In Ermangelung latenter Modellierungsverfahren bzw. Verfahren der probabilistischen Testtheorie bietet die Konfigurationsfrequenzanalyse als Verfahren der Dimensionsreduktion eine Möglichkeit „den Begriff des Typus statistisch, anstatt ihn intuitiv und deskriptiv zu definieren“ (Lienert, 1971, S. 110; Stemmler, Lautsch & Martinke, 2008). Dabei werden die tatsächlichen Auftretenshäufigkeiten mit den zu erwartenden Häufigkeiten aller möglichen Konfigurationen bzw. Merkmalskombinationen abgeglichen (Eye & Brandstädter, 1997; Eye, 2001; Krauth, 1993; Lienert, 1971; Stemmler, 2014). Werden beispielsweise die Merkmale Geschlecht und Schuhgröße untersucht, wobei letztere dichotomisiert wurde, sodass „größer als 40“ und „kleiner oder gleich 40“ als Merkmale auftreten, ergeben sich also insgesamt ($2 \times 2 =$) vier mögliche Konfigurationen.

Konfigurationen, welche in der Stichprobe deutlich häufiger auftreten als nach zugrundeliegendem statistischem Modell erwartet, werden als *Typen* bezeichnet; Konfigurationen, die deutlich seltener auftreten, als *Antitypen* (Eye, 1998). Entsprechend werden lediglich diese Konfigurationen zur Interpretation herangezogen, für die die Nullhypothese der totalen Unabhängigkeit aller Merkmale verworfen werden muss (Lienert, 1971). Für die Berechnung wurde das Paket *confreq* (Heine, 2016) für das Programm *R* (Hatzinger et al., 2014; Luhmann, 2011) eingesetzt.

Kritisch anzumerken ist bei diesem Verfahren jedoch, dass die vermeintlich metrischen Variablen von Skalenmittelwerten⁶² für die Analyse in diskrete Kategorien unterteilt werden müssen (Krauth, 1993). Im Falle der vorgenommenen Auswertungen

⁶² Für die Berechnung von Skalenmittelwerten wird im Vorfeld bereits angenommen, dass die mit Likert-Skalen erfragten Items Intervall-Niveau aufweisen, was vielfach bezweifelt wird, sondern eher von Ordinalskalen ausgegangen wird. Für Ordinalskalen wäre jedoch die Berechnung von Mittelwerten nicht zulässig (Bortz & Döring, 2006; Eid, Gollwitzer & Schmitt, 2011).

Teil 3 – Methode

Analysemethoden

werden die Merkmale in zwei Merkmalsstufen unterteilt: gesplittet 1) am empirischen Median sowie 2) am semantischen Median. In diesem Zusammenhang diskutieren beispielsweise Maxwell und Delaney (1993) Probleme bei der Dichotomisierung von stetigen Merkmalen. Eine weitere Einschränkung ergibt sich aus der recht geringen Stichprobengröße. Als Datengrundlage dienen die beschriebenen sechs Skalen zu den Lernstandserhebungen. Werden diese dichotomisiert liegen also $2^6 = 64$ mögliche Konfigurationen vor, sodass Größe der Stichprobe und Anzahl möglicher Konfigurationen fast identisch sind. Daher werden zwei Konfigurationsfrequenzanalysen durchgeführt: Eine mit den Skalen zur *Akzeptanz*, der wahrgenommenen *Lehrplanvalidität* und der resultierenden *Belastung* eher bezüglich der Wahrnehmung der Lernstandserhebungen sowie eine mit den eher nutzungsorientierten Skalen zum *unterrichtsplanerischen*, der *förderdiagnostischen* und der *selektionsdiagnostischen Nutzen* der Lernstandserhebungen. Da jeweils lediglich drei Skalen vorliegen, ergeben sich entsprechend $2^3 = 8$ mögliche Konfigurationen.

Im Anschluss an die Typenbildung wird mit Hilfe von Varianzanalysen sowie χ^2 -Tests versucht, die Typen voneinander statistisch abgrenzend zu beschreiben.

3.5.2 Online-Fragebogen

Einzelne Items des Online-Fragebogens werden zunächst mit Hilfe von deskriptiven Statistiken betrachtet und anschließend werden auf der Suche nach Unterschieden zwischen den Fächern mit Hilfe von t -Tests, Varianzanalysen sowie χ^2 -Tests näher betrachtet. Für die Hypothesen, die eine Vorhersage betreffen werden lineare Regressionen eingesetzt.

4 Ergebnisse

In diesem Teil werden die Ergebnisse der Auswertungen beschrieben. Dabei behandelt der erste Abschnitt 4.1 die Hypothesen 1 bis 22, die der Evaluation des Projekts VELM-8 dienen (Forschungsfrage 2). Im Anschluss werden die Ergebnisse der Typenbildung und deren Interpretation vorgestellt (4.2; Forschungsfrage 3), bevor zuletzt die Hypothesen 23 bis 29 bearbeitet werden, die sich mit der projektunabhängigen Einschätzung und Wahrnehmung der LSE beschäftigen (4.3, Forschungsfrage 1). Abschließend wird eine Übersicht aller Hypothesen bzw. Ergebnisse präsentiert (4.4). Diese Zusammenfassung sowie die vorherige detaillierte Darstellung der Ergebnisse in 4.1 und 4.3 sind dabei möglichst deskriptiv und ohne direkte Interpretation oder Einordnung zu verstehen. Diese folgen in der anschließenden Diskussion.

4.1 VELM-8

Im ersten Abschnitt zur Projektevaluation wird der Vergleich zwischen Projekt- und Kontrollgruppe berichtet und gliedert sich in erster Ebene inhaltlich (Abschnitt 4.1.1 zu den Einstellungen gegenüber den LSE, Hypothesen 1 bis 6; Abschnitt 4.1.2 zu den Einschätzungen der Unterrichtsqualität, Hypothesen 7 bis 21; Abschnitt 4.1.3 zur Leistungsentwicklung der Schülerinnen und Schüler, Hypothese 22). In der nächsttieferen Ebene sind die Abschnitte nach den eingesetzten Verfahren (jeweils parametrische und rangtransformierte messwiederholte Varianzanalyse sowie Kovarianzanalyse) gegliedert. In dieser Gliederung liegt begründet, dass Hypothesen mehrfach aufgegriffen werden, so beispielsweise in der mANOVA und der rmANOVA.

4.1.1 Einstellungsskalen

Zunächst soll betrachtet werden, ob sich die Einstellungen gegenüber den Lernstandserhebungen im Verlauf des Projekts VELM-8 zwischen Projekt- und Kontrollgruppe in der Entwicklung unterscheiden (Hypothesen 1 bis 6). Dabei werden als erste Analysen die messwiederholten Varianzanalysen der Panel-Stichprobe 1 berichtet (Abschnitt 4.1.1.1). Direkt im Anschluss werden diese Ergebnisse vor dem Hintergrund von Prätestunterschieden durch Kovarianzanalysen der Stichprobe 1 ergänzt (Abschnitt 4.1.1.2). Hieran schließen sich die Ergebnisse der non-parametrischen Analysen an (Abschnitt 4.1.1.3).

4.1.1.1 Messwiederholte Varianzanalysen zu den Einstellungsskalen

Zur Durchführung der messwiederholten Varianzanalyse müssen aufgrund von fehlenden Werten einige Fälle ausgeschlossen werden. Dabei wurden die Skalen jeweils separat behandelt, sodass die jeweiligen Stichproben leicht variieren. Die deskriptiven Ergebnisse legen bereits durch bloße Sichtkontrolle eine gering eingeschätzte *Belastung durch die Lernstandserhebungen*, eine hoch eingeschätzte *Lehrplanvalidität der Lernstandserhebungen* sowie eine zustimmende Haltung gegenüber dem *förderdiagnostischen Nutzen* und eine eher ablehnende Haltung gegenüber dem *selektionsdiagnostischen Nutzen* nahe (vgl. Tabelle 41). Die Skalen *Akzeptanz* und *unterrichtsplanerischer Nutzen* bewegen sich vergleichsweise nahe dem semantischen Median von 3. Ebenfalls erkennbar ist, dass sich die Projektgruppe zwischen Prä- und Posttest durch eine wünschenswerte Entwicklung kennzeichnet: Zum Posttest liegen die Mittelwerte der Skalen *Akzeptanz*, *Lehrplanvalidität*, *förderdiagnostischer Nutzen* und *unterrichtsplanerischer Nutzen* trotz des z. T. bereits anfänglich hohen Niveaus über den Prätest-Mittelwerten. Diese Tendenz lässt sich für die Kontrollgruppe nicht durchgängig erkennen.

Tabelle 41: Deskriptive Statistiken der Skalen zu beiden Messzeitpunkten (Stichprobe 1)

Skala	Gruppe	MZP	N	M	SD	FLSD ⁶³
Akzeptanz (H1a)	Kontrollgruppe	Prätest	9	2.86	0.77	0.41
		Posttest	9	3.04	0.88	0.41
	Projektgruppe	Prätest	16	3.26	0.94	0.41
		Posttest	16	3.61	0.65	0.41
Validität⁶⁴ (H2)	Kontrollgruppe	Prätest	8	3.69	0.46	0.41
		Posttest	8	3.75	0.94	0.41
	Projektgruppe	Prätest	12	3.81	0.55	0.41
		Posttest	12	4.12	0.35	0.41
Belastung (H3a)	Kontrollgruppe	Prätest	9	2.06	0.77	0.48
		Posttest	9	2.29	0.74	0.48
	Projektgruppe	Prätest	16	2.23	0.66	0.48
		Posttest	16	1.84	0.57	0.48
Förderung (H4a)	Kontrollgruppe	Prätest	9	3.18	0.73	0.39
		Posttest	9	3.22	0.76	0.39
	Projektgruppe	Prätest	15	3.48	0.69	0.39
		Posttest	15	3.8	0.53	0.39
Planung (H5a)	Kontrollgruppe	Prätest	9	3.24	0.52	0.49
		Posttest	9	3.11	0.87	0.49
	Projektgruppe	Prätest	15	3.12	0.85	0.49
		Posttest	15	3.57	0.66	0.49
Selektion (H6)	Kontrollgruppe	Prätest	9	2.47	0.6	0.43
		Posttest	9	2.73	0.93	0.43
	Projektgruppe	Prätest	15	2.49	0.81	0.43
		Posttest	15	2.68	0.64	0.43

In der nachfolgenden Tabelle 42 sind die Ergebnisse der messwiederholten Varianzanalyse für die Hypothesen 1a bis 6a dargestellt. Im Anschluss an die Tabelle sind die Mittelwerte grafisch dargestellt (Abbildung 28 bis Abbildung 32). Auf dem gängigen 5%-Niveau zeigen sich weder Haupteffekt der Gruppe oder Zeit noch Interaktionseffekte. Folgende Effekte sollen dennoch hervorgehoben werden:

- Für die *Akzeptanz* zeigt sich ein marginal signifikanter mittlerer Effekt der Zeit ($F(1,23) = 4.18, p = .05, \eta^2 = .03$). Der Einfluss der Gruppe scheint – ausgehend von der mittleren Effektstärke – zwar bedeutsam, wird jedoch nicht signifikant ($F(1,23) = 2.57, p = .12, \eta^2 = .09$).

⁶³ Fisher Least Significant Difference; nach Luhmann (2011, S. 211) α -Fehler kontrollierte Möglichkeit für Paarvergleiche bei genau drei Mittelwerten. Dies wird hier der Vollständigkeit mitberichtet, da es als Indiz für signifikante Differenzen genutzt werden kann.

⁶⁴ Die Ergebnisse werden der Vollständigkeit halber berichtet, jedoch nicht interpretiert.

Teil 4 – Ergebnisse

Projektevaluation

- Für die *Belastung durch die Lernstandserhebung* zeigt sich ein marginal signifikanter mittlerer Interaktionseffekt Gruppe x Zeit ($F(1,23) = 3.35$, $p = .08$, $\eta^2 = .05$).
- Für den *förderdiagnostischen Nutzen* ergibt sich ein marginal signifikanter mittlerer Effekt der Gruppenzugehörigkeit ($F(1,22) = 3.26$, $p = .08$, $\eta^2 = .10$). Der Effekt für die Zeit verfehlt das marginale Signifikanzniveau nur knapp ($F(1,22) = 2.70$, $p = .11$, $\eta^2 = .03$).
- Für den *unterrichtsplanerischen Nutzen* verfehlt der Interaktionsterm das marginale Signifikanzniveau ebenfalls nur knapp ($F(1, 22) = 2.85$, $p = .11$, $\eta^2 = .03$).

Lediglich für die *Belastung* kann ein marginal signifikanter Interaktionsterm identifiziert werden. Für den *unterrichtsplanerischen Nutzen* verfehlt der Interaktionsterm das marginale Signifikanzniveau knapp. Dementsprechend lassen sich an dieser Stelle Effekte der Fortbildung auf einem 5%-igen Signifikanzniveau nicht bestätigen und die Hypothesen 1a, 3a, 4a, 5a und 6 müssen zunächst abgelehnt werden. Ihre grafische Entsprechung finden diese beiden Effekte (Hypothese 3a und 5a) in den Abbildungen 31 und 33 durch die sich kreuzenden Linien, also darin, dass die Entwicklungen der beiden Gruppen im Verlauf der Zeit gegenläufig ist: Während die Projektgruppe beispielsweise im Projektverlauf von einer abnehmenden Belastung berichtet, fallen die Belastungswerte der Kontrollgruppe zu Projektabschluss höher aus als zu Projektbeginn. Für den *unterrichtsplanerischen Nutzen* zeigen sich im Projektverlauf höhere Werte für die Projekt- und niedrigere Werte für die Kontrollgruppe. Als grafische Entsprechung der Haupteffekte Zeit und Gruppe zeigen sich in den Abbildungen eine gleichförmige Entwicklung beider Gruppen (Haupteffekt Zeit, etwa bei der Skala *Akzeptanz*; Abbildung 30) bzw. große Abstände zwischen den Gruppenwerten (Haupteffekt Gruppe, etwa bei der Skala *förderdiagnostischer Nutzen*; Abbildung 32).

Tabelle 42: Ergebnisse mANOVA Skalen zu Lernstandserhebungen (H1a bis H6)

Skala	Effekt	d/fn	d/d	Quadratsumme Effekt	Quadratsumme Fehler	F	p	p < .05	generalisiertes η^2
Akzeptanz (H1a)	(Intercept)	1	23	532.36	24.78	494.14	<.001	*	.95
	Gruppe	1	23	2.77	24.78	2.57	.12		.08
	Zeit	1	23	1.03	5.68	4.18	.05		.03
	Gruppe x Zeit	1	23	0.09	5.68	0.35	.56		<.01
Belastung (H3a)	(Intercept)	1	23	217.78	12.78	391.89	<.001	*	.91
	Gruppe	1	23	0.20	12.78	0.37	.55		.01
	Zeit	1	23	0.34	7.70	1.03	.32		.02
	Gruppe x Zeit	1	23	1.12	7.70	3.35	.08		.05
Förderung (H4a)	(Intercept)	1	22	578.75	14.73	864.61	<.001	*	.97
	Gruppe	1	22	2.18	14.73	3.26	.08		.10
	Zeit	1	22	0.58	4.69	2.70	.11		.03
	Gruppe x Zeit	1	22	0.22	4.69	1.05	.32		.01
Planung (H5a)	(Intercept)	1	22	516.67	17.06	666.36	<.001	*	.95
	Gruppe	1	22	0.33	17.06	0.42	.52		.01
	Zeit	1	22	0.67	7.33	2.00	.17		.03
	Gruppe x Zeit	1	22	0.95	7.33	2.85	.11		.04
Selektion (H6)	(Intercept)	1	22	322.75	18.95	374.63	<.001	*	.93
	Gruppe	1	22	0.00	18.95	0.00	.95		<.01
	Zeit	1	22	0.55	5.66	2.13	.16		.02
	Gruppe x Zeit	1	22	0.01	5.66	0.06	.81		<.01

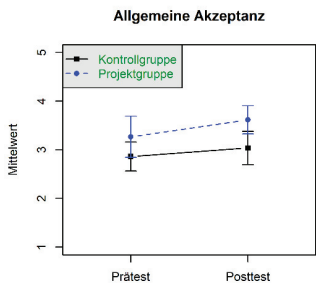


Abbildung 28: Allgemeine Akzeptanz in Prä- und Posttest getrennt nach Gruppen (H1a)

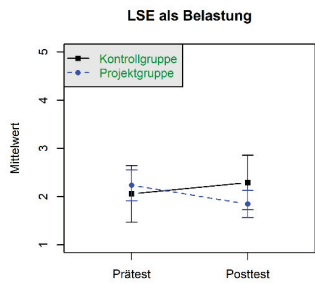


Abbildung 29: LSE als Belastung in Prä- und Posttest getrennt nach Gruppen (H3a)

Teil 4 – Ergebnisse

Projektevaluation

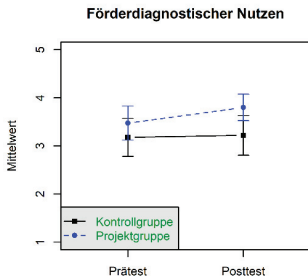


Abbildung 30: Förderdiagnostischer Nutzen in Prä- und Posttest nach Gruppen (H4a)

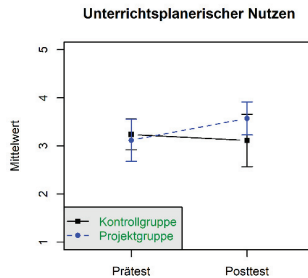


Abbildung 31: Unterrichtsplanerischer Nutzen in Prä- und Posttest nach Gruppen (H5a)

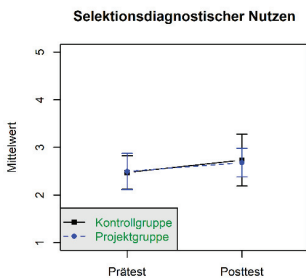


Abbildung 32: Selektionsdiagnostischer Nutzen in Prä- und Posttest nach Gruppen (H6)

4.1.1.2 Kovarianzanalysen zu den Einstellungsskalen

Folgend werden die Ergebnisse der Kovarianzanalyse zu den in Prätestunterschieden auffälligen Skalen als Absicherung der messwiederholten Varianzanalyse dargestellt (Hypothesen 1b, 3b, 4b und 5b).

Für die Skala *Akzeptanz* zeigt sich ein signifikanter Einfluss der Interaktion ($b = -0.62$, $p = .04$). Dabei klärt das Modell gut die Hälfte an Varianz auf ($R^2 = .52$). Der negative Regressionskoeffizient des Interaktionsterms bedeutet also, dass für die Projektgruppe der Einfluss der Prätestleistung geringer ausfällt als für die Kontrollgruppe. Ein Haupteffekt der Gruppe, wie in Hypothese 1b postuliert, lässt sich jedoch nicht bestätigen.

Tabelle 43: ANCOVA für *Allgemeine Akzeptanz (H1b)*⁶⁵

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	3.34	0.19	17.23	<.001
Prätest-Einstellung	0.99	0.25	4.00	<.001
Gruppe	0.24	0.24	1.02	.32
Prätest-Einstellung * Gruppe	-0.62	0.29	-2.16	.04

adjustiertes $R^2 = .52$

Die Skala *LSE als Belastung* zeigt bei der kovarianzanalytischen Betrachtung einen marginal signifikanten negativen Effekt auf ($b = -0.49$, $p = .08$) der Gruppe bei einer sehr geringen Varianzaufklärung von lediglich 9%: Insofern lässt sich die Hypothese 3b unter Vorbehalt des nur marginalen Signifikanzniveaus bestätigen.

Tabelle 44: ANCOVA für *LSE als Belastung (H3b)*⁶⁶

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	2.31	0.21	10.98	<.001
Prätest-Einstellung	0.23	0.19	1.21	.24
Gruppe	-0.49	0.26	-1.84	.08

adjustiertes $R^2 = .09$

Tabelle 45: ANCOVA für *förderdiagnostischen Nutzen (H4b)*⁶⁷

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	3.30	0.18	17.91	<.001
Prätest-Einstellung	0.46	0.17	2.80	.01
Gruppe	0.44	0.24	1.88	.07

adjustiertes $R^2 = .35$

In der Betrachtung der Skala *förderdiagnostischer Nutzen* zeigt sich entsprechend der messwiederholten Varianzanalyse ein positiver marginal signifikanter Einfluss der Gruppe auf die Einschätzung zum zweiten Messzeitpunkt ($b = 0.44$, $p < .10$) bei einer Varianzaufklärung von $R^2 = .35$. Mit dem Wissen des lediglich marginal signifikanten Ergebnisses lässt sich Hypothese 4b jedoch bestätigen.

Eine vergleichbar geringe Varianzaufklärung wie für die *Belastung* weist das Ergebnis für die Skala *unterrichtsplanerischer Nutzen* mit lediglich 16% auf. Der positive Einfluss der Gruppe wird in dieser Auswertung mit einem p -Wert von .10 knapp nicht

⁶⁵ Der Verzicht auf den Interaktionsterm im Modell führt zu einem signifikanten Verlust an Erklärungskraft: $F = 4.65$, $p < .05$.

⁶⁶ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.55$, $p = .46$.

⁶⁷ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 1.84$, $p = .19$.

marginal signifikant ($b = 0.51$). Entsprechend muss die Hypothese 5b verworfen werden.

Tabelle 46: ANCOVA für *unterrichtsplanerischer Nutzen* (H5b)⁶⁸

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	3.10	0.23	13.33	<.01
Prätest-Einstellung	0.40	0.20	1.99	.06
Gruppe	0.51	0.30	1.72	.10

adjustiertes $R^2 = .16$

4.1.1.3 Rangtransformierte, messwiederholte Varianzanalysen (rmANOVA)
zu den Einstellungsskalen

Die Skala *Lehrplanvalidität* fällt in der Prüfung der Voraussetzung der parametrischen Verfahren negativ auf⁶⁹. Daher analysiere ich Hypothese 2 zur Validität mittels einer rangtransformierten, messwiederholten Varianzanalyse (non-parametrisch). Hierbei wurde die Rangtransformation so vorgenommen, dass hohe (zustimmende) Werte auf den Skalen auch für hohe Ränge sprechen, damit die Orientierung beibehalten wird.

Tabelle 47: Deskriptive Beschreibung der Rangtransformation der Skala *Lehrplanvalidität*

Skala	Gruppe	MZP	N	M	SD	FLSD
Validität (H2)	Kontrollgruppe	Prätest	9	11.75	6.23	5.79
	Kontrollgruppe	Posttest	9	14.00	9.01	5.79
	Projektgruppe	Prätest	15	12.83	7.47	5.79
	Projektgruppe	Posttest	15	19.00	4.84	5.79

Analog zum Vorgehen der parametrischen mANOVA zeigt Tabelle 47 die deskriptiven Statistiken und Tabelle 48 die Ergebnisse der rangtransformierten mANOVA: Es zeigt sich ein signifikanter Effekt der Zeit ($F(1,18) = 5.57, p = .03, \eta^2 = .11$; vgl. auch Abbildung 33). Die *Lehrplanvalidität* nimmt also gruppenübergreifend im Verlauf des Projektzeitraumes zu, was gleichzeitig bedeutet, dass Hypothese 2 abgelehnt werden muss. Anhand der Abbildung lässt sich erkennen, dass die Projektgruppe zwar eine deutlichere Entwicklung zeigt, allerdings lässt sich dieser Unterschied nicht statistisch

⁶⁸ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 1.94, p = .18$.

⁶⁹ Die anderen Skalen zu Einstellungen gegenüber den Lernstandserhebungen wurden ebenfalls non-parametrisch analysiert, um die Ergebnisse der parametrischen Verfahren zu prüfen. Die Ergebnisse finden sich im Anhang (Anhang 7.3.1.). Hierbei unterstreichen die Ergebnisse im Wesentlichen die parametrischen Befunde. Abweichungen gibt es z. T. im Signifikanzniveau. Ein bemerkenswerter Unterschied zeigt sich für die Skala *Allgemeine Akzeptanz*, für die ein marginal signifikant Interaktionseffekt deutlich wird.

vom Zufall unterscheiden und eine Interpretation wäre unangebracht (Interaktions-
effekt nicht signifikant).

Tabelle 48: Ergebnisse der rangtransformierten, messwiederholten Varianzanalyse für die Skala *Lehrplanvalidität* (H2)

Skala	Effekt	dfn	dfd	Quadrat- summe Effekt	Quadrat- summe Fehler	F	p	p < .05	generalisier- tes η^2
Validität (H2)	(Intercept)	1	18	8643.60	1027.58	151.41	<.01	*	.83
	Gruppe	1	18	88.82	1027.58	1.56	.23		.05
	Zeit	1	18	211.60	683.58	5.57	.03	*	.11
	Gruppe x Zeit	1	18						
				36.82	683.58	0.97	.34		.02

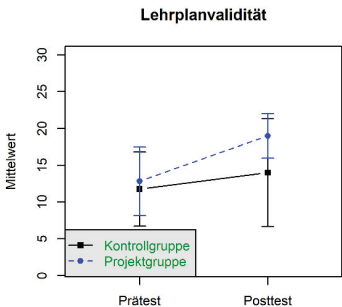


Abbildung 33: Rangmittelwerte *Lehrplanvalidität* im Prä- und Posttest nach Gruppen (H2)

4.1.2 Auswertungen zur Unterrichtsqualität

Auf den folgenden Seiten werden die Hypothesen zur Unterrichtsqualität betrachtet. Dabei werden die Ergebnisse und Beschreibungen jeweils für alle eingesetzten Skalen zusammen präsentiert, weshalb sich vergleichsweise lange Tabellen anschließen. Die Abbildungen zu den Entwicklungen der Projekt- und Kontrollgruppe im Projektverlauf folgen ebenfalls ohne zwischenstehende Erläuterungen aufeinander. Der Abschnitt gliedert sich dabei analog zum vorherigen nach den eingesetzten Analyseverfahren, sodass zunächst alle Hypothesen untersucht werden, deren Auswertung mittels messwiederholter Varianzanalyse erfolgt (vgl. Tabelle 35). Nachfolgend werden die Ergebnisse zur ANCOVA (Kapitel 4.1.2.2) präsentiert, woran sich die non-parametrischen Auswertungsergebnisse anschließen (Kapitel 4.1.2.3 und 4.1.2.4). Explizit beschreibend wird nur auf relevante Ergebnisse eingegangen, die beispielsweise besonders erwartungskonform oder überraschend ausfallen. Auf die Darstellung

Teil 4 – Ergebnisse Projektevaluation

deskriptiver Statistiken wird aus Gründen der Lesefreundlichkeit verzichtet und auf den Anhang verwiesen (vgl. Abschnitt Anhang 7.3.2).

4.1.2.1 Messwiederholte Varianzanalysen zu den Skalen der Unterrichtsqualität

Tabelle 49: Ergebnisse der mANOVA zu Skalen der Unterrichtsqualität (Hypothesen 7, 8, 9, 10, 11a und 18)

Skala	Effekt	dfn	dfd	Quadratsumme Effekt	Quadratsumme Fehler	F	p	p < .05	generalisiertes η^2
Begründungen (H7)	(Intercept)	1	21	420.02	11.99	735.55	<.01	*	.97
	Gruppe	1	21	0.17	11.99	0.30	.59		.01
	Zeit	1	21	0.09	2.47	0.74	.40		.01
	Gruppe x Zeit	1	21	0.51	2.47	4.33	.05	*	.03
Variation Schwierigkeit (H8)	(Intercept)	1	21	347.33	9.79	745.29	<.01	*	.97
	Gruppe	1	21	0.00	9.79	0.00	.98		<.01
	Zeit	1	21	0.00	1.91	0.00	.97		<.01
	Gruppe x Zeit	1	21	0.01	1.91	0.12	.73		<.01
kA Aufgaben (H9)	(Intercept)	1	24	389.04	11.35	822.63	<.01	*	.97
	Gruppe	1	24	0.06	11.35	0.12	.73		<.01
	Zeit	1	24	0.02	1.20	0.40	.53		<.01
	Gruppe x Zeit	1	24	0.02	1.20	0.41	.53		<.01
kA Üben (H10)	(Intercept)	1	21	387.11	5.88	1383.48	<.01	*	.98
	Gruppe	1	21	0.00	5.88	0.01	.92		<.01
	Zeit	1	21	0.16	1.23	2.67	.12		.02
	Gruppe x Zeit	1	21	0.00	1.23	0.01	.90		<.01
Repetitives Üben (H11a)	(Intercept)	1	21	269.30	9.78	577.97	<.01	*	.96
	Gruppe	1	21	5.66	9.78	12.15	<.01	*	.35
	Zeit	1	21	0.03	0.94	0.59	.45		<.01
	Gruppe x Zeit	1	21	0.02	0.94	0.37	.55		<.01
Lernpartnerschaften (H18)	(Intercept)	1	21	389.96	11.45	714.99	<.01	*	.97
	Gruppe	1	21	0.21	11.45	0.39	.54		.02
	Zeit	1	21	0.06	1.26	0.93	.35		<.01
	Gruppe x Zeit	1	21	0.47	1.26	7.83	.01	*	.04

Die Ergebnisse aus Tabelle 49 zeigen an, dass keine der betrachteten Skalen einen signifikanten Haupteffekt der Zeit aufweist. Signifikante Effekte bzgl. der Gruppenzugehörigkeit zeigen sich bei der Skala *repetitives Üben* (vgl. Abbildung 38; $F(1,21) = 12.15$, $p = .002$, $\eta^2 = .35$). Die Kontrollgruppe berichtet, dieses Üben deutlich häufiger einzusetzen als die Projektgruppe. Insofern ist jedoch kein Effekt der Fortbildung zu erkennen. Die Skalen *Variationen in der Aufgabenschwierigkeit*, *kognitive aktivierende Aufgaben* und *kognitiv aktivierendes Üben* zeigen keinerlei

signifikante Effekte, sodass die Hypothesen H8, H9, 10 und 11a abgelehnt werden müssen. Bezüglich der formulierten Hypothesen lassen sich bei zwei Skalen signifikante Interaktionseffekte zwischen Zeit und Gruppenzugehörigkeit erkennen:

- Für die Skala *Insistieren auf Begründungen* deutet Abbildung 34 eine positive Entwicklung der Projektgruppe bei gleichzeitiger negativer Entwicklung der Kontrollgruppe an. Diese Entwicklung wird in der Varianzanalyse signifikant mit einem kleinen Effekt: $F(1,21) = 4.33, p < .05, \eta^2 = .03$). Entsprechend lässt sich die in der Hypothese 7 postuliert Entwicklung bestätigen.
- Ebenfalls zwar nur mit einem kleinen, aber deutlich signifikanten Effekt unterscheiden sich die Entwicklungen der Projekt- und Kontrollgruppe bzgl. der *Tutoriellen Unterstützung in Lernpartnerschaften und Gruppen* ($F(1,21) = 7.83, p = .01, \eta^2 = .04$; Abbildung 39). Hier zeigt die Kontrollgruppe eine negative Entwicklung bei leicht positiver Entwicklung der Projektgruppe. Insofern kann auch die Hypothese 18 vorläufig bestätigt werden.

4.1.2.2 Kovarianzanalysen zu den Skalen der Unterrichtsqualität

Betrachtet man die Skala, welche die Voraussetzungen parametrischer Verfahren erfüllt, jedoch bedeutsamen Unterschiede im Prätest aufweist, so zeigt sich in der Kovarianzanalyse folgendes Ergebnis: Trotz der Kontrolle der Einschätzung im Prätest zur Nutzung von *einschleifendem, repetitivem Üben* zeigt sich der Einfluss der Gruppe dennoch hochsignifikant bei einer enormen Varianzaufklärung von fast 90% ($b = -.35, p = .002$; vgl. Tabelle 50). Die Projektgruppe weist geringere Werte auf als die Kontrollgruppe (negativer Regressionskoeffizient). Insofern kann die Hypothese 11b bestätigt werden. Aus der hohen Varianzaufklärung wird deutlich, dass die Einschätzungen zum Posttest innerhalb der Gruppen sehr homogen sind ($SD_{\text{Projektgruppe}} = .31; SD_{\text{Kontrollgruppe}} = .49$).

Tabelle 50: ANCOVA für *Einschleifendes repetitives Üben* (H11b)⁷⁰

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	2.68	0.08	35.59	<.01
Prätest-Einstellung	0.60	0.07	8.51	<.01
Gruppe	-0.36	0.10	-3.65	<.01

adjustiertes $R^2 = .88$

⁷⁰ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 1.40; p = .25$.

Teil 4 – Ergebnisse Projektevaluation

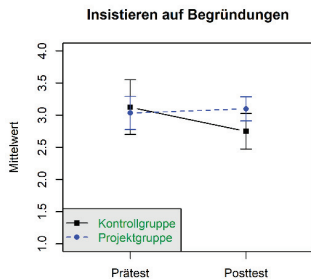


Abbildung 34: Insistieren auf Begründungen im Prä- und Posttest nach Gruppen (H7)

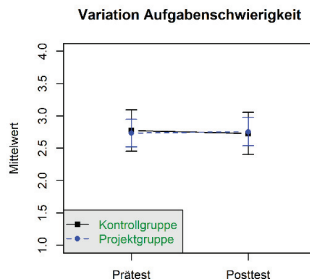


Abbildung 35: Variation der Aufgabenschwierigkeit im Prä- und Posttest nach Gruppen (H8)

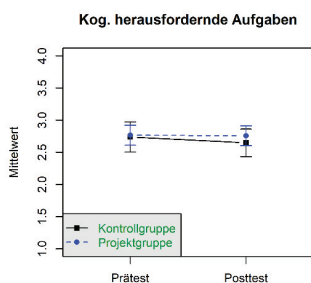


Abbildung 36: Kognitiv herausfordernde Aufgaben im Prä- und Posttest nach Gruppen (H9)

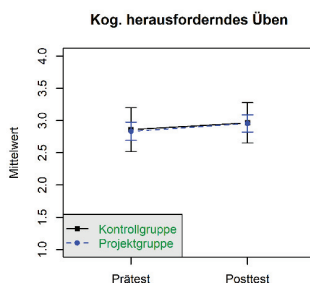


Abbildung 37: Kognitiv herausforderndes Üben im Prä- und Posttest nach Gruppen (H10)

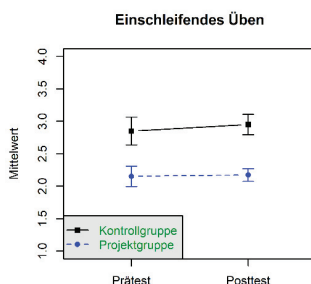


Abbildung 38: Einschleifendes Üben im Prä- und Posttest Gruppen (H11a)

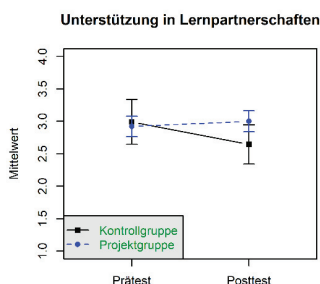


Abbildung 39: Tutorielle Unterstützung in Lernpartnerschaften und Gruppen im Prä- und Posttest nach Gruppen (H18)

4.1.2.3 Rangtransformierte, messwiederholte Varianzanalysen zu den Skalen der Unterrichtsqualität

Für Skalen, die die Voraussetzungen parametrischer Verfahren nicht erfüllen, wird nun auf non-parametrische Analysen zurückgegriffen. Analog zum bisherigen

Vorgehen liefert Tabelle 51 einen Überblick über die Ergebnisse der rmANOVA (für deskriptive Statistiken vgl. Anhang). Aus dieser Tabelle lassen sich für einige Skalen signifikante Effekte ablesen. Auf dem Faktor Zeit ergeben sich für folgende Skalen (marginal) signifikante Effekte (vgl. auch Abbildung 40 bis Abbildung 46):

- *Fehlervermeidende Erarbeitung durch den Lehrer selbst*: $F(1,21) = 2.77$, $p = .11$, $\eta^2 = .04$; Abnahme über die Zeit
- *Persönliches Vertrauen*: $F(1,21) = 3.98$, $p = .06$, $\eta^2 = .06$; Abnahme über die Zeit

Für den Effekt der Gruppe zeigen sich bei zwei Skalen signifikante große Effekte. In beiden Fällen schneidet die Projektgruppe wünschenswerter ab (also weniger Erarbeitung durch die Lehrkraft und mehr Unterstützung kognitiver Selbstständigkeit):

- *Fehlervermeidende Erarbeitung durch den Lehrer selbst*: $F(1,21) = 10.73$, $p < .01$, $\eta^2 = .27$ (Projektgruppe mit geringeren Rängen als Kontrollgruppe)
- *Unterstützung kognitiver Selbstständigkeit*: $F(1,21) = 6.56$, $p = .02$, $\eta^2 = .18$ (Projektgruppe mit höheren Rängen als Kontrollgruppe)

Unter Berücksichtigung der Hypothesen liegt ein besonderes Interesse auf den Interaktionseffekten der Varianzanalyse: Für die Skala *Diagnose im sozialen Bereich* zeigt sich ein marginal signifikanter Effekt Gruppe x Zeit mit $F(1,21) = 3.62$, $p = .07$, $\eta^2 = .04$, welcher für die gegenläufige Entwicklung der beiden Gruppen zugunsten der Projektgruppe spricht (vgl. Abbildung 46). Hypothese 20 kann demzufolge bestätigt werden.

Teil 4 – Ergebnisse
Projektevaluation

Tabelle 51: Ergebnisse der rangtransformierten, messwiederholten Varianzanalyse

Skala									
	Effekt	dfn	dfd	Quadratsumme Effekt	Quadratsumme Fehler	F	p		generalisiertes η^2
Schülerbeiträge (H12a)	(Intercept)	1	21	11776	1362.4	181.51	<.01	*	0.82
	Gruppe	1	21	34.6	1362.4	0.53	.47		0.01
	Zeit	1	21	1.39	1245.8	0.02	.88		0
	Gruppe x Zeit	1	21	31.8	1245.8	0.54	.47		0.01
Erarbeitung LPS (H13a)	(Intercept)	1	21	9156.54	1363.3	141.05	<.01	*	0.83
	Gruppe	1	21	696.65	1363.3	10.73	<.01	*	0.27
	Zeit	1	21	70.63	535.1	2.77	.11		0.04
	Gruppe x Zeit	1	21	0.77	535.1	0.03	.86		0
Fürsorgliche Anleitung (H14a)	(Intercept)	1	21	9527.04	1911.47	104.67	<.01	*	0.81
	Gruppe	1	21	63.49	1911.47	0.7	.41		0.03
	Zeit	1	21	5.57	318.55	0.37	.55		0
	Gruppe x Zeit	1	21	25.88	318.55	1.71	.21		0.01
Kognitive Selbststän- digkeit (H15a)	(Intercept)	1	21	10927.85	1528.87	150.1	<.01	*	0.83
	Gruppe	1	21	477.79	1528.87	6.56	.02	*	0.18
	Zeit	1	21	33.07	681.42	1.02	.32		0.01
	Gruppe x Zeit	1	21	0.02	681.42	0	.98		0
Soziale Konstruktion (H17)	(Intercept)	1	21	8600.89	2082.9	86.71	<.01	*	0.79
	Gruppe	1	21	3.7	2082.9	0.04	.85		0
	Zeit	1	21	2.63	226.24	0.24	.63		0
	Gruppe x Zeit	1	21	16.63	226.24	1.54	.23		0.01
Persönliches Vertrauen (H19a)	(Intercept)	1	21	12755.57	1740.42	153.91	<.01	*	0.83
	Gruppe	1	21	40.02	1740.42	0.48	.49		0.01
	Zeit	1	21	168.35	909.47	3.89	.06		0.06
	Gruppe x Zeit	1	21	0.19	909.47	0	.95		0
Soziale Diagnose (H20)	(Intercept)	1	21	10260.2	1710.47	125.97	<.01	*	0.82
	Gruppe	1	21	91.84	1710.47	1.13	.30		0.04
	Zeit	1	21	29.76	471.42	1.33	.26		0.01
	Gruppe x Zeit	1	21	81.32	471.42	3.62	.07		0.04

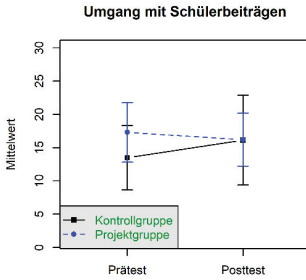


Abbildung 40: Rangmittelwerte *Umgang mit Schülerbeiträgen* im Prä- und Posttest nach Gruppen (H12a)

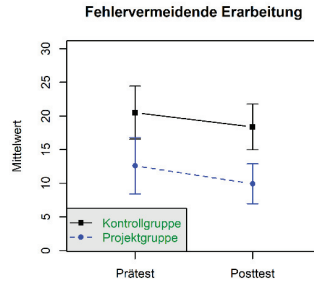


Abbildung 41: Rangmittelwerte *Fehlervermeidende Erarbeitung* im Prä- und Posttest nach Gruppen (H13a)

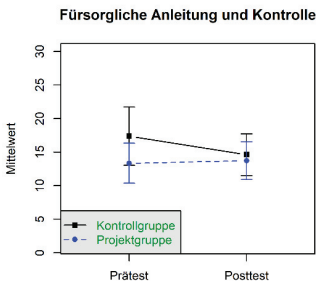


Abbildung 42: Rangmittelwerte *Fürsorgliche Anleitung und Kontrolle* im Prä- und Posttest nach Gruppen (H14a)

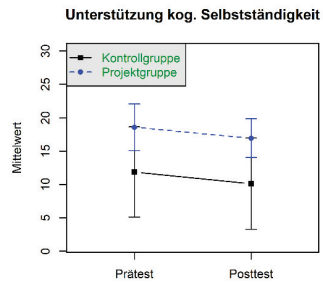


Abbildung 43: Rangmittelwerte *Unterstützung kognitiver Selbstständigkeit* im Prä- und Posttest nach Gruppen (H15a)

Teil 4 – Ergebnisse

Projektevaluation

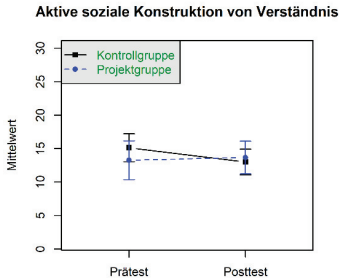


Abbildung 44: Rangmittelwerte Soziale Konstruktion von Verständnis im Prä- und Posttest nach Gruppen (H17)

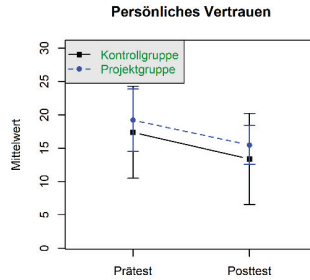


Abbildung 45: Rangmittelwerte Persönliches Vertrauen im Prä- und Posttest nach Gruppen (H19a)

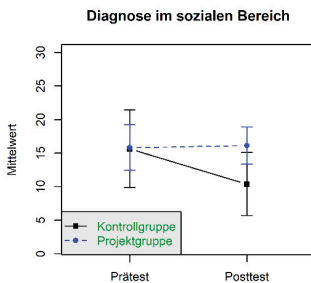


Abbildung 46: Rangmittelwerte Diagnose im sozialen Bereich im Prä- und Posttest nach Gruppen (H20)

4.1.2.4 Rangtransformierte Kovarianzanalysen zu den Skalen der Unterrichtsqualität

Im Nachgang werden die Ergebnisse der Skalen/Hypothesen präsentiert, die sowohl erhebliche Prätest-Unterschiede aufweisen als auch die Voraussetzungen der parametrischen Verfahren verletzen. Um die Hypothesen zu bestätigen oder zu verwerfen wird hier nun eine rangtransformierte ANCOVA (RANCOVA) herangezogen.

Für die Skala *Fehlervermeidende Erarbeitung* zeigt sich sowohl für die Prätest-Einstellung als auch die Gruppe eine signifikante Vorhersagekraft auf die Posttest-Einstellung (Tabelle 53). Die Posttestleistung unterscheidet sich also zwischen den Gruppen signifikant, insofern lässt sich die Hypothese 13b bestätigen, wobei das adjustierte R^2 allerdings die Interpretierbarkeit stark einschränkt. Sowohl die Skala *fürsorgliche*

Anleitung und Kontrolle als auch *Unterstützung kognitiver Selbstständigkeit* weisen lediglich einen signifikanten Einfluss der Prätest-Einstellungen auf. Die postulierten Gruppenunterschiede lassen sich nicht bestätigen und konsequent müssen die Hypothesen 14b und 15b abgelehnt werden (Tabelle 54 und Tabelle 55). Für die Hypothesen 12b und 19b zeigen Tabelle 52 und Tabelle 56 ebenfalls keinen signifikanten Einfluss der Gruppe auf die Posttest-Einstellung. Die Hypothesen 12b und 19b müssen daher abgelehnt werden.

Tabelle 52: rANCOVA für Schülerbeiträge (Ideen/Fehler) als Ausgangspunkt für weiteren Unterricht (H12b)⁷¹

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	12.89	2.32	5.56	<.01
Prätest-Rang	0.10	0.19	0.53	.61
Gruppe	4.18	2.91	1.44	.17

adjustiertes $R^2 = -.10$

Tabelle 53: rANCOVA für Fehlervermeidende Erarbeitung durch den Lehrer selbst (H13b)⁷²

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	16.53	2.00	8.25	<.01
Prätest-Rang	0.34	0.15	2.29	.03
Gruppe	-5.76	2.56	-2.25	.04

adjustiertes $R^2 = .44$

Tabelle 54: rANCOVA für Fürsorgliche Anleitung und Kontrolle (H14b)⁷³

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	12.97	1.77	7.34	<.01
Prätest-Rang	0.64	0.14	4.62	<.01
Gruppe	1.71	2.21	0.77	.45

adjustiertes $R^2 = .47$

⁷¹ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.56$, $p = .46$.

⁷² Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.70$, $p = .41$.

⁷³ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.63$, $p = .44$.

Tabelle 55: rANCOVA für *Unterstützung kognitiver Selbstständigkeit (H15b)*⁷⁴

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	11.65	2.46	4.74	<.01
Prätest-Rang	0.35	0.19	1.87	.08
Gruppe	4.44	3.14	1.41	.17

adjustiertes $R^2 = .24$

Tabelle 56: rANCOVA für *Persönliches Vertrauen (H19b)*⁷⁵

	<i>b</i>	<i>SE</i>	<i>t</i>	<i>p</i>
Intercept	14.17	2.43	5.82	<.01
Prätest-Rang	0.42	0.29	1.44	.17
Gruppe	-0.34	3.02	-0.11	.91

adjustiertes $R^2 = .04$

4.1.3 Schülerleistung

Die Hypothese 22 postuliert eine bessere Leistung der Schülerinnen und Schüler, deren Lehrkräfte an VELM-8 teilgenommen haben. Nach der Darstellung der deskriptiven Werte (Tabelle 57) zeigt Tabelle 58, dass die Leistung signifikant und zu großem Maße von der Gruppe sowie hochsignifikant in mittlerem Maße von der Zeit abhängig ist. Allerdings ergibt sich keine Interaktion zwischen Zeit und Gruppe, sodass die Entwicklungen in beiden Gruppen (nahezu) parallel verlaufen (vgl. Abbildung 47). Auch die Untersuchung mit Hilfe der Kovarianzanalyse ergibt einen hochsignifikanten Effekt der Prätestleistung, jedoch keinen signifikanten Effekt für die Gruppenzugehörigkeit (vgl. Tabelle 59).

Tabelle 57: Deskriptive Statistiken der Skalen zu beiden Messzeitpunkten mit reduzierter Stichprobe 1

Skala	Gruppe	MZP	<i>N</i>	<i>M</i>	<i>SD</i>	FLSD
Leistung (H22)	Kontrollgruppe	Prätest	13	-0.83	0.67	0.24
		Posttest	13	-0.35	0.98	0.24
	Projektgruppe	Prätest	21	-0.10	0.93	0.24
		Posttest	21	0.40	1.19	0.24

⁷⁴ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.00, p = 1.00$.

⁷⁵ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.85, p = .37$.

Tabelle 58: Ergebnisse mANOVA (Hypothese 22a)

Skala	Effekt	dfn	dfd	Quadratsumme Effekt	Quadratsumme Fehler	F	p	p < .05	generalisiertes η^2
Leistung (H22)	(Intercept)	1	32	1.19	59.09	0.65	.43		.02
	Gruppe	1	32	8.74	59.09	4.73	.04	*	.12
	Zeit	1	32	4.21	3.66	36.74	<.01	*	.06
	Gruppe x Zeit	1	32	0.00	3.66	0.03	.87		<.01

Tabelle 59: ANCOVA Schülerleistung (H22b)⁷⁶

	b	SE	t	p
Intercept	0.22	0.14	1.56	.13
Prätest-Leistung	1.26	0.10	12.70	<.01
Gruppe	-0.14	0.18	-0.78	.44

adjustiertes $R^2 = .85$

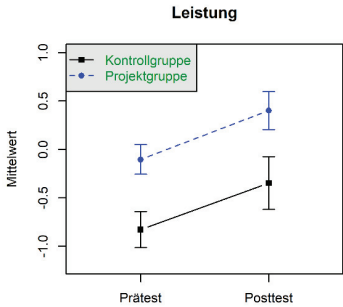


Abbildung 47: Leistung der Schülerinnen und Schüler in Prä- und Posttest nach Gruppen (H22a)

⁷⁶ Der Verzicht auf den Interaktionsterm im Modell führt nicht zu einem signifikanten Verlust an Erklärungskraft: $F = 0.46$, $p = .50$.

4.2 Typenfindung

4.2.1 Clusteranalyse

Bei der Durchführung der Clusteranalyse zur Typenfindung liegt ausgehend von Stichprobe 2 nach Ausschluss der fehlenden Werte ein N von 47 Lehrkräften vor. Die Clusteranalyse mit dem Ward-Verfahren und der Manhattan-Metrik liefern einen sehr guten Wert von .87 für den Agglomerations-Koeffizient⁷⁷. Abbildung 48 zeigt einen Streuplot der Verschmelzungshöhen der Cluster in Abhängigkeit der Clusteranzahl (*Elbow-Kriterium*). Hieraus lässt sich, analog zum Scree-Plot der Faktorenanalyse, deutlich erkennen, dass beim fünften Datenpunkt ein „Knick“ vorhanden ist. Die Abstände von Einzelpersonen bzw. Clustern vor der Verschmelzung liegen unterhalb (also bei steigender Clusterzahl) nur noch minimal auseinander. Zwar legen zwei der genannten Indizes (*Gplus* und *Gamma*) sechs Cluster als beste Lösung nahe und die anderen drei Indizes (*Calinski/Harabasz*, *PTbiserial Index* und *C-Index*) schlagen drei Cluster vor, doch scheinen nach Sichtkontrolle des Dendrogramms 6 Cluster gegenüber der nächsten Verschmelzungsebene (sieben Cluster) einen nur minimalen Abstand zu haben. Der Blick auf das Dendrogramm unterstützt ebenfalls nicht eindeutig eine Nutzung von drei anstelle von vier oder zwei Clustern. Daher entscheide ich mich vor dem Hintergrund der Verschmelzungsebenen und der inhaltlichen Interpretierbarkeit für die inhaltliche Arbeit für die 5-Cluster-Lösung⁷⁸. Diese Aufteilung wird auch aus dem Dendrogramm in Abbildung 49 klar ersichtlich, wenn man exemplarisch bei einem Abstand (*height*) von 3 das Dendrogramm „abschneidet“. Der resultierende Dunn-Index ergibt .23⁷⁹ (Brock et al., 2008; Dunn, 1974). Die resultierenden Cluster weisen dabei folgende Probandenzahlen auf:

$$N_{\text{Cluster1}} = 6$$

$$N_{\text{Cluster4}} = 12$$

$$N_{\text{Cluster2}} = 9$$

$$N_{\text{Cluster5}} = 8$$

$$N_{\text{Cluster3}} = 12$$

⁷⁷ Der Agglomerationskoeffizient kann Werte von 0 bis 1 annehmen, wobei 1 einer perfekten Einteilung entspricht.

⁷⁸ Für die Ward-Methode bei euklidischem Abstand ergeben sich z. T. andere Vorschläge zur optimalen Clusteranzahl. Insbesondere empfehlen die Indizes *Gplus* und *Calinski/Harabasz* eine Clusterzahl von vier. Bei dieser Clusterlösung sind im Vergleich zur 5-Clusterlösung die Cluster 2 und 3 verschmolzen. Da sich diese Cluster jedoch in Bezug auf die Akzeptanz voneinander unterscheiden, wird hier die Lösung mit fünf Clustern präsentiert. Eine Mittelwertübersicht der 4-Cluster-Lösung ist dem Anhang zu entnehmen (vgl. Anhang 7.3.4).

⁷⁹ Zwar fällt dieser Wert nicht besonders zufriedenstellend aus, jedoch zeigen alternative Clusterlösungen und Clusterverfahren ähnliche Indizes.

Diese Clusterlösung ähnelt im Wesentlichen anderen Clusterberechnungen etwa dem *Complete*- oder dem *Average*-Ansatz und zugrundeliegender Manhattan- bzw. euklidischer Metrik (auf deren Bericht hier jedoch mangels Unterschieden verzichtet wird). Der Agglomerationskoeffizient als Kennwert der Güte der Clusteranalyse fällt für alternative Skalierungen mit anderen Verfahren geringer aus als beim favorisierten und für kleine Stichproben empfohlenen Ward-Verfahren (Bacher et al., 2010; Hatzinger et al., 2014). Die Clusteranzahl bzw. die Verschmelzungsniveaus der Cluster sind dabei vergleichsweise ähnlich. Aus Platzgründen sind die entsprechenden Dendrogramme im Anhang zu finden (vgl. Anhang 7.3.4).

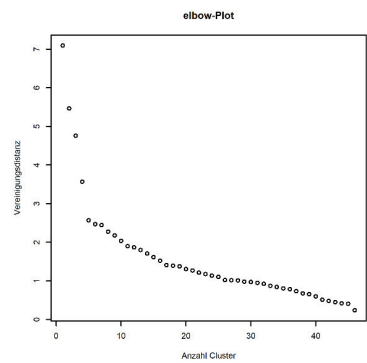


Abbildung 48: Elbow-Plot der Verschmelzungshöhen

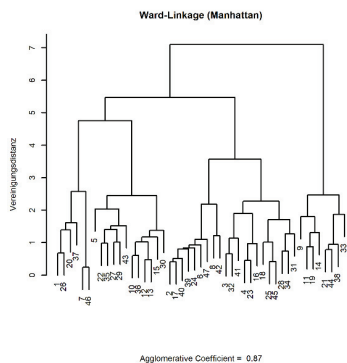


Abbildung 49: Dendrogramm Clusteranalyse (Ward-Verfahren; Manhattan-Metrik)

Tabelle 60: Mittelwerte, Standardabweichung und ANOVA für LSE-Skalen nach Clustern

	Akzeptanz		Belastung		Validität		Förderung		Planung		Selektion	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Cluster 1	2.66	0.24	3.58	0.25	3.75	0.50	3.17	0.23	3.26	0.40	3.25	0.76
Cluster 2	3.13	0.39	1.94	0.42	3.99	0.23	3.27	0.22	3.73	0.32	1.99	0.32
Cluster 3	2.63	0.29	2.40	0.47	3.43	0.41	3.31	0.36	3.10	0.30	2.20	0.35
Cluster 4	3.11	0.37	2.10	0.44	3.92	0.39	3.70	0.24	3.59	0.58	3.23	0.32
Cluster 5	2.60	0.28	2.81	0.31	3.04	0.69	2.47	0.27	2.25	0.56	1.75	0.41
F(4,42)	6.72		19.06		6.07 ⁸⁰		24.17		14.64		28.53 ⁸¹	
p	<.001		<.001		.003		<.001		<.001		<.001	

⁸⁰ Da die Voraussetzung der Varianzhomogenität nicht vorliegt, wurde für die Skala „Lehrplanvalidität“ der Welch-Test durchgeführt: $F(4, 17.64) = 6.07, p = .003$.

⁸¹ Da die Voraussetzung der Varianzhomogenität nicht vorliegt, wurde für die Skala *selektionsdiagnostischer Nutzen* der Welch-Test durchgeführt: $F(4, 17.47) = 28.53, p < .001$.

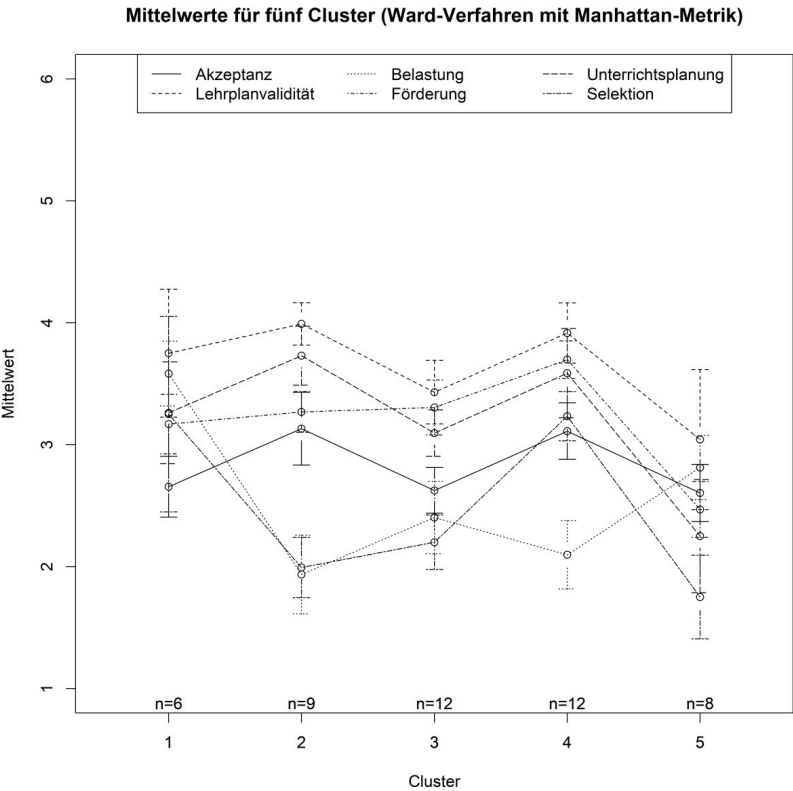


Abbildung 50: Mittelwerte der Skalen zu Lernstandserhebungen nach Clustern (Ward-Verfahren mit Manhattan-Metrik)

Tabelle 61: t-Test auf Abweichungen der Mittelwerte der Cluster vom semantischen Median von 3

		Cluster				
		1	2	3	4	5
Akzeptanz	<i>t</i>	-3.56	1.02	-4.42	1.06	-3.99
	<i>df</i>	5	8	11	11	7
	<i>p</i>	.016	.339	.001	.314	.005
Belastung	<i>t</i>	5.65	-7.60	-4.42	-7.10	-1.69
	<i>df</i>	5	8	11	11	7
	<i>p</i>	.002	<.001	.001	<.001	.135
Validität	<i>t</i>	3.67	13.15	3.64	8.16	0.17
	<i>df</i>	5	8	11	11	7
	<i>p</i>	.014	<.001	.004	<.001	.869
Förderung	<i>t</i>	1.79	3.70	2.98	10.03	-5.49
	<i>df</i>	5	8	11	11	7
	<i>p</i>	.134	.006	.013	<.001	.001
Planung	<i>t</i>	1.61	6.95	1.10	3.53	-3.82
	<i>df</i>	5	8	11	11	7
	<i>p</i>	.168	<.001	.296	.005	.007
Selektion	<i>t</i>	0.80	-9.38	-7.88	2.55	-8.61
	<i>df</i>	5	8	11	11	7
	<i>p</i>	.459	<.001	<.001	.027	<.001

In Bezug auf die Skala *Akzeptanz* zeigen sich im Wesentlichen zwei Niveaus: Cluster 1, 3 und 5 weisen eine vergleichbar niedrige *Akzeptanz* auf – insbesondere mit einem Mittelwert, der signifikant abweichend unterhalb des semantischen Medians von 3 liegt (vgl. Tabelle 61) –, während die Cluster 2 und 4 Mittelwerte der *Akzeptanz* zeigen, die leicht über dem semantischen Median liegen. Die Mittelwerte der Cluster 2 und 4 weichen dabei jedoch nicht signifikant von 3 ab. Entsprechend der Aufteilung in Cluster mit niedriger Akzeptanz (Cluster 1, 3, 5) und eher hoher Akzeptanz (Cluster 2 und 4) zeigen sich auch die Signifikanzen in den Post-Hoc-Tests (Tabelle 62) im Anschluss an die Varianzanalyse nach Mittelwert-Gruppenunterschieden (letzte Zeile in Tabelle 60). Cluster 1, 3 und 5 zeigen untereinander keine signifikanten Abweichungen, jedoch mindestens marginal signifikante ($p < .08$) bonferroni-korrigierte Paarvergleiche zu den Clustern 2 und 4 (Tabelle 62).

Tabelle 62: *p*-Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für *Akzeptanz der LSE*

Cluster	1	2	3	4
2	.08			
3	1.00	.01		
4	.08	1.00	.01	
5	1.00	.02	1.00	.014

Teil 4 – Ergebnisse
Typenfindung

Die *Belastung* wird insbesondere in Cluster 1 anders eingeschätzt als in den restlichen Clustern. Cluster 1 liegt deutlich über dem semantischen Median und zeugt entsprechend von einer zustimmenden Einschätzung zur Belastung. Die Einschätzung der Belastung fällt in diesem Cluster (hoch-)signifikant höher als in den anderen Clustern aus ($p < .011$; Tabelle 63). Die Cluster 2, 3 und 4 unterscheiden sich bzgl. des Belastungsempfindens nicht signifikant voneinander. Cluster 5 nimmt in Bezug auf die Belastung wiederum einen Platz zwischen dem „belasteten“ Cluster 1 und den wenig „belasteten“ Clustern 2, 3 und 4 ein und unterscheidet sich auch signifikant von den Mittelwerten der Cluster 2 und 4 ($p < .004$).

Tabelle 63: *p*-Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für *LSE als Belastung*

Cluster	1	2	3	4
2	<.001			
3	<.001	.13		
4	<.001	1.00	.73	
5	.01	<.01	.33	<.01

In puncto *Lehrplanvalidität* fällt besonders Cluster 5 deutlich heraus. Der Mittelwert liegt fast exakt übereinstimmend mit dem semantischen Median, während die restlichen Cluster eine zustimmende Haltung zeigen (und dabei auch alle signifikant von 3 abweichen). Vor allem Cluster 1 und 3 schätzen die *Lehrplanvalidität* mit Mittelwerten nahe 4 sehr hoch ein. Unter dem Gesichtspunkt der Unterschiede zwischen den einzelnen Clustern fällt vor allem der signifikante Paarvergleich der Cluster 2 und 3 auf ($p < .01$; Tabelle 64). Darüber hinaus zeigen sich für den indifferenten Mittelwert des Clusters 5 marginal signifikante Abweichungen zu den beiden höchsten Mittelwerten aus Cluster 2 und 4. Grund für die nur marginal signifikante Abweichungen dürfte die vergleichsweise große Standardabweichung des Clusters 5 sein (vgl. Tabelle 60).

Tabelle 64: *p*-Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für *Lehrplanvalidität*

Cluster	1	2	3	4
2	1.00			
3	1.00	.01		
4	1.00	1.00	.07	
5	.45	.05	1.00	.08

Betrachtet man die Skala *förderdiagnostischer Nutzen*, so heben sich vor allem Cluster 4 und 5 ab: Cluster 5 weicht hochsignifikant in negativer Richtung vom semantischen Median ab und Cluster 4 hochsignifikant in positiver Richtung (wobei diese Abweichungen für die Cluster 2 und 3 ebenfalls signifikant ausfallen). Entsprechend

scheinen die Lehrkräfte des Clusters 5 die Lernstandserhebungen für individuelle Förderung kaum zu nutzen, während Lehrkräfte der Cluster 1, 2, 3 und 4 dies in ansteigender Reihenfolge tun. Cluster 5 zeigt den mit Abstand geringsten Mittelwert für diese Skala und entsprechend fallen alle Paarvergleiche hochsignifikant aus ($p < .001$; Tabelle 65). Den höchsten Wert weist hier Cluster 4 auf, welches sich ebenfalls signifikant von den restlichen unterscheidet. Die Cluster 1, 2 und 3 weichen untereinander jedoch nicht signifikant voneinander ab.

Tabelle 65: p -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für förderdiagnostischen Nutzen

Cluster	1	2	3	4
2	1.00			
3	1.00	1.00		
4	<.01	.01	.01	
5	<.001	<.001	<.001	<.001

Beim *unterrichtsplanerischen Nutzen* zeigt wiederum Cluster 5 eine signifikant ablehnende Haltung, während die restlichen Cluster eher zustimmend antworten. Hierbei erweisen sich Cluster 2 und 4 als diejenigen, die dieser Nutzungsperspektive am meisten zustimmen und dabei signifikant vom semantischen Median abweichen. Die Cluster 1 und 3 geben tendenzielle Zustimmung gegenüber diesem Nutzen gemäß der Dichotomisierung am semantischen Median an (vgl. Tabelle 61). Wie bei der zuvor beschriebenen Nutzungsskala liegt auch für den *unterrichtsplanerischen Nutzen* der Mittelwert von Cluster 5 hochsignifikant unterhalb der anderen Cluster (vgl. Tabelle 66). Der Unterschied zwischen Cluster 2 und 3 fällt ebenfalls signifikant aus ($p < .05$). Verglichen mit den Standardabweichungen der anderen Skalen zeugen vergleichsweise hohe Werte bei dieser Skala u. a. für das Cluster 4 von einer gewissen Heterogenität, worin die Ursache der fehlenden Signifikanz der Paarvergleiche von Cluster 4 gegenüber den weniger deutlich zustimmenden Clustern 1 und 3 liegen dürfte.

Tabelle 66: p -Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für unterrichtsplanerischen Nutzen

Cluster	1	2	3	4
2	.53			
3	1.00	.02		
4	1.00	1.00	.10	
5	<.01	<.001	<.01	<.001

Auf der letzten Skala, *selektionsdiagnostischer Nutzen*, zeigen die Cluster 2, 3 und 5 eine signifikant ablehnende Haltung, die sich untereinander nicht signifikant

Teil 4 – Ergebnisse
 Typenfindung

unterscheidet. Der Mittelwert des Clusters 4 lässt sich als Zustimmung gegenüber dem *selektionsdiagnostischen Nutzen* interpretieren. Der Mittelwert des Clusters 1 weicht nicht signifikant vom semantischen Median ab. Analog zu den anderen Nutzungsskalen hat Cluster 5 den geringsten Mittelwert, der signifikant geringer ist als bei den Clustern 1 und 4 (vgl. Tabelle 67). Cluster 4 seinerseits unterscheidet sich hochsignifikant von den beiden Clustern 2 und 3 ($p < .001$).

Tabelle 67: *p*-Werte der Paarvergleiche (Post-Hoc; bonferroni-korrigiert) der Cluster für *selektionsdiagnostischen Nutzen*

Cluster	1	2	3	4
2	.08			
3	.18	1.00		
4	1.00	<.001	<.001	
5	.03	1.00	.24	<.001

Aus diesen skalenbezogenen Analysen lässt sich folgende qualitative Zusammenfassung kondensieren, dabei bedeuten „+“ bzw. „-“ eine positive respektive negative Abweichung vom semantischen Median (Klammern deuten nicht vorhandene Signifikanz des Unterschieds vom semantischen Median an). Dabei wurde für jedes Cluster ein plakativ umschreibender Name gewählt, der im weiteren Verlauf sowie in der folgenden Zusammenfassung genauer erläutert wird:

Tabelle 68: „Dichotomisierte“ Beschreibungen der Abweichungen der Cluster-Skalenmittelwerte aufbauend auf t-Tests (vgl. Tabelle 61)

	Beschreibung ⁸²						Umschreibender Name
	A	LV	B	F	P	S	
Cluster 1	-	+	+	(+)	(+)	(+)	Zurückhaltend-uneingeschränkter Typ
Cluster 2	(+)	+	-	+	+	-	Planend-fördernder Nutzer
Cluster 3	-	+	-	+	(+)	-	Planend-fördernder Skeptiker
Cluster 4	(+)	+	-	+	+	+	Uneingeschränkter Nutzer
Cluster 5	-	⁸³	(-)	-	-	-	Ablehnender Typ

Um die gefundenen Cluster etwas genauer beschreiben zu können, wird auf die wenigen zur Verfügung stehenden Hintergrundvariablen zurückgegriffen: Als Heuristik für das aus Datenschutz-Belangen nicht direkt erfasste Alter der Lehrkräfte kann die bisherige Unterrichtserfahrung herangezogen werden (Tabelle 69). Dabei deutet sich an, dass Cluster 1 mit lediglich knapp 6 Jahren durchschnittlicher Unterrichtserfahrung am unerfahrensten ist. Die Mittelwertsunterschiede werden jedoch

⁸² A: Akzeptanz; LV: Lehrplanvalidität; B: Belastung; F: Förderung; P: Planung; S: Selektion.
⁸³ Diese Zelle bleibt bei einem Mittelwert von $M = 3.04$ bewusst frei, da hier keine erwähnenswerte Abweichung von 3.00 vorliegt.

varianzanalytisch nicht signifikant. Tendenziell lässt sich vermuten, dass die indifferente Haltung der Lehrkräfte des Clusters 1 u. a. von der allgemein geringen Erfahrung im Schulalltag, der bisher noch nicht vorhandenen Möglichkeit zur Teilnahme an den LSE bzw. den generellen Herausforderungen im Berufseinstieg rühren.

Ebenso wenig zeigen sich bei der Geschlechtsverteilung statistisch abgesicherte Unterschiede. Cluster 2 und Cluster 5 zeigen genau umgedrehte Verteilungen, mit 2 Frauen und 6 Männern in Cluster 2 bzw. 6 Frauen und 2 Männern in Cluster 5, doch fehlt dieser deskriptiven Beobachtung die inferenzstatistische Absicherung ($\chi^2 = 5.10$, $p = .28$; vgl. Tabelle 70). In Bezug auf das Geschlecht zeigt sich keines der Cluster augenscheinlich auffällig.

Betrachtet man die Projektteilnahme, so zeigen sich wiederum keine signifikanten Unterschiede zwischen den einzelnen Clustern. Der zurückhaltend-uneingeschränkte Typ (Cluster 1) sticht jedoch qualitativ aus den anderen Clustern heraus, deren Verteilung von Projektteilnahme und Nicht-Teilnahme recht gleichmäßig ist. Aus Cluster 1 hat lediglich eine einzelne Lehrkraft am Projekt teilgenommen, während 5 Lehrkräfte dies nicht taten.

Wie bereits bei der Unterrichtserfahrung vermutet, bestätigt sich der Zusammenhang zwischen geringer Unterrichtserfahrung in Cluster 1 und fehlenden Vorerfahrungen mit den Lernstandserhebungen, während für Cluster 2 „planend-fördernder Nutzer“ die meisten Vorerfahrungen im Mittel vorliegen (vgl. Tabelle 72). Für die Cluster 3, 4 und 5 lässt sich bereits vor einer statistischen Prüfung kein Unterschied erwarten. Jedoch zeugen die Cluster 1 und 2 – in der Beschreibung „zurückhaltend-uneingeschränkter Typ“ und „planend-fördernder Nutzer“ – von sich unterscheidenden Vorerfahrungen, die zumindest marginal signifikant im Paarvergleich werden ($p = .07$).

Als weitere Hintergrundvariable wurde die Schulform der Lehrkräfte erfragt, bei denen sich jedoch keine clusterspezifischen Unterschiede zeigen. Wird die inhaltliche Analyse auf die klarer zu umreißen Schulformen Gymnasium und Realschule (gegenüber eher schwer zu interpretierender Gesamtschule) beschränkt, fällt auf, dass der „zurückhaltend-uneingeschränkter Typ“ des Clusters 1 vorrangig aus dem gymnasialen Bereich kommt und damit schulformbezogen das homogenste Cluster darstellt (vgl. Tabelle 73).

Teil 4 – Ergebnisse

Typenfindung

Die einzige Hintergrundvariable, die zumindest marginal signifikante Unterschiede zwischen den Clustern aufweist, ist die Ausbildung. Hierbei wird u. a. zwischen 2. Staatsexamen im gymnasialen sowie Haupt-/Realschul-Bereich und dem ersten Staatsexamen unterschieden. Dabei fällt insbesondere Cluster 2, der „planend-fördernde Nutzer“, auf, dem zu einem Großteil Lehrkräfte mit zweitem Staatsexamen im HR-Bereich zugeordnet werden.

Tabelle 69: Deskriptive Statistiken nach Clustern getrennt für Unterrichtserfahrung der Lehrkräfte

	<i>N</i>	<i>M</i>	<i>SD</i>	Median	Minimum	Maximum	<i>SE</i>
1	6	5.83	4.96	5.50	1.00	15.00	2.02
2	9	15.67	12.18	12.00	2.00	36.00	4.06
3	12	14.92	8.87	15.00	2.00	33.00	2.56
4	12	14.00	13.13	6.50	1.00	35.00	3.79
5	8	11.00	9.77	6.50	3.00	30.00	3.45

ANOVA: $F(1,45) = 0.20, p = .66$

Tabelle 70: Häufigkeitstabelle nach Clustern getrennt für Geschlecht

	weiblich	männlich
1	4	2
2	2	6
3	7	5
4	8	4
5	6	2

Geschlecht nicht signifikant: $\chi^2(4) = 5.10, p = .28$

Tabelle 71: Häufigkeitstabelle nach Cluster getrennt für Projektteilnahme

	1	2	3	4	5
Teilgenommen	1	4	5	7	4
Nicht teilgenommen	5	5	7	5	4

Projektteilnahme nicht signifikant: $\chi^2(4) = 2.95, p = .57$

Tabelle 72: Deskriptive Statistiken nach Clustern getrennt für LSE-Vorerfahrungen (als Summenscore der Jahrgänge 2009, 2010, 2011, 2012)

	<i>N</i>	<i>M</i>	<i>SD</i>	Median	Minimum	Maximum	<i>SE</i>
1	6	0.17	0.41	0.00	0.00	1.00	0.17
2	9	1.44	1.13	1.00	0.00	3.00	0.38
3	12	0.50	0.67	0.00	0.00	2.00	0.19
4	12	0.50	0.80	0.00	0.00	2.00	0.23
5	8	0.75	1.04	0.50	0.00	3.00	0.37

LSE-Vorerfahrungen nicht signifikant: $F(1,45) = 0.05, p = .83$

Tabelle 73: Häufigkeitstabelle nach Clustern getrennt für aktuelle Schulform

	Gymnasium	Realschule	KGS	IGS
1	5	0	1	0
2	1	1	4	3
3	4	1	6	1
4	4	0	4	3
5	2	0	4	2

Schulform: nicht signifikant: $\chi^2(12) = 12.83, p = .38$

Tabelle 74: Häufigkeitstabelle nach Clustern getrennt für Ausbildung

	2. Staatsexamen Gymnasium	2. Staatsexamen HR	2. Staatsexamen Primar	1. Staatsexamen	Sonstiges
1	2	1	0	1	1
2	1	7	0	1	0
3	7	5	0	0	0
4	7	5	0	0	0
5	5	3	0	0	0

Ausbildung marginal signifikant: $\chi^2(12) = 19.83, p = .07^{84}$

4.2.2 Zusammenfassende Clusterbeschreibung

Cluster 1 lässt sich als eher skeptische Gruppe beschreiben, deren Angaben auf den nutzungsorientierten Skalen nicht signifikant vom semantischen Median abweichen, die jedoch eine deutliche Zustimmung für die Skala *Belastung* sowie eine deutliche Ablehnung bei der Skala *Akzeptanz* zeigt. Dabei weist dieses Cluster sowohl die mit mehr als einer Standabweichung entfernte, geringste Unterrichtserfahrung als auch die geringste Vorerfahrung mit den Lernstandserhebungen auf. Insofern lässt sich die „Skepsis“ bzw. „Zurückhaltung“ dieser Gruppe u. U. dadurch erklären, dass die Lernstandserhebungen zum Großteil noch nicht selbst durchgeführt wurden und somit lediglich auf externe Erfahrungen und Berichte zurückgegriffen werden konnte. Als Ausdruck der Skepsis dieses Clusters gegenüber den Lernstandserhebungen lässt sich die sehr verhaltene Projektteilnahme von nur einer Lehrkraft werten. Auffällig ist,

⁸⁴ Bei der Berechnung wurde in der Häufigkeitstabelle die Spalte „2. Staatsexamen Primar“ entfernt, da keine Lehrkräfte diese Angabe gemacht haben.

Teil 4 – Ergebnisse

Typenfindung

dass wiederum fünf der sechs Lehrkräfte dieses Clusters am Gymnasium unterrichten.

Cluster 2 beschreibt den „planend-fördernden Nutzer“: Ausgehend von einer hohen Lehrplanvalidität und geringer Belastung zeigt dieser Typ eine positive Einstellung zum *förderdiagnostischen* und insbesondere dem *unterrichtsplanerischen Nutzen*, ohne jedoch die Lernstandserhebung zur *Selektion* heranzuziehen. Erstaunlich ist, dass die *Akzeptanz* – wenn auch im Clustervergleich deutlich und signifikant positiver – absolut betrachtet nicht signifikant vom semantischen Median unterscheidbar ist. Diese Gruppe weist den höchsten Mittelwert für die Anzahl an bereits durchgeführten Lernstandserhebungen auf und kann somit auf eigene Erfahrungswerte zurückgreifen. Bezogen auf die Ausbildung handelt es sich bei Lehrkräften dieses Clusters vorrangig um Haupt- und Realschullehrkräfte. Cluster 2, 3 und 4 zeigen eine hohe Unterrichtserfahrung, die sind insbesondere von Cluster 1 abhebt.

Cluster 3 weist ähnliche Mittelwerte wie Cluster 1 auf, wobei die Skalen *Belastung* und *selektionsdiagnostischer Nutzen* deutlich ablehnender eingeschätzt werden. Gemessen am semantischen Median fallen die Mittelwerte für die nutzungsorientierten Skalen positiv aus, jedoch im Clustervergleich geringer als für Cluster 2 und 4. Insofern ist dieser Typ als eine andere Form des Skeptikers zu beschreiben, der jedoch scheinbar die Lernstandserhebungen eher in intendierter Art nutzt (im Sinne der vorgestellten Ziele).

Cluster 4 bezeichne ich als „uneingeschränkten Nutzer“. Dieser Typ zeichnet sich dadurch aus, dass alle Möglichkeiten der Nutzung gewählt werden. Damit liegt der essentielle Unterschied zum „planend-fördernden Nutzer“ darin, dass auch *selektionsdiagnostischer Nutzen* aus den Lernstandserhebungen gezogen wird. Im *förderdiagnostischen Nutzen* hebt sich dieses Cluster in wünschenswerter Richtung signifikant von allen anderen Clustern ab (Paarvergleiche $p \leq .01$).

Cluster 5 zeigt eine ablehnende Haltung gegenüber den nutzungsorientierten Skalen und hat geringe Werte in Bezug auf die *Akzeptanz*. Sogar die von allen restlichen Clustern als positiv eingeschätzte *Lehrplanvalidität* wird indifferent beurteilt. Mit dem Wissen der eher ablehnenden Haltung gegenüber der Nutzung bezogen auf den semantischen Median zeigt sich entsprechend eine gering eingeschätzte *Belastung*, da mit den Lernstandserhebungen nicht oder nur sehr eingeschränkt gearbeitet wird.

4.2.3 Konfigurationsfrequenzanalyse

Wie im Methodenteil erläutert, fehlt der Clusteranalyse eine Absicherung, dass die gefundenen Gruppen neben der interpretatorisch-inhaltlichen auch eine statistische Relevanz haben. Mit der Konfigurationsfrequenzanalyse wird diese Grundlage gelegt, indem die beobachteten Häufigkeiten von Konfigurationen (Kombinationen von Merkmalsausprägungen) mit den stochastisch erwarteten Häufigkeiten verglichen werden. Tabelle 75 und Tabelle 76 zeigen die Ergebnisse der durchgeführten Konfigurationsfrequenzanalyse mit Mediansplit („+“ und „-“ stehen jeweils in Relation zum Median).⁸⁵

Tabelle 75: Ergebnisse der Konfigurationsfrequenzanalyse für die Wahrnehmungsskalen *Akzeptanz*, *Belastung* und *Lehrplanvalidität* (Mediansplit)

Konfiguration ⁸⁶			Beobachtungen	Erwartete Beobachtungen	χ	df	p
A	B	LV					
+	+	+	2	6.99	3.56	1	.06
+	+	-	4	5.18	0.27	1	.60
+	-	+	17	7.94	10.32	1	<.01
+	-	-	3	5.89	1.41	1	.23
-	+	+	5	5.65	0.07	1	.79
-	+	-	11	4.18	11.11	1	<.01
-	-	+	3	6.42	1.82	1	.18
-	-	-	2	4.75	1.59	1	.21

Tabelle 76: Ergebnisse der Konfigurationsfrequenzanalyse für die Nutzungsskalen *Förderung*, *Planung* und *Selektion* (Mediansplit)

Konfiguration ⁸⁷			Beobachtungen	Erwartete Beobachtungen	χ	df	p
F	UP	S					
+	+	+	13	7.25	4.57	1	.03
+	+	-	7	7.51	0.03	1	.85
+	-	+	4	7.00	1.28	1	.26
+	-	-	5	7.25	0.70	1	.40
-	+	+	4	7.00	1.28	1	.26
-	+	-	5	7.25	0.70	1	.40
-	-	+	7	6.76	0.01	1	.93
-	-	-	12	7.00	3.58	1	.06

⁸⁵ In der Methode wird erläutert, weshalb zwei getrennte Konfigurationsfrequenzanalysen anstelle einer gemeinsamen mit allen sechs Lernstandserhebungsskalen durchgeführt werden.

⁸⁶ Jeweils als Abkürzungen die Skalen: „A“ – Akzeptanz; „B“ – Belastung; „LV“ – Lehrplanvalidität

⁸⁷ Jeweils als Abkürzungen die Skalen: „F“ – Förderung; „UP“ – Unterrichtsplanung; „S“ – Selektion

Teil 4 – Ergebnisse

Typenfindung

Für die wahrnehmungsbezogenen Skalen *Akzeptanz*, *Lehrplanvalidität* und *Belastung* zeigen sich folgende Konfigurationen als sog. *Typen* in der Nomenklatur der Konfigurationsfrequenzanalyse, also Konfigurationen, die (marginal) signifikant häufiger auftreten als erwartet:

- Typ 1 (marginal signifikant; $p = .06$):
Akzeptanz: Hohe Ausprägung
Belastung: Hohe Ausprägung
Validität: Hohe Ausprägung
- Typ 2 (hoch signifikant; $p < .01$):
Akzeptanz: Hohe Ausprägung
Belastung: Niedrige Ausprägung
Validität: Hohe Ausprägung
- Typ 3 (hoch signifikant; $p < .01$):
Akzeptanz: Niedrige Ausprägung
Belastung: Hohe Ausprägung
Validität: Niedrige Ausprägung

Die Typen 2 und 3 sind der jeweilige Gegenpol, wobei insbesondere Typ 2 einen Bezug zu Cluster 2 bzw. 4 ermöglicht. In beiden Clustern liegen geringe *Belastungsmittelwerte* bei hoher Ausprägung der *Akzeptanz* und *Validität* vor. Soweit lässt sich zumindest feststellen, dass die „erste Hälfte“ der Skalen dieser beiden Cluster durch die Konfigurationsfrequenzanalyse unterstützt wird und damit als „überzufällig“ anzusehen ist. Cluster 5 zeigt als Gegenpol geringe *Akzeptanz* und *Lehrplanvalidität* bei gleichzeitig hoher *Belastung* und entspricht damit Typ 3 der KFA. Somit ist auch für dieses Cluster die erste Hälfte der Skalen bestätigt.

Die nutzungsorientierten Skalen *Förderung*, *Planung* und *Selektion* liefern hingegen ein weniger differenziertes Bild, sodass sich lediglich zwei Typen zeigen: Typ 1 weist über alle diese Skalen hohe Mittelwerte (signifikant; $p = .03$), Typ 2 durchweg niedrige Werte auf (marginal signifikant; $p = .06$).

Da der Mediansplit einige Schwierigkeiten birgt (vgl. Abschnitt 5.1.2) und zusätzlich stichprobenspezifisch ist, wird zusätzlich eine Dichotomisierung am semantischen Median vorgenommen. Dadurch können die Konfigurationen besser als Zustimmung oder Ablehnung interpretiert werden. Bei dieser Betrachtung zeigen sich für die

Wahrnehmungsskalen lediglich eine auffällige Konfiguration mit durchgängiger Ablehnung aller drei Skalen (vgl. letzte Konfiguration in Tabelle 77). Bei den nutzungsorientierten Skalen zeigt sich keine der Konfigurationen als signifikant häufiger bzw. seltener gegenüber der Erwartung des Modells (Tabelle 78). Einschränkung für diese Erweiterung muss hier angeführt werden, dass der Split am semantischen Median bei den vorliegenden Verteilungen zu ungleichen Gruppen je Skala führt: Für die Skala *Lehrplanvalidität* teilt sich die Stichprobe in 42 Lehrkräfte mit Skalenmittelwerten größer als der semantische Median und nur 5 Personen mit einem kleiner oder gleichen Skalenmittelwert bezogen auf den semantischen Median.

Tabelle 77: Ergebnisse der Konfigurationsfrequenzanalyse für die Wahrnehmungsskalen (Split am semantischen Median)

Konfiguration ⁸⁸			Beobachtungen	Erwartete Beobachtungen	χ	df	p
A	B	LV					
+	+	+	11	7.52	1.61	1	.20
+	+	-	22	21.05	0.04	1	.84
+	-	+	4	4.06	0.00	1	.98
+	-	-	7	11.38	1.68	1	.19
-	+	+	0	2.22	2.22	1	.14
-	+	-	4	6.22	0.79	1	.37
-	-	+	0	1.20	1.20	1	.27
-	-	-	9	3.36	9.46	1	<.01

Tabelle 78: Ergebnisse der Konfigurationsfrequenzanalyse für die Nutzungsskalen (Split am semantischen Median)

Konfiguration ⁸⁹			Beobachtungen	Erwartete Beobachtungen	χ	df	p
F	UP	S					
+	+	+	1	3.56	1.84	1	.18
+	+	-	0	0.42	0.42	1	.52
+	-	+	16	11.64	1.64	1	.20
+	-	-	0	1.39	1.39	1	.24
-	+	+	8	6.27	0.47	1	.49
-	+	-	2	0.75	2.10	1	.15
-	-	+	17	20.53	0.61	1	.44
-	-	-	3	2.44	0.13	1	.72

⁸⁸ Jeweils als Abkürzungen die Skalen: „A“ – Akzeptanz; „B“ – Belastung; „LV“ – Lehrplanvalidität

⁸⁹ Jeweils als Abkürzungen die Skalen: „F“ – Förderung; „UP“ – Unterrichtsplanung; „S“ – Selektion

4.3 Ist-Stand zu Einstellungen von Mathematiklehrkräften zu den Lernstandserhebungen in Hessen

Die folgenden Ergebnisse sollen den Ist-Stand für Hessen beschreiben und stellen damit einen ersten Versuch dar, die Wahrnehmung und Einschätzung der Lernstandserhebungen für Hessen zu beschreiben (Forschungsfrage 3). Grundlage vieler Auswertungen (sofern nicht anders erläutert) stellt dabei die Online-Stichprobe dar (vgl. Abschnitt 3.3.2). Es sei daran erinnert, dass das Erhebungsinstrument als Online-Fragebogen von der Hessischen Lehrkräfteakademie eingesetzt wurde und die Auswertung sich z. T. auf Einzelitems bezieht.

4.3.1 Unterrichts-Validität der Lernstandserhebungen

Tabelle 79: Ergebnisse der t-Tests für einzelne Stichproben⁹⁰ der Unterrichtsvalidität (operationalisiert durch die Ungewohntheit der Aufgaben) zwischen Mathematik und Deutsch sowie Englisch (bei Kontrolle der Teilnahme)⁹¹

		Deutsch	Englisch
		$M = 3.04$ $SD = 0.98$	$M = 3.47$ $SD = 0.94$
Mathematik	$M = 3.27$ $SD = 0.92$	$t(40) = 1.60$ $p = .12$	$t(40) = -1.44$ $p = .16$

Basierend auf den Daten der Online-Befragung lässt sich Folgendes festhalten: Der Mittelwert der Unterrichtsvalidität weicht für Mathematik in Jahrgang 8 mit $M = 3.27$ marginal signifikant vom semantischen Median 3 ab ($t(40) = 1.86, p = .07$). Für die Operationalisierung der Unterrichtsvalidität wird hier die Ungewohntheit der Aufgaben im Selbstbericht der Lehrkräfte herangezogen. Es zeigt sich also ein indifferentes Bild ohne klare Zustimmung oder Ablehnung, da die marginale Signifikanz nicht überinterpretiert werden darf. Das bedeutet entsprechend, dass die (Aufgaben der) Lernstandserhebungen dem Unterricht bzw. den im Unterricht eingesetzten Aufgaben in der Wahrnehmung der Lehrkräfte sowohl in Mathematik als auch in Deutsch und Englisch weder widersprechen noch sehr ähneln. Tendenziell lässt sich an den Mittelwerten der Ungewohntheit erkennen, dass Mathematiklehrkräfte der Ungewohntheit im Mittel durch den höheren Wert eher widersprechen als dies Deutsch-Lehrkräfte tun, während Lehrkräfte im Fach Englisch der Ungewohntheit noch deutlicher

⁹⁰ Untersucht wurde mittels eines t-Tests für einzelne Stichproben, ob der Mittelwert der Mathematiklehrkräfte vom Mittelwert für Deutsch- bzw. Englischlehrkräfte abweicht.

⁹¹ Für alle drei Fächer muss die Normalverteilungsannahme verworfen werden.

Antwortformat: 1 = „++“ bis 5 = „--“

als Mathematiklehrkräfte widersprechen. Die Mittelwertsunterschiede von Mathematik zu den beiden anderen Fächern ist in beiden Fällen nicht signifikant (vgl. Tabelle 79). Interpretiert heißt das, dass die Passung zwischen Lernstandserhebungen und Unterricht scheinbar in Mathematik und Englisch höher eingeschätzt wird als dies in Deutsch der Fall ist, diese Unterschiede jedoch auch zufällig sein können.

Kontrastierend zu diesen Ergebnissen zeigen sich im Prätest des Projekts VELM-8 als unbeeinflusster Ist-Stand (vgl. Abschnitt 4.1) hohe Mittelwerte von $M_{Projektgruppe} = 3.81$ ($SD = 0.55$) bzw. $M_{Kontrollgruppe} = 3.69$ ($SD = .46$) bei fünfstufigem Antwortformat (1: „stimme nicht zu“ bis 5: „stimme voll und ganz zu“), die wiederum eher eine hohe Lehrplanvalidität nahelegen. Etwas geringer fällt die Zustimmung der größeren Stichprobe 2 mit $M = 3.64$ ($SD = 0.55$) und einer signifikanten und bedeutsamen Mittelwertdifferenz vom semantischen Median aus ($t(46) = 7.92$, $p < .001$, $Cohens\ d = 1.16$).

Zusammenfassend lässt sich festhalten, dass ein postulierter Fachunterschied in der Unterrichtsvalidität (operationalisiert durch die Ungewohntheit der Aufgaben eingeschätzt durch die Lehrkräfte) zugunsten der Mathematik nicht bestätigt werden kann (Hypothese 23b). Für die fachunabhängige Einschätzung zeigen sich uneinheitliche Ergebnisse, sodass Hypothese 23a weder klar abgelehnt noch bestätigt werden kann.

4.3.2 Auswertungszeit

Tabelle 80: Ergebnisse der t-Tests der Auswertungszeit (in Minuten) zwischen Mathematik und Deutsch sowie Englisch⁹²

		Deutsch	Englisch
		$M = 368.29$ $SD = 177.02$	$M = 275.66$ $SD = 135.11$
Mathematik	$M = 188.00$ $SD = 103.15$	$t(39) = -11.100$ $p < .001$	$t(39) = -5.37$ $p < .001$

Da eine bewusste Trennung zwischen Eingabe- und Korrekturzeit aufgrund der Selbstauskünfte und damit verbundenen Schätzwerte für diese Items kaum möglich und sinnvoll erscheint, wird die Summe aus angegebener Korrektur- und Eingabezeit als Indikator für die Auswertungszeit betrachtet. Hypothese 24a formuliert die Annahme, dass die Auswertungszeit in Mathematik geringer ausfällt als in den sprachlichen Fächern. Die paarweisen t-Tests weisen hoch signifikante Unterschiede

⁹² Für die Auswertungszeit in Deutsch und Englisch muss die Normalverteilungsannahme verworfen werden, was jedoch aufgrund der großen Stichprobengröße hier ignoriert wird.

zwischen Mathematik und Deutsch bzw. Englisch auf (vgl. Tabelle 80), sodass diese Hypothese als bestätigt angesehen werden kann.

Die Einschätzung der Angemessenheit der Auswertungszeit (hier wiederum differenziert nach Korrektur- und Eingabezeit) fällt in den Fächern Deutsch und Englisch signifikant geringer aus als in Mathematik (vgl. Tabelle 81). Die zugehörigen *t*-Tests weisen alle eine sehr hohe Signifikanz auf ($p < .001$). Insbesondere fällt auf, dass die Mittelwerte der sprachlichen Fächer größer als der semantische Median 3 sind und somit als ablehnende Haltung aufgefasst werden können, während die Mittelwerte für Mathematik Zustimmung zur Angemessenheit der Korrektur- und Eingabezeit andeuten. Insofern lässt sich Hypothese 24b mit den vorliegenden Daten stützen und sie kann vorläufig als bestätigt angesehen werden.

Tabelle 81: Angemessenheit der Auswertungszeit im Fächervergleich (mit Kontrolle der Teilnahme)⁹³

		<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i> -Test ggn. Mathematik
Deutsch	Korrekturzeit	105	3.83	1.03	$t(38) = -10.65, p < .001$
	Eingabezeit	102	3.48	1.21	$t(38) = -8.08, p < .001$
Englisch	Korrekturzeit	102	3.38	1.15	$t(38) = -6.36, p < .001$
	Eingabezeit	105	3.40	1.23	$t(38) = -7.49, p < .001$
Mathematik	Korrekturzeit	39	2.38	0.85	
	Eingabezeit	40	2.35	0.86	

4.3.3 Nützlichkeit

Aus einer geringeren Auswertungszeit sollte prinzipiell eine geringere Belastung gefolgert werden können, sodass die Auswertungszeit als Proxy des durch die Lernstandserhebungen entstehenden Aufwandes gesehen werden kann. Je geringer die Auswertungszeit ausfällt, umso höher sollten nach theoretischen Annahmen Lehrkräfte die Nützlichkeit des Instrumentes einschätzen. Betrachtet man die regressionsanalytischen Analysen für die Auswertungszeit auf die drei Nützlichkeitsperspektiven, so zeigt sich ein leicht positiver, allerdings hochsignifikanter Regressionskoeffizient ($b_{max} = 0.0006, p < .001$) bei einer nur sehr gering ausfallenden Varianzaufklärung von $R^2_{max} = .02$ (fokussiert auf Mathematik ergeben sich ähnliche Werte: $b_{max} = 0.001, p_{max} = .01; R^2_{max} = .03$; vgl. Tabelle 82). Aufgrund der minimalen Größe des Regressionskoeffizienten und der aufgeklärten Varianz ist davon auszugehen, dass trotz Signifikanz dieses Ergebnis inhaltlich nicht bedeutsam ist. Für Hypothese

⁹³ Antwortformat: 1 = „++“ bis 5 = „--“

25a lässt sich zwar ein prädiktiver Effekt überzufällig nachweisen, jedoch kann aufgrund der geringen Varianzaufklärung die Hypothese nicht als bestätigt angesehen, sondern eher als vorerst nicht widerlegt betrachtet werden.

Tabelle 82: Regressionsanalyse der Auswertungszeit (allgemein bzw. mathematikbezogen) auf die Einschätzung der Nützlichkeit bzgl. der Lernstandsdiagnose, Unterrichts- und Schulentwicklung

		<i>b</i>	<i>p</i>	<i>R</i> ²
Lernstandsdiagnose	allgemein	0.0005	< .001	.01
	Mathematik	0.0010	.0040	.02
Unterrichtsentwicklung	allgemein	0.0006	< .001	.01
	Mathematik	0.0011	< .001	.03
Schulentwicklung	allgemein	0.0006	< .001	.01
	Mathematik	0.0009	.0100	.01

Tabelle 83: Einschätzung der Nützlichkeit nach Fächern (mit Kontrolle der Teilnahme)

	<i>N</i>	<i>M</i>	<i>SD</i>	
Mathematik	35	2.58	0.87	$F(2,206) = 8.81$ $p < .01$
Deutsch	81	3.37	1.03	
Englisch	95	3.28	0.91	

Tabelle 84: Nützlichkeit der LSE nach Fächern (mit Kontrolle der Teilnahme)

		<i>N</i>	<i>M</i>	<i>SD</i>	
Lernstandsdiagnose	Mathematik	34	1.97	0.97	$F(2,103.66) = 11.13$ $p < .01^{94}$
	Deutsch	81	2.73	1.16	
	Englisch	94	1.20	1.20	
Unterrichtsentwicklung	Mathematik	34	2.50	0.99	$F(2,206) = 6.80$ $p < .001$
	Deutsch	81	3.19	1.13	
	Englisch	94	3.19	1.06	
Schulentwicklung	Mathematik	33	2.79	0.89	$F(2,202) = 4.71$ $p < .01$
	Deutsch	78	3.47	1.08	
	Englisch	94	3.30	1.07	

Die Nützlichkeit wird von Mathematiklehrkräften gegenüber ihren Kolleginnen und Kollegen der sprachlichen Fächer positiver eingeschätzt (paarweise *t*-Tests mit Bonferroni-Alpha-Korrektur zeigen hochsignifikante Unterschiede zwischen Mathematik und Deutsch sowie Englisch; $p < .001$; vgl. Tabelle 83). Hierbei fällt auf, dass über den reinen Unterschied hinaus zwischen den Fächern wiederum auch ein Bedeutungsunterschied vorliegt: Nur Mathematik zeigt eine zustimmende Haltung ($M_{\text{Mathematik}} < 3$). Analog zum postulierten Fachunterschied in der Auswertungszeit kann auch Hypothese 25b hier als bestätigt gelten. Differenziert man die Auswertung nun nach verschiedenen Nützlichkeitsfacetten (Tabelle 84), so ist auch hier die Hypothese durch signifikante Varianzanalysen bestätigt (die bonferroni-korrigierten Paarvergleiche gegenüber Mathematik fallen signifikant auf dem 5%-Niveau aus; lediglich die

⁹⁴ Varianzgleichheit wurde geprüft (Levene-Test) und ist nicht vorhanden: $F(2,206) = 3.50$, $p = .03$. Entsprechend wird auf die Varianzanalyse nach Welch zurückgegriffen (vgl. Luhmann, 2011).

Teil 4 – Ergebnisse

Ist-Stand in Hessen

Einschätzung zur Schulentwicklung zwischen Englisch und Mathematik erreicht nur marginales Signifikanzniveau). Generell fällt jedoch auf, dass Mathematik als einziges Fach für alle drei Nutzungsmöglichkeiten zustimmende Mittelwerte zeigt, während Deutsch- und Englisch-Lehrkräfte die Nutzungsmöglichkeiten für Unterrichtsentwicklung und Schulentwicklung eher ablehnen. Dabei sinkt jedoch die Zustimmung für alle Fächer, je entfernter die Entwicklung von der einzelnen Schülerin/vom einzelnen Schüler ist: Die niedrigsten Werte (also die höchste Zustimmung) ergeben sich mit der Lernstandsdiagnose auf Individualebene, die höchsten (am wenigsten zustimmenden Werte) bei der Schulentwicklung.

4.3.4 Besprechung der Ergebnisse und Intensität der Auseinandersetzung mit den Ergebnissen

Naheliegend ist die Vermutung, dass aus einer höher eingeschätzten Nützlichkeit eine tiefere Auseinandersetzung mit den Ergebnissen bzw. eine intensivere Besprechung folgt, die auch mit mehr Austauschgruppen angestrebt wird. Dem Modell von Helmke (Abbildung 10) folgend liegt vor der Reflexion der Ergebnisse zwingend die Rezeption. Entsprechend wird der Effekt der Nützlichkeitseinschätzung⁹⁵ auf die Durchsicht⁹⁶ untersucht (Hypothese 26a; Tabelle 87 und

Tabelle 88) und in einem nächsten Schritt (Hypothese 26b) der Effekt der Durchsicht auf die Besprechung mit verschiedenen Personengruppen (vgl. Tabelle 89 und Tabelle 90)⁹⁷.

Als Gedankenstütze sei an dieser Stelle an die verschiedenen Rückmeldungen erinnert:

- Sofortbericht: Unmittelbar nach der Eingabe der 0/1-Kodierung der Tests erhalten die Lehrkräfte eine mit Balkendiagrammen visualisierte Übersicht

⁹⁵ Die Nützlichkeitseinschätzung ist entsprechend dem vorangegangenen Kapitel in die drei Einzelitems zu den Nützlichkeitsaspekten „Lernstandsdiagnose“, „Unterrichtsentwicklung“ und „Schulentwicklung“ und zusätzlich durch die Nützlichkeits-Skala (vgl. Instrumentenbeschreibung Teil 3) repräsentiert.

⁹⁶ Die Selbsteinschätzungen der Durchsicht wurden von den teilnehmenden Lehrkräften auf einer dreistufigen Skala von 1: „Systematisch, vollständig“ über 2: „punktuell, teilweise“ bis 3: „gar nicht“ eingeschätzt. Als Zusatzkategorie gab es 4: „keine Angabe möglich“. Antworten der Kategorie 4 wurden aus den Auswertungen ausgeschlossen.

⁹⁷ Die postulierte Wirkrichtung für diese beiden Hypothesen leitet sich aus dem Modell von Helmke ab. Letztlich lassen sich aus den vorliegenden Daten jedoch eher ungerichtete Zusammenhänge ableiten und, da es sich um einfache Regressionsanalysen handelt, ist der Regressionskoeffizient dem Korrelationskoeffizient gleichzusetzen.

entsprechend der Merkmale des Kompetenzstrukturmodells (Leitideen, prozessbezogenen Kompetenzen, Anforderungsbereiche) sowie nach Kompetenzstufen differenziert, wie hoch der Lösungsanteil der eigenen Lerngruppe ist.

- Tabellarische Übersicht: Unmittelbar nach der Eingabe der 0/1-Kodierung der Tests erhalten die Lehrkräfte eine tabellarische Zusammenstellung, aus der für jede Schülerin und jeden Schüler deutlich wird, ob die einzelnen Aufgaben korrekt (Code 1) oder falsch (Code 0) gelöst wurden. Zusätzlich enthält die Tabelle für jede Schülerin und jeden Schüler die Zusammenfassung der Lösungshäufigkeiten nach Kompetenzstrukturmodell und Kompetenzstufen.
- Ergebnisbericht: Mit einigem zeitlichen Abstand zur Dateneingabe werden die Informationen des Sofortberichts für die Lehrkräfte durch einen sozialen Vergleichswert, das korrigierte Landesmittel, ergänzt.
- Schulbericht: Dieser Bericht steht zeitgleich mit dem Ergebnisbericht zur Verfügung und umfasst die Ergebnisse aller Teilnahmen der letzten Jahre und in allen Fächern (Deutsch, Mathematik sowie 1. Fremdsprache), sodass ein gesamtschulischer Blick und der zeitliche Verlauf betrachtet werden können.

Bevor nun wie beschrieben auf Grundlage des Rahmenmodells von Helmke (2004; Abbildung 10) die prädiktiven Zusammenhänge betrachtet werden, ist der tatsächliche Ist-Stand der Durchsicht und Besprechung der Rückmeldungen ebenfalls von großem Interesse (Hypothesen 26c und d). Die Durchsicht der zur Verfügung stehenden Rückmeldungen wurde von den Lehrkräften auf einer Skala von 1: „systematisch, vollständig“ bis 3: „gar nicht“ eingeschätzt (Tabelle 85). Hierbei zeigt sich, dass für die eigentlichen, schriftlichen Berichte keine Fächerunterschiede zu erkennen sind, jedoch die tabellarische Zusammenfassung signifikant häufiger von Mathematik-Lehrkräften genutzt wird. Zusätzlich interessant ist daran, dass die tabellarische Übersicht den absolut niedrigsten Wert (also höchste Zustimmung/Systematik in der Durchsicht) von allen Fächern und Berichten aufweist. Eine pauschale Ablehnung oder Zustimmung ist für Hypothese 26c demnach nicht möglich. Wenig verwundert die geringere Häufigkeit in der Nutzung/Durchsicht der Schulberichte gegenüber dem Sofort- und Ergebnisbericht, da diese nur bedingt und mittelbar für den Unterricht der einzelnen Lehrkraft nutzbar sind.

Teil 4 – Ergebnisse

Ist-Stand in Hessen

Tabelle 85: Durchsicht der Rückmeldungen

	<i>N</i>	<i>M</i>	<i>SD</i>	
Durchsicht Sofortbericht				
Mathematik	34	1.38	0.55	$F(2,196) = 0.88$ $p = .42$
Deutsch	74	1.51	0.58	
Englisch	91	1.54	0.62	
Durchsicht Ergebnisbericht				
Mathematik	33	1.48	0.67	$F(2,196) = 0.86$ $p = .42$
Deutsch	73	1.56	0.60	
Englisch	92	1.52	0.65	
Durchsicht tabellarische Auswertung				
Mathematik	34	1.26	0.45	$F(2,104.08) = 4.51$ $p = .01^{98}$
Deutsch	75	1.49	0.60	
Englisch	92	1.57	0.65	
Durchsicht Schulbericht				
Mathematik	32	1.84	0.81	$F(2,176) = 2.09$ $p = .13$
Deutsch	67	2.03	0.76	
Englisch	80	2.15	0.66	

Die ersten Ergebnismrückmeldungen liegen den Lehrkräften unmittelbar nach Eingabe der Daten vor und werden nach einigen Wochen um elaborierte Rückmeldungen (etwa mit dem korrigierten Landesmittelwert) ergänzt. Diese Rückmeldungen sollten ihrer Intention nach als Grundlage der Diskussion verschiedenen Personengruppen dienen. In den Ergebnissen in Tabelle 86 lässt sich ablesen, dass sich für die Besprechung zwischen den Fächern unabhängig von der Durchsicht keine vom Zufall zu trennenden Unterschiede zeigen (Hypothese 26d ist abzulehnen). Es gibt jedoch fachübergreifend eine deutliche Tendenz, dass Lehrkräfte die Ergebnisse deutlich häufiger in ihrer eigenen Klasse oder mit einzelnen Schülerinnen und Schülern besprechen, als sie dies mit der Schulleitung oder Eltern tun. Es liegt also der Schluss nahe, dass eine Besprechung mit abnehmendem Kontakt zum eigenen Unterricht seltener vorgenommen wird. Interessant ist, dass die Lehrkräfte die unmittelbar zur Verfügung stehenden Vergleichsmöglichkeiten mit anderen Klassen oder dem korrigierten Landesmittelwert lediglich *punktuell/teilweise* (dieses Label entspricht dem Code 2) nutzen.

⁹⁸ Varianzgleichheit wurde geprüft (Levene-Test) und ist nicht vorhanden: $F(2,195) = 3.09, p < .05$.

Tabelle 86: Besprechung der Ergebnisse der Lernstandserhebungen differenziert nach Fächern

	N	M	SD	
Besprechung der Ergebnisse mit der ganzen Klasse				
Mathematik	34	1.74	0.67	$F(2,195) = 0.77$ $p = .47$
Deutsch	75	1.85	0.67	
Englisch	89	1.73	0.67	
Besprechung der Ergebnisse mit einzelnen Schülerinnen/Schülern				
Mathematik	34	1.94	0.69	$F(2,193) = 1.57$ $p = .21$
Deutsch	74	2.18	0.67	
Englisch	88	2.05	0.68	
Besprechung der Ergebnisse mit den Eltern				
Mathematik	34	2.50	0.51	$F(2,188) = 0.40$ $p = .67$
Deutsch	71	2.59	0.58	
Englisch	86	2.59	0.54	
Besprechung der Ergebnisse Klassenkonferenz/Team				
Mathematik	32	2.34	0.75	$F(2,191) = 1.43$ $p = .24$
Deutsch	73	2.30	0.72	
Englisch	89	2.15	0.67	
Besprechung mit der Fachschaft				
Mathematik	34	2.09	0.62	$F(2,91.92) = 1.46$ $p = .24^{99}$
Deutsch	70	2.26	0.74	
Englisch	85	2.07	0.65	
Besprechung der Ergebnisse im gesamten Kollegium				
Mathematik	34	2.62	0.60	$F(2,194) = 0.43$ $p = .65$
Deutsch	75	2.51	0.58	
Englisch	88	2.55	0.57	
Besprechung der Ergebnisse mit der Schulleitung				
Mathematik	34	2.35	0.69	$F(2,190) = 0.24$ $p = .79$
Deutsch	72	2.44	0.71	
Englisch	87	2.39	0.65	
Vergleich der Ergebnisse (z. B. mit anderen Klassen, dem korrigierten Landesmittelwert...)				
Mathematik	34	2.12	0.60	$F(2,188) = 0.04$ $p = .96$
Deutsch	72	2.14	0.63	
Englisch	85	2.15	0.59	

Untersucht man nun die postulierten Zusammenhänge zwischen Nützlichkeitseinschätzungen und der Durchsicht, so zeigen sich folgende Ergebnisse:

Tabelle 87: Regression der Nützlichkeit auf die Durchsicht der verfügbaren Berichte

	Sofortbericht			Ergebnisbericht			Tabellarische Übersicht			Schulbericht		
	<i>b</i>	<i>p</i>	<i>R</i> ²	<i>b</i>	<i>p</i>	<i>R</i> ²	<i>b</i>	<i>p</i>	<i>R</i> ²	<i>b</i>	<i>p</i>	<i>R</i> ²
Nützlichkeits-Skala	0.00	.284	.00	0.00	.155	.00	0.00	.375	.00	0.00	.003	.01
Lernstandsdiagnose	0.15	<.001	.07	0.11	<.001	.03	0.14	<.001	.06	0.04	.172	.00
Unterrichtsentwicklung	0.15	<.001	.07	0.13	<.001	.05	0.13	<.001	.05	0.09	.003	.02
Schulentwicklung	0.12	<.001	.04	0.12	<.001	.03	0.12	<.001	.04	0.12	<.001	.02

⁹⁹ Varianzgleichheit wurde geprüft (Levene-Test) und ist nicht vorhanden: $F(2,186) = 3.27$, $p = .04$.

Tabelle 88: Regression der Nützlichkeit auf die Durchsicht der verfügbaren Berichte (nur Mathematik)

	Sofortbericht			Ergebnisbericht			Tabellarische Übersicht			Schulbericht		
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²
Nützlichkeits-Skala	0.00	.199	.02	0.00	.340	.00	0.00	.245	.01	0.00	.092	.05
Lernstandsdiagnose	0.30	.002	.20	0.21	.081	.06	0.16	.069	.06	0.09	.537	-.02
Unterrichtsentwicklung	0.21	.022	.11	0.22	.037	.09	0.20	.012	.14	0.18	.160	.03
Schulentwicklung	0.27	.006	.17	0.24	.041	.09	0.22	.009	.15	0.15	.278	.01

Für die fachunabhängige Betrachtung fällt auf, dass alle drei per Einzelitem erfragten Perspektiven der Nützlichkeit einen hochsignifikanten Effekt auf die Durchsicht von Sofort- und Ergebnisbericht wie auch der tabellarischen Übersicht zeigen (vgl. Tabelle 89). Für den Schulbericht zeigt sich der Erwartung entsprechend eine steigende Vorhersagekraft je abstrakter der Nützlichkeitsaspekt ist ($b_{\text{Lernstandsdiagnose}} < b_{\text{Unterrichtsentwicklung}} < b_{\text{Schulentwicklung}}$, bei gleichzeitig steigender Signifikanz; vgl. jeweils letzten drei Spalten). Markant ist, dass die Ergebnisse der Skala von diesen Ergebnissen abweichen und dabei insbesondere die Nützlichkeit keine Vorhersagekraft für die Durchsicht des Sofort- und Ergebnisberichts sowie der tabellarischen Übersicht aufweist. Hypothese 26a lässt sich somit ebenfalls nur differenzierter betrachten. Die Einzelitemauswertung legt die Bestätigung der Hypothese nahe, während aus der Auswertung der Nützlichkeitsskala eher ein Ablehnen folgen müsste.

Die gleichen Analysen lediglich mit Mathematiklehrkräften führen zu leicht abweichenden Ergebnissen (Tabelle 88). Wiederum unterscheidet sich die Skala Nützlichkeit von den Ergebnissen der Einzelitems. Für den Prädiktor Lernstandsdiagnose ist ein signifikanter Effekt auf die Durchsicht des Sofortberichts erkennbar. Dies ist gut nachvollziehbar: Ergebnis- und Schulbericht werden in großem zeitlichen Abstand zur diagnostischen Situation bereitgestellt (meist etwa acht Wochen danach) und sind damit nur sehr eingeschränkt unmittelbar für Individualförderung und Lernstandsdiagnose nutzbar. Sowohl auf die Durchsicht von Sofort- als auch Ergebnisbericht und tabellarische Übersicht zeigen die Nützlichkeitsperspektiven Unterrichts- und Schulentwicklung einen positiven signifikanten Effekt. Durch diese Ergänzung erscheint das Bild für Hypothese 26a nochmals komplexer. Die Unterschiede in den Ergebnissen zwischen der Einzelitem- und der Skalenauswertung spiegeln sich wider, allerdings sich auf Einzelitemebene ein wie dargestellt leicht abweichendes Bild.

Nach der Rezeption erfolgt im Modell von Helmke die Reflexion und dabei gleichzeitig auch der Austausch mit anderen Personengruppen. Folglich sollte die Durchsicht der

Rückmeldungen die Besprechung der Ergebnisse im Sinne einer Reflexion positiv vorhersagen (Hypothese 26b). Die Durchsicht aller verfügbaren Berichte wirkt sich fachunabhängig hochsignifikant auf die Besprechung mit allen etwaigen Gruppen aus (vgl. Tabelle 89). Die Varianzaufklärung R^2 fällt für alle Effekte eher gering aus ($R^2_{max} = .16$). Bemerkenswert ist, dass gerade dieser maximale Wert für den Effekt der Durchsicht des Schulberichtes auf die Besprechung mit der Schulleitung gefunden wird (vorletzte „Dreierspalte“ in Tabelle 89). Es lässt sich also ein gerichteter Zusammenhang zwischen den beiden Bereichen Durchsicht von Berichten und Besprechung der Lernstandserhebungen mit verschiedenen Gruppen finden und statistisch absichern, allerdings ist die Bedeutung dieser Prädiktion eher gering einzuschätzen.

Differenzierter erscheint das Bild bei einer auf Mathematik eingeschränkten Stichprobe (Tabelle 90). Hierbei zeigt sich vor allem der Schulbericht als vergleichsweise entscheidend für Besprechungen mit Eltern ($b = 0.20$, $p = .07$) sowie verschiedenen Gremien innerhalb der Schule (Team, Fachschaft, Kollegium und Schulleitung).

Interessant sind zwei weitere Effekte:

- Für die Besprechung mit der Fachschaft fällt die Durchsicht des Sofortberichts marginal signifikant aus ($b = 0.33$, $p = .08$). Lehrkräfte, die ein stärkeres Gewicht auf die Durchsicht des Sofortberichtes legen und damit die nach Leitideen, Anforderungsbereichen, Kompetenzen etc. aggregierten Zusammenfassungen rezipieren, nutzen also häufiger die Möglichkeit, diese Ergebnisse in der Fachschaft zu besprechen.
- Die Durchsicht des Ergebnisberichts wirkt sich marginal signifikant positiv auf die Nutzung von Vergleichen der Ergebnisse (u. a. mit korrigiertem Landesmittelwert) aus ($b = 0.29$, $p = .05$). Die angesprochenen Vergleichsmaßstäbe werden in Gänze erst mit dem einige Wochen nach Durchführung bereitstehenden Ergebnisbericht geliefert. Insofern lässt sich dieses Ergebnis als eine Art Validierung der Messungen und des Verständnisses der Lehrkräfte für die gestellten Fragen werten.

Bezugsnehmend auf die Hypothese lässt sich wiederum kein klares Fazit formulieren.

Tabelle 89: Ergebnisse Regressionsanalyse Durchschnitt auf Besprechung

	Klasse			Schülerinnen und Schüler			Eltern			Team			Fachschaft			Kollegium			Schulleitung			Vergleich		
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²
Durchschnitt Sofort-rückmeldung	0.40	<.001	.11	0.35	<.001	.08	0.20	<.001	.03	0.29	<.001	.06	0.27	<.001	.05	0.15	.001	.02	0.20	<.001	.03	0.34	<.001	.09
Durchschnitt tabellarische Übersicht	0.36	<.001	.09	0.33	<.001	.08	0.24	<.001	.05	0.17	<.001	.02	0.24	<.001	.04	0.11	.013	.01	0.19	<.001	.02	0.35	<.001	.10
Durchschnitt Ergebnisbericht	0.39	<.001	.12	0.32	<.001	.08	0.23	<.001	.05	0.27	<.001	.06	0.30	<.001	.08	0.15	<.001	.02	0.24	<.001	.04	0.36	<.001	.12
Durchschnitt Schulbericht	0.22	<.001	.05	0.23	<.001	.05	0.32	<.001	.13	0.33	<.001	.12	0.33	<.001	.13	0.28	<.001	.10	0.39	<.001	.16	0.31	<.001	.12

Tabelle 90: Ergebnisse Regressionsanalyse Durchschnitt auf Besprechung nur Mathematiklehrkräfte

	Klasse			Schülerinnen und Schüler			Eltern			Team			Fachschaft			Kollegium			Schulleitung			Vergleich		
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²
Durchschnitt Sofort-rückmeldung	0.18	.37	.00	0.02	.92	-.03	-0.05	.72	-.02	0.28	.21	.02	0.33	.08	.06	0.01	.95	-.03	-.01	.96	-.03	0.29	.12	.04
Durchschnitt tabellarische Übersicht	0.27	.25	.01	0.00	1.00	-.03	0.23	.18	.02	-0.04	.87	-.03	0.23	.30	.00	-0.11	.60	-.02	0.23	.36	.00	0.29	.18	.02
Durchschnitt Ergebnisbericht	0.15	.41	-.01	0.13	.46	-.01	0.19	.14	.03	0.30	.14	.04	0.17	.28	.01	0.11	.47	-.01	0.12	.50	-.02	0.29	.05	.08
Durchschnitt Schulbericht	-0.08	.62	-.02	0.17	.26	.01	0.20	.07	.07	0.43	.01	.19	0.33	.01	.16	0.31	.02	.13	0.29	.06	.08	0.17	.21	.02

4.3.5 Nutzung der LSE

Als Konsequenz einer differenzierten *Rezeption* und *Reflexion* – also im Sinne der zuvor geschilderten Ergebnisse einer differenzierten Durchsicht und Besprechung – ist im Bezug gebenden Modell der Schritt *Aktion* anzunehmen; als *Aktion* werden die Ergebnisse zur Nutzung berichtet, wobei wiederum zunächst die im Modell postulierte Vorhersage geprüft wird. Wie bereits bei der Intensität der Auseinandersetzung bzw. der Besprechung der Ergebnismrückmeldungen fallen auch die regressionsanalytischen Betrachtungen der prädiktiven Wirkung der Nützlichkeit in allen vier Ausprägungen auf die abhängige Variable *Nutzung* hochsignifikant aus, ohne dass durch den jeweiligen Prädiktor eine größere Varianzaufklärung erreicht würde ($R^2_{max} = .06$; vgl. Tabelle 91). Reduziert man die Betrachtung lediglich auf Mathematiklehrkräfte, so zeigen sich kontrastierend keine signifikanten Effekte. Diese Ergebnisse werden aufgrund der geringen Bedeutsamkeit jedoch nicht weiter interpretiert ($R^2_{max} = .06$). Bei der Betrachtung der Einzelitems der Nutzung als abhängige Variable (anstelle der aus den Einzelitems gebildeten Skala) ergeben sich sehr ähnliche Werte, daher wird hierauf nicht näher eingegangen. Aus Platzgründen befindet sich die Ergebnistabelle im Anhang (vgl. 7.3.5). Für Mathematiklehrkräfte lässt sich daraus ableiten, dass die Hypothese 27a abgelehnt werden muss, während aufgrund der geringen Varianzaufklärung bei gleichzeitiger hoher Signifikanz eine Entscheidung im fächerübergreifenden Fall nicht möglich ist.

Tabelle 91: Regressionsanalyse der Nützlichkeit auf die Gesamtskala „Nutzung“

	allgemein			Mathematik		
	<i>b</i>	<i>p</i>	R^2	<i>b</i>	<i>p</i>	R^2
Lernstandsdiagnose	0.07	<.001	.03	0.11	.20	.02
Unterrichtsentwicklung	0.09	<.001	.04	0.04	.58	-.02
Schulentwicklung	0.10	<.001	.06	0.03	.77	-.03
Nützlichkeit	0.11	<.001	.05	0.06	.51	-.02

Neben der *a-priori*-Nützlichkeit sollte auch die Intensität der Auseinandersetzung einen positiven Einfluss auf die Nutzung haben. Operationalisiert als Durchsicht sowie als Besprechung der Ergebnisse sind die regressionsanalytischen Ergebnisse fachunabhängig und nur für Mathematiklehrkräfte in Tabelle 92 dargestellt. Dabei wird wie schon häufiger ersichtlich, dass es für die fachunabhängige Betrachtung zwar durchweg hohe Signifikanzen gibt, dabei jedoch die Effektstärke R^2 und der Regressionskoeffizient eher gering ausfallen. Auf Mathematik beschränkt zeigen sich für die Nutzung folgende Aspekte der Auseinandersetzung mit den Ergebnissen als relevant:

Teil 4 – Ergebnisse

Ist-Stand in Hessen

- Durchsicht tabellarische Übersicht ($b = 0.34, p = .03$)
- Besprechung mit der Klasse ($b = 0.28, p = .01$)
- Besprechung mit einzelnen Schülerinnen und Schülern ($b = 0.24, p = .03$)
- Besprechung mit den Eltern ($b = 0.39, p < .01$)

Daraus lässt sich folgern, dass Lehrkräfte, die Lernstandserhebungen eher nutzen, wenn die Ergebnisse der Lernstandserhebungen mit der Klasse, einzelnen Schülerinnen und Schülern und/oder den Eltern besprochen werden. Ebenfalls einen positiven Effekt hat die Nutzung der tabellarischen Übersicht als systematische Zusammenstellung der Ergebnisse für die Lehrkräfte. Insofern bleibt wie auch bei Hypothese 27a für die fachunabhängige Betrachtung eine Aussage über Hypothese 27b undeutlich, während für die ausgewählte Rückmeldeformate sowie ausgewählte Besprechungsgruppen die Hypothese für Mathematiklehrkräfte bestätigt werden kann.

Tabelle 92: Regressionsanalyse der Durchsicht bzw. Besprechung auf die Nutzung (als Skala)

	allgemein			Mathematik		
	<i>b</i>	<i>p</i>	<i>R</i> ²	<i>b</i>	<i>p</i>	<i>R</i> ²
Durchsicht Sofortbericht	0.14	<.001	.03	0.04	.78	-.03
Durchsicht Ergebnisbericht	0.11	<.010	.02	0.04	.77	-.03
Durchsicht tabellarische Übersicht	0.09	<.010	.01	0.34	.03	.10
Durchsicht Schulbericht	0.11	<.001	.03	-0.01	.89	-.03
Besprechung Klasse	0.09	<.010	.02	0.28	.01	.13
Besprechung Schülerinnen und Schüler	0.08	<.010	.01	0.24	.03	.10
Besprechung Eltern	0.11	<.001	.02	0.39	<.01	.15
Besprechung Team	0.11	<.001	.02	0.02	.83	-.03
Besprechung Fachschaft	0.10	<.010	.02	0.01	.91	-.03
Besprechung Kollegium	0.07	.030	.01	0.11	.38	-.01
Besprechung Schulleitung	0.12	<.001	.03	0.10	.34	.00
Vergleich Ergebnisse	0.12	<.001	.03	0.05	.68	-.02

Hypothese 27c postuliert Fachunterschiede in der Nutzung zugunsten der Mathematiklehrkräfte. In Tabelle 93 lassen sich jedoch varianzanalytisch für die Gesamtskala sowie die meisten Einzelitems keine signifikante Fachunterschiede nachweisen. Lediglich die Auswirkungen auf Leistungskontrollen zeigen eine marginale Signifikanz ($F(2,166) = 0.33, p = .07$) mit einem marginal signifikanten Paarvergleich zwischen Mathematik und Englisch ($p = .09$). Konsequenterweise muss die Hypothese 27c abgelehnt werden. Generell fallen die Angaben zur Nutzung der Ergebnisse für alle Fächer im Vergleich höher aus als die zuvor berichteten Werte für die Besprechung der Ergebnisse. Obwohl die Rückmeldungen also nicht durchgehend *punktuell/teilweise* besprochen werden, geben die Lehrkräfte bzgl. der meisten aufgeführten Aspekte zumindest an, die Ergebnisse *punktuell/teilweise* (entsprechend einem Mittelwert um

2) zu nutzen. Positiv fällt auf, dass die Lernstandserhebungen Mathematik dabei laut Aussagen der Lehrkräfte v. a. zum Einsatz kompetenzorientierter Unterrichtsmethoden, zum Nachdenken über den eigenen Unterricht sowie einer stärkeren Koordination innerhalb der Jahrgangsstufen führt (jeweils Mittelwerte unter 2).

Tabelle 93: Ergebnisse zur Nutzung der Lernstandserhebungen differenziert nach Fächern

	N	M	SD	
Gesamtskala				
Mathematik	33	1.97	0.47	$F(2,177) = 0.24$ $p = .79$
Deutsch	67	2.02	0.45	
Englisch	80	2.04	0.43	
Veränderung der Lernschwerpunkte für den aktuellen Unterricht				
Mathematik	31	2.13	0.56	$F(2,163) = 1.8$ $p = .16$
Deutsch	61	2.21	0.85	
Englisch	74	2.34	0.50	
Änderung d. Anordnung der Unterrichtsthemen im schuleigenen Stoffverteilungsplan/Fachcurriculum				
Mathematik	32	2.06	0.70	$F(2,165) = 2.00$ $p = .14$
Deutsch	62	2.26	0.57	
Englisch	74	2.31	0.57	
Veränderungen der Aufgabenformate in Leistungskontrollen				
Mathematik	32	2.09	0.47	$F(2,166) = 2.73$ $p = .07$
Deutsch	62	2.21	0.60	
Englisch	75	2.36	0.58	
Einsatz kompetenzorientierter Unterrichtsmethoden				
Mathematik	33	1.85	0.67	$F(2,171) = 0.33$ $p = .72$
Deutsch	64	1.73	0.70	
Englisch	77	1.77	0.63	
Methodisch-didaktische Überlegungen zum eigenen Unterricht				
Mathematik	33	1.82	0.58	$F(2,169) = 0.80$ $p = .45$
Deutsch	63	1.94	0.67	
Englisch	76	1.80	0.65	
Veränderung der Differenzierungsstruktur in meiner Klasse				
Mathematik	33	2.03	0.64	$F(2,167) = 0.43$ $p = .65$
Deutsch	61	1.90	0.70	
Englisch	76	1.91	0.72	
Neuanschaffung moderner Lehrwerke, die kompetenzorientiertes Lernen fördern				
Mathematik	33	1.97	0.73	$F(2,170) = 0.07$ $p = .94$
Deutsch	63	1.95	0.63	
Englisch	77	1.92	0.72	
Besuch von Fortbildungsveranstaltungen				
Mathematik	32	2.00	0.67	$F(2,166) = 0.07$ $p = .94$
Deutsch	59	2.00	0.67	
Englisch	78	1.96	0.73	
Stärkere Koordination der Arbeit in den Jahrgangsstufen ¹⁰⁰				
Mathematik	31	1.87	0.62	$F(2,81.952) = 0.85$ $p = .43$
Deutsch	61	1.98	0.53	
Englisch	78	2.05	0.72	

¹⁰⁰ Varianzgleichheit wurde geprüft (Levene-Test) und ist nicht vorhanden: $F(2,167) = 4.04$, $p = .02$. Entsprechend wird auf den Welch-Test zugegriffen (Luhmann, 2011).

4.3.6 Belastung der Lehrkräfte im Zusammenhang mit Alter sowie Klassen-
größe

Als Voraussetzung einer bewussten Nutzung bzw. eines bewussten Umgangs mit den Lernstandserhebungen wird auf Seiten der Lehrkraft die wahrgenommene Belastung herangezogen, die laut Hypothese 28 von der Klassengröße sowie laut Hypothese 29 vom Alter beeinflusst wird. Äußerst kontraintuitiv fällt der Zusammenhang zwischen der Klassengröße und der Belastung aus: Intuitiv wäre ein positiver Zusammenhang zu vermuten, also bei mehr Schülerinnen und Schülern und folglich mehr Korrektur- und Eingabeaufwand eine höhere Belastung. Tabelle 94 zeigt an, dass sowohl im Prä- als auch im Posttest die Korrelationen negativ ausfallen, also je größer die Lerngruppe ist, desto weniger Belastung wird berichtet. Es ist jedoch zu fragen, ob dies vielleicht eine zufällige Konstellation ist, da der Koeffizient für den Prätest lediglich marginal signifikant wird und im Posttest keine Signifikanz vorhanden ist. Dabei liegt die Stichprobe 1 zugrunde mit einem N von 27 Mathematiklehrkräften, da anstelle der nicht direkt erfragten Anzahl der Schülerinnen und Schüler die mittlere Anzahl der Prä- und Postschülerleistungstests der jeweiligen Lehrkräfte herangezogen wurde. Demnach sind in den Lerngruppen im Mittel $M = 21.34$ Schülerinnen und Schüler ($SD = 5.80$). Hypothese 28 lässt sich mit den vorliegenden Daten jedoch in keinem Falle stützen und muss verworfen werden.

Tabelle 94: Korrelationen zwischen Klassengröße und der Belastung von Mathematik-Lehrkräften der Stichprobe 1 in Prä- und Posttest

	<i>r</i>	<i>p</i>
Belastung Prätest	-.35	.07
Belastung Posttest	-.22	.27

Neben der Klassengröße wurde des Weiteren das Alter bzw. aus Datenschutzgründen die Unterrichtserfahrung im Zusammenhang mit der *Belastung*, dem *förderdiagnostischen* sowie dem *unterrichtsplanerischen Nutzen* untersucht. In Tabelle 95 sind die Korrelationskoeffizienten für die Unterrichtserfahrung (getrennt nach Stichproben) zu den genannten Skalen aufgeführt. Einen direkten Zusammenhang mit den nutzungsorientierten Skalen *Förderung* und *Planung* lässt sich nicht belegen. Für die Belastung fällt allerdings der negative Korrelationskoeffizient von $r = -.41$, $p = .03$ für den Posttest der Stichprobe 1 auf: Dementsprechend schätzen Lehrkräfte mit mehr Unterrichtserfahrung zum Ende des Schuljahres die Belastung durch die Lernstandserhebungen geringer ein. Betrachtet man die Differenz aus Belastungs-Posttestwert und Belastungs-Prätestwert in Korrelation zur Unterrichtserfahrung, so ergibt sich für

Stichprobe 1 keine signifikante Korrelation ($r = -.22$, $p = .28$). Daraus lässt sich also schließen, dass es keinen Zusammenhang in der Entwicklung von Prä- zu Posttest mit dem Alter gibt, jedoch zum Schuljahres- und damit auch Projektende von VELM-8 erfahrene Lehrkräfte weniger Belastung erlebten. Zu vermuten ist, dass beispielsweise jüngere Lehrkräfte zum Ende des Schuljahres wegen der geringeren Erfahrung und weniger vorhandenen Materials generell mehr Belastung erfahren (vgl. Schaar-schmidt, 2005, 2007). Die nutzungsorientierten Hypothesen 29b und 29c müssen demnach abgelehnt werden. Hypothese 29a müsste konkretisiert werden, da Indizien zeigen, dass ein Zusammenhang bestehen könnte (beispielsweise auch der marginal signifikante Korrelationskoeffizient von $r = -.23$ für Stichprobe 2).

Tabelle 95: Korrelationen zwischen Unterrichtserfahrung und Einstellungen zu den Lernstandserhebungen (getrennt nach Stichproben)

	Belastung				Förderung				Planung			
	Prä		Post		Prä		Post		Prä		Post	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
Stichprobe 1	-.16	.41	-.41	.03	.06	.76	.03	.89	.11	.61	-.02	.92
Stichprobe 2	-.23	.08			.00	.98			.02	.87		

4.3.7 Deskriptive Statistik zur Lage in Hessen

Nachfolgend werden explorative deskriptive Ergebnisse aus der Onlinebefragung hessischer Lehrkräfte berichtet. Diesen Ergebnissen liegen dabei ausschließlich Einzelitems zu Grunde.

Einen wesentlichen Beitrag zur Nutzung der Lernstandserhebungen sollen wie dargelegt die *didaktischen Kommentare* leisten. Daher ist von großem Interesse, wie diese Materialien von den Lehrkräften aufgenommen werden. Dabei zeigt sich, dass lediglich etwas mehr als die Hälfte der teilnehmenden Lehrkräfte von den *didaktischen Kommentaren* wissen (vgl. Tabelle 96), die Bewertung der Lehrkräfte allerdings positiv ausfällt (Tabelle 97). Zumindest im Fach Mathematik vergibt keine der bewertenden Lehrkräfte die Schulnote „Mangelhaft“ oder „Ungenügend“ und der Mittelwert liegt mit 2.29 (0.83) im Bereich der Schulnote „Gut“. Die bereits im hypothesengeleiteten Vorgehen beschriebene Sonderstellung von Mathematik (bessere Bewertung, weniger Probleme etc.) zeigt sich auch bei der Bewertung des *didaktischen Materials* mit einer signifikanten Varianzanalyse ($F(2,78) = 3.60$, $p < .05$). mit marginal signifikanten Paarvergleichen von Deutsch ($p = .08$) und Englisch ($p = .06$). Trotz der m. E. besseren Bewertung sind bei der Nutzung des *didaktischen Materials* keine fachbezogenen Unterschiede zu erkennen (Tabelle 98) und die Lehrkräfte aller Fächer geben

Teil 4 – Ergebnisse

Ist-Stand in Hessen

an, das Material im Mittel noch „punktuell/teilweise“ zu nutzen. Auch die nachfolgend gestellten differenzierteren Fragen zur Nutzung des *didaktischen Materials* offenbaren keine Unterschiede zwischen den Fächern (vgl. Tabellen 99 bis 102). Auffallend ist jedoch, dass die differenzierteren Fragen durchweg negativer eingeschätzt werden als die allgemeine Frage.

Tabelle 96: Deskriptive Statistik zur Frage: „Kennen Sie das didaktische Material?“

	D	E	M
<i>N</i>	81	95	33
Material bekannt	58%	65%	58%

Tabelle 97: Deskriptive Statistik zur Bewertung des didaktischen Materials mit Schulnoten

	D	E	M
<i>N</i>	34	33	14
<i>M</i>	2.85	2.36	2.29
<i>SD</i>	0.99	0.70	0.83
theoretisches Minimum	1.00	1.00	1.00
Minimum	2.00	1.00	1.00
1. Quartil	2.00	2.00	2.00
Median	3.00	2.00	2.00
3. Quartil	3.00	3.00	3.00
Maximum	6.00	4.00	4.00
theoretisches Maximum	6.00	6.00	6.00
	$F(2,78) = 3.60, p < .05$		

Tabelle 99: Deskriptive Statistik zur Frage: „Die Zusatzinformationen zur Aufgabe helfen mir die Klassenergebnisse richtig einzuordnen.“¹⁰²

	D	E	M
<i>N</i>	28	31	14
<i>M</i>	2.57	2.42	2.43
<i>SD</i>	0.69	0.81	0.65
theoretisches Minimum	1.00	1.00	1.00
Minimum	2.00	2.00	1.00
1. Quartil	2.00	2.00	2.00
Median	2.50	2.00	2.50
3. Quartil	3.00	3.00	3.00
Maximum	5.00	5.00	3.00
theoretisches Maximum	5.00	5.00	5.00
	$F(2,75) = 1.30, p = .28$		

Tabelle 98: Deskriptive Statistik zur Frage: „Wie haben Sie mit dem didaktischen Material gearbeitet?“¹⁰¹

	D	E	M
<i>N</i>	30	33	14
<i>M</i>	2.17	2.21	2.07
<i>SD</i>	0.38	0.42	0.47
theoretisches Minimum	1.00	1.00	1.00
Minimum	2.00	2.00	1.00
1. Quartil	2.00	2.00	2.00
Median	2.00	2.00	2.00
3. Quartil	2.00	2.00	2.00
Maximum	3.00	3.00	3.00
theoretisches Maximum	3.00	3.00	3.00
	$F(2,74) = 0.57, p = .57$		

Tabelle 100: Deskriptive Statistik zur Frage: „Die Anregungen zur Weiterarbeit finde ich hilfreich für meine eigene Arbeit.“¹⁰¹

	D	E	M
<i>N</i>	32	32	14
<i>M</i>	2.84	2.56	2.43
<i>SD</i>	0.85	0.8	0.76
theoretisches Minimum	1.00	1.00	1.00
Minimum	2.00	1.00	1.00
1. Quartil	2.00	2.00	2.00
Median	3.00	2.00	2.00
3. Quartil	3.00	3.00	3.00
Maximum	5.00	5.00	4.00
theoretisches Maximum	5.00	5.00	5.00
	$F(2,75) = 1.61, p = .21$		

¹⁰¹ Antwortformat: 1: „systematisch, vollständig“; 2: „punktuell, teilweise“; 3: „gar nicht“

¹⁰² Antwortformat: 1: „++“ bis 5: „--“

Tabelle 101: Deskriptive Statistik zur Frage: „Ich nutze die didaktischen Materialien zur Information über didaktische Konzepte und Entwicklungen.“¹⁰³

	D	E	M
N	31	31	13
M	2.52	2.32	2.54
SD	0.51	0.54	0.66
theoretisches Minimum	1.00	1.00	1.00
Minimum	2.00	1.00	1.00
1. Quartil	2.00	2.00	2.00
Median	3.00	2.00	3.00
3. Quartil	3.00	3.00	3.00
Maximum	3.00	3.00	3.00
theoretisches Maximum	3.00	3.00	3.00
	$F(2,72) = 1.21, p = .30$		

Tabelle 102: Deskriptive Ergebnisse zur Frage: „Ich nutze die aktuellen didaktischen Materialien als Übungsmaterial nach dem Lernstand.“¹⁰²

	D	E	M
N	30	32	13
M	2.43	2.31	2.31
SD	0.57	0.54	0.63
theoretisches Minimum	1.00	1.00	1.00
Minimum	1.00	1.00	1.00
1. Quartil	2.00	2.00	2.00
Median	2.00	2.00	2.00
3. Quartil	3.00	3.00	3.00
Maximum	3.00	3.00	3.00
theoretisches Maximum	3.00	3.00	3.00
	$F(2,72) = 0.42, p = .66$		

4.4 Zusammenfassung der Ergebnisse nach Hypothesen

Der vorliegende Ergebnisteil orientiert sich nicht konsequent an der Reihenfolge der Hypothesen, da dies u. U. mit den verschiedenen Verfahren und Stichproben unübersichtlich würde. Daher soll nun eine kurze Übersicht die Ergebnisse zu den einzelnen Hypothesen zusammenfassen (ein Resümee der Clusteranalyse ist in Abschnitt 4.2.2 zu finden).

In Tabelle 103 lassen sich für die ersten sechs Haupt-Hypothesen die Ergebnisse in einer Übersicht erkennen. Wie auch für die folgenden Tabellen gilt hierbei, dass Effekte nur dann angegeben sind, wenn sie einen p -Wert $< .15$ aufweisen. „+“ und „-“ geben die Richtung des Effektes an. Schwarze Zellen stehen für nicht durchgeführte Analysen. Generell gilt, dass sich die verschiedenen Verfahren weitestgehend gegenseitig stützen.¹⁰⁴ Nach dem harten statistischen Kriterium ($p < .05$) lässt sich aus den messwiederholten Varianzanalysen ablesen, dass es keine signifikanten Effekte der Interaktion zwischen Zeit und Gruppe gibt und damit die Hypothesen H1a, H2, H3a, H4a, H5a und H6 abgelehnt werden müssen. Folglich lässt sich keine der vermuteten Entwicklungen in den Einstellungen der Lehrkräfte gegenüber den

¹⁰³ Antwortformat: 1: „systematisch; vollständig“; 2: „punktuell, teilweise“; 3: „gar nicht“

¹⁰⁴ An dieser Stelle sei kurz darauf verwiesen, dass unter dem Namen *Lords Paradoxon* das Phänomen beschrieben ist, dass sich Ergebnisse von Varianz- und Kovarianzanalysen bei nicht randomisierten Prä-Post-Designs widersprechen können. Die Gründe hierfür liegen u. a. in Messfehlern und der unterschiedlichen Modellierung der Verfahren (Eid et al., 2011; Lord, 1967, 1969; Tu, Gunnell & Gilthorpe, 2008). Daher greife ich bei allen Auswertungen auch auf die Sichtkontrolle der Abbildungen zurück.

Teil 4 – Ergebnisse

Zusammenfassung

Lernstandserhebungen statistisch abgesichert nachweisen. Legt man etwas „weitere“ Kriterien bzgl. der Signifikanz an, so zeigen sich zumindest marginal signifikante Einflüsse der Gruppe (z. B. *förderdiagnostischen Nutzen*). Daraus lässt sich schließen, dass die Gruppen sich unabhängig von der Zeit (und somit auch unabhängig von der Fortbildungsteilnahme) per se im mittleren Antwortniveau unterscheiden. In Abbildung 30 zeigt sich die höhere Einschätzung der Projektgruppe unabhängig vom Messzeitpunkt. Dabei zeigt sich in der mANOVA zusätzlich ein (knapp nicht) marginal signifikanter Effekt der Zeit, sodass eine Veränderung im Laufe der Zeit weiterhin vermutet werden kann. Die Ergebnisse der ANCOVA unterstützen für den *förderdiagnostischen Nutzen* den Einfluss der Gruppe signifikant (H4b). Für die *Akzeptanz* ergibt neben einem positiven Zeiteffekt in der mANOVA in der ANCOVA ein marginal signifikant negativer Interaktionsterm, der bedeutet, dass die Vorhersage der Posttestleistung für die Kontrollgruppe in höherem Maße durch die Prätestakzeptanz erfolgt als für die Projektgruppe. Demnach liegt bei der Projektgruppe eine stärkere Veränderung über die Zeit vor. Diese Entwicklung lässt sich jedoch in der mANOVA in Form eines Interaktionsterms nicht bestätigen (H1a). Entsprechend können die Hypothesen 1(a+b) und 4a zunächst nicht als bestätigt gelten, während H4b angenommen werden kann. Es lassen sich jedoch Indizien finden, die eine weitere Prüfung der Hypothesen nahelegen, so beispielsweise ergänzend zu den berichteten Ergebnissen ein marginal signifikanter Interaktionsterm der rangtransformierten messwiederholten Varianzanalyse für die *Akzeptanz* (vgl. Anhang 7.3.1)¹⁰⁵. Für die Skalen *Lehrplanvalidität* und *selektionsdiagnostischer Nutzen* ergeben sich in der (r)mANOVA keine Effekte mit Gruppenbezug (Haupteffekt der Gruppe oder Interaktion). In diesem Sinne muss auch H6 zum *selektionsdiagnostischen Nutzen* abgelehnt werden. Für die Skala *Lehrplanvalidität* zeigt die rmANOVA einen signifikanten Effekt der Zeit. Da jedoch neben der positiven zeitlichen Veränderung im Laufe des Schuljahres bzw. der Projektteilnahme kein Effekt der Gruppe zu verzeichnen ist, muss die Hypothese 2 ebenfalls abgelehnt werden. Bei der Skala *Belastung durch die Lernstandserhebungen* lassen sich in der mANOVA ein marginal signifikanter Interaktionseffekt (H3a) und bei der ANCOVA ein marginal signifikanter Effekt der Gruppe (H3b) ablesen. Die Belastung sinkt innerhalb der Projektgruppe, während die Kontrollgruppe im Laufe des

¹⁰⁵ Diese Ergebnisse befinden sich lediglich im Anhang, da die Voraussetzungen für die parametrischen Verfahren bei der Skala *Akzeptanz* erfüllt sind. Die non-parametrischen Verfahren bieten jedoch aufgrund der geringen Stichprobengröße eine gute Ergänzung und Unterstützung der Ergebnisse parametrischer Verfahren.

Schuljahres von höherer Belastung durch die LSE berichtet. Da jedoch nur marginale Signifikanz ($p < .10$) erreicht wird, kann die Hypothese 3 nur als vorläufig bestätigt angesehen werden. In Bezug auf den *unterrichtsplanerischen Nutzen* fallen die Effekte mit Gruppeneinfluss (mANOVA: Interaktion; ANCOVA: Haupteffekt der Gruppe) durchweg nicht marginal signifikant aus, verfehlen dieses Niveau jedoch alle lediglich vergleichsweise knapp ($p \leq .15$). Insofern können die Hypothesen 5a und 5b nicht bestätigt werden, mit Blick auf Abbildung 31 und die nur knapp verfehlten signifikanten Effekten scheint es dennoch ratsam, diese Hypothese erneut zu prüfen.

Insbesondere für die Skalen zu den Einstellungen und den Nutzen der Lernstandserhebungen scheint es sinnvoll den Einfluss der Fortbildungsveranstaltungsteilnahme¹⁰⁶ in der Panel-Stichprobe 1 für die Projektgruppe zu kontrollieren: Dabei zeigt sich lediglich für die Skala *förderdiagnostischer Nutzen* ein signifikanter Effekt ($b = .14$, $t = 2.28$, $p = .03$): Je mehr Fortbildungsveranstaltungen die Lehrkräfte im Rahmen des Projekts VELM-8 besucht haben, desto höher wird der *förderdiagnostische Nutzen* der LSE eingeschätzt. Allerdings ergeben die Auswertungen zu H4a und H4b keine Effekte.

Die Hypothesen 7 bis 21 drehen sich um ausgewählte Aspekte der Unterrichtsqualität. Wie im Methodenteil bereits erläutert (vgl. Abschnitt 3.4.1.2), wird auf die Auswertung der Hypothesen 16 und 21 verzichtet, da die betreffenden Skalen leider keine ausreichende Reliabilität aufweisen. Für die restlichen Hypothesen wurden die Ergebnisse in Tabelle 104 und Tabelle 105 zusammengefasst.

Für einige der Skalen zeigen sich keine signifikanten Effekte mit Gruppenbezug. Die zu diesen Skalen gehörenden Hypothesen müssen entsprechend basierend auf den vorliegenden Daten abgelehnt werden:

H8	Variation der Aufgabenschwierigkeit	keine Effekte
H9	Kognitiv anspruchsvolle Aufgaben	keine Effekte
H10	Kognitiv herausforderndes Üben	keine Effekte
H12(a+b)	Umgang mit Schülerbeiträgen	keine Effekte
H14(a+b)	Fürsorgliche Anleitung und Kontrolle	Effekt der Zeit (rANCOVA)
H17	Gemeinsame Wissenskonstruktion	Keine Effekte
H19(a+b)	Persönliches Vertrauen	Effekt der Zeit (rmANOVA)

¹⁰⁶ theoretisches Minimum: 0; theoretisches Maximum: 4;
Minimum: 0; 1. Quartil: 0; Mittelwert: 1.85; 3. Quartil: 3.75; Maximum: 4; Standardabweichung: 1.76

Teil 4 – Ergebnisse

Zusammenfassung

Drei weitere Hypothesen lassen sich in ihrem Ergebnis ebenfalls zusammenfassen: Hypothese 11(a+b), Hypothese 13(a+b) und Hypothese 15(a+b) zeigen im Wesentlichen in den (rangtransformierten) messwiederholten Varianzanalysen Haupteffekte der Gruppe, für die Hypothesen 11 und 13 bestätigt durch Haupteffekte der (rangtransformierten) Kovarianzanalysen. Unter Berücksichtigung der zugehörigen Abbildungen (repetitives Üben: Abbildung 38; Erarbeitung durch LPS: Abbildung 41; kognitive Selbstständigkeit: Abbildung 43) fallen sehr deutlich die zu beiden Messzeitpunkten vorhandenen Niveauunterschiede zwischen Kontroll- und Projektgruppe auf. Die Lehrkräfte der Projektgruppe berichten demnach einen signifikant selteneren Einsatz von *einschleifendem Üben* sowie *fehlervermeidender Erarbeitung* durch sich als Lehrkraft selbst, geben jedoch an, Schülerinnen und Schüler signifikant eher in deren *kognitiven Selbstständigkeit* zu unterstützen. Trotz dieser Einschätzungen der Lehrkräfte der Projektgruppe zum eigenen Unterricht folgt aus diesen Ergebnissen, dass die Hypothesen 11a, 13a und 15a abgelehnt werden müssen, da durch die Projektteilnahme sich nicht die postulierten Entwicklungen ergeben haben. Ein Gruppen-Unterschied zum Posttest ist in den Einschätzungen jedoch zu verzeichnen, so dass die Hypothesen 11b und 13b bestätigt werden können. In diesem Sinne sollten die Hypothesen ggf. unter strengeren methodischen Vorgaben erneut geprüft werden, um einen Fortbildungseinfluss sicher ausschließen oder bestätigen zu können.

Für die verbleibenden drei Hypothesen (7, 18 und 20) lässt sich aus den Ergebnissen ableiten, dass diese (vorläufig) bestätigt werden können. Für die Hypothese 7 (vgl. auch Abbildung 34) und die Hypothese 18 (vgl. auch Abbildung 39) zeigen sich bei den messwiederholten Varianzanalysen Interaktionseffekte. Für die Hypothese 20 zur *Diagnose im sozialen Bereich* fällt der Interaktionseffekt der rangtransformierten Varianzanalyse marginal signifikant aus, sodass in allen Fällen empirisch abgesichert die vermutete Entwicklung nachgewiesen ist.

Als letzte Zeile in Tabelle 105 wird die Hypothese 22 zur Leistungsentwicklung der Schülerinnen und Schüler aggregiert auf Lehrkraft-Ebene zusammengefasst. Hierbei zeigt sich erwartbar eine enorme Veränderung im Laufe des Schuljahres mit einem korrespondierenden Haupteffekt der Zeit in der messwiederholten Varianzanalyse. Der Haupteffekt der Gruppe fällt ebenfalls signifikant aus, was insbesondere nach Inspektion von Abbildung 47 eindeutig einen massiven Niveauunterschied zwischen Projekt- und Kontrollgruppe zugunsten der erstgenannten bedeutet. In der ANCOVA

zeigt sich hingegen lediglich ein Einfluss der Prätest-Leistung. Die Hypothesen 22a (mangels eines Interaktionseffekts in der mANOVA) und 22b (mangels Einflusses der Gruppe in ANCOVA) müssen dementsprechend abgelehnt werden.

Der dritte Hypothesenblock dreht sich schließlich weniger um das Projekt VELM-8 als um die allgemeine Wahrnehmung und Einschätzung der hessischen Lehrkräfte zu den Lernstandserhebungen. In Bezug auf die Lehrplanvalidität zeigt sich bereits in Hypothese 2 ein hohes Ausgangsniveau für die Lehrkräfte aus VELM-8. Sowohl Projekt- als auch Kontrollgruppe weichen signifikant vom semantischen Median ab. Stichprobe 2 liegt ebenfalls mit einem großen Effekt von *Cohens d* = 1.16 signifikant über dem semantischen Median. In der Online-Abfrage wurde die Lehrplanvalidität nur mittelbar über den Grad der Ungewohntheit der Aufgaben erfragt. Unter Kontrolle der Teilnahme (da nur so eine aussagekräftige Einschätzung möglich erscheint) ergibt sich jedoch für die Online-Stichprobe lediglich eine marginal signifikante Abweichung vom semantischen Median für Mathematik. Unterschiede in der Einschätzung der drei Fächer Deutsch, Englisch und Mathematik lassen sich dabei jedoch nicht erkennen. Insofern lässt sich die Hypothese 23a tendenziell für Mathematiklehrkräfte annehmen, wobei keine bedeutsamen Fachunterschiede vorliegen und somit Hypothese 23b abgelehnt werden muss.

In den Paarvergleichen der Auswertungszeit der Fächer lassen sich hochsignifikante Unterschiede erkennen, sodass die Hypothese 24a eindeutig bestätigt wird. Die Lehrkräfte wurden im Online-Fragebogen sowohl gebeten, ihre Einschätzung zur Nützlichkeit der Lernstandserhebungen als Instrument zur Lernstandsdiagnose, zur Unterrichtsentwicklung und zur Schulentwicklung (jeweils erhoben als Einzelitem) als auch in Form einer Skala anzugeben. Sowohl für Mathematik als auch fachübergreifend zeigen sich hochsignifikante Effekte der Auswertungszeit für alle drei separat erfragten Nützlichkeitsperspektiven als auch für die Skala, welche jedoch in ihrer Bedeutung äußerst gering ausfallen ($b_{max} = .006$, $p < .001$ fachunabhängig; $b_{max} = .001$, $p < .001$ für Mathematik). In konsequenter Folge aus Fachunterschieden der Auswertungszeit und prädiktiven Zusammenhängen selbiger mit der Nützlichkeit, ergeben sich auch für die Nützlichkeit signifikante Fachunterschiede (Tabelle 83). Herauszustellen ist hierbei, dass – zusätzlich zu den Paarvergleichen der Fächer – Mathematik als einziges Fach in Bezug auf alle vier Nützlichkeitsaspekte Mittelwerte aufweist, die gemessen am semantischen Median von 3 als Zustimmung zu werten sind (vgl.

Tabellen 85 und 86). Insofern lässt sich Hypothese 25a zwar als statistisch abgesichert ansehen, dies jedoch nur mit zu vernachlässigender Bedeutsamkeit, und Hypothese 25b kann als bestätigt gelten.

Betrachtet man nun unabhängig voneinander Durchsicht und Besprechung der Rückmeldungen vor dem Hintergrund von Fachunterschieden, fällt zunächst das generell hohe Niveau der Durchsicht auf: Lediglich die Mittelwerte zur Durchsicht des Schulberichts fallen für Deutsch und Englisch über dem semantischen Median von 2 aus¹⁰⁷. Abgesehen von der tabellarischen Übersicht ergeben sich jedoch keine signifikanten Unterschiede zwischen den Fächern. Dass Mathematiklehrkräfte angeben, gerade die tabellarische Übersicht am häufigsten gegenüber den anderen Rückmeldungen und gegenüber den anderen Fächern zu nutzen, ist hierbei bemerkenswert. Für die Besprechung der Ergebnisse zeigen sich keine signifikanten Unterschiede zwischen den Fächern, allerdings fällt auf, dass tendenziell die Besprechungsmittelwerte mit Distanz zum eigenen Unterricht abnehmen: Die Besprechung in der Klasse und mit einzelnen Schülerinnen und Schülern fallen gegenüber der Besprechung im Kollegium bzw. mit der Schulleitung höher aus. Bezogen auf die Hypothesen 26c und 26d lässt sich jedoch vermerken, dass diese abgelehnt werden müssen, wobei die tabellarische Übersicht eine Ausnahme darzustellen scheint.

Die per Skala erfasste Einschätzung der Nützlichkeit hat in Bezug auf die Durchsicht der verfügbaren Berichte lediglich für den Schulbericht eine signifikante Bedeutung, bei einer allerdings minimaler Varianzaufklärung und nicht von Null unterscheidbarem Regressionskoeffizient. Für die per Einzelitem erhobene Nützlichkeitseinschätzung zeigen sich umgedreht für die anderen Berichte signifikante Prädiktionen. Dieses Bild ist auch für Mathematik ähnlich mit jedoch geringer ausfallenden p -Werten bei gleichzeitig höheren R^2 -Werten. Eine mögliche Erklärung hierfür könnten in der Formulierung der Items der Nützlichkeitsskala liegen, die eher auf gesamtschulische Entwicklungen fokussiert ist (vgl. Tabelle 32). Insofern lässt sich die Hypothese 26a (Einfluss der Nützlichkeit auf die Durchsicht der Rückmeldungen) zunächst nur unter Vorbehalt für die unmittelbar verfügbaren Rückmeldungen bestätigen.

Im Modell folgt der Rezeption die Reflexion, sodass ein gerichteter Zusammenhang der Durchsicht von Rückmeldungen auf die Besprechung der Ergebnisse anzunehmen

¹⁰⁷ Die Einschätzung wurde dreistufig von 1: „systematisch/vollständig“ bis 3: „gar nicht“ vorgenommen.

ist. Dieser Zusammenhang ist fachunabhängig für alle vier Durchsichten (Sofortbericht, Ergebnisbericht, tabellarische Übersicht, Schulbericht) hochsignifikant. Für Mathematiklehrkräfte fallen die signifikanten Ergebnisse in ihrer Anzahl geringer aus: Die Durchsicht des Schulberichts führt demnach zu einer systematischeren Besprechung mit den Eltern sowie den Schulgremien (Team, Fachschaft und Kollegium; Schulleitung nur marginal signifikant; Tabelle 90). Die Durchsicht des Sofortberichtes wirkt sich auf eine häufigere Besprechung innerhalb der Fachschaft aus, sodass hier wohlwollend die Anregung didaktischer Diskussionen innerhalb der Mathematik-Fachschaft auf Grundlage der Lernstandserhebungen vermutet werden kann. Zusammenfassend lässt sich ein positiver Zusammenhang der Durchsicht der verfügbaren Rückmeldungen und der Besprechung in verschiedenen Gruppen/Gremien zunächst bestätigen (Hypothese 26b).

Letztlich interessiert vor allem die Nutzung der Lernstandserhebungen zur Unterrichts- und Schulentwicklung. Die Werte der Nutzungsskala werden fachunabhängig hochsignifikant aber mit geringer Bedeutsamkeit durch alle Aspekte der Nützlichkeit vorhergesagt ($b_{max} = .11$, $p < .001$). Für Mathematiklehrkräfte zeigt sich kontrastierend diese Vorhersagekraft der Nützlichkeit jedoch nicht. Allgemein lässt sich Hypothese 27a also bestätigen, mathematikbezogen muss sie verworfen werden. Ebenfalls hochsignifikant, jedoch gering bedeutsame Effekte weist die Besprechung der Ergebnisse auf die Nutzung in der fachunabhängigen Betrachtung auf. Für Mathematik zeigen sich hier signifikante Effekte für die tabellarische Übersicht, die Besprechung der Ergebnisse mit der Klasse, einzelnen Schülerinnen und Schülern sowie den Eltern. Generell lässt sich Hypothese 27b also ebenfalls als bestätigt ansehen, wobei die Ergebnisse in Mathematik etwas differenzierter ausfallen: Vor allem unmittelbar auf den Unterricht bezogene Besprechungsperspektiven weisen Effekte auf die Nutzung auf und eine systematische Übersicht unterstützt die Nutzung ebenfalls. Die in Hypothese 27c vermuteten Fachunterschiede lassen sich hingegen nicht finden. Lediglich die Veränderung von Lernkontrollen wird in Mathematik marginal signifikant häufiger angegeben als in den sprachlichen Fächern ($p = .07$), jedoch weisen Mathematiklehrkräfte generell eine sehr positive Einschätzung der genannten Nutzungsaspekte auf keiner der Mittelwerte liegt über 2.13. Dies lässt die Interpretation zu, dass Mathematiklehrkräfte keinen der genannten Nutzungsaspekte im Mittel *gar nicht* nutzen.

Teil 4 – Ergebnisse

Zusammenfassung

Durch die Intervention im Projekt VELM-8 hat sich das Belastungsempfinden wünschenswert entwickelt (Hypothese 3a+b). Dabei wird in Hypothese 28 vermutet, dass die Klassengröße ein Prädiktor für die Belastung darstellt. Dies lässt sich lediglich an der Evaluationsstichprobe untersuchen, zeigt jedoch einen marginal signifikanten und insbesondere negativen Zusammenhang im Prätest ($r = -.35$, $p = .07$). Entsprechend muss diese Hypothese abgelehnt werden.

Abschließend wurde das Alter (operationalisiert durch die Unterrichtserfahrung) als einflussreich auf die Belastung und den Nutzen der Lernstandserhebungen untersucht. Diese Hypothesen wurden ebenfalls an der VELM-Stichprobe untersucht. Für den *förderdiagnostischen* und den *unterrichtsplanerischen Nutzen* ist jedoch kein signifikanter Zusammenhang mit der Berufserfahrung festzustellen. Lediglich bei der wahrgenommenen Belastung zeigt sich für die Lehrkräfte der Stichprobe 2 ein marginal signifikanter Effekt ($r = -.23$, $p = .08$). Insofern müssen die Hypothesen 29b und 29c zunächst abgelehnt werden und Hypothese 29a kann auch nur unter Vorbehalt angenommen werden.

Tabelle 103: Hypothesenübersicht H1 - H6¹⁰⁸

H	Formulierung	mANOVA	ANCOVA	rmANOVA	rANCOVA
1a	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, eine höhere Akzeptanz gegenüber den Lernstandserhebungen zu haben als Lehrkräfte der Kontrollgruppe.	G: $p = .12$ + Z: $p = .05$ + I:			
1b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einstellung eine höhere Akzeptanz der Lernstandserhebungen.		G: $p < .01$ + Z: $p = .04$ - I:		
2	Im Verlauf der Projektteilnahme schätzen Lehrkräfte der Projektgruppe die wahrgenommene Lehrplanvalidität der Lernstandserhebungen positiver ein als Lehrkräfte der Kontrollgruppe.			G: $p = .03$ + Z: $p = .03$ + I:	
3a	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe eine geringere wahrgenommene Belastung durch die Lernstandserhebungen an als Lehrkräfte der Kontrollgruppe.	G: Z: I: $p = .08$ -			
3b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Belastung eine geringere Belastung durch die Lernstandserhebungen.		G: $p = .08$ - Z: I:		
4a	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des förderdiagnostischen Nutzens der Lernstandserhebungen für die Lehrkräfte der Projektgruppe positiver als für die Kontrollgruppe.	G: $p = .08$ + Z: $p = .11$ + I:			
4b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einschätzung eine höhere Einschätzung des förderdiagnostischen Nutzens.		G: $p < .07$ + Z: $p = .01$ + I:		
5a	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des unterrichtsplanerischen Nutzens der Lernstandserhebungen für die Lehrkräfte der Projektgruppe positiver als für die Kontrollgruppe.	G: Z: I: $p = .11$ +			
5b	Nach Abschluss des Projektes zeigt die Projektgruppe gegenüber der Kontrollgruppe nach Kontrolle der Prätest-Einschätzung eine höhere Einschätzung des unterrichtsplanerischen Nutzens.		G: $p = .10$ + Z: $p = .06$ + I:		
6	Im Verlauf der Projektteilnahme entwickelt sich die Einschätzung des selektionsdiagnostischen Nutzens der Lernstandserhebungen für die Lehrkräfte der Projektgruppe negativer als für die Kontrollgruppe.	G: Z: I:			

¹⁰⁸ Analysen für schwarze Zellen wurden nicht durchgeführt. G: Gruppe; Z: Zeit; I: Interaktion

Tabelle 104: Hypothesenübersicht H7 – H13¹⁰⁹

H	Formulierung	mANOVA	ANCOVA	rmANOVA	rANCOVA
7	Im Verlauf der Projektteilnahme entwickelt sich die selbst wahrgenommene Häufigkeit des <i>Insistierens auf Begründungen und Erklärungen</i> für Projektgruppenlehrkräfte positiver als für Kontrollgruppenlehrkräfte.	G: Z: I: $p = .05$ +			
8	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger Aufgaben mit <i>Variation in der Aufgabenschwierigkeit</i> einzusetzen als Lehrkräfte der Kontrollgruppe.	G: Z: I:			
9	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>kognitiv anspruchsvolle Aufgaben</i> zu nutzen als Lehrkräfte der Kontrollgruppe.	G: Z: I:			
10	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>kognitiv herausforderndes Üben</i> einzusetzen als Lehrkräfte der Kontrollgruppe.	G: $p = .12$ + Z: I:			
11a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, Ihre Schülerinnen und Schüler seltener <i>repetitiv Üben</i> zu lassen als Lehrkräfte der Kontrollgruppe.	G: Z: I: $p < .01$ -			
11b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, weniger <i>repetitives Üben</i> zu nutzen.		G: $p < .01$ - Z: $p < .01$ + I:		
12a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, häufiger <i>Schülerbeiträge (Ideen/Fehler)</i> als Ausgangspunkt für den weiteren Unterricht zu nutzen als Lehrkräfte der Kontrollgruppe.		G: Z: I:		
12b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, häufiger <i>Schülerbeiträge (Ideen/Fehler)</i> als Ausgangspunkt für den weiteren Unterricht zu nutzen.			G: Z: I:	
13a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, seltener <i>fehlervermeidende Erarbeitungsphasen</i> durch sich selbst zu gestalten als Lehrkräfte der Kontrollgruppe.		G: $p < .01$ - Z: $p = .11$ - I:		
13b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, seltener <i>fehlervermeidende Erarbeitungsphasen</i> durch die Lehrkraft selbst zu gestalten.			G: $p = .04$ - Z: $p = .03$ + I:	

¹⁰⁹ Analysen für schwarze Zellen wurden nicht durchgeführt. G: Gruppe; Z: Zeit; I: Interaktion

Tabelle 105: Hypothesenübersicht H14 - H22¹⁰⁷

H	Formulierung	mANOVA	ANCOVA	rmANOVA	rANCOVA
14a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, weniger bzw. seltener kleinschrittige, <i>fürsorgliche Anleitung und Kontrolle</i> zu geben als Lehrkräfte der Kontrollgruppe.			G: Z: I:	
14b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, seltener kleinschrittige, <i>fürsorgliche Anleitung und Kontrolle</i> zu geben.				G: Z: $p < .01$ + I:
15a	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, Schülerinnen und Schüler stärker in der <i>kognitiven Selbstständigkeit zu unterstützen</i> als Lehrkräfte der Kontrollgruppe.			G: $p = .02$ + Z: I:	
15b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, Schülerinnen und Schüler stärker in deren <i>kognitiven Selbstständigkeit zu unterstützen</i> .				G: Z: $p = .08$ + I:
17	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, größeren Wert auf eine <i>gemeinsame Wissenskonstruktion</i> zu legen als Lehrkräfte der Kontrollgruppe.			G: Z: I:	
18	Im Verlauf der Projektteilnahme geben die Lehrkräfte der Projektgruppe an, <i>Lernpartnerschaften</i> mehr zu unterstützen als Lehrkräfte der Kontrollgruppe.	G: Z: $p = .01$ + I:			
19a	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, größeren Wert auf <i>persönliches Vertrauen</i> zu legen als Lehrkräfte der Kontrollgruppe.			G: Z: $p = .06$ - I:	
19b	Nach Abschluss des Projektes gibt die Projektgruppe im Vergleich zur Kontrollgruppe nach Kontrolle der Prätest-Einschätzung an, größeren Wert auf <i>persönliches Vertrauen</i> zu legen.				G: Z: I:
20	Im Verlauf der Projektteilnahme geben Lehrkräfte der Projektgruppe an, häufiger <i>Diagnosen im sozialen Bereich</i> durchzuführen als Lehrkräfte der Kontrollgruppe.			G: Z: I: $p = .07$ +	
22a	Die Schülerinnen und Schüler der Projektgruppenlehrkräfte zeigen einen größeren <i>Leistungszuwachs</i> im Verlauf des Projektes gegenüber den Schülerinnen und Schülern der Kontrollgruppenlehrkräfte.	G: $p = .04$ + Z: $p < .01$ + I:			
22b	Nach Abschluss des Projektes zeigen die Schülerinnen und Schüler der Projektgruppenlehrkräfte gegenüber denen der Kontrollgruppenlehrkräfte nach Kontrolle der Prätestleistung eine höhere Leistung.		G: Z: $p < .01$ + I:		

5 Diskussion

5.1 Reflexion der Ergebnisse

Am Ende des vorangegangenen Ergebnisteils findet sich eine kurze Zusammenfassung aller Ergebnisse sortiert nach den Hypothesen. Auf eine derartige Darstellung verzichte ich daher im Diskussionsteil. Vielmehr sollen die Ergebnisse nun kritisch in den Blick genommen werden; einerseits vor dem Hintergrund des im Teil 1 geschilderten Forschungsstands, andererseits vor dem Hintergrund des Forschungsdesigns (vgl. hierzu insbesondere auch die Beschreibung der Fortbildungsinhalte von VELM-8, Teil 2). Anschließend betrachte ich, welche Schlüsse aus den Ergebnissen gezogen werden können und welche Anschlussforschungen und -fragen aus den Ergebnissen resultieren sowie welche Fragen in dieser Arbeit nicht (zufriedenstellend) beantwortet werden konnten.

5.1.1 VELM-8 Projektevaluation

Zunächst möchte ich auf den Kern der Arbeit, das Projekt VELM-8, eingehen. Wie bereits dargestellt, war das Projekt VELM-8 weniger ein reines Forschungsprojekt als vielmehr zur Fortbildung von Lehrkräften mit einer ergänzenden Evaluation gedacht. Aus rein evaluationsbezogener Sicht muss zusammenfassend mit statistisch hartem Standpunkt (des üblichen 5%-Signifikanzniveaus) der Großteil der Hypothesen abgelehnt werden. Werden die statistischen Kriterien etwas lockerer gefasst (10%iges Signifikanzniveau), zeigen sich einige erwartete Effekte, die sich im Sinne der Projektevaluation m. E. als Fortbildungseffekte interpretieren lassen könnten.

Lehrkräfte berichten nach Projektteilnahme tendenziell also

- eine geringere *Belastung* durch die LSE (H3a+b),
- einen höheren *förderdiagnostischen Nutzen der LSE* (H4b),
- einen höheren *unterrichtsplanerischen Nutzen der LSE* (H5b; H5a verfehlt das 10%-Signifikanzniveau minimal),
- häufigeres *Einfordern von Begründungen und Erklärungen* (H7),
- seltener *repetitives Üben* zu nutzen (H11b),
- seltener *fehlervermeidende Erarbeitungsphasen* zu nutzen (H13b),
- eine vermehrte Unterstützung von *Lernpartnerschaften* (H18) und
- häufiger durchgeführte *Diagnosen im sozialen Bereich* (H20).

Unter anderem aus dem höher eingeschätzten Nutzen der LSE in Bezug auf *Förderdiagnostik* und *Unterrichtsplanung* könnte eine geringere *Belastung* resultieren. Dieser Zusammenhang müsste in entsprechender Anschlussforschung genauer untersucht werden, da eine Überprüfung mit den vorliegenden Daten wegen des $N = 16$ für die Projektgruppe nicht sinnvoll erscheint. Der direkte Zusammenhang eines höheren wahrgenommenen Nutzens und geringerer Belastung könnte jedoch ein Erklärungsansatz sein, da die Belastung der Lehrkräfte durch die LSE kein expliziter Fortbildungsbestandteil war. In der Terminologie von Visscher und Coe (2003) wie auch Schildkamp und Kuiper (2010) fallen beide Nutzungsaspekte als *intended use* der LSE bzw. anvisierten Konsequenzen aus den LSE positiv auf (vgl. Abschnitt 1.6.1.3). Auch eine belastungsreduzierende Entwicklung (durch die Fortbildungen) ist vor dem Hintergrund der Belastungsforschung gut nachvollziehbar (Heymann, 2005; Schaarschmidt, 2005, 2007), da ohne Intervention eine schuljahresbegleitend steigende Belastung zu erwarten wäre, die sich u. a. durch Depersonalisierung (fehlende Fähigkeit zur Perspektivübernahme und verringerte Empathiefähigkeit) als eines von drei Merkmalen/Symptomen von Burnout nach Maslach, Schaufeli und Leiter (2001) erkennen lässt (Klusmann, Kunter, Trautwein & Baumert, 2006; Klusmann, 2011; Kunter et al., 2011; Schaarschmidt, 2005, 2007). Als Indiz für geringe Depersonalisierung und damit einer belastungsmindernden Wirkung der Fortbildung lässt der positive Interaktionseffekt in der rangtransformierten messwiederholten Varianzanalyse für die Nutzung *diagnostischer Situationen mit sozialem Fokus* (H20) interpretieren. Ebenfalls auffällig zeigt sich das *persönliche Vertrauen* (H19a): Über die Erwartung hinaus (vgl. Schaarschmidt, 2005, 2007) nimmt das *persönliche Vertrauen* allerdings nicht nur im zeitlichen Verlauf signifikant ab, sondern unterscheidet sich zum Posttest auch signifikant von der COACTIV-Stichprobe und zwar mit einer negativen Abweichung (vgl. Tabelle 20). Die unterschiedliche Entwicklung der verschiedenen Belastungsskalen spricht dafür, dass die Lehrkräfte die Angaben zur *Belastung durch die LSE* von der allgemeinen Belastung im Unterrichtsalltag mit der Folge der Depersonalisierung – hier operationalisiert als *persönliches Vertrauen* – trennen.

Kontraintuitiv fällt hingegen der Zusammenhang zwischen der *Belastung durch die LSE* und der Klassengröße aus (Hypothese 28): Es zeigt sich ein negativer signifikanter Korrelationskoeffizient. Damit widersprechen diese Ergebnisse der bestehenden Forschung (Maier, 2008b; Maier et al., 2012) und könnten ein Indiz dafür sein, dass Lehrkräfte insbesondere bei großen Lerngruppen nach Unterstützung in der

Lernstandsdiagnose und Unterrichtsentwicklung suchen. Eine profanere Erklärung wäre darin zu sehen, dass die aus den LSE erwachsende Hauptbelastung für die Lehrkräfte in der Korrektur der Testhefte liegt: Da die Einarbeitung in die sonst bei Lehrkräften eher unübliche richtig/falsch-Kodierung vergleichsweise viel Zeit in Anspruch nimmt, wirkt damit die wahrgenommene Belastung bei einem größeren Klassensatz geringer, weil sich die Einarbeitung eher „gelohnt“ hat. Die *Belastung* zeigt des Weiteren eine signifikante negative Korrelation mit der Unterrichtserfahrung, also bei höherer Unterrichtserfahrung eine geringere *Belastung*. Auch dieser Zusammenhang ist eher unerwartet und entgegen den bestehenden Befunden (Ditton & Arnoldt, 2004; Korngiebel, 2009). Neben den Daten der Projektevaluation legt auch die Clusteranalyse einen Zusammenhang dieser beiden Merkmale nahe, da sich mit Cluster 2 ein relativ unterrichtserfahrenes Cluster bildet, welches die LSE umfassend in anvisierter Form nutzt. Unter Umständen hängt diese Korrelation jedoch auch mit dem Bedarf an Fortbildungen und „akuten Problemen“ von Lehrkräften zusammen, da Kast (2010) beispielsweise berichtet, dass insbesondere jüngere Lehrkräfte Probleme v. a. in alltäglichen Bereichen wie etwa Klassenführung etc. hätten. Durch diese Schwierigkeiten in der Ausübung des Unterrichts stellen die LSE eine zusätzliche *Belastung* dar, die mit steigender Unterrichtserfahrung und Routine abnimmt.

Die beiden letztgenannten Effekte (H18: *Lernpartnerschaften* und H20: *Diagnose im sozialen Bereich*) lassen sich eindeutig einer effektiven und schülerorientierten Unterrichtsführung zuordnen (vgl. Qualitätskriterien für Unterricht; Abschnitt 1.5.4.2). Dabei erstaunen diese Effekte durchaus, da die Bedeutung der sozialen Interaktion etwa durch Lernpartnerschaften in den Fortbildungen nicht im Fokus stand. Zum Erstaunen trägt bei, dass für die *Unterstützung von Lernpartnerschaften* zu beiden Messzeitpunkten signifikant höhere Werte vorliegen als in der Referenzstichprobe aus COACTIV (vgl. Tabelle 20) und somit schon ein vergleichsweise hohes Niveau zum Prätest vorliegt. Als mögliche Erklärung kann hier angeführt werden, dass im Rahmen der Fortbildungen wesentlich auf eine respektvolle, freundliche und an den Lehrkräften sowie ihren Vorkenntnissen orientierte Atmosphäre geachtet wurde. Damit wurde eine schülerorientierte Unterrichtsführung nach deren expliziter Thematisierung als Aspekt der Unterrichtsqualität (vgl. Fortbildung 2) implizit illustriert. Diese positive Entwicklung in der Einschätzung dieser Basisdimension lässt sich jedoch nicht für alle Indikatoren statistisch absichern. Bei den Skalen *fürsorgliche Anleitung und Kontrolle* (H14a) und *gemeinsame Wissenskonstruktion* (H17) zeigen sich keine

signifikanten Interaktionsterme in der rmANOVA, allerdings lässt die grafische Darstellung erkennen, dass sich die Gruppen entgegengesetzt entwickeln: Die Projektgruppe gibt an, im zeitlichen Verlauf des Projekts häufiger fürsorglich anzuleiten und gemeinsame, soziale Wissenskonstruktion zu ermöglichen.

Das essentielle, jedoch nur sehr bedingt fachspezifische Unterrichtskriterium „effektive und schülerorientierte Unterrichtsführung“ wurde in der Fortbildungsreihe (konkret Fortbildung 2) zwar thematisiert, doch bei späteren Analysen von erarbeiteten Unterrichtssequenzen weniger in den Fokus gestellt als etwa die fachspezifische „kognitive Aktivierung“ oder die Nutzung der LSE-spezifischen Materialien/Informationen. Daher lassen sich die fehlenden Effekte der Fortbildungsteilnahme vermutlich auf diese eher implizite Thematisierung zurückführen.

Wie angedeutet, lag der Fokus der Fortbildungen auf der Unterrichts-Basisdimension „kognitive Aktivierung“: Ein erhöhter Anspruch an *Begründungen und Erklärungen* einzufordern lässt sich einer erhöhten kognitiven Aktivierung der Lernenden zuordnen (vgl. auch Bloomsche Taxonomie). Analog hierzu passen auch die negativen Indikatoren fehlender kognitiver Aktivierung in H11b, *repetitives Üben*, und H13b, *fehlervermeidende Erarbeitung durch die Lehrkraft selbst*. Die Projektgruppe gibt an, im Laufe der Fortbildungszeit diese Aspekte häufiger (H7) respektive seltener (H11b und H13b) eingesetzt zu haben und in diesem Sinne einen qualitativ höherwertigen Unterricht angeboten zu haben. Weitere Indikatoren kognitiver Aktivierung zeigen jedoch in den Auswertungen keine signifikanten Ergebnisse mit Hypothesenbezug auf (wenngleich z. T. andere Effekte vorhanden sind):

H8	Variation der Aufgabenschwierigkeit	keine Effekte
H9	Einsatz kognitiv anspruchsvoller Aufgaben	keine Effekte
H10	Kognitiv anspruchsvolles Üben	keine Effekte
H12	Umgang mit Schülerbeiträgen (Ideen/Fehler)	keine Effekte
H15	Unterstützung zur kognitiven Selbstständigkeit	
H15a		positiver Haupteffekt der Zeit
H15b		(positiver Effekt der Prätest-Einstellung)
H17	Gemeinsame Wissenskonstruktion	Keine Effekte

Dementsprechend ergeben sich für die kognitive Aktivierung, einem der Kernbestandteile der Fortbildungen, lediglich auf drei von neun Skalen Effekte mit Hypothesen- bzw. Fortbildungsbezug. Die Relevanz der kognitiven Aktivierung der Schülerinnen und Schüler wurde als Referenz bei der Diskussion der in den Arbeitsphasen entwickelten Stundenvorschlägen der Lehrkräfte herangezogen. Erfreulich ist vor dem Hintergrund der Schwerpunktsetzung auf die prozessbezogenen Kompetenzen K1

„Mathematisch argumentieren“ und K2 „Probleme mathematisch lösen“ der Effekt bzgl. des Einsatzes von *Begründungen und Erklärungen* zu werten. Die in den Fortbildungen genutzten Aufgaben und die in den Arbeitsphasen entwickelten Stunden der Lehrkräfte spiegelten diese Schwerpunktsetzung wider. Folglich liegt es nahe, diese Schwerpunktsetzung innerhalb der Fortbildungen als Grund für das verstärkte Einfordern seitens der Lehrkräfte im Unterricht zu vermuten. Als mögliche Erklärung für die mangelnden Effekte auf anderen Skalen könnten aufgrund des ohnehin extremen Ausgangsniveaus der Einschätzungen Deckeneffekte in zweierlei Richtung herangezogen werden:

- Die angegebenen Skalen ohne hypothesenbezogene Effekte weisen mit Ausnahme der Skala zum *Umgang mit Schülerbeiträgen* signifikante positive Abweichungen von der Referenzstichprobe aus COACTIV auf (zumindest im Vergleich Prätest VELM-8 zum ersten Messzeitpunkt in COACTIV; vgl. Tabelle 20). Aus dieser Vermutung lassen sich zwei Folgerungen ableiten: Entweder handelt es sich bei der Stichprobe aus VELM-8 um eine Positiv-Selektion gegenüber der deutlich größeren Stichprobe aus COACTIV oder die aus COACTIV übernommenen Instrumente sind zur Erfassung in diesem Fall nur bedingt geeignet. Für Anschlussforschung sollte daher auf eine genaue Passung der Instrumente geachtet werden bzw. vorab eine Pilotierung der Instrumente durchgeführt werden.
- Darüber hinaus zeigen sich beispielsweise für die Skalen *fehlervermeidende Erarbeitung* (H13a) und *Unterstützung kognitiver Selbstständigkeit* (H15a) signifikante Haupteffekte der Gruppen, sodass die Projektgruppe deutlich geringere (H13a) bzw. höhere (H15a) Angaben macht als die Kontrollgruppe. Aus dieser Vermutung ergibt sich die Frage nach der Vergleichbarkeit von Projekt- und Kontrollgruppe für VELM-8. Die Gruppeneffekte wie auch die in Teil 3 durchgeführten *t*-Tests zu Prätest-Unterschieden deuten an, dass sich die beiden Gruppen generell in ihren Einschätzungen unterscheiden. Diesem Einwand könnte in einer Replikation des Projekts etwa durch Randomisierung der Gruppen begegnet werden.

Ein weiterer wesentlicher Fortbildungsinhalt war es, Informationen zu den LSE und deren Nutzung vorzustellen. Dabei ging es darum, die Entstehung, den Ablauf und die vorhandenen Rückmeldungen/Handreichungen sowie die Intentionen, die mit der

Einführung der LSE seitens Fachdidaktik und Politik verbunden sind, den Lehrkräften zu präsentieren. Zu den Intentionen der LSE gehören v. a. Unterrichtsentwicklung und Förderung, worauf letztlich zwei der drei auffälligen Einstellungs-Skalen/ -Hypothesen abzielen: *förderdiagnostischer Nutzen* (H4b) und *unterrichtsplanerischer Nutzen* (H5a+b). In Projekt- und Kontrollgruppe verändert sich die Einschätzung in Abhängigkeit von der Gruppenzugehörigkeit im Laufe des Projektjahres unterschiedlich; dabei insbesondere günstiger für die Projektgruppe. Ob sich diese Veränderungen in den Einstellungen der Lehrkräfte tatsächlich auch im Unterricht widerspiegeln (vgl. Ebene 3 des Fortbildungserfolges in Abbildung 11), lässt sich mit den vorliegenden Daten leider nicht klären. Es kann allerdings sicher gesagt werden, dass die Veränderungen der Einstellungen der Lehrkräfte keine Veränderungen in der Schülerleistung nach sich ziehen. Inhaltlich spielte in den Fortbildungen eine Nutzung der LSE(-Rückmeldungen) zu *selektionsdiagnostischen* Zwecken keine Rolle. Wie bereits in der Theorie angesprochen, ist eine Selektion durch die LSE nicht intendiert; deutlicher noch: eine Benotung der LSE ist in Hessen nach Amtsblatt untersagt (Hessisches Kultusministerium, 2011). Die zugehörige Hypothese 6 muss zwar abgelehnt werden, da sich die Projektgruppe gegenüber der Kontrollgruppe nicht anders entwickelt hat, doch zeigt sich generell eine eher ablehnende Haltung in beiden Gruppen zu beiden Messzeitpunkten. Diese Ergebnisse sind damit erwartungskonform aus Sicht der hessischen Situation sowie aus Perspektive anderer Untersuchungen (etwa Kuper & Diemer, 2012).

Neben den drei eher nutzungsorientierten Skalen zu den LSE wurden die *Akzeptanz*, die *Lehrplanvalidität* und die *Belastung* (vgl. oben) untersucht. Während H1 (eine höhere Akzeptanz in der Projektgruppe durch die Fortbildungen) auf Basis der mANOVA abgelehnt werden muss, zeigen Abbildung 28 die erwartete Entwicklung für die mANOVA und die Ergebnisse der ANCOVA einen signifikanten Interaktionsterm (H1b), der aussagt, dass die Einstellung zum Posttest für die Projektgruppe signifikant weniger durch die Vortest-Einschätzung bestimmt werden als für die Kontrollgruppe. Dieser Effekt deutet an, dass die zeitliche Veränderung in der Projektgruppe anders verläuft als in der Kontrollgruppe. Diese Unterschiede können vermutlich durch die höhere Informiertheit erklärt werden. Dieses Erklärungsmuster wurde auch in anderen Studien angeführt (Maag Merki, 2010; Maier, 2010; Rolf, 2001). Deutlich wird hier auch ein Effekt der Zeit (marginal signifikant). Ebenfalls anschlussfähig an andere Untersuchungen zeigt sich in der Projektevaluation wie auch in der Clusteranalyse ein

indifferenten Mittelwert nahe dem semantischen Median (vgl. Abschnitt 1.6.4.2; u. a. Groß-Ophoff & Koch, 2007b). In Verbindung mit der vergleichsweise geringen Vorerfahrung mit den LSE ($M = 0.56$, $SD = 0.79$; vgl. Kapitel 3.3 zur Stichprobe) liegt die Vermutung nahe, dass durch die erstmalige Durchführung eine eher skeptische Haltung gegenüber den LSE vorliegt, die sich durch die tatsächliche Teilnahme an den LSE auflöst – frei nach dem Sprichwort: „Was der Bauer nicht kennt, frisst er nicht“.

Gleiches gilt für die Einschätzung der *Lehrplanvalidität*: Zwar lässt sich H2 nicht bestätigen, jedoch ist hinter dem positiven Haupteffekt der Zeit in der rmANOVA ebenfalls ein Einfluss der erstmaligen tatsächlichen Durchführung der LSE zu vermuten. Die rangtransformierte messwiederholte Varianzanalyse ergibt einen signifikanten positiven Haupteffekt der Zeit. Generell fallen die Werte der *Lehrplanvalidität* sehr hoch aus (vgl. Kapitel 3.4). Mit einem Mittelwert von $M_{\text{Projektgruppe}} = 4.12$ im Posttest (vgl. Tabelle 42) liegt ein Deckeneffekt nahe. Nach Roberts (1978, S. 7) ist eine bedeutsame Schiefe (*skewness*) als Indiz für einen Decken- bzw. Bodeneffekt anzunehmen, wenn Median und Mittelwert der Stichprobe zu einem Fünftel der Standardabweichung verschoben sind. Im Falle der *Lehrplanvalidität der LSE* liegt dies bereits zum Prätest vor, sodass die Ergebnisse im Posttest verzerrt sein könnten: Bei einer Standardabweichung im Prätest von $SD_{\text{Prätest}} = 0.49$ liegen Mittelwert und Median mehr als ein Fünftel hiervon auseinander, wobei der Median als Zeichen einer negativen Schiefe größer ist als der Mittelwert ($M_{\text{Prätest}} = 3.79$, $\text{Median}_{\text{Prätest}} = 4.00$, $\Delta = 0.21$; vgl. auch Vogel et al., 2016). Diese hohe *Validität* ist hingegen Kennzeichen dafür, dass die LSE als Verfahren *close-to-the-job* sind. Diese Einschätzung deckt sich auch mit den Befunden aus der Online-Stichprobe bzgl. der Ungewohntheit der Aufgabe (H23).

Ebenfalls keinen direkten Effekt mit Verbindung zur Fortbildungsteilnahme lässt sich für die Leistungsentwicklung der Schülerinnen und Schüler finden. In der Analyse zeigen sich ein enormer Effekt der Zeit sowie ein Gruppenunterschied zugunsten der Projektgruppe. Die Entwicklung wird dabei nicht von der Gruppenzugehörigkeit beeinflusst ($\Delta M_{\text{Kontrollgruppe}} = 0.46$, $\Delta M_{\text{Projektgruppe}} = 0.48$). Da jedoch die Fortbildungsdauer unter dem Yoon et al. (2007) angeführten Richtwert von 14 Stunden liegt, wurde bereits bei der Formulierung der Hypothesen ein Effekt als unwahrscheinlich eingeschätzt. Bemerkenswert ist jedoch die absolute Entwicklung zwischen Prä- und Posttest. Es zeigt sich – unabhängig von der Gruppenzugehörigkeit – eine Veränderung

von $M_{\text{Prätest}} = -0.33$ (0.89) zu $M_{\text{Posttest}} = 0.11$ (1.16)¹¹⁰. Dementsprechend liegt eine Veränderung von Prä- zu Posttest um etwa eine halbe Standardabweichung (des Prätests) vor. Dieser absolute Leistungszuwachs innerhalb eines knappen Schuljahres ist im Vergleich zu anderen längsschnittlichen Leistungserhebungen etwa in PALMA (Projekt zur Analyse der Leistungsentwicklung in Mathematik) insofern als hoch einzuschätzen, da über die Hälfte der Schülerinnen und Schüler der vorliegenden Stichprobe aus dem gymnasialen Bildungszweig stammen (respektive zur Leistungsmessung das entsprechende Testheft der LSE erhalten haben) und Vom Hofe et al. (2006, S. 93) für diese Schulform einen Leistungszuwachs pro Schuljahr von 0.29 Standardabweichungen berichten. Dieser durchschnittliche Zuwachs liegt in der Realschule mit 0.49 etwas höher. Der Zuwachs der VELM-8-Schülerinnen und Schüler entspricht mit 0.49 Standardabweichungen genau diesem Wert. Für PISA werden geringere Werte für die Leistungsentwicklung innerhalb eines Schuljahres angegeben. Prenzel, Baumert, Blum, Lehmann, Leutner, Neubrand, Pekrun, Rolff, Rost und Schiefele (2004b) beschreiben den Abstand Deutschlands zu den PISA-führenden Staaten mit einem Abstand von maximal 41 Punkten auf der zugrundeliegenden Skala als Entwicklungsrückstand von bis zu einem ganzen Schuljahr. 41 Punkte auf der üblichen PISA Skala mit Mittelwert $M = 500$ und Standardabweichung $SD = 100$ bedeuten also nur knapp die Hälfte einer Standardabweichung. Vor dem Hintergrund, dass weder Prätest direkt zu Beginn des Schuljahres noch Posttest unmittelbar vor Ende des Schuljahres lagen und somit überspitzt formuliert die Entwicklung eines Dreiviertel-Schuljahres betrachtet wird, liegt die Posttest-Leistung in Bezug auf die Prätestleistung über dem erwarteten Bereich. Diese Ergebnisse lassen sich insofern validieren, als dass beim Leistungstest davon auszugehen ist, dass die zugrundeliegende Stichprobe aus Schülerinnen und Schülern in VELM-8 sich nicht deutlich von der Gesamtpopulation bzw. einer repräsentativen Stichprobe unterscheidet: Werden bei der Skalierung des Leistungstests die Itemparameter aus der repräsentativen Pilotierungsstichprobe der Lernstandserhebungen vorgegeben, zeigen sich kaum Abweichungen in den resultierenden Personenfähigkeiten von der Skalierung mit frei zu schätzenden Itemparametern (vgl. Tabelle 106): Entsprechend bedeutet dies, dass die zugrundeliegende Schülerstichprobe von VELM-8 die Leistungstests in ähnlicher Weise bearbeitet hat wie die repräsentative Pilotierungsstichprobe der

¹¹⁰ Standardabweichung in Klammern

Lernstandserhebungen. Diese Gesamtleistungsentwicklung unterstützt die Vermutung, dass es sich bei den am Projekt teilnehmenden Lehrkräften (sowohl in Projekt- als auch Kontrollgruppe) um Lehrkräfte handelt, die lernwirksamen Unterricht anbieten und in diesem Sinne eine Positiv-Selektion darstellen. Mit Bezug auf die in Kapitel 1.5.2 genannten Hypothesen zum Fortbildungsbesuch bestärken diese Befunde die *Neigungshypothese*.

Tabelle 106: Vergleich der Itemparameter aus der freien Skalierung und Skalierung mit vorgegebenen Itemparametern aus der Pilotierung VerA2010/2011

	freie Skalierung	Itemparameter aus Pilotierung
Minimum	-5.57	-5.37
1. Quartil	-0.85	-1.51
Median	-0.03	-0.66
Mittelwert	0.01	-0.62
3. Quartil	0.92	0.31
Maximum	6.15	5.21
Standardabweichung	1.36	1.32

Reflektiert man die insbesondere nach Betrachtung der Abbildungen augenscheinlich vielfach positiven Entwicklungsrichtungen, welche sich jedoch nur selten statistisch abgesichert darstellen, lässt sich als Resümee ziehen, dass VELM-8 als Fortbildungsprojekt Lehrkräfte der Region unterstützt und dabei zu z. T. merklichen Veränderungen geführt hat. Diese Ergebnisse sind allerdings nur sehr bedingt zu verallgemeinern oder auf andere Regionen/Gruppen von Lehrkräften zu übertragen, da sich die vorliegende Stichprobe 1 in diversen Fällen als speziell herausgestellt hat, insbesondere durch die deutlichen und statistisch abgesicherten Abweichungen der Prätest-Einstellungen von den Vergleichsstichproben aus Maier (2008a) und Baumert et al. (2008; Kunter et al., 2011; vgl. Tabelle 18 und Tabelle 20; Vogel, Achmetli et al., 2013; Vogel et al., 2016). Dabei lässt sich nur spekulieren, dass im Großraum Nordhessen insbesondere durch das Engagement von Werner Blum bereits viele u. a. bildungsstandardorientierte Projekte durchgeführt wurden (etwa SINUS, DISUM, Co²CA) und die Fortbildungsteilnahme wie vielfach bestätigt der *Neigungshypothese* folgt (vgl. 1.5.2.1; Desimone, 2006; Richter, 2011).

5.1.2 Clusteranalyse/Typenbildung

In der Untersuchung konnten fünf Typen von Lehrkräften gefunden werden, die sich z. T. signifikant voneinander unterscheiden. Zweifelsfrei eignet sich diese Analyse nicht als belastbare Klassifizierung von Lehrkräften, die sich breit verallgemeinern lässt. Allerdings geben die Ergebnisse und gefundenen Typen Hinweise auf mögliche Muster bzw. werden durch die Konfigurationsanalyse bereits in dieser Arbeit z. T.

bestärkt. Sowohl die Suche nach Typen von Lehrkräften von Groß-Ophoff et al. (2007) als auch diejenige von A. Hosenfeld (2010) weisen je drei verschiedene Typen auf. Bei Betrachtung des Dendrogramms (Abbildung 49) ist erkennbar, dass eine 3-Clusterlösung für die vorliegende Untersuchung ggf. ebenfalls möglich gewesen wäre. Interessant hierbei ist, dass das ablehnende und durch KFA in Teilen bestätigte Cluster 5 auch in der 3-Clusterlösung bereits auftritt; mehr noch zeigt das Dendrogramm, dass dieses Cluster sich bereits im ersten Clusterschritt ausbildet. Diese Tatsache ist Zeugnis, dass dieses Cluster sich deutlich von den anderen unterscheidet. Einen ablehnenden Typ findet auch A. Hosenfeld (2010) in ihrer Analyse (Typ 3): Dieser kennzeichnet sich dadurch, dass die im modellhaften Durchlauf der Lernstandserhebungen postulierten Schritte deutlich seltener durchgeführt werden als von den beiden anderen Typen. Auch bei Groß-Ophoff et al. (2007) stellt sich ein skeptischer Typ heraus: Zwar wird von Lehrkräften dieses Typs die Intensität der Auseinandersetzung vergleichsweise hoch angegeben, jedoch weicht die eingeschätzte Nützlichkeit negativ von den anderen beiden Typen ab. Ditton und Arnoldt (2004) berichten in ihrer Clusteranalyse von einem eher zustimmenden und einem skeptischen, pessimistischen Cluster. Für diese Übereinstimmung ergeben sich zwei Erklärungsmuster: Die Abgrenzung eines Typen/Clusters im Sinne von Niveau-Unterschieden könnte ein Artefakt der eingesetzten Methode *Clusteranalyse* sein oder – etwas optimistischer – könnte die Erklärung darin liegen, dass es unabhängig von der Stichprobe, Untersuchungsort und -zeitpunkt zu erwarten ist, Lehrkräfte zu finden, die den LSE, deren Nützlichkeit bzw. Nutzung eher skeptisch gegenüberstehen. Während sich für die vorliegende Untersuchung keine Clusterunterschiede bzgl. des Geschlechts finden, deuten Ditton und Arnoldt (2004) an, dass diese ablehnende, skeptische Haltung eher bei Männern und im Fach Mathematik zu finden ist. Diesen Ergebnissen widersprechend werden jedoch in den Online-Daten gerade von den Mathematik-Lehrkräften für die LSE wohlwollendere Urteile getroffen, wobei hier aufgrund des zeitlichen Abstandes der beiden Befragungen/Analysen ein Vergleich nur sehr eingeschränkt möglich ist.

Einen pauschal zustimmenden Typen wie bei Ditton und Arnoldt (2004) weisen die Untersuchungen von Groß-Ophoff et al. (2007) und A. Hosenfeld (2010) ebenfalls auf. Diese Typen zeigen entweder eine intensive Auseinandersetzung bei hoher Verständlichkeit und hoher Nützlichkeit oder in allen Phasen des Nutzungsmodells (Rezeption, Reflexion, Aktion; Abbildung 10) hohe Werte, welche jeweils in weitreichenden Maßnahmen der Unterrichtsentwicklung münden. In der Clustereinteilung der

vorliegenden Arbeit gibt es mit Cluster 2 einen korrespondierenden Typ, der die LSE bei mittlerer (aber positiv eingeschätzter) *Akzeptanz* mit hoher *Lehrplanvalidität* nicht als *Belastung* erkennt und dabei noch in der intendierten Art (*Förderung* und *Planung*, jedoch ohne *Selektion*) nutzt und in Bezug auf die *Akzeptanz*, *Belastung* und *Lehrplanvalidität* bereits statistisch durch die Konfigurationsfrequenzanalyse abgesichert ist.

Hinsichtlich des verbleibenden Typs in den zitierten Arbeiten zeigt sich der Typ 2 in A. Hosenfeld (2010) eher als Skeptiker: Nach eigenen Angaben wird zwar rezipiert, jedoch nur eingeschränkt reflektiert und kaum Aktionen aus den Ergebnissen ableitet. Einen vergleichbaren Typ in der vorliegenden Arbeit zu finden, ist aufgrund der anderen Operationalisierung nur schwer möglich. Weniger gilt dies gegenüber Groß-Ophoff et al. (2007), deren zweiter Typ die Ergebnisse zwar weniger intensiv rezipiert und ihnen auch eine geringere Verständlichkeit attestiert, sie jedoch in großem Umfang nutzt, insbesondere für Wiederholungen. Diese eher unintendierte, „blinde“ Nutzung findet sich in den Clustern 1 und 4 der präsentierten Ergebnisse mit der Nutzung der LSE-Rückmeldungen u. a. zu *selektionsdiagnostischen* Zwecken.

Interessant in diesem Zusammenhang ist, dass sich auch aus dem Dendrogramm eine Nähe zwischen den Clustern 1 und 4 sowie zwischen Clustern 2 und 3 ergibt (vgl. Abbildung 49). Betrachtet man die 3-Clusterlösung, so verschmelzen die beiden genannten Clusterpaare zu je einem größeren. In der Betrachtung der Mittelwerte fällt auf, dass sich die beiden größeren Cluster vor allem im *selektionsdiagnostischen Nutzen* unterscheiden.

5.1.3 Ist-Stand in Hessen

Ditton und Arnoldt (2004) beschreiben ein eher skeptisches, ablehnendes Cluster gegenüber vergleichender Leistungsmessung, welches zu einem signifikant größeren Teil aus Mathematiklehrkräften bestand als das zustimmende Cluster. Diesem etwas älteren Befund widersprechend zeugen die aktuellen Ergebnisse der Online-Befragung im Vergleich der drei Fächer Deutsch, Englisch und Mathematik generell weniger kritische Einschätzung der Mathematiklehrkräfte und sind damit anschlussfähig an jüngere Untersuchungen (Bonsen et al., 2006; Kühle & Peek, 2007; beispielsweise eine signifikant eher zustimmende Haltung bei der Nützlichkeit). Dabei fällt insbesondere auf, dass anschlussfähig an bisherige Forschung sich eine vermutete Schwierigkeit zu bestätigen scheint: Mathematik-Lehrkräfte nutzen die tabellarische

Übersicht signifikant häufiger als Lehrkräfte der sprachlichen Fächer. Der Grund hierfür könnte etwa in etwaigen Interpretationsschwierigkeiten der Lehrkräfte sprachlicher Fächer liegen, während eine matrixähnliche Darstellung Mathematiklehrkräften vertraut ist und somit nicht die angesprochenen Rezeptionsschwierigkeiten entstehen (Kühle & Peek, 2007; Rolff, 2001). Die Auswertung einer Datentabelle und das Verständnis von Mittelwerten ist Voraussetzung für den bewussten Umgang mit der tabellarischen Übersicht. Hieraus lässt sich ableiten, dass Mathematiklehrkräfte tendenziell eher in der Lage sind, ihre Schlüsse aus Rohdaten selbst abzuleiten; insbesondere kann vermutet werden, dass unbearbeitete Rohdaten von ihnen eher in individueller Weise gelesen und angepasst werden können. Weitere Fachunterschiede zeigen sich in der Auswertungszeit (Korrektur- und Eingabezeit zusammen betrachtet). Neben der geringeren Dauer für die Korrektur und die Eingabe in Mathematik wird diese Zeit für Mathematik auch signifikant angemessener eingeschätzt. Diese Befunde sind sowohl erwartbar als auch anschlussfähig an andere Studien etwa aus Nordrhein-Westfalen, wenngleich die hier berichteten Auswertungszeiten deutlich geringer ausfallen: etwa 31 Minuten pro Testheft bei Bensen et al. (2006) gegenüber im Mittel 190 Minuten für den gesamten Klassensatz in der vorliegenden Untersuchung. Da jedoch die Erhebungsmethoden unterschiedlich sind und ein enormer zeitlicher Abstand der Befragungen vorliegt, ist ein direkter Vergleich der beiden Werte nur eingeschränkt möglich.

Mit den LSE soll unter anderem Schulentwicklung angestoßen werden. Wie dargelegt, sollen die Schulen zu *pädagogischen Handlungseinheiten* werden. Dabei ist Schulentwicklung jedoch als Trias aus Organisations-, Personal- und Unterrichtsentwicklung zu verstehen (Rolff, 2007) und die Durchführung der LSE steht lediglich am Anfang der Entwicklung. Aus einer positiven Wahrnehmung sollten sich entsprechend weitere Schritte ableiten, so wie Helmke (2004) modellhaft postuliert. Eine positive Wahrnehmung in Form von *Nützlichkeit* zeigt sich dabei interessanterweise abhängig von der Teilnahme am Verfahren: Lehrkräfte, die aktiv die LSE durchführen, schätzen die *Nützlichkeit* (sowohl als Skala als auch in den Einzeldimensionen *Lernstandsdiagnose*, *Unterrichtsentwicklung* und *Schulentwicklung*) positiver ein. Dieses Ergebnis lässt sich vor dem Hintergrund der Bedingungen effektiver Nutzung sowohl mit einem positiven Einfluss der Vorerfahrungen auf die Aufwand-Ertrags-Balance als auch durch einen höheren Grad an Informiertheit über die Rückmeldungen und Handreichungen erklären und unterstützt die Vermutung aus den VELM-8-Daten,

dass die erstmalige Teilnahme an den LSE die Einschätzung des Verfahrens positiv beeinflusst. Vor dem Hintergrund der berichteten akzeptanzsteigernden Wirkung von Partizipation ist dieses Ergebnis erwartungskonform (Rolf, 2007; Watermann & Stanat, 2008). Die Vorhersagekraft der eingeschätzten Nützlichkeit auf die Durchsicht (im Modell *Rezeption/Reflexion*) fällt jedoch fachübergreifend recht gering aus ($b \leq .15$, $R^2 \leq .06$). Für Mathematik stellt sich der Sachverhalt etwas differenzierter dar, indem die Nützlichkeitsperspektive *Lernstandsdiagnose* signifikant ($b = .30$, $p = .002$) die Durchsicht des Sofortberichts vorhersagt ($R^2 = .20$). Von den Nützlichkeitsperspektiven *Unterrichtsentwicklung* und *Schulentwicklung* werden ebenfalls die Durchsicht des Ergebnisberichtes und der tabellarischen Übersicht vorhergesagt. Dies entspricht den Erwartungen, da zur Diagnose in zeitlicher Nähe zur Durchführung zwecks Lernstandsdiagnose eher der Sofortbericht genutzt wird. Sowohl die Varianzaufklärung als auch die Regressionskoeffizienten und die p -Werte der Prädiktion der Nützlichkeitsperspektiven auf die Durchsicht von Ergebnis- und Schulbericht fallen dagegen deutlich schwächer aus. Der Grund für diesen Unterschied lässt sich in der Verfügbarkeit der Rückmeldungen vermuten. Während der Sofortbericht und die tabellarische Übersicht unmittelbar nach der Dateneingabe bereitstehen, sind die umfassenderen Analysen des Ergebnis- und Schulberichtes erst mit einigem zeitlichen Abstand zur LSE-Durchführung verfügbar. Zum Zeitpunkt der Verfügbarkeit ist eine unterrichtliche Nutzung nicht mehr im Fokus der Lehrkräfte und damit eine Durchsicht der Berichte – unabhängig von der Nützlichkeitsperspektive – weniger gegeben. Dabei tritt eine weitere Besonderheit auf: Während die Durchsicht des Ergebnisberichtes mit nur geringfügig schlechter abschneidet als die direkt verfügbaren Berichte, fällt der Schulbericht deutlicher ab¹¹¹.

Der Trend, dass mit zunehmender Distanz zum unmittelbaren Unterrichtsgeschehen auch die Nutzung/Nützlichkeit der LSE in der Wahrnehmung der Lehrkräfte geringer wird, zeigt sich auch in der Besprechung der Ergebnisse mit anderen sowie der Nutzung der Ergebnisse und bestätigt in Summe den Befund, dass der Unterricht das Kerngeschäft von Lehrkräften ist (Schildkamp & Kuiper, 2010, S. 484). Ein beachtlicher Unterschied zeigt sich im Bundeslandvergleich zu Thüringen. Nachtigall (2014) berichtet von Mittelwerten zur Nützlichkeit in den Facetten *Lernstandsdiagnostik*, *Unterrichtsentwicklung* und *Schulentwicklung*, die nahe dem semantischen Median

¹¹¹ M_{Ergebnisbericht} = 1.48 (0.45); M_{tabellarische Übersicht} = 1.26 (0.45); M_{Sofortbericht} = 1.38 (0.55); M_{Schulbericht} = 1.84 (0.81)

liegen (vgl. Tabelle 6). Von dieser Einschätzung weichen die Angaben der hessischen Lehrkräfte ab. Die Englisch-Lehrkräfte unterscheiden sich dabei nicht statistisch bedeutsam, während die Mathematik- und Deutschlehrkräfte in Hessen der *Lernstandsdiagnostik* signifikant deutlicher zustimmen ($t < -2.76$, $p < .007$). Bei den restlichen Einschätzungen fehlt der positiveren Einschätzung der hessischen Lehrkräfte die statistische Signifikanz. Zwar lassen sich die beiden Stichproben nicht unmittelbar in Bezug setzen oder gar durch Drittvariablen matchen, jedoch handelt es sich in beiden Befragungen um Lehrkräfte, die die LSE selbst im jeweiligen Jahr durchgeführt haben und ihre Selbst-Einschätzung berichten. Die Befragungen stammen somit aus sehr ähnlichen Rahmenbedingungen, ohne eine direkte Vergleichbarkeit zu gewährleisten.

Während es keine Fachunterschiede in der Besprechung gibt, zeigt sich dennoch die eindeutige Tendenz, dass die Besprechung mit dem gesamten Kollegium am seltensten erfolgt, während Besprechungen mit der Klasse, einzelnen Schülerinnen und Schülern oder der jeweiligen Fachschaft vergleichsweise häufig genannt werden. Diese Tendenz ist vor dem Hintergrund der *Unterrichtsentwicklung* nachvollziehbar, da die LSE lediglich auf die sprachlichen Fächer (Deutsch, Englisch/Französisch) und Mathematik beschränkt sind, und sich entsprechend wenig unterrichtsleitende Anhaltspunkte für andere Fächer finden lassen. Bezogen auf die Trias der *Schulentwicklung* lässt sich daraus mutmaßen, dass die LSE zwar eine *Unterrichtsentwicklung* und ggf. auch eine direkte *Personalentwicklung* (im Sinne eines erhöhten Fortbildungsbedarfes etc.) bewirken können, eine auf das Ganze bezogene *organisations- bzw. schulentwickelnde* Wirkung jedoch nicht unmittelbar entfalten können.

Interessanterweise fällt auf, dass die Prädiktion der Durchsicht bzw. der Nützlichkeit auf die tatsächliche Nutzung für Mathematik-Lehrkräfte zum Großteil nicht statistisch zu belegen ist. Scheinbar ist die Nutzung der LSE unabhängig von deren *a-priori*-Einschätzung der Lehrkräfte. Diese Tatsache lässt sich in zwei Richtungen interpretieren:

1. Diese vermuteten Zusammenhänge sind empirisch tatsächlich nicht zu finden bzw. in der untersuchten Stichprobe (aus welchen Gründen auch immer) nicht existent: Für Lehrkräfte hängen Nützlichkeit und Nutzung nicht zusammen.
2. Die Lehrkräfte nutzen die LSE nach dem Prinzip „Ich mache das Beste draus.“ Unabhängig von der *a-priori*-Einschätzung zu Nützlichkeit oder der

Ergebnisdurchsicht/-besprechung nutzt jede Lehrkraft die Ergebnisse für sich in individueller Weise.

Möglichkeit 1 würde stark verwundern bzw. die Validität des eingesetzten Instruments stark in Zweifel ziehen. Möglichkeit 2 scheint hier die plausiblere Variante zu sein. Die Nutzung der Ergebnisse erfolgt letztlich wiederum weitestgehend hinter verschlossener Klassentür und jede Lehrkraft hat die Umsetzung selbst in der Hand. Während also eine Durchsicht oder Besprechung von Ergebnismeldungen eine von Schulleitung, Fachvorsitzenden o. ä. aufoktroyierte Handlung sein kann, erfolgt die tatsächliche Nutzung „in Ruhe“.¹¹² Analog fällt der Zusammenhang zwischen der Nutzung und der Besprechung der Ergebnisse mit der Schulleitung im Vergleich zur Besprechung mit der Klasse und den Eltern weniger stark aus. Eine resultierende, in weiterer Forschung zu prüfende Vermutung hierzu wäre, dass die Lehrkräfte eine Art externe soziale Verpflichtung zur Nutzung der Lernstandserhebungen verspüren, wenn eine Besprechung mit der außerkollegialen Öffentlichkeit vorgenommen wird und damit ein Rechtfertigungsbedarf entsteht, wenn die Ergebnisse zwar besprochen werden, jedoch nicht Grundlage von Entwicklungen darstellen.

5.2 Grenzen der Arbeit

Ungeachtet der inhaltlichen Interpretation der Ergebnisse gilt es trotz aller Ansprüche kritisch reflektierend einige Aspekte in den Blick zu nehmen. Sowohl VELM-8 als auch die Online-Befragung der Lehrkräfteakademie sollten den Unterrichtsalltag so wenig wie möglich behindern: Entsprechend war in beiden Untersuchungen die Teilnahme der Lehrkräfte an Befragungen freiwillig. Aufgrund der Erhebungssituation handelt es sich des Weiteren in beiden Fällen lediglich um Gelegenheitsstichproben, die *per se* (also ohne Kontrolle/Beachtung von Hintergrundvariablen) keinen Anspruch auf Repräsentativität haben. Zweifelsfrei ist die Stichprobe der Online-Befragung sehr groß (sofern alle Jahrgangsstufen und Fächer eingeschlossen werden, liegen mehr als 6000 Fälle vor, sodass die hohe Signifikanz einiger Ergebnisse bei gleichzeitiger geringer Bedeutsamkeit, etwa H25a, durch die Stichprobengröße erklärt werden könnten), jedoch reduziert sich die Stichprobe bei Analyse von Mathematik in Jahrgang 8. Leider liegen bei dieser Erhebung im Grunde keine Hintergrundvariablen vor, sodass die Repräsentativität der Stichprobe vollkommen verborgen bleibt. Für

¹¹² Diese Interpretation konterkariert zweifelsfrei die Funktion der *Organisationsentwicklung* bei den LSE.

VELM-8 wurde im Rahmen der Projektteilnahme auf eine Randomisierung verzichtet und im Laufe des Projektes kam es zu Mortalität der Teilnehmenden bzw. um die Fortbildung nicht zu behindern, zum Nachrücken von vertretenden Kolleginnen und Kollegen. Diese Tatsachen verdeutlichen, dass die Ergebnisse von VELM-8, der aus diesen Prätestdaten berechneten Clusteranalyse und der Online-Befragung nur eingeschränkt als Referenzwerte für andere Untersuchungen genutzt werden können. Vielmehr lassen sich die Ergebnisse dieser Untersuchungen als ersten Standpunkt für die Wahrnehmung und Nutzung der LSE in Hessen verstehen. Von diesem Standpunkt aus sollten sich weitere Untersuchungen anschließen, um ein umfassendes und verlässliches Bild für Hessen und die Bundesländer zu schaffen, für die noch keine oder wenige Untersuchungen vorliegen.

Ebenfalls für alle drei Analysebereiche bzw. die beiden Befragungen identisch ist, dass die eingesetzten Instrumente weder selbst noch gezielt entwickelt, sondern weitestgehend unverändert übernommen wurden. Die eingesetzten Skalen aus COACTIV (Baumert et al., 2008) beispielsweise stammen dabei aus den Jahren 2003 und 2004, sind also zum Erhebungszeitpunkt etwa zehn Jahre alt. Aufgrund des zeitlichen Abstandes und/oder der föderal bedingten Unterschiede zwischen den Erhebungen in Baden-Württemberg sowie Thüringen (Maier, 2010) und Hessen könnte die Validität einiger Items eingeschränkt sein. Dies betrifft im Speziellen die Skala zum *selektionsdiagnostischen Nutzen*, da diese Nutzung in Hessen untersagt ist. Gleiches gilt für Aussagen über den Unterricht, die u. U. von Lehrkräften unterschiedlich interpretiert werden könnten: Beispielsweise werden in der Skala *kognitiv herausfordernder Umgang mit Schülerbeiträgen (Ideen/Fehler)* Lehrkräfte nach der Zustimmung zu folgender Aussage gefragt: „Ich lasse meine Schüler/innen auch einmal bewusst in die Irre laufen, bis sie sehen, dass etwas nicht stimmen kann.“ Die schlechte Trennschärfe dieses Items deutet möglicherweise bereits an, dass Lehrkräften dieses Item in unterschiedlicher Weise verstehen. Bei COACTIV war mit diesem Item sicher keinesfalls intendiert, zu erfragen, ob Lehrkräfte böswillig Schülerinnen und Schüler stundenlang ins „Leere“ laufen lassen, um ihnen am Ende zu sagen: „Das war wohl nichts...nochmal von vorne!“. Allerdings kann diese Interpretation durch Lehrkräfte nicht ausgeschlossen werden. Ähnliche Schwierigkeiten zeigen sich bei anderen Items ebenfalls. Entsprechend wäre es ratsam bei einer erneuten Untersuchung der Situation in Hessen den Fokus deutlicher auf die Forschung zu legen. Durch diese Fokussierung könnte ausreichend Zeit in die Instrumentenentwicklung investiert werden, sodass

die Validität der eingesetzten Items (wie selbstverständlich auch die Objektivität und Reliabilität als deren Voraussetzung) nicht in Frage steht. Für die Online-Erhebung wurde auf die eingesetzten Items und Skalen der Hessischen Lehrkräfteakademie zurückgegriffen. Diese Items wurden nicht gezielt unter *wissenschaftlichen* Kriterien entwickelt. Dadurch ist eine fehlende Validität der Items nicht auszuschließen. Insbesondere bleibt fraglich, ob die Lehrkräfte bewusst zwischen *Nützlichkeit* und *Nutzung* der Lernstandserhebungen unterscheiden (vgl. Abschnitt 3.4 zu eingesetzten Instrumenten bzw. Abschnitte 4.3.3 und 4.3.5 im Vergleich). Ebenfalls in Bezug auf die Validität von Items bzw. Skalen verwundern die z. T. vorhandenen Abweichungen zwischen der Skala *Nützlichkeit* und den per Einzelitem erfragten Nützlichkeits-Dimensionen *Lernstandsdiagnose*, *Unterrichtsentwicklung* und *Schulentwicklung*, die jedoch ihrerseits untereinander jeweils hochsignifikant miteinander in Zusammenhang stehen (fachübergreifend mit $r > .75$, $p < .01$ wie auch für Mathematik mit $r > .69$, $p < .01$; vgl. auch Abschnitte 3.4.2, 4.3.3 und 4.3.5).

Neben der Validität der Items steckt eine weitere Gefahr in den Likert-Skalen zur Erfassung der Zustimmung. Diese drei- bis fünfstufigen Skalen wurden in den Erhebungen fast ausschließlich eingesetzt. Einige Items im Online-Fragebogen wurden mittels Schulnoten bewertet, welche letztlich aber ebenfalls „nur“ ordinales Skalenniveau haben (Bortz & Döring, 2006; Eid et al., 2011). Die vieldiskutierte Problematik besteht nun zum einen darin, Ordinalskalen zu einem Skalen-„Mittelwert“ zusammenzufassen und damit eine aus mathematischer Sicht nicht zulässige Operation durchzuführen. Andererseits ist die Validität der einzelnen Ausprägungen der Likert-Skalen fraglich, so ist nicht eindeutig festzustellen, „ab wann“ eine Lehrkraft in den Fragen zur Nutzung (Online-Fragebogen) statt 2 (= *punktuell/teilweise*) eine 3 (= *systematisch/vollständig*) ankreuzt und was für die einzelne Lehrkraft jeweils *punktuell/teilweise* bzw. *systematisch/vollständig* bedeutet.

Nicht zuletzt wird die soziale Erwünschtheit ebenfalls einen gewissen Einfluss haben, insbesondere wegen der Selbstauskünfte, die Grundlage aller Berechnungen sind. Dieser sowie der zuvor genannte Punkt sind jedoch Einwände, die an jede Erhebung mit Hilfe von Fragebögen inkl. Likert-Skalen gerichtet werden muss. Auf dem Stand der Forschung ist der likert-skalierte Fragebogen das vermutlich mit am häufigsten genutzte Erhebungsinstrument der empirischen Bildungsforschung und Fachdidaktik und es liegen aktuell keine vergleichbaren Alternativen vor, die die Erfassung von

Einstellungen von Lehrkräften in adäquatem Maße zulassen würden. Als Erweiterung dieses Punktes muss ergänzt werden, dass für die teilnehmenden Lehrkräfte eine Trennung zwischen evaluierendem und fortbildendem Personal innerhalb des Projektes nicht zwingend erkennbar war. Insbesondere waren sowohl in der Projekt- als auch der Kontrollgruppe persönliche Bekanntschaften zwischen Projektteam und teilnehmenden Lehrkräfte nicht auszuschließen. Dieser Fakt könnte u. U. die soziale Erwünschtheit erhöhen, wobei selbstverständlich auf Anonymität und Vertraulichkeit der Fragebögen geachtet wurde, sodass eine Rückverfolgung der Itemantworten zu den beteiligten Lehrkräften nicht möglich war.¹¹³

Konkret auf die Studie VELM-8 bezogen muss an dieser Stelle nochmals erwähnt werden, dass die Stichprobe mit zu Beginn weniger als 30 Personen für den Längsschnitt abzüglich Drop-Outs z. T. nur bei einem N von 20 liegt und zusätzlich nach den Ergebnissen der Mittelwertvergleiche zu Maier (2010) sowie Baumert et al. (2008) vermutlich eine Positiv-Selektion darstellt (vgl. Tabelle 18 und Tabelle 20). Wie bereits bei den allgemeinen Ausführungen zu den Stichproben erwähnt, schmälert dies jedoch zunächst nicht die Aussagekraft der Ergebnisse, sondern nur die Möglichkeit sie zu verallgemeinern. Gegebenenfalls könnten Deckeneffekte die Entwicklung der Gruppen bzw. die Einschätzung der Stichprobe verzerren, da aufgrund der hohen Werte keine Differenzierung möglich ist (vgl. u. a. Skala *Lehrplanvalidität*). Als Ausgangspunkt für weitere Analysen und als Grundlage weiterer Fortbildungen eignet sich VELM-8 und die zugehörige Evaluation jedoch in jedem Falle.

Im Rahmen des Projektablaufes und -designs wurde bereits erwähnt, dass sich das Projekt VELM-8 über ein gesamtes Schuljahr erstreckte. Die Auftaktveranstaltung, auf der gleichzeitig auch die Datengrundlage für die Clusteranalyse erhoben wurde, lag bereits kurz nach den Sommerferien und die Abschlussbefragung erst kurz vor den nächsten Sommerferien. Daraus resultiert, dass die Fortbildungen in verschiedensten schuljahresbedingten Situationen auf die Lehrkräfte zukamen. Obwohl bei der Durchführung und Terminwahl für die Fortbildungen die Arbeitsgruppe durch erfahrene Lehrkräfte beraten wurde, ist nicht auszuschließen, dass die ein oder andere Fortbildung terminlich schlecht gewählt wurde, wodurch eine geringe

¹¹³ Hierfür wurden innerhalb des Projektteams Zuständigkeiten verteilt, die jedoch für die teilnehmenden Lehrkräfte u. U. nur bedingt erkennbar waren.

Fortbildungsbeteiligung resultierte, denn nicht alle Lehrkräfte der Projektgruppe waren zu allen Fortbildungen anwesend.

Wird das Modell der Fortbildungswirksamkeit nach Lipowsky (2014; Abbildung 11) herangezogen, so muss ergänzt werden, dass aufgrund der langen Projektlaufzeit keine Möglichkeit bestand, den Unterricht der teilnehmenden Lehrkräfte zu beobachten bzw. zu kontrollieren. Damit bleibt für den Moment verborgen, ob die berichteten Effekte zu veränderten Kognitionen (Ebene 2) oder sogar veränderter Unterrichtspraxis (Ebene 3) geführt haben. Messbare Unterschiede in der Leistung der Schülerinnen und Schüler resultierten jedoch in dieser Untersuchung nicht. In den theoretischen Ausführungen wurde das Feedback-Modell von Hattie und Timperley (2007) angeführt, nach dem wirksames Feedback Antworten auf drei Fragen geben sollte: (1) Where am I going?, (2) How am I going? und (3) Where to next?. Postuliert wurde, dass das für Feedback an Schülerinnen und Schüler konzipierte Modell auch für die Rückmeldungen (das Feedback) aus Lernstandserhebungen für Lehrkräfte plausibel ist: Zweifelsfrei liefern die Rückmeldungen aus den LSE Informationen, die korrekt interpretiert eine Antwort auf Frage (2) im Modell liefern. Die Antwort auf Frage (1) ist für Lehrkräfte letztlich im Lehrplan/Kerncurriculum festgehalten, sodass es hierzu nur eingeschränkt Rückmeldungen bedarf. Die Antwort auf Frage (3) liegt hingegen vor allem in der professionellen Kompetenz der Lehrkräfte selbst – und weniger in den LSE oder deren Rückmeldungen. Als ergänzende und konkret handlungsleitende Vorschläge zur Unterrichtsgestaltung nach den LSE bieten die *didaktischen Kommentare* Anregungen, wie mit erwartbaren Schwierigkeiten umgegangen werden könnte. Insofern bieten die Rückmeldungen aus den LSE das Potenzial, im Sinne Hattie und Timperleys als wirksames Feedback zu gelten. Allerdings bedarf es über die reine Rückmeldung hinaus, zwingend einer bewussten Auseinandersetzung seitens der Lehrkräfte, um dies zu erreichen. Hinzukommt, dass – wie in Fußnote 11 angemerkt – der „Lerngegenstand“ bei Lehrkräften, also eine Unterrichtsveränderung oder Erweiterung der eigenen professionellen Kompetenz, um ein Vielfaches komplexer ist als in dem Zusammenhang, für den das Feedback-Modell entwickelt wurde. In diesem Sinne stellt das Modell als Inhalt der Fortbildungen für die Lehrkräfte eine Orientierung zur Gestaltung von Feedback an die Schülerinnen und Schüler dar sowie eine Heuristik zur Strukturierung von Feedback durch die Lernstandserhebungen an sich selbst.

Ähnlich wie die Erwartung an die direkte Übertragbarkeit des Feedback-Modells im Rahmen von VELM-8 sind die formulierten Hypothesen ebenfalls sehr weitreichend formuliert. Dieser Umstand wurde bereits in Teil 3 angeführt, soll hier allerdings nochmals verdeutlicht werden: VELM-8 wurde als Fortbildungs- und nicht als Forschungsprojekt konzipiert und unterlag dabei wesentlichen Einschränkungen, die eine aus theoretischer Sicht wirksame bzw. wirksamere Fortbildungsgestaltung erschwerten. Dennoch wurden weitreichende Hypothesen artikuliert, die kritisch angemerkt, nicht immer realistisch erscheinen. Insbesondere erwartete Wirkungen auf Unterrichtsqualitätskriterien, die in der Fortbildung nur implizit thematisiert wurden, stellen eher wünschenswerte Entwicklungen denn erwartbare Veränderungen durch die Fortbildung dar. Wie in Abschnitt 1.5 dargelegt bedarf es zur Veränderung von Einstellungen und besonders von Haltungen und Verhaltensweisen der Lehrkräfte wiederholter Fortbildungen mit einer gewissen Dauer. Diese Voraussetzungen sind bei VELM-8 mit vier Einzeltermine ohne Begleitung der Lehrkräfte in der Zwischenzeit letztlich nicht gegeben gewesen.

Aus methodischer Sicht lassen sich zwei Punkte anführen: Für die Auswertungen der Projektevaluation und der Typenbildung wären elaborierte Verfahren an einigen Stellen hilfreich gewesen. So könnte eine latente Modellierung im Rahmen von Strukturgleichungsmodellen (etwa als Cross-Lagged-Panel zur Untersuchung von Beziehungen zwischen Belastung und Nutzen) oder aber Mixed-Rasch-Modelle bzw. Latent-Class-Analyses für die Typenbildung ein statistisch akkurateres Bild erbringen. Die eingesetzten Analysemethoden sind vermutlich nicht auf dem aktuellsten Stand der methodischen Forschung, stellen allerdings geringere Anforderungen an die Datenqualität und die Stichprobengröße als dies elaborierte Verfahren mit vielen zu schätzenden Parametern tun. Für den Einsatz von latenten Analyseverfahren war letztlich die ausreichende Stichprobengröße nicht vorhanden. Ein weiterer Aspekt betrifft die Konfigurationsfrequenzanalyse (KFA), die die Clusteranalyse statistisch erweitert, jedoch durch den Mediansplit und die aufgrund der Stichprobengröße notwendige Aufteilung auf zwei KFAs ebenfalls problembehaftet ist. Durch die getrennte Betrachtung der Skalen *Akzeptanz*, *Belastung* und *Lehrplanvalidität* sowie der Skalen *förderdiagnostischer Nutzen*, *unterrichtsplanerischer Nutzen* und *selektionsdiagnostischer Nutzen* besteht keine Möglichkeit die gefundenen Typen miteinander direkt in Bezug zu setzen und so einen direkten Abgleich mit der Clusteranalyse zu erreichen. Daher stehen letztlich die Ergebnisse dieser beiden KFAs zunächst

unverbunden nebeneinander. Maxwell und Delaney (1993) weisen überdies auf die Problematik des Mediansplits hin, da dieser prinzipiell eine direkte Abhängigkeit von der Stichprobe aufweist sowie u. U. zu Verzerrungen in Beziehungen einzelner Variablen zueinander führen kann.

5.3 Forschungsdesiderate

Der Einsatz einer Fortbildungsreihe in Nordhessen zur Nutzung der LSE hat dazu geführt, dass teilnehmende Lehrkräfte die LSE weniger als Belastung einschätzen, die LSE mehr nutzen und eine z. T. verbesserte Unterrichtsqualität berichten. Dieses positive Bild der Projektevaluation soll nicht darüber hinwegtäuschen, dass die Veränderungen keinesfalls kausal mit den Fortbildungen in Verbindung zu bringen sind. Das Design von VELM-8 macht deutlich, dass der Fortbildungsaspekt größer geschrieben wurde als der Forschungsaspekt und resultierend keine kausalen Aussagen möglich sind. Hierfür müssten striktere Regeln befolgt werden, wie etwa eine randomisierte Zuweisung zu Projekt- und Kontrollgruppe, eine nach Möglichkeit kürzere Projektlaufzeit, um das Treatment deutlicher fassen zu können, eine größere Kontrolle von Einflussfaktoren sowie die Erhebung weiterer Hintergrundvariablen. Mit diesen Punkten ist bereits angesprochen, dass die Projektevaluation keine Aussagen auf Ebene 2 und 3 des Fortbildungserfolges zulässt. Weder die Kognitionen von Lehrkräften noch der tatsächliche Unterricht wurden evaluiert und können somit in die Beurteilung nicht miteingeschlossen werden. Bei einer Replikation von VELM-8 sollte demnach darauf geachtet werden, die methodische Qualität der Evaluation zu steigern und weitere Ebenen von Fortbildungserfolg etwa durch einen Kompetenztest für Lehrkräfte in den Blick zu nehmen, dafür ggf. eher auf die Untersuchung von Ebene 4 (Leistung der Schülerinnen und Schüler) zunächst zu verzichten. Sollten Effekte auf Ebene 2 und 3 nachgewiesen sein, kann weitere Anschlussforschung wiederum Ebene 4 in den Blick nehmen. Diese Forderung gewinnt an Bedeutung, da die bisherigen Forschungsergebnisse fast ausschließlich auf Selbstauskünften der Lehrkräfte beruhen (Maier et al., 2012; Ramm, 2014).

Spannend wäre eine Verknüpfung des Projektes VELM-8 mit der Clusteranalyse, die der Frage nachgehen könnte, ob verschiedene Typen von Lehrkräften in ähnlichem oder unterschiedlichem Maße Fortbildungen zu LSE besuchen und von diesen profitieren, modelliert etwa durch Latent-Growth-Analysen. Aus dem Bildungstrend bzw. Ländervergleich ist bekannt, dass beispielsweise fachfremd unterrichtende

Lehrkräfte weniger Fortbildungen zum Fachwissen besuchen und wünschen (Richter et al., 2013). Setzt sich dies für die Nutzung der und Einstellung zu den LSE fort? Besuchen und/oder profitieren gerade solche Lehrkräfte von gezielten LSE-Fortbildungen, die diesen ohnehin bereits optimistisch gegenüberstehen (vgl. *Neigungshypothese*)? Ist es möglich, mit Hilfe von Fortbildungen die ablehnende Haltung einiger Lehrkräfte gegenüber den LSE abzubauen (vgl. Cluster 5)? Letztlich führt dieser Gedanke zu einer allgemeineren Frage: Welche Lehrkräfte profitieren von Fortbildung und auf welche Weise profitieren sie? In diesem größeren Sinne wären meta-analytische Untersuchungen notwendig, um ein aussagekräftiges Bild zu erstellen. Insbesondere vor dem Hintergrund der jüngsten medienpädagogischen und mediendidaktischen Entwicklungen wäre auch sinnvoll, in diesem Zuge die Rolle von Fortbildungsformaten (Präsenzveranstaltung gegenüber Webinar o. ä.) zu untersuchen.

Als Desiderat lässt diese Arbeit im Speziellen für die Situation der LSE in Hessen jedoch zwei weitere offene Punkte zurück: Zum einen kann die Arbeit (aufgrund der genannten Einschränkungen) nur einen ersten Versuch eines Ist-Standes darstellen, den es zu ergänzen und repräsentativ auszubauen gilt. Zum anderen ist der Auflösungsgrad der Arbeit bei weitem zu groß, um die tatsächliche Arbeit der Lehrkräfte mit den LSE am heimischen Schreibtisch, in der Unterrichtsplanung und im alltäglichen Unterricht zu erforschen. In diesem Sinne stellt die vorliegende Arbeit eher eine Momentaufnahme von einer mittleren Ebene dar. Diesen Ausschnitt gilt es nun in zwei Richtungen zu ergänzen: einerseits einen allumfassenderen Blick, der die Gesamtsituation repräsentativ analysiert, und andererseits einen Blick „mit Lupe“ auf die konkrete Arbeit mit den LSE, deren Rückmeldungen etc. Diese beiden Perspektiven könnten dringende Fragen beantworten, wie etwa:

- Welche zusätzlichen Hilfen benötigen Lehrkräfte?
- Welche der verfügbaren Rückmeldungen sind redundant?
- Welche zusätzlichen Informationen wünschen sich Lehrkräfte?
- Wie können die *didaktischen Kommentare* und Aufgabenkommentare noch nützlicher gestaltet werden?
- ...

Im Rahmen einer solchen Untersuchung sollte genauer betrachtet werden, ob Lehrkräfte die Ergebnisse der LSE im Sinne von Visscher und Coe (2003) in intendierter oder unintendierter Art nutzen bzw. welche von den zugehörigen Unterarten

auftreten. Hierfür wäre im Idealfall eine Panel-Untersuchung zu ermöglichen, die Lehrkräfte insbesondere auf Ebene 2 und 3 des Fortbildungserfolges begleitet/evaluiert. Ein allgemeinerer Blick könnte darüber hinaus bereits aufgeworfene Fragen klären, wie etwa die Wirkrichtung von Verständlichkeits- und Nützlichkeits Einschätzungen (vgl. auch Koch, 2011) oder die Hypothese sozialen Drucks bei der Nutzung nach Besprechung der Ergebnisse mit Eltern und/oder Schülerinnen und Schülern.

Die Erweiterung des Ausschnittes „mit der Lupe“ würde dabei zusätzlich die Grundlage legen, alternative Unterstützungsformate für die datengestützte Schulentwicklung zu erproben. Wesentlicher Kritikpunkt an VELM-8 ist die fehlende Begleitung in den Distanzphasen zwischen den Präsenzterminen der Fortbildungsreihe. Mit dem Wissen, welche Hilfen, Information und/oder Unterstützung Lehrkräfte für die Nutzung der LSE-Rückmeldungen benötigen, könnten für die Distanzphasen zielgerichtete Hilfen angeboten werden. Denkbar wären hierbei etwa eine dezentrale Moderation der Auswertungen an den einzelnen Schulen, wie beispielsweise in der Bildungsstandardevaluation in Österreich erfolgt (Rieß & Zuber, 2014) oder eine „Auswertungssprechstunde“. Ebenfalls denkbar wäre der Aufbau eines innerschulischen „LSE-Teams“ im Sinne einer *Professionellen Lerngemeinschaft* aus den Fortbildungsteilnehmenden und Fortbildenden, von denen die Auswertung der LSE-Rückmeldungen gemeinsam erfolgt.

Insbesondere vor dem Hintergrund der aktuellen Entwicklungen im Bereich digitaler Medien (KMK, 2016a) wäre eine Online-Variante der Lernstandserhebungen ein interessantes Untersuchungsfeld (Ramm & Rogalski, 2013), sofern durch die Interaktivität Lehrkräfte die Möglichkeit erhalten, auf verschiedene Rückmeldungen zuzugreifen und so den Übersetzungsschritt aus tabellarischer Übersicht hin zur eigenen Analyse abgenommen bekämen. Auf diese Weise würden Rezeptionsschwierigkeiten vermieden (Rolff, 2001) und Lehrkräfte hätten Zugriff auf diverse Rückmeldungen.

5.4 Fazit

Den Ergebnissen aus der Evaluation von VELM-8, einer Typisierung von Lehrkräften und einem Ist-Stand aus Online-Daten stehen diverse Einschränkungen und ausbleibende Antworten auf Fragen bzw. neu entstandene Forschungsfragen gegenüber. Die Menge der Einschränkungen ist dabei sicherlich fast erdrückend hoch, doch stellt die vorliegende Arbeit zumindest eine Etappe auf dem Weg zur Weiterentwicklung und Effektivierung der Lernstandserhebungen dar.

Bisher gibt es leider keine Datengrundlage für Hessen. Diese wird mit den Analysen aus VELM-8 und der Online-Daten erstmals gelegt. Dies kann weiteren Analysen und Untersuchungen als *Baseline* dienen. Ebenfalls als Referenz sollte die Typenbildung herangezogen werden können. Die Fortbildungsevaluation zeigt vereinzelt Effekte der Fortbildungen oder lässt diese zumindest z. T. vermuten. Dabei sollen einige besonders interessante Befunde nochmals verdeutlicht werden:

- Im Verlauf von VELM-8 berichten die Lehrkräfte über eine abnehmende *Belastung* und dies vor allem bei den unterrichtserfahrenen Lehrkräften. Überdies scheinen die Lehrkräfte in der Wahrnehmung der eigenen Belastung zwischen allgemeiner Belastung und der konkreten Belastung durch die LSE zu unterscheiden, da erstgenannte erwartungskonform im Laufe des Schuljahres ansteigt.
- Die Lehrkräfte berichten sowohl in VELM-8 als auch in den Online-Daten von einer hohen *Validität* des Instrumentes. Daraus lässt sich ableiten, dass die LSE ein Instrument darstellen, welches als Spiegel des alltäglichen Unterrichts von den Lehrkräften eingeschätzt wird.
- Nicht zu vernachlässigen ist vermutlich die Wirkung der *erstmaligen LSE-Teilnahme*. In den Online-Daten zeigt sich höhere Nützlichkeit bei Teilnahme an den LSE und die Einstellungen der Lehrkräfte in VELM-8 entwickeln sich ebenfalls positiv bei einer initial sehr geringen Vorerfahrung mit den LSE.
- In Bezug auf *Typen* zeigt sich offensichtlich studienübergreifend ein ablehnender, sehr kritischer und ein zustimmender, optimistischer Typ. Hierbei stellt sich die Frage nach Anschlussforschung, wie diese Typen tatsächlich mit den LSE umgehen bzw. ob sich ihre Einstellungen verändern/verbessern lassen. Selbstverständlich gelten diese Fragen ebenso für die nicht direkt an andere Studien anknüpfenden Typen.

Trotz aller Einschränkungen ist es somit gelungen, mit VELM-8 einen ersten Vorschlag für eine in Teilen wirksame Fortbildungskonzeption zu machen. Wiederum eignet sich dieser Vorschlag als Referenz, der für eigene Vorhaben angepasst, erweitert, reduziert werden kann, um so die spezifischen regionalen oder föderalen Besonderheiten zu berücksichtigen.

Es bleibt zu hoffen, dass die Mühen und die investierte Arbeit in eine effektivere Nutzung der LSE für die weitere Gestaltung von Fortbildungen, die Konzeption und

Entwicklung der LSE sowie von Informationen, Unterrichtsvorschlägen, Fördermaterial etc. zu den LSE genutzt werden. Dieser Fortbildungsvorschlag sollte in folgenden Studien und Projekten fortentwickelt und angepasst werden, um weitere Lehrkräfte in Hessen über die Lernstandserhebungen bzw. in ganz Deutschland über die Vergleichsarbeiten, Kompetenztests etc. zu informieren, ihnen Unterstützung bei der Nutzung zu geben und so letztlich eine Unterrichtsentwicklung mit einer Personalentwicklung zu verbinden. Bezugnehmend auf das Eingangszitat zum Abschnitt 1.5.4.1 zu Zieldimensionen von Unterrichtsentwicklung muss nicht nur jede Unterrichtsstunde vor dem Hintergrund der vermittelten und durch die Schülerinnen und Schüler erworbenen Kompetenzen reflektiert werden, sondern in gleichem Maße gilt dies auch für Fortbildungen, welche Kompetenzen bei Lehrkräften durch die Fortbildungsveranstaltung bzw. -reihe entwickelt wurden.

Durch die Ergebnisse der Evaluation von VELM-8, den Online-Daten und durch die Fortbildungskonzeption VELM-8 liegen diverse Informationen vor, die über diese Arbeit hinaus für die Effektivierung der Lernstandserhebungen genutzt werden können, um so letztlich Entwicklungen im Mathematikunterricht anzustoßen.

6 Literaturverzeichnis

- Abbott, S. E. (2014, 18. August). *High-stakes test*, The Glossary of Education Reform. Zugriff am 05.09.2016. Verfügbar unter <http://edglossary.org/high-stakes-testing/>
- AC Del Re. (2015). *Package ‚compute.es‘. Compute effect sizes*, CRAN. Zugriff am 05.06.2017. Verfügbar unter <https://cran.r-project.org/web/packages/compute.es/compute.es.pdf>
- Adams, R. J. & Wu, M. L. (2002, 1. Januar). *PISA 2000 technical report*. Paris: OECD. Verfügbar unter <http://www.oecd.org/edu/school/programmeforinternational-studentassessmentpisa/33688233.pdf>
- Ahrling, I. & Brömer, B. (1999). *Schule machen. Das pädagogische Konzept der Offenen Schule Waldau/Kassel*. Baltmannsweiler: Schneider-Verl. Hohengehren.
- Aldenderfer, M. S. & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, Calif.: SAGE Publications.
- Alfieri, L., Brooks, P. J., Aldrich, N. J. & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18.
- Altrichter, H. (1998). Reflexion und Evaluation in Schulentwicklungsprozessen. In H. Altrichter, W. Schley & M. Schratz (Hrsg.), *Handbuch zur Schulentwicklung* (S. 263–335). Innsbruck: StudienVerlag.
- Altrichter, H. (2010). Konzepte der Fremdevaluation. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 170–175). Stuttgart: UTB.
- Ames, C. & Archer, J. (1988). Achievement goals in the classroom. Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80(3), 260–267.
- Anderson, J. C. & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14(2), 115–129.
- Arnold, K.-H. (1999). *Fairness bei Schulsystemvergleichen. Diagnostische Konsequenzen von Schulleistungsstudien für die unterrichtliche Leistungsbewertung und binenschulische Evaluation* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 9). Münster: Waxmann.

- Arnold, K.-H. (2007). Bildungspolitische, diagnostische und didaktische Bedingungen und Wirkungen von Schulleistungsevaluationen. *Empirische Pädagogik*, 21(4), 448–457.
- Artelt, C. (1999). Lernstrategien und Lernerfolg - Eine handlungsnaher Studie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 31(2), 86–96.
- Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, 64(6), 359–372.
- Autorengruppe Bildungsberichtserstattung. (2016). *Bildung in Deutschland 2016. Ein indikatorengestützter Bericht mit einer Analyse zu Bildung und Migration*. Bielefeld: W. Bertelsmann Verlag GmbH et Co. KG.
- Bacher, J., Pöge, A. & Wenzig, K. (2010). *Clusteranalyse. Anwendungsorientierte Einführung in Klassifikationsverfahren*. München: Oldenbourg.
- Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung*. Berlin, Heidelberg: Springer Gabler.
- Baguley, T. (2012). *Serious stats. A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave Macmillan.
- Bakeman, R. & Gottman, J. M. (2010). *Observing interaction. An introduction to sequential analysis*. Cambridge: Cambridge University Press.
- Baker, F. B. & Hubert, L. J. (1975). Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*, 70(349), 31.
- Ball, D. L., Thames, M. H. & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, 59(5), 389–407.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84(2), 191–215.
- Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist*, 37(2), 122–147.
- Baptist, P. (1998). *Elemente einer neuen Aufgabenkultur. Anregungen zu den Modulen 1 und 5*. Bayreuth. Zugriff am 19.12.2019. Verfügbar unter <http://sinus-transfer.uni-bayreuth.de/fileadmin/MaterialienDB/21/aufgabenkultur.pdf>
- Baptist, P. (2001). Aus der Praxis des Modellprogramms: Mathematikunterricht verändern - Verständnis fördern. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 67–73). Bonn.

- Baptist, P. & Raab, D. (2007, 1. Januar). *Auf dem Weg zu einem veränderten Mathematikunterricht* (Zentrum zur Förderung des mathematisch-naturwissenschaftlichen Unterrichts, Hrsg.). Bayreuth. Zugriff am 19.12.2019. Verfügbar unter <http://sinus-transfer.uni-bayreuth.de/fileadmin/MaterialienDB/465/sinus-transfer.pdf>
- Barfknecht, T. & Saldern, M. von (2010). Evaluation und Feedback der Lehrkräfte. In M. Demmer & M. v. Saldern (Hrsg.), *„Helden des Alltags“. Erste Ergebnisse der Schulleitungs- und Lehrkräftebefragung (TALIS) in Deutschland* (Die deutsche Schule Beiheft, Bd. 11). Münster: Waxmann.
- Bartnitzky, H. (2005). Leistungen feststellen - Fremdkörper oder Teil der pädagogischen Leistungskultur? *Grundschule aktuell*, (89), 3–6.
- Bauch, W., Maitzen, C. & Katzenbach, M. (2011). *Auf dem Weg zum Kompetenzorientierten Unterricht - Lehr- und Lernprozesse gestalten. Ein Prozessmodell zur Unterstützung der Unterrichtsentwicklung* (Amt für Lehrerbildung, Hrsg.). Wiesbaden: Amt für Lehrerbildung.
- Baumert, J. (2001a). Evaluationsmaßnahmen im Bildungsbereich. Eine kritische Sicht auf mögliche Zugänge. *ÖFEB-Newsletter*, 1(2), 3–15. Zugriff am 19. Februar 2015. Verfügbar unter <http://www.oefeb.at/>
- Baumert, J. (2001b). Vergleichende Leistungsmessung im Bildungsbereich. *Zeitschrift für Pädagogik. Beiheft*, (43), 13–36.
- Baumert, J., Blum, W., Brunner, M., Dubberke, T., Jordan, A., Klusmann, U. et al. (2008). *Professionswissen von Lehrkräften, kognitiv aktivierender Mathematikunterricht und die Entwicklung von mathematischer Kompetenz (COACTIV) Dokumentation der Erhebungsinstrumente*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Baumert, J., Klieme, E. & Bos, W. (2001). Mathematische und naturwissenschaftliche Bildung am Ende der Schullaufbahn - Die Herausforderung von TIMSS für die Weiterentwicklung des mathematischen und naturwissenschaftlichen Unterrichts. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 11–42). Bonn.
- Baumert, J. & Kunter, M. (2011). Das Kompetenzmodell von COACTIV. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.),

- Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 29–53). Münster: Waxmann.
- Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A. et al. (2009). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal*, 47(1), 133–180.
- Baumert, J., Kunter, M., Brunner, M., Krauss, S., Blum, W. & Neubrand, M. (2004). Mathematikunterricht aus Sicht der PISA-Schülerinnen und Schüler und ihrer Lehrkräfte. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland: Ergebnisse des zweiten internationalen Vergleichs* (S. 314–354). Band 1. Münster: Waxmann.
- Baumert, J. & Lehmann, R. (1997). *TIMSS - Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich. Deskriptive Befunde*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Baur, N. & Blasius, J. (Hrsg.). (2019). *Handbuch Methoden der empirischen Sozialforschung*. Wiesbaden: Springer VS.
- Besser, M., Blum, W. & Klimczak, M. (2013). Formative assessment in everyday teaching of mathematical modelling. Implementation of written and oral feedback to competency-oriented tasks. In G. Stillman, G. Kaiser & W. Blum (Eds.), *Teaching mathematical modelling. Connecting to research and practice* (pp. 469–478). Dordrecht: Springer.
- Biermann, M., Wiegand, B. & Blum, W. (2003). Nicht „irgendwie“, sondern zielgerichtet Aufgaben verändern. *Friedrich Jahresheft XXI*, 32–35.
- Bischof, L. M., Hochweber, J., Hartig, J. & Klieme, E. (2013). Schulentwicklung im Verlauf eines Jahrzehnts. Erste Ergebnisse des PISA-Schulpanels. In E. Klieme & N. Jude (Hrsg.), *PISA 2009 - Impulse für die Schul- und Unterrichtsforschung* (S. 172–199). Weinheim: Beltz.
- Blömeke, S. (2013). *Deutsches Zentrum für Lehrerbildung Mathematik (DZLM): Theoretischer Rahmen*. Deutsches Zentrum für Lehrerbildung Mathematik (DZLM). Zugriff am 24.04.2014. Verfügbar unter http://www.dzlm.de/files/uploads/DZLM_Theorierahmen.pdf
- Blömeke, S., Kaiser, G. & Lehmann, R. (2010). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich*. Münster: Waxmann.

- Bloom, B. S. (1976). *Taxonomie von Lernzielen im kognitiven Bereich* (Beltz-Studienbuch, Bd. 35, 5. Aufl.). Weinheim: Beltz.
- Blum, W. (2001). Was folgt aus TIMSS für Mathematikunterricht und Mathematiklehrausbildung? In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 75–83). Bonn.
- Blum, W. (2006). Modellierungsaufgaben im Mathematikunterricht - Herausforderung für Schüler und Lehrer. In A. Büchter, H. Humenberger, S. Hußmann & S. Prediger (Hrsg.), *Realitätsnaher Mathematikunterricht - Vom Fach aus und für die Praxis. Festschrift für Hans-Wolfgang Henn zum 60. Geburtstag* (S. 8–23). Hildesheim [u.a.]: Franzbecker.
- Blum, W. (2012). Einführung. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM* (6. Auflage, S. 14–32). Berlin: Cornelsen Scriptor.
- Blum, W. (2015). Quality teaching of mathematical modelling: What do we know, what can we do? In S. J. Cho (Hrsg.), *The Proceedings of the 12th international congress on mathematical education* (S. 73–96). Cham: Springer International Publishing.
- Blum, W., Drüke-Noe, C., Hartung, R. & Köller, O. (Hrsg.). (2012). *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM*. Berlin: Cornelsen Scriptor.
- Blum, W., Drüke-Noe, C., Leiß, D. & Wiegand, B. (2005). Unterrichtsqualität durch Bildungsstandards - Utopie oder Wirklichkeit. In G. Graumann (Hrsg.), *Beiträge zum Mathematikunterricht. Vorträge auf der 39. Tagung für Didaktik der Mathematik vom 28.2. bis 4.3.2005 in Bielefeld*. Hildesheim: Franz Becker. Zugriff am 01.02.2017. Verfügbar unter <http://www.mathematik.tu-dortmund.de/ieem/cms/media/BzMU/BzMU2005/Beitraege/blum-druekenoe-leiss-wiegand-gdm05.pdf>
- Blum, W., Drüke-Noe, C., Leiß, D., Wiegand, B. & Jordan, A. (2005). Zur Rolle von Bildungsstandards für die Qualitätsentwicklung im Mathematikunterricht. *Zentralblatt der Didaktik der Mathematik (ZDM)*, 37(4), 267–274.
- Blum, W., Hammann, M., Höfer, D., Katzenbach, M., Messner, R., Stäudel, L. et al. (2006). *PISA macht Schule - Konzeption und Praxisbeispiele zur neuen*

- Aufgabenkultur* (1. Auflage) (Steffens, U. & Messner, R., Hrsg.) (Folgerungen aus PISA für Schule und Unterricht Nr. 3). Wiesbaden: Institut für Qualitätsentwicklung (IQ).
- Blum, W., Krauss, S. & Neubrand, M. (2008). Zusammenhänge des Professionswissens mit Lehrermerkmalen, Unterrichtsqualität und Leistungszuwächsen der Schüler. In E. Vásárhelyi (Hrsg.), *Beiträge zum Mathematikunterricht 2008. Vorträge auf der 42. Tagung für Didaktik der Mathematik vom 13.03. bis 18.03.2008 in Budapest* (S. 62–65). Münster: Martin Stein Verlag.
- Blum, W. & Leiß, D. (2003). Diagnose- und Interventionsformen für einen selbstständigkeitsorientierten Unterricht am Beispiel Mathematik - Vorstellung des Projekts DISUM. In H.-W. Henn (Hrsg.), *Beiträge zum Mathematikunterricht 2003. Vorträge auf der 37. Tagung für Didaktik der Mathematik vom 3. bis 7. März 2003 in Dortmund* (S. 129–132). Hildesheim und Berlin: Franzbecker.
- Blum, W. & Leiß, D. (2005). Modellieren im Unterricht mit der „Tanken“-Aufgabe. *mathematik lehren*, 128, 18–21.
- Blum, W. & Leiß, D. (2006). „Filling up“ - the problem of independence-preserving teacher interventions on lessons with demanding modelling tasks. In M. Bosch (Hrsg.), *European Research in Mathematics Education IV. Proceedings of the fourth congress of the European Society for Research in Mathematics Education, Sant Feliu de Guíxols, Spain - 17-21 February 2005* (S. 1623–1633). Universitat Ramon Llull.
- Blum, W. & Leiß, D. (2007a). How do students and teachers deal with modelling problems? The Example „Sugarloaf“ and the DISUM Project. In C. Haines, P. Galbraith, W. Blum & S. Khan (Hrsg.), *Mathematical modelling. Education, engineering and economics - ICTMA 12* (S. 222–231). San Diego, CA, USA: Elsevier Science.
- Blum, W. & Leiß, D. (2007b). Investigating quality mathematics teaching - the DISUM project. In C. Bergsten & B. Grevholm (Hrsg.), *Developing and researching quality in mathematics teaching and learning. Proceedings of MADIF-5* (S. 3–16). Linköping: SMDF.
- Blum, W. & Neubrand, M. (Hrsg.). (1998). *TIMSS und der Mathematikunterricht: Informationen, Analysen, Konsequenzen*. Hannover: Schroedel.
- Blum, W., Vogel, S., Drüke-Noe, C. & Roppelt, A. (Hrsg.). (2015). *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II* (Bildungsstandards aktuell). Braunschweig: Schroedel.

- Boekaerts, M. & Minnaert, A. (1999). Self-regulation with respect to informal learning. *International Journal of Educational Research*, 31(6), 533–544.
- Bong, M. & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15(1), 1–40.
- Bonsen, M. (2010). Die Bedeutung der Schulleitung für die Schulentwicklung. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 199–203). Stuttgart: UTB.
- Bonsen, M., Büchter, A. & Peek, R. (2006). Datengestützte Schul- und Unterrichtsentwicklung. In W. Bos, H. Pfeiffer, H.-G. Rolff, R. Schulz-Zander & H. G. Holtappels (Hrsg.), *Jahrbuch der Schulentwicklung Band 14. Daten, Beispiele und Perspektiven* (S. 125–148). Weinheim [u.a.]: Juventa Verlag.
- Bonsen, M. & Frey, K. A. (2014). Lernen im Kontext des eigenen Unterrichts. Von der Fortbildungsforschung zu neuen Formen wirksamer Lehrerprofessionalisierung. *Lernende Schule*, 17(68), 13–15.
- Bonsen, M. & Rolff, H.-G. (2006). Professionelle Lerngemeinschaften von Lehrerinnen und Lehrern. *Zeitschrift für Pädagogik*, 52(2), 167–184.
- Borich, G. D. (2007). *Effective teaching methods*. Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.
- Borich, G. D. & Martin, D. B. (2008). *Observation skills for effective teaching*. Upper Saddle River, N.J.: Pearson Merrill Prentice Hall.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. Berlin: Springer.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation. Für Human- und Sozialwissenschaftler*. Heidelberg: Springer.
- Bortz, J., Lienert, G. A. & Boehnke, K. (2008). *Verteilungsfreie Methoden in der Biostatistik*. Heidelberg: Springer.
- Bosch, S. (2006). *Lineare Algebra* (Springer-Lehrbuch, 3. Aufl.). Berlin: Springer.
- Böttcher, W. & Keune, M. (2011). Wirkungsanalyse der Schulinspektionen. Empirische Ergebnisse aus Deutschland. In C. Quesel, V. Husfeldt, N. Landwehr & P. Steiner (Hrsg.), *Wirkungen und Wirksamkeit der externen Schulevaluation* (S. 123–143). Bern: Hep.
- Böttcher, W. & Rürup, M. (2010). Landesspezifische Steuerungskonzepte. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 56–62). Stuttgart: UTB.

- Brackbill, Y., Bravos, A. & Starr, R. H. (1962). Delay-improved retention of a difficult task. *Comparative and Physiological Psychology*, 55(6), 947–952.
- Brackbill, Y. & Lintz, L. M. (1967). The delay-retention effect in children: Roles of reward and knowledge of result. *Journal of Psychology*, 66(1), 149–155.
- Bremerich-Vos, A., Groß-Ophoff, J., Helmke, A., Hosenfeld, I. & Wagner, S. (2005). Stellungnahme zur Kritik an VERA in „Grundschule aktuell“, Heft 89. *Grundschule Aktuell*, 90, 3–6.
- Brock, G., Pihur, V., Datta, S. & Datta, S. (2008). clValid. An R package for cluster validation. *Journal of Statistical Software*, 25(4). <https://doi.org/10.18637/jss.v025.i04>
- Bromme, R., Seeger, F. & Steinbring, H. (1990). *Aufgaben als Anforderungen an Lehrer und Schüler*. Köln: Aulis-Verlag Deubner.
- Brophy, J. E. & Good, T. L. (1984). *Teacher behavior and student achievement* (Occasional Paper Nr. 73). Washington, DC: Michigan State University; East Lansing. Inst. for Research on Teaching.
- Bruder, R. (2003). Konstruieren - auswählen - begleiten. Über den Umgang mit Aufgaben. *Friedrich Jahresheft XXI*, 12–15.
- Bruder, R. (2006). *Modul 1: Weiterentwicklung der Aufgabenkultur im Mathematikunterricht*. Zugriff am 08.09.2017. Verfügbar unter http://sinus-transfer.uni-bayreuth.de/fileadmin/MaterialienBT/Bruder_Modul1.pdf
- Bruder, R. (2012). Langfristiger Kompetenzaufbau. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM* (6. Auflage, S. 135–151). Berlin: Cornelsen Scriptor.
- Bruder, R. & Collet, C. (2011). *Problemlösen lernen im Mathematikunterricht*. Berlin: Cornelsen Scriptor.
- Brügelmann, H. (2005a). Thesen zu den Bildungsstandards. *Grundschule aktuell*, (90), 8.
- Brügelmann, H. (2005b). Wahrheit durch VERA? Anmerkungen zum ersten Durchgang der landesweiten Leistungstests in sieben Bundesländern. *Grundschule aktuell*, 89, 7–9.
- Brummelman, E., Thomaes, S., Orobio de Castro, B., Overbeek, G. & Bushman, B. J. (2014). „That’s not just beautiful - that’s incredibly beautiful!“. The adverse

- impact of inflated praise on children with low self-esteem. *Psychological Science*, 25(3), 728–735.
- Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T. et al. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? *Zeitschrift für Erziehungswissenschaft*, 9(4), 521–544.
- Büchter, A. & Leuders, T. (2005a). From students' achievement to the development of teaching: Requirement for feedback in comparative tests. *Zentralblatt der Didaktik der Mathematik (ZDM)*, 37(4), 324–336.
- Büchter, A. & Leuders, T. (2005b). Zentrale Tests und Unterrichtsentwicklung... ...bei guten Aufgaben und gehaltvollen Rückmeldungen kein Widerspruch. *Pädagogik*, 57(5), 14–18.
- Büchter, A. & Leuders, T. (2007). *Mathematikaufgaben selbst entwickeln. Lernen fördern - Leistung überprüfen* (3. Aufl.). Berlin: Cornelsen Scriptor.
- Büchter, A. & Leuders, T. (2008). Leistungen verstehensorientiert überprüfen. Gute Aufgaben für Klassenarbeiten entwickeln. In R. Bruder, T. Leuders & A. Büchter (Hrsg.), *Mathematikunterricht entwickeln. Bausteine für kompetenzorientiertes Unterrichten* (S. 155–191). Berlin: Cornelsen Scriptor.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson Studium.
- Bühner, M. & Ziegler, M. (2012). *Statistik für Psychologen und Sozialwissenschaftler*. München: Pearson Studium.
- Bundesministerium für Bildung und Forschung (Hrsg.). (2001). *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente*. Bonn. Zugriff am 18.09.2014. Verfügbar unter <http://www.ssg-bildung.ub.uni-erlangen.de/timss.pdf>
- Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung. (1997). *Gutachten zur Vorbereitung des Programms „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“*. Bonn.
- Burchard, A. (2014). *Lehrer wollen Vergleichstests abschaffen*. Zugriff am 04.08.2015. Verfügbar unter <http://www.tagesspiegel.de/wissen/schulvergleich-vera-lehrer-wollen-vergleichstests-abschaffen/9847244.html>
- Bürgermeister, A., Klimczak, M., Klieme, E., Rakoczy, K., Blum, W., Leiß, D. et al. (2011). Leistungsbeurteilung im Mathematikunterricht - Eine Darstellung des

- Projekts „Nutzung und Auswirkungen der Kompetenzmessung in mathematischen Lehr-Lernprozessen“. In A. Fächter & K. Moegling (Hrsg.), *Theorie und Praxis der Schulpädagogik* (S. 28–51). Immenhausen: Prolog-Verlag.
- Burkard, C. & Orth, G. (2005). Standardsetzung und Standardüberprüfung in Nordrhein-Westfalen. In J. Reus (Hrsg.), *Bildungsstandards, Kerncurricula und die Aufgabe der Schule* (S. 136–148). Münster: Aschendorff.
- Burmeister, E., Eckert, U., Kropp, A., Metscher, M., Schmidt, B., Schüttrumpf, R. et al. (Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien (ThILLM), Hrsg.). (2005). *Kompetenztests als Beitrag zur Schul- und Unterrichtsentwicklung*, Thüringer Institut für Lehrerfortbildung, Lehrplanentwicklung und Medien. Materialien: 115.
- Calinski, T. & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1), 1–27.
- Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1), 67–90.
- Cardelle-Elawar, M. (1995). Effects of metacognitive instruction on low achievers in mathematics problems. *Teaching and Teacher Education*, 11(1), 81–95.
- Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C.-P. & Loeef, M. (1989). Using knowledge of children's mathematics thinking in classroom teaching: An experimental study. *American Educational Research Journal*, 26(4), 499–531.
- Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6). <https://doi.org/10.18637/jss.v061.i06>
- Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. (2015). *Package 'NbClust'*, CRAN. Zugriff am 12.07.2017. Verfügbar unter <https://cran.r-project.org/web/packages/NbClust/NbClust.pdf>
- Christiansen, B. & Walther, G. (1986). Task and activity. In B. Christiansen, A. G. Howson & M. Otte (Hrsg.), *Perspectives on mathematics education* (S. 243–308). Dordrecht: D. Reidel Publishing Company.
- Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg & M. E. Sobel (Hrsg.), *Handbook of statistical modeling for the social and behavioral sciences* (S. 311–359). New York: Plenum Press.

- Cobb, P., Wood, T., Yackel, E., Nicholls, J., Wheatley, G., Trigatti, B. et al. (1991). Assessment of a problem-centered second-grade mathematics project. *Journal for Research in Mathematics Education*, 22(1), 3–29.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Cohen, J. (1988). *Statistical power Analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.
- Cohen, M. D., March, J. G. & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly*, 17(1), 1.
- Colquitt, J. A., LePine, J. A. & Noe, R. A. (2000). Toward an integrative theory of training motivation. A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85(5), 678–707.
- Conover, W. J. & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, 35(3), 124.
- Cookson, P. S. (1986). A framework for theory and research on adult education participation. *Adult Education Quarterly*, 36(3), 130–141.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, 78(1), 98–104.
- Cortina, J. M. & Nouri, H. (2000). *Effect size for ANOVA design*. Thousand Oaks, London, New Dehli: SAGE Publications.
- Crawley, M. J. (2013). *The R book*. Chichester, West Sussex, U.K.: John Wiley & Sons.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- Cuban, L. (1984). Transforming the frog into a prince: Effective school research, policy and practice at the district level. *Harvard Educational Review*, 54(2), 129–151.
- Deci, E. L. (1972). Intrinsic motivation, extrinsic reinforcement and inequity. *Journal of Personality and Social Psychology*, 22(1), 113–120.
- Deci, E. L. & Ryan, R. M. (1993). Die Selbstbestimmungstheorie der Motivation und ihre Bedeutung für die Pädagogik. *Zeitschrift für Pädagogik*, 39(2), 223–239.
- Deci, E. L. & Ryan, R. M. (Hrsg.). (2004). *Handbook of self-determination research*. Rochester, N.Y.: University of Rochester Press.
- Dedering, K. (2011). Hat Feedback eine positive Wirkung? Zur Verarbeitung extern erhobener Leistungsdaten in Schulen. *Unterrichtswissenschaften*, 39(1), 63–83.

- Demmer, M. & Schweitzer, J. (2005). Es fährt ein Zug nach nirgendwo... Zwischenbilanz einer unaufhaltsamen (?) Entwicklung. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 68–69). Seelze: Friedrich Verlag.
- Desimone, L. M. (2006). Are teachers who need sustained, content-focused professional development getting it? An administrator's dilemma. *Educational Administration Quarterly*, 42(2), 179–215.
- Desimone, L. M. (2009). Improving impact studies of teachers' professional development. Toward better conceptualizations and measures. *Educational Researcher*, 38(3), 181–199.
- Diel, E. (Dezember 2011). *Hessischer Referenzrahmen Schulqualität. Qualitätsbereiche, Qualitätsdimensionen und Qualitätskriterien* (Institut für Qualitätsentwicklung (IQ), Hrsg.). Hattersheim.
- Ditton, H. (2000). Qualitätskontrolle und -sicherung in Schule und Unterricht - ein Überblick zum Stand der empirischen Forschung. In A. Helmke, W. Hornstein & E. Terhart (Hrsg.), *Qualität und Qualitätssicherung im Bildungsbereich. Schule, Sozialpädagogik, Hochschule* (S. 73–92). Weinheim: Beltz.
- Ditton, H. & Arnoldt, B. (2004). Schülerbefragungen zum Fachunterricht - Feedback an Lehrkräfte. *Empirische Pädagogik*, 18(18), 115–139.
- Dobbelstein, P. & Peek, R. (2008). Lernstandserhebungen und zentrale Prüfungen im Kontext der Qualitätsanalyse. In S. Müller, K. Dederich & W. Bos (Hrsg.), *Schulische Qualitätsanalyse in Nordrhein-Westfalen. Konzepte, erste Erfahrungen, Perspektiven* (S. 17–35). Kronach: Link Luchterhand.
- Drücke-Noe, C. (2011). *Basiskompetenzen Mathematik für Alltag und Berufseinstieg am Ende der allgemeinen Schulpflicht*. Berlin: Cornelsen.
- Drücke-Noe, C. (2012). Basiskompetenzen. Was sollte jeder am Ende der allgemeinen Schulpflicht in Mathematik können? In M. Ludwig (Hrsg.), *Beiträge zum Mathematikunterricht 2012. Vorträge auf der 46. Tagung für Didaktik der Mathematik vom 05.03.2012 bis 09.03.2012 in Weingarten Bd. 1* (209–121). Münster: WTM, Verl. für wiss. Texte u. Medien.
- Drücke-Noe, C. (2014). *Aufgabenkultur in Klassenarbeiten im Fach Mathematik. Empirische Untersuchungen in neunten und zehnten Klassen* (Perspektiven der Mathematikdidaktik). Wiesbaden: Springer Spektrum.

- Dubberke, T., Kunter, M., McElvany, N., Brunner, M. & Baumert, J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften: Einflüsse auf die Unterrichtsgestaltung und den Lernerfolg von Schülerinnen und Schülern. *Zeitschrift für Pädagogische Psychologie*, 22(3), 193–206.
- Duden. (2016). *null-acht-fünf-zehn*. Zugriff am 24.09.2016. Verfügbar unter http://www.duden.de/rechtschreibung/_nullachtfuenfzehn
- Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, 4(1), 95–104.
- Ehren, M. C. M. & Visscher, A. J. (2006). Towards a theory on the impact of school inspections. *British Journal of Educational Studies*, 54(1), 51–72.
- Eid, M., Gollwitzer, M. & Schmitt, M. (2011). *Statistik und Forschungsmethoden*. Weinheim: Beltz.
- Evertson, C. M., Emmer, E. T. & Worsham, M. E. (2006). *Classroom management for elementary teachers*. Boston, MA.: Pearson Education.
- Eye, A. von. (1998). Configural Frequency Analysis program for 32-bit operation system. *Methods of Psychological Research Online*, 3(1). Zugriff am 15.06.2016. Verfügbar unter <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue4/art1/voneye.pdf>
- Eye, A. von. (2001). Configural Frequency Analysis - version 2000. A program for 32 Bit Windows operating systems. *Methods of Psychological Research Online*, 6(2). Zugriff am 15.06.2016. Verfügbar unter <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue14/art1/cfa.pdf>
- Eye, A. von & Brandstädter, J. (1997). Configural Frequency Analysis as a searching devise for possible causal relationships. *Methods of Psychological Research Online*, 2(2). Zugriff am 15.06.2016. Verfügbar unter <http://www.dgps.de/fachgruppen/methoden/mpr-online/issue3/art3/voneye.pdf>
- Fend, H. (1986). Gute Schulen - schlechte Schulen. Die einzelne Schule als pädagogische Handlungseinheit. *Die Deutsche Schule*, 78(3), 275–293.
- Fend, H. (2008). *Schule gestalten. Systemsteuerung, Schulentwicklung und Unterrichtsqualität*. Wiesbaden: VS, Verlag für Sozialwissenschaften.
- Fitz-Gibbon, C. (1996). Monitoring school effectiveness: Simplicity and complexity. In J. Gray (Hrsg.), *Merging traditions. The future of research on school effectiveness and school improvement* (S. 74–92). London: Cassell.

- Fleiss, J. L. & Cohen, J. (1973). The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3), 613–619.
- Forsa Politik- und Sozialforschung GmbH. (2017). *Qualität der MINT-Lehrerfortbildung in Deutschland. Ergebnisse einer repräsentativen Befragung von MINT-Lehrern*, Deutsche Telekom Stiftung. Zugriff am 31.05.2017. Verfügbar unter https://www.telekom-stiftung.de/sites/default/files//dts-library/materialien/pdf/umfrage_qualitaet_mint-lehrerfortbildung_ergebnisbericht.pdf
- Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S. et al. (2016). *Package 'car'. Companion to applied regression*. Zugriff am 05.06.2017. Verfügbar unter <https://cran.r-project.org/web/packages/car/car.pdf>
- Frindte, W. (2001). *Einführung in die Kommunikationspsychologie* (Beltz-Studium). Weinheim: Beltz.
- Furrer, C. & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95(1), 148–162.
- Gage, N. L., Berliner, D. C. & Bach, G. (1996). *Pädagogische Psychologie*. Weinheim: Psychologische Verlags-Union.
- Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W. et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement*. Zugriff am 09.03.2017. Verfügbar unter <http://files.eric.ed.gov/fulltext/ED502700.pdf>
- Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S. & Santoro, L. E. (2010). Teacher study group. Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal*, 47(3), 694–739.
- Gesellschaft für Evaluation e.V. (2002). *Standards für Evaluation*. Köln: Gesellschaft für Evaluation e.V. (DeGEval). Zugriff am 29.10.2016. Verfügbar unter http://www.degeval.de/images/stories/Publikationen/DeGEval_-_Standards.pdf
- Gibbons, J. D. (1993). *Nonparametric statistics. An introduction*. Newbury Park, London, New Dehli: SAGE Publications.
- Girden, E. R. (1992). *ANOVA. Repeated measures*. Newbury Park, London, New Dehli.
- Girmes, R. (2003). Die Welt als Aufgaben?! *Friedrich Jahresheft*, 6–11.

- Goddar, J. (Süddeutsche Zeitung, Hrsg.). (2010). *Bildungsreform – Einmal Lehrer, immer Lehrer*. Zugriff am 23.11.2016. Verfügbar unter <http://www.sueddeutsche.de/karriere/bildungsreform-einmal-lehrer-immer-lehrer-1.544740>
- Goodenow, C. (1993). Classroom belonging among early adolescent students. Relationships to motivation and achievement. *The Journal of Early Adolescence*, 13(1), 21–43.
- Grabensberger, E., Freudenthaler, H. H. & Specht, W. (2008). *Bildungsstandards: Testungen und Ergebnismeldungen auf der achten Schulstufe aus der Sicht der Praxis* (BIFIE Report). Graz: Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens.
- Gräsel, C., Fußangel, K. & Pröbstel, C. (2006). Lehrkräfte zur Kooperation anregen - eine Aufgabe für Sisypchos. *Zeitschrift für Pädagogik*, 52(2), 205–219.
- Gross, J. & Ligges, U. (2015). *Package ‚nortest‘. Tests for normality*, CRAN. Zugriff am 05.06.2017. Verfügbar unter <https://cran.r-project.org/web/packages/nortest/nortest.pdf>
- Groß-Ophoff, J., Hosenfeld, I. & Koch, U. (2007). Formen der Ergebnisrezeption und damit verbundene Schul- und Unterrichtsentwicklung. *Empirische Pädagogik*, 21(4), 411–427.
- Groß-Ophoff, J. & Koch, U. (2007a). *Erfahrungen aus Schulrückmeldungen im Projekt VERA*, Universität Koblenz-Landau. Zugriff am 16.10.2014. Verfügbar unter http://www.emse-netzwerk.de/uploads/Main/EMSE07_Gross_Ophoff_Koch_VERA.pdf
- Groß-Ophoff, J. & Koch, U. (2007b). *Erfahrungen aus Schulrückmeldungen im Projekt VERA. Vortrag auf der 7. Tagung des Netzwerk Empiriegestützte Schulentwicklung (EMSE) am 06./07.12.2007 in Mainz*. Mainz. Zugriff am 28.03.2017. Verfügbar unter https://www.emse-netzwerk.de/uploads/Main/EMSE07_Gross_Ophoff_Koch_VERA.pdf
- Groß-Ophoff, J., Koch, U., Hosenfeld, I. & Helmke, A. (2006). Ergebnismeldung und ihre Rezeption im Projekt VERA. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 19–40). Münster: Waxmann.
- Gruehn, S. (2000). *Unterricht und schulisches Lernen. Schüler als Quellen der Unterrichtsbeschreibung*. Münster: Waxmann.

- Gruschka, A., Herrmann, U., Radtke, F.-O., Rauin, U., Ruhloff, J., Rumpf, H. et al. (2005). *Frankfurter Erklärung*. Zugriff am 16.09.2014. Verfügbar unter <https://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Sozialwissenschaften/BF/Lehre/WiSe0910/HK/frankfurtererklrung.pdf>
- Haag, L. (2012). *Klassenführung. Erfolgreich unterrichten mit Classroom Management*. Weinheim, Bergstr: Beltz.
- Hagenaars, J. A. & McCutcheon, A. L. (Hrsg.). (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press.
- Hammer, S. & Ufer, S. (2015). Aufgaben in einem kompetenzorientierten Mathematikunterricht. In W. Blum, S. Vogel, C. Drüke-Noe & A. Roppelt (Hrsg.), *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II* (S. 160–170). Braunschweig: Schroedel.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M. & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94(3), 562–575.
- Hardy, I., Jonen, A., Möller, K. & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of „Floating and Sinking.“. *Journal of Educational Psychology*, 98(2), 307–326.
- Harks, B., Rakoczy, K., Hattie, J., Besser, M. & Klieme, E. (2013). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology*, 34(3), 1–22.
- Harris, D. & Sass, T. R. (2006). *The effect of teacher training on teacher value-added*. Florida State University. Zugriff am 09.03.2017. Verfügbar unter ftp://econpapers.fsu.edu/RePEc/backups/wpaper/wp2006_03_01.pdf
- Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London: Routledge.
- Hattie, J., Beywl, W. & Zierer, K. (2013). *Lernen sichtbar machen. Überarbeitete deutschsprachige Ausgabe von „Visible Learning“*. Baltmannsweiler: Schneider Hohengehren.
- Hattie, J., Biggs, J. & Purdie, N. (1996). Effects of learning skills interventions on student learning. A meta-analysis. *Review of Educational Research*, 66(2), 99–136.

- Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hatzinger, R., Hornik, K., Nagel, H. & Maier, M. J. (2014). *R. Einführung durch angewandte Statistik*. Hallbergmoos: Pearson.
- Hecht, M., Roppelt, A. & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs. 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. 391–402). Münster: Waxmann.
- Heine, J.-H. (2016). *Package ‚confreq‘. Configural Frequencies Analysis using log-linear modeling*. Zugriff am 10.05.2017. Verfügbar unter <https://cran.r-project.org/web/packages/confreq/confreq.pdf>
- Helmke, A. (2004). Von der Evaluation zur Innovation. Pädagogische Nutzbarmachung von Vergleichsarbeiten in der Grundschule. *Seminar*, 2, 90–112.
- Helmke, A. (2009). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze-Velber: Kallmeyer.
- Helmke, A., Helmke, T., Heyne, N., Hosenfeld, A., Kleinbub, I., Schrader, F.-W. et al. (2007). Erfassung, Bewertung und Verbesserung des Grundschulunterrichts: Forschungsstand, Probleme und Perspektiven. In K. Möller (Hrsg.), *Qualität von Grundschulunterricht. Entwickeln, erfassen und bewerten* (S. 17–34). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Helmke, A. & Hosenfeld, I. (2003). Vergleichsarbeiten (VERA): eine Standortbestimmung zur Sicherung schulischer Kompetenzen. Teil I: Ziele, Konzepte und Organisation. *Schulverwaltung HRS*, (1), 10–13.
- Helmke, A. & Hosenfeld, I. (2005). Leistungsmessung in der Grundschule. Vorbereitung und Nutzung von Vergleichs- und Orientierungsarbeiten. *klexer*, 14, 3–5.
- Helmke, A., Hosenfeld, I. & Schrader, F.-W. (2003). Diagnosekompetenz in Ausbildung und Beruf entwickeln. *Karlsruher Pädagogische Beiträge*, 55, 15–34.
- Helmke, A. & Weinert, F. E. (1989). The impact of instructional quality on student cognitive motivational development. The result of the munich study. *Reprint/Max-Planck-Institut für Psychologische Forschung*, 26, 132–141.
- Hentig, H. v. (1994). *Die Schule neu denken. Eine Übung in praktischer Vernunft; eine zornige, aber nicht eifernde, eine radikale, aber nicht utopische Antwort auf Hoyerswerda und Mölln, Rostock und Solingen*. München: Hanser.

- Herzog, W. (2006). *Bildungsstandards - Aussichten einer Selbstverständlichkeit*. Zugriff am 03.03.2015. Verfügbar unter http://www.forum-allgemeinbildung.ch/files/Bildungsstandards_-_Aussichten_einer_Selbstverstaendlichkeit.pdf
- Hessisches Kultusministerium. (Hrsg.) (2009). *Lernstandserhebungen*. Wiesbaden: Hessisches Kultusministerium. Zugriff am 07.05.2014. Verfügbar unter http://qualitaetsentwicklung.lsa.hessen.de/irj/servlet/prt/portal/prt-root/slimp.CMReader/HKM_15/IQ_Internet/med/f99/f991bc68-18e4-931f-012f-312b417c0cf4,22222222-2222-2222-2222-222222222222
- Hessisches Kultusministerium (Hrsg.). (2011). *Amtsblatt des Hessisches Kultusministeriums 8/11*, Hessisches Kultusministerium. 8/11. Zugriff am 01.10.2014. Verfügbar unter http://www.hessisches-amtsblatt.de/download/pdf_2011/alle_user/08_2011.pdf
- Hessisches Kultusministerium. (2016). Hessisches Schulgesetz. HSchG. Zugriff am 24.01.2017. Verfügbar unter http://www.rv.hessenrecht.hessen.de/lexsoft/default/hessenrecht_rv.html?p1=0&eventSubmit_doNavigate=searchInSubtreeTOC&showdoc-case=1&doc.hl=0&doc.id=jlr-SchulGHE2005rahmen&doc.part=R&toc.pos-key=#docid:169561,1,20150401
- Hessisches Kultusministerium & Institut für Qualitätsentwicklung. (2011). *Hessischer Referenzrahmen Schulqualität. Qualitätsbereiche, Qualitätsdimensionen und Qualitätskriterien*. Wiesbaden: Hessisches Kultusministerium; Institut für Qualitätsentwicklung (IQ).
- Heymann, K. (2005). Lernstandserhebungen. Erste Erfahrungen im Umgang mit einem neuen Instrument. *Pädagogik*, 57(5), 10–13.
- Hiebert, J. & Wearne, D. (1993). Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic. *American Educational Research Journal*, 30(2), 393–425.
- Hoffmann, L. & Richter, D. (2016). Aspekte der Aus- und Fortbildung von Deutsch- und Englischlehrkräften im Ländervergleich. In P. Stanat, K. Böhme, S. Schipolowski & N. Haag (Hrsg.), *IQB-Bildungstrend 2015. Sprachliche Kompetenzen am Ende der 9. Jahrgangsstufe im zweiten Ländervergleich* (S. 481–508). Münster: Waxmann.
- Hofmann-Göttig, J., Eschmann, W. & Daumen, C. (2005). Und sie bewegt sich doch... Vom Umgang mit den Ergebnissen externer Evaluation aus der Sicht von

- Bildungspolitik und Schulaufsicht. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 32–36). Seelze: Friedrich Verlag.
- Holtappels, H. G. (Hrsg.). (2014a). *Schulentwicklung und Schulwirksamkeit als Forschungsfeld. Theorieansätze und Forschungserkenntnisse zum schulischen Wandel*. Münster: Waxmann.
- Holtappels, H. G. (2014b). Schulentwicklung und Schulwirksamkeit. Erkenntnisse aus der Perspektive von Schulentwicklungstheorie und -forschung. In H. G. Holtappels (Hrsg.), *Schulentwicklung und Schulwirksamkeit als Forschungsfeld. Theorieansätze und Forschungserkenntnisse zum schulischen Wandel* (S. 11–47). Münster: Waxmann.
- Holtappels, H. G. & Feldhoff, T. (2010). Einführung: Change Management. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 159–166). Stuttgart: UTB.
- Holtappels, H. G. & Rolff, H.-G. (2010). Einführung: Theorien der Schulentwicklung. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 73–79). Stuttgart: UTB.
- Hord, S. M. (2005). Professional learning communities: An overview. In S. M. Hord (Hrsg.), *Learning together, leading together. Changing schools through professional learning communities* (S. 5–14). New York: Teachers College Press.
- Horster, L. & Rolff, H.-G. (2001). *Unterrichtsentwicklung. Grundlagen, Praxis, Steuerungsprozesse*. Weinheim: Beltz.
- Hosenfeld, A. (2010). *Führt Unterrichtsrückmeldung zu Unterrichtsentwicklung? Die Wirkung von videographischer und schriftlicher Rückmeldung bei Lehrkräften der vierten Jahrgangsstufe*. Münster: Waxmann.
- Hosenfeld, I. (2005). Rezeption - Reflexion - Aktion. Wie lassen sich Lernstandserhebungen und Vergleichsarbeiten pädagogisch nutzen? In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 112–114). Seelze: Friedrich Verlag.
- Hosenfeld, I., Groß-Ophoff, J. & Bittins, P. (2006). *Vergleichsarbeiten und Schulentwicklung. [Fakten, Gründe, Praxistipps]* (Schulmanagement-Handbuch, Bd. 118). München: Oldenbourg.

- Hosenfeld, I., Koch, U., Groß-Ophoff, J. & Scherthan, F. (2008). Projekt VERA: Ergebnisorientierte Unterrichtsentwicklung durch internetgestützte externe Evaluation? In A. Breiter, A. Lange & E. Stauke (Hrsg.), *School information systems and data-based decision-making* (S. 49–60). Frankfurt am Main: P. Lang.
- Hothorn, T., Hornik, K., van de Wiel, Mark, A., Winell, H. & Zeileis, A. (2017). *Package 'coin'. Conditional inference procedures in a permutation test framework*, CRAN. Zugriff am 03.07.2017. Verfügbar unter <https://cran.r-project.org/web/packages/coin/coin.pdf>
- Hothorn, T., Hornik, K., van de Wiel, Mark, A. & Zeileis, A. (2008). Implementing a class of permutation tests. The coin Package. *Journal of Statistical Software*, 28(8).
- Ingenkamp, K. (1985). *Lehrbuch der pädagogischen Diagnostik*. Weinheim: Beltz.
- Ingenkamp, K. (1995). Pädagogische Diagnostik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik. Ein Lehrbuch* (S. 495–510). Weinheim: Beltz, PsychologieVerlagsUnion.
- Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik*. Weinheim: Beltz.
- Institut für Sozialwissenschaftliche Studien (ISS) e.V. & Gewerkschaft Erziehung und Wissenschaft (GEW). (2009). *Wirksame Lehr- und Lernumgebungen schaffen. Erste Ergebnisse von TALIS-GEW (Deutschland)*, Institut für Sozialwissenschaftliche Studien (ISS) e.V.; Gewerkschaft Erziehung und Wissenschaft (GEW). Zugriff am 16.10.2014. Verfügbar unter http://www.gew.de/Binaries/Binary50253/3_GEW_First_Results_deck.pdf
- Isaac, K., Halt, A. C., Hosenfeld, I., Helmke, A. & Groß-Ophoff, J. (2006). VERA: Qualitätsentwicklung und Lehrerprofession durch Vergleichsarbeiten. Kommentar zur Diskussion um VERA. *Die Deutsche Schule. Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis*, 98(1), 107–111.
- Jäger, R. S. (1995). Der diagnostische Prozeß. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik. Ein Lehrbuch* (S. 450–455). Weinheim: Beltz, PsychologieVerlagsUnion.
- Jahnke, T. & Meyerhöfer, W. (Hrsg.). (2008). *PISA & Co. Kritik eines Programms*. Hildesheim: Franzbecker.

- Jerald, C. D. (2006). „*Teach to the test? Just say no.*“, Center for Comprehensive School Reform and Improvement. Zugriff am 03.01.2017. Verfügbar unter <http://files.eric.ed.gov/fulltext/ED494086.pdf>
- Jerusalem, M. & Hopf, D. (Hrsg.). (2002). *Selbstwirksamkeit und Motivationsprozesse in Bildungsinstitutionen*. Zugriff am 24.01.2017. Verfügbar unter http://www.pe-docs.de/volltexte/2013/7863/pdf/ZfPaed_44.Beiheft.pdf
- Jordan, A., Ross, N., Krauss, S., Baumert, J., Blum, W., Neubrand, M. et al. (2006). *Klassifikationsschema für Mathematikaufgaben: Dokumentation der Aufgabekategorisierung im COACTIV-Projekt*. Berlin: Max-Planck-Institut für Bildungsforschung.
- Kaiser, G. & Stender, P. (2015). Die Kompetenz mathematisch Modellieren. In W. Blum, S. Vogel, C. Drüke-Noe & A. Roppelt (Hrsg.), *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II* (S. 95–106). Braunschweig: Schroedel.
- Kao, C.-P., Wu, Y.-T. & Tsai, C.-C. (2011). Elementary school teachers' motivation toward web-based professional development, and the relationship with internet self-efficacy and belief about web-based learning. *Teaching and Teacher Education*, 27(2), 406–415.
- Kast, F. (2010). Fortbildungsbedarf: Disparitäten in Abhängigkeit von Schulart, Alter und Geschlecht der Lehrer/innen. In J. Schmich & C. Schreiner (Hrsg.), *Talis 2008: Schule als Lernumfeld und Arbeitsplatz. Vertiefende Analysen aus österreichischer Perspektive* (S. 27–50). Graz: Leykam.
- Katzenbach, M. (2011a). Unterstützung der Lehrkräfte bei der individuellen Förderung. Anregungen aus einem neuseeländischen Projekt. *SchulVerwaltung NRW*, 7-8, 194–197. Zugriff am 03.12.2014. Verfügbar unter http://mediathek.bildung.hessen.de/material/themen/dia_foe/mathematik/Katzenbach_nrw11-07.pdf
- Katzenbach, M. (2011b). *Vom Interview zur Förderung. Beispielmaterialien aus dem neuseeländischen Numeracy Project*. Zugriff am 03.12.2014. Verfügbar unter http://mediathek.bildung.hessen.de/material/themen/dia_foe/mathematik/mathematik_5_10/Ma5-10_17_Katzenbach.pdf
- Katzenbach, M., Köller, O., Roppelt, A., Blum, W., Drüke-Noe, C., Keller, K. et al. (2009a). *Bildungsstandards: Kompetenzen überprüfen Mathematik (Heft 1). Sekundarstufe I*. Berlin: Cornelsen.

- Katzenbach, M., Köller, O., Roppelt, A., Blum, W., Drüke-Noe, C., Keller, K. et al. (2009b). *Bildungsstandards: Kompetenzen überprüfen Mathematik (Heft 2). Sekundarstufe I*. Berlin: Cornelsen.
- Kaufman, L. & Rousseeuw, P. J. (1990). *Finding groups in data. An introduction to cluster analysis*. New York: Wiley.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B. et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386.
- Kiefer, T., Robitzsch, A. & Wu, M. (2013, 7. Juli). *TAM Tutorials*. Zugriff am 30.09.2016. Verfügbar unter <http://www.edmeasurementsurveys.com/TAM/Tutorials/index.htm>
- Kiefer, T., Robitzsch, A. & Wu, M. (2016, 31. Mai). *Package ‚TAM‘. Test Analysis Modules*. Zugriff am 12.06.2016. Verfügbar unter <http://www.edmeasurementsurveys.com/TAM/Tutorials/>
- Kiel, E., Frey, A. & Weiß, S. (2013). *Trainingsbuch Klassenführung*. Bad Heilbrunn: UTB.
- Klein, H. P. & Jahnke, T. (2012). Die Folgen der Kompetenzorientierung im Fach Mathematik. *Journal der Didaktik der Biowissenschaften*, 2, 1–9.
- Kleine, M. (2005). Latent-Class-Analyse. Ein Bindeglied zwischen Empirie und Theorie zur quantitativen Erfassung mathematischer Leistungen. *Journal für Mathematik-Didaktik*, 26(2), 97–113.
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Expertise*. Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2009). *Zur Entwicklung nationaler Bildungsstandards. Expertise*. Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Klieme, E. & Baumert, J. (2001). TIMSS als Startpunkt für Qualitätssicherung und Qualitätsentwicklung im Bildungswesen. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbe-funde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 5–9). Bonn.
- 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-Nägel

- (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 127–146). Münster: Waxmann.
- Klieme, E., Schlümer, G. & Knoll, S. (2001). Mathematikunterricht in der Sekundarstufe I: „Aufgabenkultur“ und Unterrichtsgestaltung. In Bundesministerium für Bildung und Forschung (BMBF) (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 43–57). Bonn.
- Klippert, H. (1997). Schule entwickeln - Unterricht neu gestalten. *Pädagogik*, (2), 12–17.
- Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary Feedback Intervention Theory. *Psychological Bulletin*, 119(2), 254–284.
- Klusmann, U. (2011). Allgemeine berufliche Motivation und Selbstregulation. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 277–294). Münster: Waxmann.
- Klusmann, U., Kunter, M., Trautwein, U. & Baumert, J. (2006). Lehrerbelastung und Unterrichtsqualität aus der Perspektive von Lehrenden und Lernenden. *Zeitschrift für Pädagogische Psychologie*, 20(3), 161–173.
- KMK (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). (1997). *Grundsätzliche Überlegungen zu Leistungsvergleichen innerhalb der Bundesrepublik Deutschland. Konstanzer Beschluss*, Kultusministerkonferenz. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/1997/1997_10_24-Konstanzer-Beschluss.pdf
- KMK. (2001). *Kultusministerkonferenz erzielt Einigung mit Lehrerverbänden über Konsequenzen aus der PISA-Studie*, KMK. Zugriff am 22.09.2014. Verfügbar unter <http://www.kmk.org/presse-und-aktuelles/pm2001/einigung-mit-lehrerverbaenden.html>
- KMK. (2002a). *298. Plenarsitzung der Kultusministerkonferenz am 23. und 24. Mai 2002 in Eisenach*. Zugriff am 15.09.2016. Verfügbar unter <https://www.kmk.org/presse/pressearchiv/mitteilung/298-plenarsitzung-der-kultusministerkonferenz-am-23-und-24mai-2002-in-eisenach.html>

- (KMK, Hrsg.). (2002b). *Bildungsstandards zur Sicherung von Qualität und Innovation im föderalen Wettbewerb der Länder*, KMK. Erklärungen und Stellungnahmen der Kultusministerkonferenz. Zugriff am 22.09.2014. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2002/2002_05_24-Bildungsstandards-Qualitaet.pdf
- KMK. (2002c). *PISA 2000 - Zentrale Handlungsfelder. Zusammenfassende Darstellung der laufenden und geplanten Maßnahmen in den Ländern*, KMK. Beschluss der Kultusministerkonferenz: 299. Zugriff am 22.09.2014. Verfügbar unter <http://www.kmk.org/fileadmin/pdf/PresseUndAktuelles/2002/massnahmen.pdf>
- KMK. (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Bildungsabschluss. Beschluss vom 4.12.2003* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.) (Beschlüsse der Kultusministerkonferenz). München: Kultusministerkonferenz. Zugriff am 29.04.2014. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mathe-Mittleren-SA.pdf
- KMK. (2004a). *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss. Beschluss vom 15.10.2004* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. Zugriff am 16.07.2014. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Haupt.pdf
- KMK. (2004b). *Bildungsstandards im Fach Mathematik für den Primarbereich. Beschluss vom 15.10.2004* (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. Zugriff am 15.09.2016. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_10_15-Bildungsstandards-Mathe-Primar.pdf
- KMK. (2004c, 1. Januar). *Standards für die Lehrerbildung: Bildungswissenschaften (Beschluss der Kultusministerkonferenz vom 16.12.2004)*. Sekretariat der Ständigen Konferenz der Kultusminister in der Bundesrepublik Deutschland. Zugriff am 22.09.2014. Verfügbar unter <http://www.kmk.org/>

- fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrer-
erziehung.pdf
- KMK. (2010). *Konzeption der Kultusministerkonferenz zur Nutzung der Bildungsstandards für die Unterrichtsentwicklung*. Köln: Wolters Kluwer. Zugriff am 14.04.2014. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2010/2010_00_00-Konzeption-Bildungsstandards.pdf
- KMK. (2012a). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife. Beschluss der Kultusministerkonferenz vom 18.10.2012*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). Zugriff am 15.09.2016. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2012/2012_10_18-Bildungsstandards-Mathe-Abi.pdf
- KMK. (2012b). *Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Mathematik*. Stand: 15. Februar 2012 (Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland, Hrsg.). Zugriff am 26.05.2014. Verfügbar unter https://www.iqb.hu-berlin.de/bista/ksm/Kompetenzstufenm_1.pdf
- KMK. (2012c, 8. März). *Medienbildung in der Schule. Beschluss der Kultusministerkonferenz vom 8. März 2012*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.) (Beschlüsse der Kultusministerkonferenz). Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland.
- KMK. (2015). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring* (Beschluss der Kultusministerkonferenz, Bd. 350.). Berlin. Zugriff am 17.12.2015. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2015/2015_06_11-Gesamtstrategie-Bildungsmonitoring.pdf
- KMK. (2016a). *Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt“*. Beschluss der Kultusministerkonferenz vom 08.12.2016. Zugriff am 03.02.2017. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/pdf/PresseUndAktuelles/2016/Bildung_digitale_Welt_Webversion.pdf
- KMK. (2016b, 16. Juni). *Vereinbarung zur Gestaltung der gymnasialen Oberstufe in der Sekundarstufe II (Beschluss der KMK vom 07.07.1972 i.d.F. vom 16.06.2016)*. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Hrsg.). Verfügbar unter

- http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1972/1972_07_07-Vereinbarung-Gestaltung-Sek2.pdf
- KMK & IQB. (2006). *Gesamtstrategie der Kultusministerkonferenz zum Bildungsmonitoring*. München: Wolters Kluwer. Zugriff am 08.03.2013. Verfügbar unter http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2006/2006_08_01-Gesamtstrategie-Bildungsmonitoring.pdf
- Koch, U. (2011). *Verstehen Lehrkräfte Rückmeldungen aus Vergleichsarbeiten?: Datenkompetenz von Lehrkräften und die Nutzung von Ergebnissrückmeldungen aus Vergleichsarbeiten*. Münster: Waxmann.
- Koch, U., Groß-Ophoff, J., Hosenfeld, I. & Helmke, A. (2006). Qualitätssicherung: Von der Evaluation zur Schul- und Unterrichtsentwicklung - Ergebnisse der Lehrerbefragung zur Auseinandersetzung mit den VERA-Rückmeldungen. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 187–199). Münster: Waxmann.
- Kohler, B. & Schrader, F.-W. (2004). Ergebnissrückmeldung und Rezeption: Von der externen Evaluation zur Entwicklung von Schule und Unterricht. *Empirische Pädagogik*, 18(1), 3–17.
- Korngiebel, J. (2009). *Kompetenztests in der Sekundarstufe I als Überprüfungsinstrument der Bildungsstandards* (Marburger Schriften zur Lehrerbildung, Bd. 2). Marburg: Tectum-Verlag.
- Kounin, J. S. (2006). *Techniken der Klassenführung*. Münster: Waxmann.
- Krainz-Dürr, M. (1999). *Wie kommt Lernen in die Schule? Zur Lernfähigkeit der Schule als Organisation*. Innsbruck: Studien-Verlag.
- Kramarski, B. & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom. The effects of cooperative learning and metacognitive training. *American Educational Research Journal*, 40(1), 281–310.
- Krapp, A. & Seidel, T. (2014). *Pädagogische Psychologie. Mit Online-Materialien*. Weinheim: Beltz.
- Krapp, A. & Weidenmann, B. (Hrsg.). (2005). *Pädagogische Psychologie*. Weinheim, Basel: Beltz Verlag.
- Krapp, S. (2006). Evaluationsbegleitende Abstimmung: Aufgaben von Auftraggebern und -nehmer im idealen Evaluationsprozess. In W. Böttcher, H. G. Holtappels & M. Brohm (Hrsg.), *Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele* (S. 79–86). Weinheim: Juventa-Verl.

- Krauss, S., Neubrand, M., Blum, W., Baumert, J., Brunner, M., Kunter, M. et al. (2008). Die Untersuchung des professionellen Wissens deutscher Mathematik-Lehrerinnen und -Lehrer im Rahmen der COACTIV-Studie. *Journal für Mathematik-Didaktik*, 29(3-4), 233–258.
- Krauth, J. (1993). *Einführung in die Konfigurationsfrequenzanalyse (KFA). Ein multivariates nichtparametrisches Verfahren zum Nachweis und zur Interpretation von Typen und Syndromen*. Weinheim: Beltz.
- Krawitz, J., Achmetli, K., Blum, W., Vogel, S. & Besser, M. (2016). *Report on the relative strengths and weaknesses of the United States in PISA 2012 mathematics* (Bd. 151). Paris: OECD Publishing.
- Kühle, B. & Peek, R. (2007). Lernstandserhebungen in Nordrhein-Westfalen. Evaluationsergebnisse zur Rezeption und zum Umgang mit Ergebnissrückmeldungen in Schulen. *Empirische Pädagogik*, 21(4), 428–447.
- Kuhn, H.-J. (2014). Anspruch, Wirklichkeit und Perspektiven der Gesamtstrategie der KMK zum Bildungsmonitoring. *Die Deutsche Schule. Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis*, 106(4), 414–426.
- Kulhavy, R. W. & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, 68(5), 505–512.
- Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research*, 47(2), 211–232.
- Kulik, J. A. & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, 58(1), 79–97.
- Kunter, M., Baumert, J., Blum, W., Klusmann, U., Krauss, S. & Neubrand, M. (Hrsg.). (2011). *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV*. Münster: 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.
- 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. Ergebnisse des Forschungsprogramms COACTIV* (S. 85–114). Münster: Waxmann.
- Kuper, H. & Diemer, T. (2012). Vergleichsarbeiten: Theoretische und empirische Betrachtungen zum Nutzen des Vergleichens. In A. Wacker, U. Maier & J. Wissinger

- (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung* (S. 225–245). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Kuper, H. & Hartung, V. (2007). Überzeugungen zur Verwendung des Wissens aus Lernstandserhebungen. *Zeitschrift für Erziehungswissenschaft*, 10(2), 214–229.
- Lambert, A. (2011). Reflektierende Unterrichtsplanung. In T. Leuders (Hrsg.), *Mathematik-Didaktik. Praxishandbuch für die Sekundarstufe 1 und 2* (S. 276–288). Berlin: Cornelsen Scriptor.
- LSA [Landesschulamt und Lehrkräfteakademie]. (2015). *Lernstandserhebungen* (Hessisches Kultusministerium, Hrsg.). Wiesbaden: Hessisches Kultusministerium.
- Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
- Landmann, M., Perels, F., Otto, B. & Schmitz, B. (2009). Selbstregulation. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 49–70). Berlin: Springer.
- Landwehr, N. (2011). Thesen zur Wirkung und Wirksamkeit der externen Schulevaluation. In C. Quesel, V. Husfeldt, N. Landwehr & P. Steiner (Hrsg.), *Wirkungen und Wirksamkeit der externen Schulevaluation* (S. 35–69). Bern: Hep.
- Lawrence, M. A. (2016). *Package ‚ez‘. Easy analysis and visualization of factorial experiments*, CRAN. Zugriff am 07.06.2017. Verfügbar unter <https://cran.r-project.org/web/packages/ez/ez.pdf>
- Leisen, J. (2006). Aufgabenkultur im mathematisch-naturwissenschaftlichen Unterricht. *MNU*, 59(5), 260–266.
- Leiß, D. (2007). „Hilf mir, es selbst zu tun“. *Lehrerinterventionen beim mathematischen Modellieren* (Texte zur mathematischen Forschung und Lehre, Bd. 57). Hildesheim: Franzbecker.
- Leiß, D. & Blum, W. (2012). Beschreibung zentraler mathematischer Kompetenzen. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM* (S. 33–80). Berlin: Cornelsen Scriptor.
- Leuders, T. (2006). Erläutere an einem Beispiel...Mathematische Kompetenzen erkennen und fördern - mit offenen Aufgaben. In G. Becker, M. Horstkemper, E. Risse, L. Stäudel, R. Werning & F. Winter (Hrsg.), *Diagnostizieren und Fördern* (S. 78–83). Seelze: Friedrich Verlag.

- Leuders, T. (2009). *Intelligent üben und Mathematik erleben*. Zugriff am 08.09.2017. Verfügbar unter https://home.ph-freiburg.de/leudersfr/preprint/2009_leuders_intelligent_ueben_mathemagische_momente.pdf
- Leuders, T. (2011). Prozessorientierter Mathematikunterricht. In T. Leuders (Hrsg.), *Mathematik-Didaktik. Praxishandbauch für die Sekundarstufe 1 und 2* (S. 265–276). Berlin: Cornelsen Scriptor.
- Leuders, T. (2012). Kompetenzorientierte Aufgaben im Unterricht. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungs-ideen; mit CD-ROM* (S. 81–95). Berlin: Cornelsen Scriptor.
- Leuders, T. (2015). Intelligentes Üben im Mathematikunterricht. In W. Blum, S. Vogel, C. Drüke-Noe & A. Roppelt (Hrsg.), *Bildungsstandards aktuell: Mathematik in der Sekundarstufe II* (Bildungsstandards aktuell, S. 192–204). Braunschweig: Schroedel.
- Leuders, T. & Leiß, D. (2012). Realitätsbezüge. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM* (6. Auflage, S. 194–206). Berlin: Cornelsen Scriptor.
- Leutner, D., Fleischer, J., Spoden, C. & Wirth, J. (2008). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (S. 149–167). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Lienert, G. A. (1971). Die Konfigurationsfrequenzanalyse. Ein neuer Weg zu Typen und Syndromen. *Zeitschrift für klinische Psychologie und Psychotherapie*, 19(2), 99–115.
- Lienert, G. A. (1978). *Verteilungsfreie Methoden in der Biostatistik. Band 2*. Meisenheim am Glan: Hain.
- Lienert, G. A. (1986). *Verteilungsfreie Methoden in der Biostatistik. Band 1*. Meisenheim am Glan: Hain.
- Linneweber-Lammerskitten, H. & Wälti, B. (2007). Wie Mindeststandards gemacht werden. Informationen und Erfahrungen aus der Schweiz. *lernchancen*, (55), 59–63.
- 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: Ausbildung und Beruf* (S. 47–70). Weinheim: Beltz.
- Lipowsky, F. (2009). Unterricht. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (Springer-Lehrbuch, S. 73–101). Berlin: Springer.
- Lipowsky, F. (2010). Lernen im Beruf. Empirische Befunde zur Wirksamkeit von Lehrerfortbildung. In F. H. Müller (Hrsg.), *Lehrerinnen und Lehrer lernen. Konzepte und Befunde zur Lehrerfortbildung* (S. 51–72). Münster: Waxmann.
- Lipowsky, F. (2013). *Lehrerfortbildung neu und weiter denken*. Zugriff am 24.02.2017. Verfügbar unter http://assets02.hessen-spd.net/docs/doc_45511_2013731128.pdf
- Lipowsky, F. (2014). Theoretische Perspektiven und empirische Befunde zur Wirksamkeit von Lehrerfort- und -weiterbildung. In E. Terhart, H. Bennewitz & M. Rothland (Hrsg.), *Handbuch der Forschung zum Lehrerberuf* (S. 470–498). Münster: Waxmann.
- Lipowsky, F., Drollinger-Vetter, B., Hartig, J. & Klieme, E. (2006). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis“*. Frankfurt am Main: DIPF.
- Lipowsky, F., Rakoczy, K., Buff, A. & Klieme, E. (2005). *Dokumentation der Erhebungs- und Auswertungsinstrumente zur schweizerisch-deutschen Videostudie „Unterrichtsqualität, Lernverhalten und mathematisches Verständnis“*. Frankfurt am Main: DIPF.
- 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.
- Lipowsky, F. & Rzejak, D. (2012). Lehrerinnen und Lehrer als Lerner - Wann gelingt der Rollentausch? Merkmale und Wirkungen wirksamer Lehrerfortbildungen. *Schulpädagogik heute*, 3(5), 1–17.
- Lipowsky, F., Rzejak, D. & Dorst, G. (2011). Lehrerfortbildung und Unterrichtsentwicklung. Oder: Wie können Wirkungen des eigenen Handelns erfahrbar gemacht werden? *Pädagogik*, 12(11), 38–41.

- Lorenz, J. H. (2005). Zentrale Lernstandsmessung in der Primarstufe—Vergleichsarbeiten Klasse 4 (VERA) in sieben Bundesländern. *ZDM*, 37(4), 317–323.
- Lortie, D. C. (1977). *Schoolteacher. A sociological study*. Chicago: Univ. of Chicago Press.
- 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: Waxmann.
- Luhmann, M. (2011). *R für Einsteiger. Einführung in die Statistiksoftware für die Sozialwissenschaften [mit Online-Materialien]* (2. Aufl.). Weinheim: Beltz.
- Lührig, H. H. (1973). *Wirtschaftsriese - Bildungszweig. Der Diskussionshintergrund zum Bildungsgesamtplan 1973: Analysen des OECD-Reports*. Reinbek bei Hamburg: Rowohlt.
- Maag Merki, K. (2010). Theoretische und empirische Analysen der Effektivität von Bildungsstandards, standardbezogenen Lernstandserhebungen und zentralen Abschlussprüfungen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 145–169). Wiesbaden: VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH.
- Maechler, M., Rousseeuw, P. J., Struyf, A., Hubert, M., Hornik, K., Studer, M. et al. (2017). *Package ‚cluster‘*, CRAN. Zugriff am 11.07.2017. Verfügbar unter <https://cran.r-project.org/web/packages/cluster/cluster.pdf>
- Maier, U. (2008a). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik*, 54(1), 95–117.
- Maier, U. (2008b). Vergleichsarbeiten im Vergleich – Akzeptanz und wahrgenommener Nutzen standardbasierter Leistungsmessungen in Baden-Württemberg und Thüringen. *Zeitschrift für Erziehungswissenschaft*, 11(3), 453–474.
- Maier, U. (2009a). Professionelle Nutzung von Vergleichsarbeiten? - Ergebnisse einer qualitativen Interviewstudie mit Lehrkräften in Baden-Württemberg. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 131–144). Bad Heilbrunn: Klinkhardt.
- Maier, U. (2009b). *Wie gehen Lehrerinnen und Lehrer mit Vergleichsarbeiten um? Eine Studie zu testbasierten Schulreformen in Baden-Württemberg und Thüringen*

- (Schul- und Unterrichtsforschung, Bd. 7). Baltmannsweiler: Schneider-Verl. Hohengehren.
- Maier, U. (2010). Vergleichsarbeiten im Spannungsfeld zwischen formativer und summativer Leistungsmessung. *Die Deutsche Schule*, 102(1), 60–69.
- Maier, U., Metz, K., Bohl, T., Kleinknecht, M. & Schymala, M. (2012). Vergleichsarbeiten als Instrument der datenbasierten Schul- und Unterrichtsentwicklung in Gymnasien. In A. Wacker, U. Maier & J. Wissinger (Hrsg.), *Schul- und Unterrichtsreform durch ergebnisorientierte Steuerung* (S. 197–224). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Margraf, J. & Fehm, L. (1996). Methoden der Klinischen Psychologie. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch quantitative Methoden* (S. 577–598). Weinheim: Beltz Psychologie Verlags Union.
- Maritzen, N. (2014). Glanz und Elend der KMK-Strategie zum Bildungsmonitoring. Versuch einer Bilanz und eines Ausblicks. *Die Deutsche Schule. Zeitschrift für Erziehungswissenschaft, Bildungspolitik und pädagogische Praxis*, 106(4), 398–413.
- Marsh, H. W., Byrne, B. M. & Shavelson, R. J. (1988). A multifaceted academic self-concept: Its hierarchical structure and its relation to academic achievement. *Journal of Educational Psychology*, 80(3), 366–380.
- Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, 23(1), 129–149.
- Maslach, C., Schaufeli, W. B. & Leiter, M. P. (2001). Job Burnout. *Annual Review of Psychology*, 52, 397–422.
- Maxwell, S. E. & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113(1), 181–190.
- Mayr, J. & Müller, F. H. (2010). Wovon hängt es ab, wie und wieviel sich Lehrerinnen und Lehrer fortbilden? In J. Schmich & C. Schreiner (Hrsg.), *Talis 2008: Schule als Lernumfeld und Arbeitsplatz. Vertiefende Analysen aus österreichischer Perspektive* (BIFIE Report, S. 11–25). Graz: Leykam.
- McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, Calif.: SAGE Publications.
- McDowall, S., Cameron, M., Dingle, R., Gilmore, A. & MacGibbon, L. (2007). *Evaluation of the Literacy Professional Development Project. Report prepared for the Ministry of Education*. Wellington, N.Z.: Ministry of Education.
- Meyer, H. (2016). *Was ist guter Unterricht?* Berlin: Cornelsen.

- Meyer, H. & Demuth, R. (2009). *Unterricht weiterentwickeln und beurteilen. Allgemeine Didaktik für Schulleitungen* (Schulmanagement-Handbuch, Bd. 132. Jg. 28). München: Oldenbourg.
- Meyerhöfer, W. (2003, 1. Januar). *Was testen Tests? Objektiv-hermeneutische Analysen am Beispiel von TIMSS und PISA*. Universität Potsdam, Potsdam. Zugriff am 10.11.2012. Verfügbar unter http://opus.kobv.de/ubp/front-door.php?source_opus=1284&la=de
- Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3), 325–342.
- Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2), 187–199.
- Milligan, G. W. & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159–179.
- Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2011). *Lernstandserhebungen als Impuls für die Unterrichtsentwicklung. Hinweise, Beispiele und Arbeitshilfen für die Praxis* (1. Auflage) (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, Hrsg.) (Schule in NRW Nr. 9046). Düsseldorf: Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen.
- Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. (2012, 25. Februar). *Zentrale Lernstandserhebungen (Vergleichsarbeiten). Runderlass des Ministeriums für Schule und Weiterbildung vom 20.12.2006* (Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen, Hrsg.). Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen. Zugriff am 24.04.2014. Verfügbar unter http://www.standardsicherung.schulministerium.nrw.de/vera3/upload/download/mat_12-13/Erlass_Zentrale_Lernstandserhebungen_Stand_25.2.2012.pdf
- Moosbrugger, H. & Kelava, A. (2012a). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Berlin: Springer.
- Moosbrugger, H. & Kelava, A. (Hrsg.). (2012b). *Testtheorie und Fragebogenkonstruktion*. Berlin: Springer.

- Nachtigall, C. (2014). *Landesbericht. Thüringer Kompetenztests 2013*. Friedrich-Schiller-Universität Jena; Thüringer Institut für Lehrerfortbildung; Thüringer Ministerium für Bildung, Wissenschaft und Kultur.
- Nachtigall, C. & Jantowski, A. (2007). Die Thüringer Kompetenztests unter besonderer Berücksichtigung der Evaluationsergebnisse zum Rezeptionsverhalten. *Empirische Pädagogik*, 21(4), 401–410.
- Nachtigall, C. & Kröhne, U. (2006). Methodische Anforderungen an schulische Leistungsvergleiche - auf dem Weg zu fairen Vergleichen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 59–74). Münster: Waxmann.
- Narcciss, S. (2006). *Informatives tutorielles Feedback. Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse* (Pädagogische Psychologie und Entwicklungspsychologie, Bd. 56). Münster: Waxmann.
- Netzwerk Empiriegestützte Schulentwicklung (2006). *Positionspapier des Netzwerks Empiriegestützte Schulentwicklung (EMSE) zu: Zentrale standardisierte Lernstandserhebungen (5. EMSE-Tagung, Berlin 08.12.2006)*, Netzwerk Empiriegestützte Schulentwicklung. Zugriff am 16.02.2015. Verfügbar unter http://www.emse-netzwerk.de/uploads/Main/EMSE_Positionsp1_Lernstandserh.pdf
- Netzwerk Empiriegestützte Schulentwicklung (2008). *Nutzung und Nutzen von Schulrückmeldungen im Rahmen standardisierter Lernstandserhebungen/Vergleichsarbeiten. Zweites Positionierungspapier des EMSE-Netzwerkes - verabschiedet auf der 9. EMSE-Fachtagung am 16.-17. Dezember 2008 in Nürnberg*, Netzwerk Empiriegestützte Schulentwicklung. Zugriff am 16.02.2015. Verfügbar unter http://www.emse-netzwerk.de/uploads/Main/EMSE_Positionsp2_Rueckmeldungen.pdf
- Neubrand, M. (2012). Multiple Lösungswege für Aufgaben: Bedeutung für Fach, Lernen, Unterricht und Leistungserfassung. In W. Blum, C. Drüke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM* (6. Auflage, S. 162–177). Berlin: Cornelsen Scriptor.
- Neubrand, M., Jordan, A., Krauss, S., Blum, W. & Löwen, K. (2011). Aufgaben im COACTIV-Projekt: Einblicke in das Potenzial für kognitive Aktivierung im

- Mathematikunterricht. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 115–132). Münster: Waxmann.
- Nieder, T. & Frühauf, S. (2012). *Bilanz der Schulinspektionen. Ergebnisse der externen Evaluation an allen hessischen Schulen.* (Institut für Qualitätsentwicklung (IQ), Hrsg.) (IQ Report Nr. 16). Wiesbaden: Institut für Qualitätsentwicklung (IQ).
- Niegemann, H. & Stadler, S. (2001). Hat noch jemand eine Frage? Systematische Unterrichtsbeobachtung zu Häufigkeit und kognitivem Niveau von Fragen im Unterricht. *Unterrichtswissenschaft*, 29(2), 171–192.
- Nir, A. E. & Bogler, R. (2008). The antecedents of teacher satisfaction with professional development programs. *Teaching and Teacher Education*, 24(2), 377–386.
- Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagatis & S. Papastravridis (Hrsg.), *3rd Mediterranean Conference on Mathematical Education* (S. 115–124). Athen: The Hellenic Mathematical Society.
- Nünning, A. & Zierold, M. (2008). *Kommunikationskompetenzen. Erfolgreich kommunizieren in Studium und Berufsleben.* Stuttgart: Klett Lerntraining.
- O'Day, J. A. (2002). Complexity, accountability, and school improvement. *Harvard Educational Review*, 72(3), 293–329.
- Oelkers, J. & Reusser, K. (2008). *Qualität entwickeln - Standards sichern - mit Differenzen umgehen. Kurzfassung* (Bildungsforschung, Bd. 27). Bonn, Berlin: Bundesministerium für Bildung und Forschung (BMBF).
- Olejnik, S. F. & Algina, J. (1985). A review of nonparametric alternatives to analysis of covariance. *Evaluation Review*, 9(1), 51–83.
- Organisation for Economic Co-operation and Development. (2014a). *PISA 2012 Ergebnisse. Was Schülerinnen und Schüler wissen und können*: W. Bertelsmann Verlag. <https://doi.org/10.1787/9789264208858-de>
- Organisation for Economic Co-operation and Development. (2014b). *PISA 2012 technical report*, Organisation for Economic Co-operation and Development. Zugriff am 05.02.2015. Verfügbar unter <http://oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>
- Organisation for Economic Co-operation and Development. (2016). *PISA 2015 Ergebnisse*. <https://doi.org/10.1787/19963793>

- Orth, B. (1995). Messtheoretische Grundlagen der Diagnostik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik. Ein Lehrbuch* (S. 286–295). Weinheim: Beltz, PsychologieVerlagsUnion.
- Orth, G. (2005). Eine Chance für die Qualität schulischer Arbeit. Standards und zentrale Leistungsüberprüfungen. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 63–65). Seelze: Friedrich Verlag.
- Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive load theory and instructional design. Recent developments. *Educational Psychologist*, 38(1), 1–4.
- Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T. & Pöhlmann, C. (Hrsg.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I*. Münster: Waxmann.
- Parr, J., Timperley, H., Reddish, P., Jesson, R. & Adams, R. J. (2007). *Literacy professional development project. Identifying effective teaching and professional development practices for enhanced student learning*. Wellington, N.Z.: Learning Media.
- Peek, R. & Dobbstein, P. (2006a). Benchmarks als Input für die Schulentwicklung - das Beispiel der Lernstandserhebungen in Nordrhein-Westfalen. In H. Kuper & J. Schneewind (Hrsg.), *Rückmeldung und Rezeption von Forschungsergebnissen. Zur Verwendung wissenschaftlichen Wissens im Bildungssystem* (S. 41–58). Münster: Waxmann.
- Peek, R. & Dobbstein, P. (2006b). Zielsetzung: Ergebnisorientierte Schul- und Unterrichtsentwicklung. Potenziale und Grenzen der nordrhein-westfälischen Lernstandserhebungen. In W. Böttcher, H. G. Holtappels & M. Brohm (Hrsg.), *Evaluation im Bildungswesen. Eine Einführung in Grundlagen und Praxisbeispiele* (S. 177–193). Weinheim: Juventa-Verl.
- Peek, R., Pallack, A., Dobbstein, P., Fleischer, J. & Leutner, D. (2006). Lernstandserhebungen 2004 in Nordrhein-Westfalen - zentrale Testergebnisse und Perspektiven für die Schul- und Unterrichtsentwicklung. In F. Eder, A. Gastager & F. Hofmann (Hrsg.), *Qualität durch Standards? Beiträge zum Schwerpunktthema der 67. Tagung der AEPF* (S. 219–233). Münster: Waxmann.
- Pekrun, R., Vom Hofe, R., Blum, W., Götz, T., Wartha, S. & Julien, S. (2006). Projekt zur Analyse der Leistungsentwicklung in Mathematik (PALMA).

- Entwicklungsverläufe, Schülervoraussetzungen und Kontextbedingungen von Mathematikleistungen bei Schülerinnen und Schülern der Sekundarstufe I. In M. Prenzel & L. Allolio-Nägel (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 21–53). Münster: Waxmann.
- Pellegrino, J. W., Chudowsky, N. & Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment*. Washington, DC: National Academy Press.
- Penk, C., Pöhlmann, C., Roppelt, A., Pietsch, E. & Pant, H. A. (2013). *Vergleichsarbeiten 2013 8. Jahrgangsstufe Mathematik. Technischer Bericht*. Berlin: Humboldt-Universität zu Berlin; Institut zur Qualitätsentwicklung im Bildungswesen.
- Penuel, W. R., Fishman, B. J., Yamaguchi, R. & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum implementation. *American Educational Research Journal*, 44(4), 921–958.
- Petersen, A. (2013). Weiß DESI mehr als VERA? Und was weiß KESS? *Begegnung*, 34(1), 29–32.
- Piaget, J. (1976). *Die Äquilibration der kognitiven Strukturen*. Stuttgart: Klett.
- Picht, G. (1964). *Die deutsche Bildungskatastrophe. Analyse und Dokumentation*. Olten, Freiburg im Breisgau: Walter-Verlag AG.
- Pintrich, P. R. & Groot, E. V. de. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82(1), 33–40.
- Polya, G. (1967). *Schule des Denkens. Vom Lösen mathematischer Probleme*. Bern: Francke.
- Popham, W. J. (2001). Teaching to the test. *Educational Leadership*, 58(6), 16–21.
- Popham, W. J. (2003). The seductive allure of data. *Educational Leadership*, 60(5), 48–51.
- Posch, P. (2009). Datengestützte Schulentwicklung und Unterstützungssysteme. *Die Deutsche Schule*, 101(2), 119–135.
- Pospeschill, M. (2006). *Statistische Methoden. Strukturen, Grundlagen, Anwendungen in Psychologie und Sozialwissenschaften* (1. Aufl.). München: Spektrum Akademischer Verlag.
- Prenzel, M. (2001). Das BLK-Modellprogramm „Steigerung der Effizienz des mathematisch-naturwissenschaftlichen Unterrichts“. In Bundesministerium für Bildung

- und Forschung (BMBF) (Hrsg.), *TIMSS - Impulse für Schule und Unterricht. Forschungsbefunde, Reforminitiativen, Praxisberichte und Video-Dokumente* (S. 59–65). Bonn.
- Prenzel, M. (2006). Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres: Die Ergebnisse von PISA-I-Plus im Überblick. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (Bd. 3, Bd. 3, S. 15–28). Münster: Waxmann.
- Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M. et al. (Hrsg.). (2004a). *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland: Ergebnisse des zweiten internationalen Vergleichs* (Bd. 1, 3 Bände). Band 1. Münster: Waxmann.
- Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., Neubrand, M. et al. (2004b). *PISA 2003. Ergebnisse des zweiten internationalen Vergleichs. Zusammenfassung*. Zugriff am 15.09.2014. Verfügbar unter <http://www.bmfsfj.de/doku/Publikationen/genderreport/01-Redaktion/PDF-Anlagen/lit-pisa-ergebnisse-2003,property%3Dpdf,bereich%3Dgenderreport,sprache%3Dde,rwb%3Dtrue.pdf>
- Prenzel, M., Drechsel, B., Carstensen, C. H. & Ramm, G. (2004). PISA 2003 - eine Einführung. In M. Prenzel, J. Baumert, W. Blum, R. Lehmann, D. Leutner, M. Neubrand et al. (Hrsg.), *PISA 2003. Der Bildungsstand der Jugendlichen in Deutschland: Ergebnisse des zweiten internationalen Vergleichs* (S. 13–46). Band 1. Münster: Waxmann.
- Projektgruppe „Innovationen im Bildungswesen“. (2005). *Abschlussbericht zum BLK-Programm „Qualitätsverbesserung in Schulen und Schulsystemen“ (QuiSS)*. Zugriff am 05.02.2015. Verfügbar unter http://www.pedocs.de/volltexte/2010/1547/pdf/abschlussbericht_quiss_projektgruppe_D.pdf
- Rakoczy, K. (2006). Motivationsunterstützung im Mathematikunterricht. Zur Bedeutung von Unterrichtsmerkmalen für die Wahrnehmung von Schülerinnen und Schüler. *Zeitschrift für Pädagogik*, 52(6), 822–843.
- Rakoczy, K. (2008). *Motivationsunterstützung im Mathematikunterricht. Unterricht aus der Perspektive von Lernenden und Beobachtern*. Münster: Waxmann.

- Rakoczy, K., Klieme, E., Drollinger-Vetter, B., Lipowsky, F., Pauli, C. & Reusser, K. (2007). Structure as a quality feature in mathematics instruction. Cognitive and motivational effects of a structured organisation of the learning environment vs. a structured presentation of learning content. In M. Prenzel (Hrsg.), *Studies on the educational quality of schools. The final report on the DFG priority programme* (S. 101–120). Münster: Waxmann.
- Ramm, G. (2014). *Gute Schulen. Wissenschaftliche Befunde, praktische Konsequenzen* (Schulmanagement-Handbuch, Bd. 149). München: Oldenbourg.
- Ramm, G. & Rogalski, K. (2013). *Mathe macht stark. Befragungsergebnisse 2012*, Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein. Zugriff am 16. Oktober. Verfügbar unter http://www.schleswig-holstein.de/IQSH/DE/SchulentwicklungITMedien/VergleichsarbeitenLSAPE/ProjektevaluationDiagnostik/Material/BefragungsergebnisseMMS__blob=publicationFile.pdf
- Redaktion Neues Deutschland. (2012). *Vom PISA-Test zur VERA-Hitparade (neues deutschland)*. Zugriff am 04.08.2015. Verfügbar unter <http://www.neues-deutschland.de/artikel/233147.vom-pisa-test-zur-vera-hitparade.html>
- Reeve, J. (2004). Self-determination theory applied to education settings. In E. L. Deci & R. M. Ryan (Hrsg.), *Handbook of self-determination research* (S. 183–203). Rochester, N.Y.: University of Rochester Press.
- Reimers, H., Ramm, G. & Klein, W. (2009). *EVIT-Landesbericht 2009*, Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein. Zugriff am 16.10.2014. Verfügbar unter http://www.schleswig-holstein.de/IQSH/DE/SchulentwicklungITMedien/SchulentwicklungInterneEvaluation/evaluation/EvitLandesbericht2009__blob=publicationFile.pdf
- Reinmann-Rothmeier, G. & Mandl, H. (2005). Unterrichten und Lernumgebungen gestalten. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 613–658). Weinheim: Beltz PVU.
- Reiss, K. (2007). Mindeststandards für den Mathematikunterricht - Wie viel Mathematik muss es sein? *lernchancen*, 55, 4–7.
- Reusser, K. (2001). Co-constructivism in educational theory and practice. In N. J. Smelser, P. Baltes & F. E. Weinert (Hrsg.), *International encyclopedia of the social & behavioral Sciences* (S. 2058–2062). Amsterdam [u.a.]: Elsevier, Pergamon.

- Revelle, W. (2017a). *How to: Use the psych package for factor analysis and data reduction*, Northwestern University. Zugriff am 08.05.2017. Verfügbar unter <http://personality-project.org/r/psych/HowTo/factor.pdf>
- Revelle, W. (2017b). *An introduction to the psych package: Part II scale construction and psychometrics*. Northwestern University. Zugriff am 08.05.2017. Verfügbar unter <https://cran.r-project.org/web/packages/psych/vignettes/overview.pdf>
- Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation* (Motivationsforschung, Bd. 8). Göttingen: Verlag für Psychologie Hogrefe.
- Rheinberg, F. (Hrsg.). (1982). *Bezugsnormen zur Schulleistungsbewertung. Analyse und Intervention*. Düsseldorf: Pädagogischer Verlag Schwann.
- Rheinberg, F. (2001). Bezugsnormen und schulische Leistungsbeurteilung. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 59–72). Weinheim: Beltz.
- Richter, D. (2011). Entwicklung professioneller Kompetenz - Lernen im Beruf. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 317–325). Münster: Waxmann.
- Richter, D., Böhme, K., Becker, M., Pant, H. A. & Stanat, P. (2014). Überzeugungen von Lehrkräften zu den Funktionen von Vergleichsarbeiten. *Zeitschrift für Pädagogik*, 60(2), 225–243.
- Richter, D., Kuhl, P., Haag, N. & Pant, H. A. (2013). Aspekte der Aus- und Fortbildung von Mathematik- und Naturwissenschaftslehrkräften im Ländervergleich. 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. 367–390). Münster: Waxmann.
- Rieß, C. & Zuber, J. (2014). *Rezeption und Nutzung von Ergebnissen der Bildungsstandardüberprüfung in Mathematik auf der 8. Schulstufe unter Berücksichtigung der Rückmeldemoderation*. Zugriff am: 20.08.2019. Verfügbar unter: https://www.bifie.at/wp-content/uploads/2017/05/E_BIST_M8_RM_RMM_20140623.pdf
- Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? *Psychologische Rundschau*, 57(2), 69–86.
- Roberts, S. J. (1978). *Test floor and ceiling effects* (Office of Education, Washington, D. C., Hrsg.). Zugriff am 28.09.2015. Verfügbar unter <http://files.eric.ed.gov/full-text/ED185106.pdf>

- Robitzsch, A. (2017). *Package ‚TAM‘. Test Analysis Modules*, CRAN. Zugriff am 07.06.2017. Verfügbar unter <https://cran.r-project.org/web/packages/TAM/TAM.pdf>
- Röhner, J. & Schütz, A. (2016). *Psychologie der Kommunikation*. Wiesbaden: Springer Fachmedien Wiesbaden.
- Rolff, H.-G. (1993). *Wandel durch Selbstorganisation. Theoretische Grundlagen und praktische Hinweise für eine bessere Schule*. Weinheim: Juventa Verlag.
- Rolff, H.-G. (2001). Was bringt die vergleichende Leistungsmessung für die pädagogische Arbeit in Schulen? In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 337–352). Weinheim: Beltz.
- Rolff, H.-G. (2002). Rückmeldung und Nutzung der Ergebnisse von großflächigen Leistungsuntersuchungen. Grenzen und Chancen. In H.-G. Rolff, H. G. Holtappels, K. Klemm, H. Pfeiffer & R. Schulz-Zander (Hrsg.), *Jahrbuch der Schulentwicklung Band 12. Daten, Beispiele und Perspektiven* (S. 75–98). Weinheim: Juventa Verlag.
- Rolff, H.-G. (2007). *Studien zu einer Theorie der Schulentwicklung*. Weinheim: Beltz.
- Rolff, H.-G. (2010). Schulentwicklung als Trias von Organisations-, Unterrichts- und Personalentwicklung. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 29–36). Stuttgart: UTB.
- Rolff, H.-G. & Buchen, H. (Hrsg.). (2016). *Professionswissen Schulleitung*. Weinheim: Beltz Verlagsgruppe.
- Roppelt, A., Katzenbach, M. & Pant, H. A. (2010, 29. Januar). *Bereitstellung eines Aufgabenpools für kompetenzbasierte Vergleichsarbeiten im Fach Mathematik in der 8. Jahrgangsstufe im Schuljahr 2009/2010. Technischer Bericht*. Berlin: Institut zur Qualitätsentwicklung im Bildungswesen; Humboldt-Universität zu Berlin.
- Rösner, E. (2005). Das doppelte Lottchen im deutschen Schulwesen. *Pädagogik*, 57(2), 36–41.
- Rost, D. H. (2007). *Handwörterbuch pädagogische Psychologie*. Weinheim [u.a.]: Beltz, PVU.
- Rost, D. H. (Hrsg.). (2010). *Handwörterbuch pädagogische Psychologie*. Weinheim: Beltz.
- Rückriem, G. (1981). *Ein Bilder-Lese-Buch über Schule und Alltag Berliner Arbeiterkin-der*. Berlin-West: Elefanten Press Verlag.

- Ryan, R. M. & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25(1), 54–67.
- Ryan, R. M., Stiller, J. D. & Lynch, J. H. (1994). Representations of relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem. *The Journal of Early Adolescence*, 14(2), 226–249.
- Rzejak, D., Küsting, J., Lipowsky, F., Fischer, E., Dezhgani, U. & Reichart, A. (2014). Facetten der Lehrerfortbildungsmotivation - eine faktorenanalytische Betrachtung. *Journal of Education Research Online*, 6(1), 139–159.
- Saxe, G. B., Gearhart, M. & Nasir, N. (2001). Enhancing student's understanding of mathematics: A study of three contrasting approaches to professional support. *Journal of Mathematics Teacher Education*, 4(1), 55–79.
- Schaarschmidt, U. (2005). *Halbtagsjobber?: psychische Gesundheit im Lehrerberuf - Analyse eines veränderungsbedürftigen Zustandes*. Weinheim: Beltz.
- Schaarschmidt, U. (2007). *Gerüstet für den Schulalltag: psychologische Unterstützungsangebote für Lehrerinnen und Lehrer*. Weinheim [u.a.]: Beltz.
- Scheerens, J. & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- Schermelleh-Engel, K. & Werner, C. S. (2012). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 119–142). Berlin: Springer.
- Schildkamp, K. (2007). *The utilization of a self-evaluation instrument for primary education*. Enschede: University of Twente.
- Schildkamp, K. & Kuiper, W. (2010). Data-informed curriculum reform. Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, 26(3), 482–496.
- Schildkamp, K. & Visscher, A. (2009). Factors influencing the utilization of a school self-evaluation instrument. *Studies in Educational Evaluation*, 35(4), 150–159.
- Schlesinger, L. & Jentsch, A. (2016). Theoretical and methodological challenges in measuring instructional quality in mathematics education using classroom observations. *ZDM mathematics Education*, 48, 29–40.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350–353.
- Schmitz, B., Landmann, M. & Perels, F. (2007). Das Selbstregulationsprozessmodell und theoretische Implikationen. In M. Landmann & B. Schmitz (Hrsg.),

- Selbstregulation erfolgreich fördern. Praxisnahe Trainingsprogramme für effektives Lernen* (S. 312–326). Stuttgart: Verl. W. Kohlhammer.
- Schneewind, J. & Kuper, H. (2009). Rückmeldeformate und Verwendungsmöglichkeiten der Ergebnisse aus zentralen Lernstandserhebungen. In T. Bohl & H. Kiper (Hrsg.), *Lernen aus Evaluationsergebnissen. Verbesserungen planen und implementieren* (S. 113–129). Bad Heilbrunn: Klinkhardt.
- Schnell, R., Hill, P. B. & Esser, E. (2014). *Methoden der empirischen Sozialforschung* (10 Aufl.): De Gruyter Oldenbourg.
- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. A. Grouws (Hrsg.), *Handbook of research on mathematics teaching and learning. A project of the National Council of Teachers of Mathematics* (S. 334–370). New York: Macmillan.
- Schönbächler, M.-T. (2008). *Klassenmanagement. Situative Gegebenheiten und personale Faktoren in Lehrpersonen- und Schülerperspektive*. Bern: P. Haupt.
- Schrader, F.-W. (2010). Leistungsmessung und Rückmeldung. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 144–146). Stuttgart: UTB.
- Schrader, F.-W. & Helmke, A. (2001). Alltägliche Leistungsbeurteilung durch Lehrer. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 45–58). Weinheim: Beltz.
- Schrader, F.-W. & Helmke, A. (2005). Überprüfte Vermutungen. Training der Diagnosefähigkeit von Lehrkräften durch die Nutzung von Vergleichsarbeiten. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (Friedrich Jahresheft, XXIII, S. 120–121). Seelze: Friedrich Verlag.
- Schukajlow, S. (2006). Schüler-Schwierigkeiten beim Lösen von Modellierungsaufgaben - Ergebnisse aus dem DISUM-Projekt. In E. Cohors-Fresenborg & I. Schwank (Hrsg.), *Beiträge zum Mathematikunterricht 2006. Vorträge auf der 40. Tagung für Didaktik der Mathematik vom 06.03. bis 10.03.2006 in Osnabrück* (S. 493–496). Hildesheim: Franzbecker.
- Schukajlow, S. & Krug, A. (2013). Considering multiple solutions for modeling problems - Design and first results from the MultiMa-project. In G. Stillman, G. Kaiser

- & W. Blum (Eds.), *Teaching mathematical modelling. Connecting to research and practice* (pp. 207–216). Dordrecht: Springer.
- Schwippert, K. (2005). Tests. Oder: Wie man Äpfel mit Birnen vergleicht. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 15–17). Seelze: Friedrich Verlag.
- Seidel, T. (2009). Klassenführung. In E. Wild & J. Möller (Hrsg.), *Pädagogische Psychologie* (S. 135–148). Berlin: Springer.
- 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.
- Selter, C. (2005a). SMS-Mathematik? Zur Stellungnahme der VERA-Gruppe zur Kritik an VERA. *Grundschule aktuell*, (90), 7.
- Selter, C. (2005b). VERA Mathematik 2004. VERbesserungsbedürftige Aufgaben! VERkapptes Ausleseinstrument? *Grundschule aktuell*, (89), 17–20.
- Sembill, D., Wuttke, E., Seifried, J., Egloffstein, M. & Rausch, A. *Selbstorganisiertes Lernen in der beruflichen Bildung - Abgrenzungen, Befunde und Konsequenzen*, Konstanzer-Online-Publikations-System. Zugriff am 02.03.2017. Verfügbar unter http://kops.uni-konstanz.de/bitstream/handle/123456789/12108/Sembill_etaal_2007_bwpat13.pdf?sequence=1&isAllowed=y
- Shavelson, R. J., Hubner, J. J. & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research*, 46(3), 407–441.
- Shavelson, R. J. & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research*, 51(4), 455–498.
- Shayer, M. & Adhami, M. (2007). Fostering cognitive development through the context of mathematics. Results of the CAME Project. *Educational Studies in Mathematics*, 64(3), 265–291.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.
- Siebenborn, R. (2005). Aufgabe: „Aufgabenentwicklerin“. Aus der Arbeit mit Vergleichsarbeiten lernen. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 46–47). Seelze: Friedrich Verlag.

- Sill, H.-D. (2005). Kritische Bemerkungen zu den aktuellen Bildungsstandards für den mittleren Schulabschluss im Fach Mathematik. In G. Graumann (Hrsg.), *Beiträge zum Mathematikunterricht. Vorträge auf der 39. Tagung für Didaktik der Mathematik vom 28.2. bis 4.3.2005 in Bielefeld*. Hildesheim: Franz Becker.
- Sill, H.-D. (2010). Probleme und Erfahrungen mit „Mindeststandards“. *Mitteilungen der Gesellschaft für Didaktik der Mathematik*, (88), 5–12.
- Six, U., Gleich, U. & Gimmler, R. (Hrsg.). (2007). *Kommunikationspsychologie—Medienpsychologie. Lehrbuch* (1. Aufl.). Weinheim: BeltzPVU.
- Sjuts, J. (2008). Aufgaben diagnostisch gestalten, Denkprozesse aufdecken und Verstehen fördern. *mathematik lehren*, (150), 58–61.
- Sjuts, J. (2012). Unterrichtliche Gestaltung und Nutzung kompetenzorientierter Aufgaben in diagnostischer Hinsicht. In W. Blum, C. Drücke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichts Anregungen, Fortbildungsideen; mit CD-ROM* (6. Auflage, S. 96–112). Berlin: Cornelsen Scriptor.
- Slavin, R. E. (1996). *Education for all*. Lisse [Netherlands]: Swets & Zeitlinger Publishers.
- Sleeter, C. (2014). Toward teacher education research that informs policy. *Educational Researcher*, 43(3), 146–153.
- Spieler, J. (Hrsg.). (1932). *Lexikon der Pädagogik der Gegenwart. 2. Kinderfürsorge bis Zwangszustände*. Freiburg im Breisgau. Verfügbar unter <http://nbn-resolving.de/urn:nbn:de:0111-bbf-spo-13899019>
- Spieß, E. (2004). Kooperation und Konflikt. In H. Schuler (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich D: Praxisgebiete: Organisationspsychologie - Gruppe und Organisation* [Neuausgabe], S. 193–250).
- Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte. *Journal für Mathematik-Didaktik*, 35(1), 79–99.
- Stamm, M. (2003). Evaluation im Spiegel ihrer Nutzung: Grande idée oder grande illusion des 21. Jahrhunderts? *Zeitschrift für Evaluation*, 2, 183–200.
- Stanat, P., Pant, H.-A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011*. Münster: Waxmann.

- Stein, M. K. & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason. An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation*, 2(1), 50–80.
- Stein, P. (2019). Forschungsdesigns für die quantitative Sozialforschung. In Baur, N. & Blasius, J. (Hrsg.), *Handbuch Methoden empirischer Sozialforschung* (S. 125-142). Wiesbaden: Springer VS.
- Stemmler, M. (1998). Nonparametrical analysis of change patterns in dependent samples. *Methods of Psychological Research Online*, 3(2), 23–38.
- Stemmler, M. (2014). *Person-centered methods. Configural Frequency Analysis (CFA) and other methods for the analysis of contingency tables*. Cham: Springer Verlag.
- Stemmler, M., Lautsch, E. & Martinke, D. (2008). *Configural Frequency Analysis (CFA) and other non-parametrical statistical methods. Gustav A. Lienert memorial issue*. Lengerich: Pabst Science Publ.
- Strähling, R. (2005). Qualitätsentwicklung statt Vergleichsarbeiten. Zu einem unfruchtbaren Verhältnis von Forschung und Schule. *Die Deutsche Schule*, 97(2), 211–221.
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80(1), 99–103.
- Stryck, T. (2005). *VERA - VERgleichArbeiten in Jahrgangsstufe 4*, Senatsverwaltung für Bildung, Jugend und Sport. Zugriff am 25.02.2015. Verfügbar unter http://www.berlin.de/imperia/md/content/sen-bildung/schulqualitaet/vergleichsarbeiten/vera2004_berlinbericht.pdf?start&ts=1227584730&file=vera2004_berlinbericht.pdf
- Stufflebeam, D. S. (1967). The use and abuse of evaluation in Title III. *Theory into Practice*, 6(3), 126–133.
- Sweller, J., van Merriënboer, J. J. G. & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296.
- Switala, B. (2005). Der Fall VERA - (meta)kritische Anmerkungen. *Grundschule aktuell*, (90), 9–11.
- Taylor, B. M., Pearson, P. D., Peterson, D. S. & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that

- encourage cognitive engagement in literacy learning. *The Elementary School Journal*, 104(1), 3–28. Zugriff am 28.02.2017.
- Terhart, E. (2002). Wie können die Ergebnisse von vergleichenden Leistungsstudien systematisch zur Qualitätsverbesserung in Schulen genutzt werden? *Zeitschrift für Pädagogik*, 48(1), 91–110.
- Thiel, F., Richter, S. G. & Ophardt, D. (2012). Steuerung von Übergängen im Unterricht. *Zeitschrift für Erziehungswissenschaft*, 15(4), 727–752.
- Thom. (2012). *Beware the Friedman test!* Zugriff am 19.06.2017. Verfügbar unter <https://www.r-bloggers.com/beware-the-friedman-test/>
- Tillmann, K.-J. (1999). Leistungsvergleichs-Untersuchungen auf nationaler und internationaler Ebene. Möglichkeiten, Grenzen, Probleme. *SchulVerwaltung spezial*, (2), 15–17.
- Tillmann, K.-J. (2005). Was fängt eine Schule mit ihren PISA-Ergebnissen an? Das Beispiel Laborschule Bielefeld. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 42–44). Seelze: Friedrich Verlag.
- Timperley, H., Wilson, A., Barrar, H. & Fung, I. (2007). *Teacher professional learning and development. Best evidence synthesis iteration (BES)*. Wellington, N.Z: Ministry of Education.
- Toothaker, L. E. & Newman, D. (1994). Nonparametric competitors to the two-Way ANOVA. *Journal of Educational and Behavioral Statistics*, 19(3), 237–273.
- Turner, J. C., Meyer, D. K., Cox, K. E., Logan, C., DiCintio, M. & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology*, 90(4), 730–745.
- Tuschel, S. (2006). Teamteaching in Integrationsklassen: fördernde und hemmende Faktoren. *Erziehung und Unterricht*, (1-2), 65–74.
- Ufer, S., Heinze, A. & Lipowsky, F. (2015). Unterrichtsmethoden und Instruktionsstrategien. In R. Bruder, L. Hefendehl-Hebeker, B. Schmidt-Thieme & H.-G. Weigand (Hrsg.), *Handbuch der Mathematikdidaktik* (S. 411–434). Berlin: Springer Spektrum.
- Uhl, S. (2006). *Die Bildungsstandards, die Outputsteuerung und ihre Kritiker* (Institut für Qualitätsentwicklung (IQ), Hrsg.) (IQ Kompakt Nr. 2). Wiesbaden: Institut für Qualitätsentwicklung (IQ).

- Van Ackeren, I. & Brauckmann, S. (2010). Internationale Diskussions-, Forschungs- und Theorieansätze zur Governance im Schulwesen. In H. Altrichter & K. Maag Merki (Hrsg.), *Handbuch Neue Steuerung im Schulsystem* (S. 41–61). Wiesbaden: VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH.
- Van Ackeren, I. & Klemm, K. (2011). *Entstehung, Struktur und Steuerung des deutschen Schulsystems. Eine Einführung*. Wiesbaden: VS, Verl. für Sozialwissenschaften.
- Vermunt, J. K. & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Hrsg.), *Applied latent class analysis* (S. 89–106). Cambridge: Cambridge University Press.
- Visscher, A. J. (2008). The utilization of school performance feedback systems for school improvement. In A. Breiter, A. Lange & E. Staake (Hrsg.), *School information systems and data-based decision-making* (S. 23–36). Frankfurt am Main: P. Lang.
- Visscher, A. J. & Coe, R. (2003). School performance feedback systems. Conceptualisation, analysis, and reflection. *School Effectiveness and School Improvement*, 14(3), 321–349.
- Vodafone Stiftung Deutschland. (2017). *Wie lernen Lehrer? Eine Umfrage unter Lehrerinnen und Lehrern zu ihrem eigenen Lernverhalten*. Düsseldorf. Zugriff am 05.05.2017. Verfügbar unter https://www.vodafone-stiftung.de/uploads/tx_newsjson/Vodafone_Stiftung_Wie_lernen_Lehrer_17_01.pdf
- Vogel, S., Achmetli, K., Krawitz, J. & Blum, W. (2013). VELM-8 - Ein Projekt zur Verbesserung der Effektivität der Lernstandserhebungen Mathematik Klasse 8. In G. Greefrath, F. Käpnick & M. Stein (Hrsg.), *Beiträge zum Mathematikunterricht 2013. Vorträge auf der 47. Tagung für Didaktik der Mathematik vom 04.03.2013 bis 08.03.2013 in Münster* (Bd. 2, S. 1042–1045). Münster: WTM.
- Vogel, S., Achmetli, K., Krawitz, J. & Blum, W. (2014). Wie können die Lernstandserhebungen in Klasse 8 effektiv genutzt werden? - Evaluation des Projekts VELM-8. In J. Roth & L. Ames (Hrsg.), *Beiträge zum Mathematikunterricht 2014. Vorträge auf der 48. Jahrestagung der Gesellschaft für Didaktik der Mathematik vom 10. bis 14. März in Koblenz* (S. 1251–1254). Münster: WTM, Verl. für Wiss. Texte und Medien.
- Vogel, S., Blum, W., Achmetli, K. & Krawitz, J. (2013). Zum Potential von Lernstandserhebungen für die Unterrichtsentwicklung - Das Projekt VELM-8. In I. Bausch, G.

- Pinkernell & O. Schmitt (Hrsg.), *Unterrichtsentwicklung und Kompetenzorientierung. Festschrift für Regina Bruder* (S. 105–118). Münster: WTM.
- Vogel, S., Blum, W., Achmetli, K. & Krawitz, J. (2016). Qualifizierung von Lehrkräften zum konstruktiven Umgang mit zentralen Lernstandserhebungen – Ergebnisse aus dem Projekt VELM-8. *Journal für Mathematik-Didaktik*, 319–348.
- Vom Hofe, R., Pekrun, R., Kleine, M. & Götz, T. (2006). Projekt zur Analyse der Leistungsentwicklungen in Mathematik (PALMA): Konstruktion des Regensburger Mathematikleistungstests für 5. bis 10. Klassen. In M. Prenzel & L. Allolio-Nägel (Hrsg.), *Untersuchungen zur Bildungsqualität von Schule. Abschlussbericht des DFG-Schwerpunktprogramms* (S. 83–100). Münster: Waxmann.
- Voss, A. & Blatt, I. (2010). Datengestützte Unterrichtsentwicklung. In T. Bohl, W. Helsper, H. G. Holtappels & C. Schelle (Hrsg.), *Handbuch Schulentwicklung. Theorie - Forschung - Praxis* (S. 175–179). Stuttgart: UTB.
- Voss, T., Kleickmann, T., Kunter, M. & Hachfeld, A. (2011). Überzeugungen von Mathematiklehrkräften. In M. Kunter, J. Baumert, W. Blum, U. Klusmann, S. Krauss & M. Neubrand (Hrsg.), *Professionelle Kompetenz von Lehrkräften. Ergebnisse des Forschungsprogramms COACTIV* (S. 235–258). Münster: Waxmann.
- Vygotskij, L. S., Cole, M., Stein, S. & Sekula, A. (1978). *Mind in society. The development of higher psychological processes*. Cambridge, Massachusetts: Harvard University Press.
- Wacker, A. & Kramer, J. (2012). Vergleichsarbeiten in Baden-Württemberg. *Zeitschrift für Erziehungswissenschaft*, 15(4), 683–706.
- Wade, R. (1984). *What makes a difference in inservice teacher education: A meta-analysis of the research*. Dissertation.
- Wade, R. (1984/1985). What makes a difference in inservice teacher education? A meta-analysis of research. *Educational Leadership*, 48–54.
- Wahl, D. (2013). *Lernumgebungen erfolgreich gestalten. Vom trägen Wissen zum kompetenten Handeln* (3. Aufl. mit Methodensammlung). Bad Heilbrunn: Klinkhardt.
- Wall, D. (2000). The Impact of high-stakes testing on teaching and learning: can this be predicted or controlled? *System*, 28(4), 499–509.
- Walther, G., van den Heuvel-Panhuizen, M., Granzer, D. & Köller, O. (2012). *Bildungsstandards für die Grundschule: Mathematik konkret. [Aufgabenbeispiele - Unterrichtsanregungen - Fortbildungsideen]*. Berlin: Cornelsen Scriptor.

- Wang, M. C., Haertel, G. D. & Walberg, H. J. (1993). Toward a knowledge base for school learning. *Review of Educational Research*, 63(3), 249–294.
- Watermann, R. & Klieme, E. (2006). Modellierung von Kompetenzstufen mit Hilfe der latenten Klassenanalyse. *Empirische Pädagogik*, 20(3), 321–336.
- Watermann, R. & Stanat, P. (2008). Schulrückmeldungen in PISA 2000: Sozialnorm- und kriteriumsorientierte Rückmeldeverfahren. *Empirische Pädagogik*, 18(1), 40–61.
- Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, 21(1), 1.
- Weiner, B. (1975). *Wirkung von Erfolg und Misserfolg auf die Leistung*. Bern: H. Huber.
- Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.
- Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (S. 17–31). Weinheim: Beltz.
- Wentzel, K. R. (1997). Student motivation in middle school. The role of perceived pedagogical caring. *Journal of Educational Psychology*, 89(3), 411–419.
- Wiegand, B. (1999). TIMSS als Spiegel für Defizite im deutschen Mathematikunterricht der Sek.II. In M. Neubrand (Hrsg.), *Beiträge zum Mathematikunterricht 1999. Vorträge auf der 33. Tagung für Didaktik der Mathematik vom 1. bis 5. März 1999 in Bern* (Beiträge zum Mathematikunterricht, S. 594–597). Franzbecker.
- Wiegand, B. (2001). TIMSS und die Folgen. *mathematik lehren*, 108, 4–8.
- Wigfield, A. & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25(1), 68–81.
- Wild, E., Hofer, M. & Pekrun, R. (2005). Psychologie des Lernens. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie* (S. 203–268). Weinheim, Basel: Beltz Verlag.
- Wild, E. & Möller, J. (Hrsg.). (2009). *Pädagogische Psychologie*. Berlin: Springer.
- Wild, E. & Möller, J. (Hrsg.). (2015). *Pädagogische Psychologie*. Berlin: Springer.
- Wild, K.-P. & Krapp, A. (2005). Pädagogisch-psychologische Diagnostik. In A. Krapp & B. Weidenmann (Hrsg.), *Pädagogische Psychologie. Ein Lehrbuch* (S. 525–574). Weinheim: Beltz PVU.

- Wild, K.-P. & Schiefele, U. (1994). Lernstrategien im Studium. Ergebnisse zur Faktorenstruktur und Reliabilität eines neuen Fragebogens. *Zeitschrift für Differentielle und diagnostische Psychologie*, 15, 185–200.
- Wildt, A. R. & Ahtola, O. T. (1978). *Analysis of covariance* (Sage university papers series. Quantitative applications in the social sciences, Bd. 12). Newbury Park, London, New Dehli: SAGE Publications.
- Wilkinson, S. S. (1981). *The relationship of teacher praise and student achievement: A meta-analysis of selected research* (Dissertations Abstracts International 41(9-A)) (S. 3998).
- Winkel, R. (2011). *Der gestörte Unterricht. Diagnostische und therapeutische Möglichkeiten*. Baltmannsweiler: Schneider Hohengehren.
- Winter, F. (2005). Standards auch von unten? Warum die Schule die Leistungsbewertung nicht delegieren darf. In G. Becker, A. Bremerich-Vos, M. Demmer, K. Maag Merki, B. Priebe, K. Schwippert et al. (Hrsg.), *Standards. Unterricht zwischen Kompetenzen, zentralen Prüfungen und Vergleichsarbeiten* (S. 76–77). Seelze: Friedrich Verlag.
- Winter, H. (1995). Mathematikunterricht und Allgemeinbildung. *Mitteilungen der Gesellschaft der Didaktik der Mathematik*, 61, 37–46.
- Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Göttingen: Hogrefe.
- Wobbrock, J. O., Findlater, L., Gergle, D., Higgins, J. J. & Kay, M. (2016, 29. Januar). *ARTool. Align-and-rank data for a nonparametric ANOVA*, University of Washington, University of Maryland, Northwestern University, Kansas State University. Zugriff am 18.08.2016. Verfügbar unter <http://depts.washington.edu/aim-group/proj/art/>
- Wood, T. & Sellers, P. (1996). Assessment of a problem-centered mathematics program: Third grade. *Journal for Research in Mathematics Education*, 27(2), 337–353. Zugriff am 25.02.2014.
- Wössmann, L., Lergetporer, P., Kugler, F. & Werner, K. (2016). Denken Lehrkräfte anders über die Bildungspolitik als die Gesamtbevölkerung? - Ergebnisse des ifo Bildungsbarometers 2016. *ifo schnelldienst*, 69(17), 19–34.
- Wright, B. & Linacre, J. M. (1994). *Reasonable mean-square fit values*. Zugriff am 08.10.2013. Verfügbar unter <http://www.rasch.org/rmt/rmt83b.htm>

- Wynands, A. (2012). Intelligentes Üben. In W. Blum, C. Drücke-Noe, R. Hartung & O. Köller (Hrsg.), *Bildungsstandards Mathematik: konkret. Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen; mit CD-ROM* (6. Auflage, S. 113–125). Berlin: Cornelsen Scriptor.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213.
- Yoon, K. S., Duncan, Lee, Silvia Wen-Yu, Scarloss, B. & Shapley, K. L. 2007. *Reviewing the evidence on how teacher professional development affects student achievement* (U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest, ed.). Washington, D.C.: Institute of Education Sciences. Zugriff am 08.11.2018. Verfügbar unter http://hub.mspnet.org/media/data/Yoon_et_al_2007_PD_effects.pdf?media_000000008039.pdf
- Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary Educational Psychology*, 25(1), 82–91.
- Zimmerman, D. W. & Zumbo, B. D. (1993). Relative power of the Wilcoxon test, the Friedman test, and repeated-measures ANOVA on ranks. *Journal of Experimental Education*, 62(1), 75–86.

7 Anhang

7.1 Skalenhandbuch

7.1.1 Allgemeine Informationen

7.1.1.1 Geschlecht

Variablenname:	sex
Frage/Instruktion:	Sind Sie weiblich oder männlich?
Anzahl Items:	1
Wertelabels:	1 = weiblich 2 = männlich
Messzeitpunkt:	Prätest

	1 ¹¹⁴	2
weiblich	15	36
männlich	11	26
missings	1	5

7.1.1.2 Alter

Variablenname:	age	
Frage/Instruktion:	Wie alt sind Sie? Kreuzen Sie bitte den entsprechenden 5-Jahre-Zeitraum an.	
Anzahl Items:	1	
Wertelabels:	01 = älter als 65 02 = 61 bis 65 03 = 56 bis 60 04 = 51 bis 55 05 = 46 bis 50	06 = 41 bis 45 07 = 36 bis 40 08 = 31 bis 35 09 = 26 bis 30 10 = jünger als 26
Bemerkung:	Zur Wahrung der Anonymität wurde das Alter lediglich in 5-Jahres-Zeiträumen erfragt.	
Messzeitpunkt:	Prätest	

	1	2
01	0	0
02	0	0
03	9	16
04	1	1
05	2	5

	1	2
07	2	5
08	4	12
09	6	15
10	0	1
missings	0	4

¹¹⁴ Die Ergebnisse werden grundsätzlich getrennt nach Stichproben und Messzeitpunkten berichtet (sofern vorliegend). Die Spalten „1“ und „2“ differenzieren nachfolgend zwischen den Stichproben.

7.1.1.3 Schulform

Variablenname:	sform
Frage/Instruktion:	An welcher Schulform unterrichten Sie?
Anzahl Items:	1
Wertelabels:	01 = Gymnasium 02 = Realschule 03 = Hauptschule 04 = KGS 05 = IGS 06 = sonstiges
Messzeitpunkt:	Prätest

	1	2
Gymnasium	4	21
Realschule	0	2
Hauptschule	0	0
KGS	16	28
IGS	6	11
sonstiges	1	1
missings	0	4

7.1.1.4 Ausbildung

Variablenname:	education
Frage/Instruktion:	Welche Ausbildung haben Sie?
Anzahl Items:	1
Wertelabels:	01 = Abgeschlossenes Studium mit 2. Staatsexamens für Gymnasium 02 = Abgeschlossenes Studium mit 2. Staatsexamens für Haupt-/Realschule 03 = Abgeschlossenes Studium mit 2. Staatsexamens für Grundschullehr- amt 04 = Abgeschlossenes Studium, bisher nur das 1. Staatsexamen 05 = sonstiges
Messzeitpunkt:	Prätest

	1	2
Abgeschlossenes Studium mit 2. Staatsexamen für Gymnasiallehramt	11	28
Abgeschlossenes Studium mit 2. Staatsexamens für Haupt-/ Realschullehramt	15	28
Abgeschlossenes Studium mit 2. Staatsexamen für Grundschullehramt	0	0
Abgeschlossenes Lehramtsstudium, bisher nur das 1. Staatsexamen	1	5
sonstiges	0	1
missing	0	5

7.1.1.5 Mathematik-Studium

Variablenname:	education_m
Frage/Instruktion:	Haben Sie Mathematik als Fach studiert?
Anzahl Items:	1
Wertelabels:	1 = nein 2 = ja
Messzeitpunkt:	Prätest

	1	2
nein	1	4
ja	26	59
missings	0	4

7.1.1.6 Unterrichtserfahrung

Variablenname:	teachage
Frage/Instruktion:	Wie viele Jahre unterrichten Sie schon Mathematik?
Anzahl Items:	1
Wertelabels:	Anzahl der Jahre (kontinuierlich)
Messzeitpunkt:	Prätest

	1	2
Minimum	1.00	0.00
1. Quartil	5.50	4.00
Median	11.00	7.00
Mittelwert	14.89	12.49
SD	12.11	10.81
3. Quartil	23.50	19.00
Maximum	37.00	37.00
missings	0	4

7.1.1.7 Lernstandserhebungen 2009, 2010, 2011, 2012

Variablenname:	LSE_2009, LSE_2010, LSE_2011, LSE_2012
Frage/Instruktion:	Haben Sie schon ein- oder mehrere Male an den Lernstandserhebungen Mathematik in Klasse 8 teilgenommen? Wenn ja, wann?
Anzahl Items:	4
Wertelabels:	1 = nein 2 = ja
Messzeitpunkt:	Prätest

	1				2			
	LSE 2009	LSE 2010	LSE 2011	LSE 2012	LSE 2009	LSE 2010	LSE 2011	LSE 2012
nein	22	24	24	23	54	54	54	54
ja	5	3	3	4	9	9	9	9
missings	0	0	0	0	4	4	4	4

7.1.1.8 Lernstandserhebungen Summenscore

Variablenname:	LSE
Frage/Instruktion:	Summenscore der bisherigen Teilnahmen
Anzahl Items:	1

	1	2
0	16	39
1	8	15
2	2	6
3	1	3
4	0	0

7.1.2 Skalen zu den Lernstandserhebungen

7.1.2.1 Allgemeine Akzeptanz von Lernstandserhebungen

Variablenname:	autove
Quelle:	Maier (2008a)
Anzahl Items:	6
Wertelabels:	1 = stimmt gar nicht 2 = stimmt kaum 3 = stimmt teilweise 4 = stimmt überwiegend 5 = stimmt genau
Bemerkungen:	Item autove6 ist invertiert. (-)

	Stichprobe 1	Prätest			Posttest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
autove1	...sollte jedes Jahr verbindlich durchgeführt werden.	.60	2.77	.89	.34	3.40	.76
autove2	...ist für die Arbeit der Schulen sehr wichtig.	.71	3.14	.87	.52	3.40	.69
autove3	...trägt dazu bei, dass man sich in den Schulen mehr bemüht.	.70	2.55	.87	.65	3.12	.65
autove4	...gibt eine objektive Basis ab, um zu sehen, wo eine Schule steht.	.63	3.00	.88	.51	3.24	.71
autove5	...ist eine wichtige Grundlage, um Unterricht weiterentwickeln zu können.	.85	3.64	.84	.68	3.88	.66
autove6	...nützt für meine eigentliche Arbeit als Lehrer wenig. (-)	.83	3.55	.85	.31	3.96	.75
Skala		$M =$	3.11		$M =$	3.50	
		$SD =$	0.90		$SD =$	0.89	
		$\alpha =$	.89		$\alpha =$	.74	
		$N =$	22		$N =$	25	

Anhang

	Stichprobe 2	Prätest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
autove1	...sollte jedes Jahr verbindlich durchgeführt werden.	.58	2.56	.80
autove2	...ist für die Arbeit der Schulen sehr wichtig.	.64	2.96	.77
autove3	...trägt dazu bei, dass man sich in den Schulen mehr bemüht.	.49	2.46	.80
autove4	...gibt eine objektive Basis ab, um zu sehen, wo eine Schule steht.	.44	2.90	.81
autove5	...ist eine wichtige Grundlage, um Unterricht weiterentwickeln zu können.	.72	3.34	.75
autove6	...nützt für meine eigentliche Arbeit als Lehrer wenig. (-)	.68	3.48	.76
Skala		$M =$	2.95	
		$SD =$	0.74	
		$\alpha =$	.81	
		$N =$	50	

7.1.2.2 Vergleichsarbeiten als Belastung

Variablenname:	verba
Quelle:	Maier (2008a) verba6 Eigenentwicklung
Anzahl Items:	6
Wertelabels:	1 = stimmt gar nicht 2 = stimmt kaum 3 = stimmt teilweise 4 = stimmt überwiegend 5 = stimmt genau

	Stichprobe 1	Prätest			Posttest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
verba1	...führt zu Konkurrenz und Missgunst innerhalb der Schulen.	.25	1.60	.73	.70	1.72	.69
verba2	...übt zusätzlichen Druck auf Lehrer und Schüler aus.	.76	2.56	.58	.60	2.64	.69
verba3	...führt dazu, dass nur noch für den Test geübt wird.	.48	1.68	.68	.62	1.48	.70
verba4	...bringt Unruhe in die Schulen.	.42	2.20	.70	.44	2.16	.73
verba5	...schafft mehr Probleme als sie nützt.	.51	1.88	.67	.40	1.72	.75
verba6	...führt zu Konkurrenz zwischen verschiedenen Schulen.	.36	2.68	.72	.40	2.44	.75
Skala		$M =$	2.10		$M =$	2.03	
		$SD =$	0.70		$SD =$	0.65	
		$\alpha =$	.72		$\alpha =$	.75	
		$N =$	25		$N =$	25	

Anhang

	Stichprobe 2	Prätest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
verba1	...führt zu Konkurrenz und Missgunst innerhalb der Schulen.	.60	2.04	.79
verba2	...übt zusätzlichen Druck auf Lehrer und Schüler aus.	.71	2.81	.77
verba3	...führt dazu, dass nur noch für den Test geübt wird.	.59	1.94	.80
verba4	...bringt Unruhe in die Schulen.	.55	2.42	.80
verba5	...schafft mehr Probleme als sie nützt.	.58	2.34	.80
verba6	...führt zu Konkurrenz zwischen verschiedenen Schulen.	0.52	2.75	0.81
Skala		$M =$	2.39	
		$SD =$	0.82	
		$\alpha =$	.85	
		$N =$	53	

7.1.2.3 Lehrplanvalidität der Lernstandserhebungen

Variablenname:	lerva
Quelle:	Maier (2008a)
Anzahl Items:	4
Wertelabels:	1 = stimmt gar nicht 2 = stimmt kaum 3 = stimmt teilweise 4 = stimmt überwiegend 5 = stimmt genau

	Stichprobe 1	Prätest			Posttest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
lerva1	...deckt mit ihren Aufgaben die im Kerncurriculum vorgegebenen Lerninhalte und Kompetenzen im Fach Mathematik gut ab.	.37	4.00	.69	.43	3.96	.80
lerva2	...stimmt in ihren Teilbereichen mit der Gewichtung der Lerninhalte und Kompetenzen im Kerncurriculum überein.	.57	3.53	.56	.74	4.04	.64
lerva3	...hat eine Punkteverteilung, die der Gewichtung der Lerninhalte und Kompetenzen im Kerncurriculum.	.41	3.65	.67	.48	3.76	.78
lerva4	...entspricht dem im Kerncurriculum geforderten Niveau.	.57	3.59	.58	.73	3.80	.65
Skala		$M =$	3.69		$M =$	3.89	
		$SD =$	0.51		$SD =$	0.62	
		$\alpha =$	.69		$\alpha =$	.78	
		$N =$	17		$N =$	25	

Anhang

	Stichprobe 2	Prätest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
lerva1	...deckt mit ihren Aufgaben die im Kerncurriculum vorgegebenen Lerninhalte und Kompetenzen im Fach Mathematik gut ab.	.43	3.75	.67
lerva2	...stimmt in ihren Teilbereichen mit der Gewichtung der Lerninhalte und Kompetenzen im Kerncurriculum überein.	.58	3.44	.57
lerva3	...hat eine Punkteverteilung, die der Gewichtung der Lerninhalte und Kompetenzen im Kerncurriculum.	.42	3.61	.68
lerva4	...entspricht dem im Kerncurriculum geforderten Niveau.	.52	3.44	.62
Skala		$M =$	3.56	
		$SD =$	0.54	
		$\alpha =$	.70	
		$N =$	36	

7.1.2.4 Lernstandserhebungen als Unterstützung für Lerndiagnose, Beratung und Förderung (förderdiagnostischer Nutzen)

Variablenname:	ulebe
Quelle:	Maier (2008a)
Anzahl Items:	8
Wertelabels:	1 = stimmt gar nicht 2 = stimmt kaum 3 = stimmt teilweise 4 = stimmt überwiegend 5 = stimmt genau
Bemerkung:	Item ulebe7 invertiert

	Stichprobe 1	Prätest			Posttest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
ulebe1	...ist ein guter Anhaltspunkt, um die Leistung einzelner Schüler einschätzen zu können.	.49	3.24	.83	.78	3.56	.79
ulebe2	...bietet eine gute Grundlage zur Planung von Fördermaßnahmen für schwächere Schüler.	.68	3.56	.80	.69	3.68	.80
ulebe3	...ist eine tragfähige Argumentationsbasis für Beratungsgespräche mit Eltern.	.67	3.36	.80	.74	3.52	.79
ulebe4	...hilft mir, die Stärken und Schwächen der Schüler deutlicher benennen zu können.	.58	3.88	.81	.81	4.08	.78
ulebe5	...gibt mir zusätzliche diagnostische Hinweise.	.72	3.68	.80	.65	3.92	.81
ulebe6	...ist eine gute Gesamtbeurteilung der Schülerleistung.	.50	2.96	.83	.54	3.36	.82
ulebe7	...spiegelt die Leistungsfähigkeit meiner Schüler nicht so gut wider, wie meine eigenen Klassenarbeiten. (-)	.30	3.12	.85	-.12	3.48	.89
ulebe8	...regt zu Diskussionen im Kollegium über den zukünftigen Unterricht an.	.61	3.44	.81	.59	3.28	.81
Skala		$M =$	3.41		$M =$	3.61	
		$SD =$	0.65		$SD =$	0.64	
		$\alpha =$	.84		$\alpha =$	.83	
		$N =$	25		$N =$	25	

Anhang

	Stichprobe 2	Prätest		
	Die Lernstandserhebung...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
ulebe1	...ist ein guter Anhaltspunkt, um die Leistung einzelner Schüler einschätzen zu können.	.56	3.14	.77
ulebe2	...bietet eine gute Grundlage zur Planung von Fördermaßnahmen für schwächere Schüler.	.58	3.39	.76
ulebe3	...ist eine tragfähige Argumentationsbasis für Beratungsgespräche mit Eltern.	.68	3.29	.75
ulebe4	...hilft mir, die Stärken und Schwächen der Schüler deutlicher benennen zu können.	.69	3.80	.75
ulebe5	...gibt mir zusätzliche diagnostische Hinweise.	.63	3.71	.76
ulebe6	...ist eine gute Gesamtbeurteilung der Schülerleistung.	.46	2.80	.78
ulebe7	...spiegelt die Leistungsfähigkeit meiner Schüler nicht so gut wider, wie meine eigenen Klassenarbeiten. (-)	.13	3.22	.83
ulebe8	...regt zu Diskussionen im Kollegium über den zukünftigen Unterricht an.	.40	3.24	.79
Skala		$M =$	3.32	
		$SD =$	0.62	
		$\alpha =$	.80	
		$N =$	49	

7.1.2.5 Lernstandserhebungen und Hinweise für zukünftige Unterrichtsgestaltung (unterrichtsplanerischer Nutzen)

Variablenname:	hizu
Quelle:	Maier (2008a)
Anzahl Items:	7
Wertelabels:	1 = stimmt gar nicht 2 = stimmt kaum 3 = stimmt teilweise 4 = stimmt überwiegend 5 = stimmt genau

	Stichprobe 1	Prätest			Posttest		
	Die Lernstandserhebung gibt mir Hinweise darüber, ...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
hizu1	...welche Inhalte in Zukunft verstärkt behandelt werden sollten.	.75	3.32	.87	.56	3.58	.85
hizu2	...welche Aufgabenstellungen in Zukunft besser geübt werden müssen.	.79	3.68	.87	.70	3.85	.83
hizu3	...welche Lerninhalte häufiger wiederholt werden müssen.	.73	3.41	.87	.81	3.85	.82
hizu4	...ob die Reihenfolge der behandelten Stoffgebiete geändert werden sollte.	.53	2.23	.90	.52	2.35	.86
hizu5	...welche neuen Aufgabenstellungen ich in meinen Unterricht einbinden muss.	.78	3.77	.88	.60	3.65	.85
hizu6	...welche Stoffgebiete neu aufgenommen werden sollten.	.77	3.36	.87	.69	3.38	.84
hizu7	...welche Lernmaterialien (z.B. Bücher, Arbeitshefte, ...) für meinen zukünftigen Unterricht gut geeignet sind.	.60	2.50	.89	.57	3.12	.85
Skala		$M =$	3.18		$M =$	3.40	
		$SD =$	0.78		$SD =$	0.74	
		$\alpha =$	.89		$\alpha =$	.86	
		$N =$	22		$N =$	26	

Anhang

	Stichprobe 2	Prätest		
	Die Lernstandserhebung gibt mir Hinweise darüber, ...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
hizu1	...welche Inhalte in Zukunft verstärkt behandelt werden sollten.	.63	3.32	.81
hizu2	...welche Aufgabenstellungen in Zukunft besser geübt werden müssen.	.71	3.57	.79
hizu3	...welche Lerninhalte häufiger wiederholt werden müssen.	.72	3.45	.80
hizu4	...ob die Reihenfolge der behandelten Stoffgebiete geändert werden sollte.	.51	2.36	.83
hizu5	...welche neuen Aufgabenstellungen ich in meinen Unterricht einbinden muss.	.42	3.70	.84
hizu6	...welche Stoffgebiete neu aufgenommen werden sollten.	.66	3.30	.80
hizu7	...welche Lernmaterialien (z.B. Bücher, Arbeitshefte, ...) für meinen zukünftigen Unterricht gut geeignet sind.	.48	2.38	.83
Skala		$M =$	3.16	
		$SD =$	0.68	
		$\alpha =$	.84	
		$N =$	47	

7.1.2.6 Lernstandserhebung zur Unterstützt der Benotung und Selektion (selektionsdiagnostischer Nutzen)

Variablenname:	besel
Quelle:	Maier (2008a)
Anzahl Items:	5
Wertelabels:	1 = stimmt gar nicht 2 = stimmt kaum 3 = stimmt teilweise 4 = stimmt überwiegend 5 = stimmt genau

	Stichprobe 1	Prätest			Posttest		
	Die Ergebnisse der Lernstandserhebung ...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
besel1	...tragen auch zur Begründung der Jahresendnote bei.	.77	1.74	.67	.43	2.27	.70
besel2	...in die Jahresendnote mit einfließen zu lassen ist sinnvoll.	.46	2.13	.77	.69	2.42	.59
besel3	...regen zum Nachdenken über den eigenen Bewertungsmaßstab an.	.36	3.61	.81	.38	3.58	.72
besel4	...geben Hinweise darauf, ob die eigenen Klassenarbeiten zu schwer oder zu leicht sind.	.56	2.96	.73	.57	3.15	.65
besel5	...sind bei unsicheren Versetzungsentscheidungen ein zusätzliches Entscheidungskriterium.	.69	1.96	.69	.37	2.31	.72
Skala		$M =$	2.48		$M =$	2.75	
		$SD =$	0.75		$SD =$	0.73	
		$\alpha =$	.78		$\alpha =$	.73	
		$N =$	23		$N =$	26	

Anhang

	Stichprobe 2	Prätest		
	Die Ergebnisse der Lernstandserhebung ...	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
besel1	...tragen auch zur Begründung der Jahresendnote bei.	.71	1.94	.66
besel2	...in die Jahresendnote mit einfließen zu lassen ist sinnvoll.	.53	2.12	.73
besel3	...regen zum Nachdenken über den eigenen Bewertungsmaßstab an.	.34	3.44	.80
besel4	...geben Hinweise darauf, ob die eigenen Klassenarbeiten zu schwer oder zu leicht sind.	.41	2.75	.77
besel5	...sind bei unsicheren Versetzungsentscheidungen ein zusätzliches Entscheidungskriterium.	.75	2.04	.65
Skala		$M =$	2.46	
		$SD =$	0.76	
		$\alpha =$	.77	
		$N =$	52	

7.1.3 Skalen zur Unterrichtsqualität

7.1.3.1 Kognitiv herausfordernde Aufgaben

Variablenname:	mauko
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie häufig verwenden Sie in Ihrer Lernstandserhebungs-Klasse die folgenden Vorgehensweisen und Aufgaben?
Anzahl Items:	8
Wertelabels:	1 = selten oder nie; 2 = manchmal; 3 = häufiger; 4 = regelmäßig
Bemerkungen:	Im Stimulus wurde „die in der fachdidaktischen Literatur gelegentlich empfohlen werden“ zur Verringerung des Risikos sozialer Erwünschtheit gestrichen.

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mauko1	Ich lasse Aufgaben bearbeiten, für die es keinen sofort erkennbaren Lösungsweg gibt.	.69	2.64	.82	.68	2.43	.80
mauko2	Ich stelle Fragen, die man nicht spontan beantworten kann, sondern die zum Nachdenken zwingen.	.60	2.84	.83	.46	2.76	.83
mauko3	Ich lasse unterschiedliche Lösungswege von Aufgaben vergleichen und bewerten.	.63	2.92	.82	.54	2.81	.82
mauko4	Ich stelle Aufgaben, bei denen es nicht allein auf das Rechnen, sondern vor allem auf den richtigen Ansatz ankommt.	.60	3.16	.83	.49	3.10	.82
mauko5	Ich stelle Aufgaben, bei denen es nicht allein auf die richtige Lösung, sondern vor allem auf das Verständnis des mathematischen Gedankengangs ankommt.	.60	2.96	.83	.67	3.00	.80
mauko6	Ich stelle Aufgaben, die keine eindeutige Lösung haben, und lasse dies erklären.	.46	2.48	.84	.66	2.33	.80
mauko7	Ich stelle Aufgaben, für deren Lösung man Zeit zum Nachdenken braucht.	.62	2.76	.83	.47	2.71	.83
mauko8	Ich stelle auch aufgaben, bei denen man mehrere Lösungswege zeigen muss.	.56	2.52	.84	.55	2.29	.82
Skala		$M =$	1.79		$M =$	2.68	
		$SD =$	0.50		$SD =$	0.47	
		$\alpha =$	.85		$\alpha =$	.83	
		$N =$	25		$N =$	21	

7.1.3.2 Insistieren auf Erklärungen und Begründungen

Variablenname:	maub
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie häufig verwenden Sie in Ihrer Lernstandserhebungs-Klasse die folgenden Vorgehensweisen und Aufgaben?
Anzahl Items:	4
Wertelabels:	1 = selten oder nie 2 = manchmal 3 = häufiger 4 = regelmäßig
Bemerkungen:	Der Stimulus wurde gegenüber der Quelle um den Zusatz „, die in der fachdidaktischen Literatur gelegentlich empfohlen werden“ verkürzt, um das Risiko sozialer Erwünschtheit in den Antworten zu verringern.

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
maub1	Ich halte Schüler/innen dazu an, ihre Gedankengänge genau zu erklären.	.74	3.08	.59	.54	3.00	.57
maub2	Ich verlange, dass Schüler/innen ihre Arbeitsschritte ausführlich begründen.	.52	3.16	.70	.54	3.04	.53
maub3	Bei mir wissen die Schüler/innen, dass sie ihre Aussagen immer begründen müssen, auch wenn ich es nicht extra dazusage.	.64	2.72	.63	.50	2.58	.56
maub4	In jeder Klassenarbeit bringe ich mindestens eine Aufgabe, in der eine Begründung gefordert ist.	.32	3.48	.80	.30	3.17	.73
Skala		$M =$	3.11		$M =$	2.96	
		$SD =$	0.63		$SD =$	0.51	
		$\alpha =$	.74		$\alpha =$	.66	
		$N =$	25		$N =$	24	

7.1.3.3 Kognitiv herausfordernder Umgang mit Schülerbeiträgen (Ideen/Fehler)

Variablenname:	mside
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie sehen Sie Ihre Rolle bei der <i>Erarbeitung eines neuen Sachverhalts?</i>
Anzahl Items:	4
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mside1	Ich gehe von den Ideen der Schüler/innen aus und spiele mit ihnen die Konsequenzen durch, bis die Schüler/innen erkennen, ob ihre Gedanken zum Ziel führen oder Ungeheimtheiten offenkundig werden.	.61	2.69	.14	.60	2.92	.28
mside2	Wenn ein/eine Schüler/in bei der Erarbeitung eines Sachverhalts einen Fehler macht, nehme ich die Vorschläge zunächst ohne Korrektur an und verfolge gemeinsam mit den Schüler/innen die Konsequenzen, bis der Fehler offenkundig wird.	.23	3.08	.52	.32	3.00	.66
mside3	Ich lasse die Schüler/innen auch einmal bewusst in die Irre laufen, bis sie sehen, dass etwas nicht stimmen kann.	.14	3.08	.59	.41	2.75	.58
mside4	Ich gehe von Vorstellungen der Schüler/innen aus und spiele mit ihnen die Konsequenzen durch, bis die Schüler/innen verblüffende Ungeheimtheiten erkennen.	.34	2.58	.44	.60	2.92	.28
Skala		$M =$	2.86		$M =$	2.17	
		$SD =$	0.43		$SD =$	0.29	
		$\alpha =$	.53		$\alpha =$	.62	
		$N =$	26		$N =$	24	

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mside1	Ich gehe von den Ideen der Schüler/innen aus und spiele mit ihnen die Konsequenzen durch, bis die Schüler/innen erkennen, ob ihre Gedanken um Ziel führen oder Ungeheimheiten offenkundig werden.	.59	2.69	.15	.60	2.92	.28
mside2	Wenn ein/eine Schüler/in bei der Erarbeitung eines Sachverhalts einen Fehler macht, nehme ich die Vorschläge zunächst ohne Korrektur an und verfolge gemeinsam mit den Schüler/innen die Konsequenzen, bis der Fehler offenkundig wird.	.15	3.08	.76	.32	3.00	.66
mside4	Ich gehe von Vorstellungen der Schüler/innen aus und spiele mit ihnen die Konsequenzen durch, bis die Schüler/innen verblüffende Ungeheimheiten erkennen.	.53	2.58	.29	.41	2.75	.58
Skala		$M =$	2.78		$M =$	2.88	
		$SD =$	0.50		$SD =$	0.39	
		$\alpha =$	.59		$\alpha =$	.62	
		$N =$	26		$N =$	23	

7.1.3.4 Unterstützung kognitiver Selbstständigkeit

Variablenname:	mkog
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Bitte lesen Sie die folgenden Aussagen zu <i>Unterrichtstempo</i> und <i>Erklären</i> durch und kreuzen Sie an, wie sehr jede einzelne Aussage auf Sie zutrifft.
Anzahl Items:	4
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mkog1	Bei mir können Schüler/innen Mathematikaufgaben so lösen, wie sie es persönlich für richtig halten.	.16	3.27	.76	.46	3.04	.66
mkog2	Bei mir dürfen die Schüler/innen zum Lösen von Aufgaben ihre eigenen Strategien einsetzen.	.48	3.85	.47	.39	3.79	.70
mkog3	Wenn im Mathematikunterricht Aufgaben bearbeitet werden, warte ich erst ab, ob die Schüler/innen selbst auf die richtige Lösung kommen.	.41	3.77	.50	.74	3.38	.44
mkog4	Wenn ich Schüler/innen bei einer Aufgabe helfe, ermuntere ich sie, erst einmal selbst die richtige Lösung zu finden.	.61	3.69	.34	.42	3.75	.68
Skala		$M =$	3.64		$M =$	3.49	
		$SD =$	0.33		$SD =$	0.38	
		$\alpha =$	.59		$\alpha =$	.70	
		$N =$	26		$N =$	24	

7.1.3.5 Kognitiv herausforderndes Üben

Variablenname:	muher
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie gehen Sie beim Durcharbeiten, Üben und Wiederholen vor?
Anzahl Items:	8
Wertelabels:	1 = selten oder nie 2 = manchmal 3 = häufiger 4 = regelmäßig

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
muher1	Ich gebe bewusst Übungsaufgaben, die die Grenzen des Erarbeiteten zeigen, auch wenn Schüler/innen dadurch verunsichert werden können.	.61	2.29	.74	.70	2.59	.74
muher2	Ich gebe regelmäßig auch Aufgaben mit wechselnder mathematischer Struktur, um Verarbeitungstiefe und Verstehenssicherheit zu verbessern.	.25	2.92	.79	.74	3.00	.73
muher3	Ich übe oft mit Aufgaben, die eine Anwendung des Gelernten auf neue Gebiete erfordern.	.69	2.83	.72	.34	2.91	.80
muher4	Beim Üben verwende ich Aufgaben, die eine Anwendung des Gelernten auf neue Gebiete erfordern.	.57	3.21	.75	.48	3.32	.78
muher5	Beim Üben verwende ich Aufgaben, bei denen die Grenzen der Gültigkeit des Erarbeiteten sichtbar werden, auch wenn Schüler/innen dadurch verunsichert werden können.	.56	2.46	.75	.69	2.68	.74
muher6	Beim Üben lege ich Aufgaben in eingekleideter Form vor.	.36	2.92	.78	.16	2.82	.81
muher7	Zum Üben verwende ich Aufgaben, die das mathematische Verständnis prüfen.	.67	3.04	.73	.41	3.14	.79
muher8	Ich verwende Aufgaben, bei denen die Schüler/innen auch beim Üben noch etwas Neues hinzulernen.	.18	3.00	.80	.54	3.18	.77
Skala		$M =$	2.83		$M =$	2.95	
		$SD =$	0.37		$SD =$	0.41	
		$\alpha =$	.78		$\alpha =$	.80	
		$N =$	24		$N =$	22	

7.1.3.6 Fehlervermeidende Erarbeitung durch den Lehrer selbst

Variablenname:	msfeh
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie sehen Sie Ihre Rolle bei der <i>Erarbeitung eines neuen Sachverhaltes?</i>
Anzahl Items:	5
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
msfeh1	Beweise und Ableitungen führe ich selbst an der Tafel vor, um für größere Klarheit zu sorgen.	.42	2.54	.88	.32	2.61	.79
msfeh2	An schwierigen Stellen trage ich selbst vor, um keine falschen Vorstellungen entstehen zu lassen.	.75	2.46	.78	.74	2.48	.65
msfeh3	Die Klärung von Sachverhalten, die für das Verständnis besonders wichtig sind, nehme ich selbst vor.	.74	2.65	.79	.39	2.57	.76
msfeh4	Wenn die Schüler/innen bei der Erarbeitung eines neuen Sachverhalts falsche Vorschläge machen oder falsche Antworten geben, korrigiere ich selbst, um keine falschen Vorstellungen entstehen zu lassen.	.80	2.23	.76	.61	2.35	.69
msfeh5	Zusammenhänge zwischen Sachverhalten stelle ich im (kurzen) Lehrervortrag dar.	.61	2.77	.82	.64	2.52	.68
Skala		$M =$	2.53		$M =$	2.50	
		$SD =$	0.57		$SD =$	0.55	
		$\alpha =$	.84		$\alpha =$	.76	
		$N =$	26		$N =$	23	

7.1.3.7 Fürsorgliche Anleitung und Kontrolle

Variablenname:	manle
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Bitte lesen Sie die folgenden Aussagen zu <i>Unterrichtstempo</i> und <i>Erklären</i> durch und kreuzen Sie an, wie sehr jede einzelne Aussage auf Sie zutrifft.
Anzahl Items:	6
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
manle1	Ich sage den Schüler/innen ausführlich, wie die Mathematikhausaufgaben zu machen sind.	.33	2.31	.53	.25	2.39	.60
manle2	Ich erkläre den Schüler/innen immer ganz genau, was sie tun müssen.	.12	3.04	.62	.51	2.96	.54
manle3	Ich informiere die Schüler/innen genau über den bevorstehenden Unterrichtsstoff.	.17	3.31	.59	.32	3.65	.58
manle4	Ich erkläre den Schüler/innen immer genau, warum sie etwas tun müssen.	.45	2.69	.48	.49	2.61	.51
manle5	Ich gehe in Mathematik im Stoff immer genau Schritt für Schritt vor.	.64	2.96	.39	.50	2.91	.50
manle6	Ich gebe den Schüler/innen immer wieder Ratschläge, wie sie den Stoff in Mathematik am besten lernen können.	.26	3.12	.57	.15	3.17	.66
Skala		$M =$	2.90		$M =$	2.95	
		$SD =$	0.39		$SD =$	0.35	
		$\alpha =$	.58		$\alpha =$	.61	
		$N =$	26		$N =$	23	

7.1.3.8 Einschleifendes, repetitives Üben

Variablenname:	murep
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie gehen Sie beim <i>Durcharbeiten</i> , <i>Üben</i> und <i>Wiederholen</i> vor?
Anzahl Items:	5
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
murep1	Ich lasse möglichst viele Aufgaben in selbstständiger Stillarbeit während der Unterrichtsstunden lösen, um die Verfahren einzuschleifen.	.80	2.29	.78	.81	2.57	.78
murep2	Ich lasse so viele Aufgaben wie möglich aus dem Lehrbuch während des Unterrichts lösen, um die Verfahren einzuschleifen.	.75	2.17	.80	.63	2.39	.83
murep3	Um das gelernte Verfahren möglichst gut zu festigen, lasse ich relativ viele Aufgaben lösen, die gegenüber der Lernsituation nur wenig geändert sind (z. B. andere Zahlen).	.73	2.29	.80	.69	2.26	.81
murep4	Ich lasse intensiv üben, auch wenn ich dadurch etwas langsamer vorankomme.	.48	2.88	.86	.63	2.78	.83
murep5	Ich wiederhole intensiv, auch wenn die schnelleren Schüler/innen murren.	.58	2.25	.84	.59	2.43	.84
Skala		$M =$	2.37		$M =$	2.49	
		$SD =$	0.65		$SD =$	0.55	
		$\alpha =$	.85		$\alpha =$	.85	
		$N =$	24		$N =$	23	

7.1.3.9 Aktive soziale Konstruktion von Verständnis

Variablenname:	msozk
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie häufig verwenden Sie die folgenden <i>Lehr-</i> und <i>Sozialformen</i> in Ihrer Lernstandserhebungs-Klasse?
Anzahl Items:	8
Wertelabels:	1 = selten oder nie 2 = manchmal 3 = häufiger 4 = regelmäßig

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
msozk1	...gliedere ich Probleme in einzelne Aufgabenstellungen auf, die ich in Gruppen bearbeiten und anschließend zusammenfassen lasse.	.69	2.04	.87	.78	2.25	.84
msozk2	...schaffe ich Gelegenheiten zu sachbezogenen Diskussionen über mathematische Problemstellungen/Themen.	.79	2.67	.86	.56	2.50	.87
msozk3	...lasse ich die Schüler/innen Sachzusammenhänge, die sie vorbereitet haben, präsentieren (z. B. Erträge aus Selbststudium, Gruppen-, Projektarbeit usw.).	.60	2.33	.88	.64	2.42	.86
msozk4	...behandle ich ein komplexes Thema auch über längere Zeit in Projekten.	.75	1.63	.87	.67	1.88	.85
msozk5	...diskutieren die Schüler/innen über ihre Lösungsideen oder Lösungswege.	.63	3.08	.88	.59	2.96	.86
msozk6	...verwende ich auch gruppenteiliges Arbeiten.	.71	2.75	.87	.76	2.67	.85
msozk7	...schaffe ich „offene Situationen“, in denen die Schüler/innen eigene Problemstellungen/Aufgaben entwickeln können.	.58	2.17	.88	.56	2.21	.87
msozk8	...mache ich Erkundungen außerhalb der Schule, um die praktische Anwendbarkeit von Mathematik zu zeigen.	.59	1.54	.88	.54	1.54	.87
Skala		$M =$	2.28		$M =$	2.30	
		$SD =$	0.61		$SD =$	0.58	
		$\alpha =$	.89		$\alpha =$	.87	
		$N =$	24		$N =$	24	

7.1.3.10 Tutorielle Unterstützung von Lernpartnerschaften und Gruppen

Variablenname:	mlern
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie häufig verwenden Sie die folgenden <i>Lehr-</i> und <i>Sozialformen</i> in Ihrer Lernstandserhebungs-Klasse?
Anzahl Items:	6
Wertelabels:	1 = selten oder nie 2 = manchmal 3 = häufiger 4 = regelmäßig

	Stichprobe 1 In meinem Mathematikunter- richt in der LSE-Klasse...	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mlern1	...setze ich Schüler/innen als Lernhelfer/innen ein.	.67	2.88	.66	.59	2.79	.73
mlern2	...führe ich Partnerarbeit durch.	.45	3.44	.73	.55	3.38	.74
mlern3	...bestimmen die Schüler/innen selbst, ob sie allein, paarweise oder in Kleingruppen arbeiten.	.63	2.76	.67	.40	2.67	.78
mlern4	...halte ich die Schüler/innen dazu an, zuerst ihre Mitschüler/innen um Hilfe zu bitten, bevor sie zu mir kommen.	.44	3.16	.73	.56	3.21	.74
mlern5	...lasse ich stärkere Schüler/innen schwächere Mitschüler/innen bei Aufgaben im Unterricht unterstützen.	.56	3.20	.72	.63	3.13	.72
mlern6	...arbeiten die Schüler/innen in festen Lernpartnerschaften.	.30	1.96	.77	.47	2.04	.77
Skala		$M =$	2.90			$M =$	2.87
		$SD =$	0.54			$SD =$	0.55
		$\alpha =$	.75			$\alpha =$	.78
		$N =$	25			$N =$	24

7.1.3.11 Differenzierung durch variierende Aufgabenschwierigkeit

Variablenname:	muvar	
Quelle:	Baumert et al. (2008)	
Frage/Instruktion:	muvar1, muvar2, muvar3, muvar4: Wie gehen Sie beim <i>Durcharbeiten</i> , <i>Üben</i> und <i>Wiederholen</i> vor? muvar5, muvar6: Wie häufig verwenden Sie beim <i>Üben</i> und <i>Wiederholen</i> die folgenden <i>Arbeitsformen</i> ?	
Anzahl Items:	6	
Wertelabels:	muvar1, muvar2, muvar3, muvar4: 1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu	muvar5, muvar6: 1 = selten oder nie 2 = manchmal 3 = häufiger 4 = regelmäßig

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
muvar1	Bei der Stillarbeit variere ich die Aufgabenstellungen, um Schüler/innen unterschiedlicher Leistungsstärke gerecht zu werden.	.51	3.04	.64	.56	3.08	.71
muvar2	Ich gebe Schüler/innen je nach Leistung unterschiedlich schwere Hausaufgaben.	.34	1.91	.68	.60	2.29	.69
muvar3	Schnellen Schüler/innen gebe ich gern Extraaufgaben, durch die sie wirklich gefordert werden.	.32	3.35	.68	.39	3.04	.75
muvar4	Von Schüler/innen mit guten Leistungen verlange ich deutlich mehr.	.29	2.96	.70	.49	2.75	.72
muvar5	Bei Gruppenarbeit unterscheide ich verschiedene Leistungsgruppen, die jeweils gesonderte Aufgaben bekommen.	.56	2.35	.60	.52	2.38	.72
muvar6	Bei der Stillarbeit gebe ich gern Aufgaben, die leistungsstarke und leistungsschwächere Schüler/innen auf unterschiedliche Weise lösen können.	.59	2.65	.60	.45	2.88	.74
Skala		$M =$	2.71		$M =$	2.74	
		$SD =$	0.47		$SD =$	0.51	
		$\alpha =$	.69		$\alpha =$	.76	
		$N =$	23		$N =$	24	

7.1.3.12 Persönliches Vertrauen

Variablenname:	mverp
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie beurteilen Sie Ihre <i>Beziehung</i> zu Ihrer Lernstandserhebungs-Klasse? Inwieweit treffen folgende Aussagen für Ihren Unterricht in dieser Klasse zu?
Anzahl Items:	4
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mverp1	Ich kümmere mich um meine Schüler/innen, wenn sie Probleme haben.	.84	3.58	.87	-.28	2.29	.65
mverp2	Ich baue Vertrauen zu meinen Schüler/innen auf.	.93	3.69	.85	.47	2.67	-.50
mverp3	Ich zeige Verständnis für meine Schüler/innen.	.64	3.65	.94	.41	3.42	.23
mverp4	Ich nehme mir Zeit, wenn meine Schüler/innen etwas mit mir bereden möchten.	.82	3.62	.88	.06	3.67	.23
Skala		$M =$	3.63		$M =$	3.01	
		$SD =$	0.47		$SD =$	0.35	
		$\alpha =$	.91		$\alpha =$	.22	
		$N =$	26		$N =$	24	

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mverp2	Ich baue Vertrauen zu meinen Schüler/innen auf.	.94	3.69	.67	.51	2.67	.47
mverp3	Ich zeige Verständnis für meine Schüler/innen.	.63	3.65	.93	.57	3.42	.40
mverp4	Ich nehme mir Zeit, wenn meine Schüler/innen etwas mit mir bereden möchten.	.74	3.62	.85	.32	3.67	.72
Skala		$M =$	3.65		$M =$	3.25	
		$SD =$	0.46		$SD =$	0.47	
		$\alpha =$	.87		$\alpha =$	.65	
		$N =$	26		$N =$	24	

7.1.3.13 *Diagnostische Kompetenz im sozialen Bereich*

Variablenname:	mdias
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie sicher sind Sie in Ihrer Schülerdiagnose?
Anzahl Items:	4
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mdias1	Ich sehe sehr schnell, wenn es zwischen Schüler/innen Streit gegeben hat.	.72	3.00	.72	.72	2.92	.21
mdias2	Ich sehe schnell, wenn zwischen Banknachbarn etwas nicht stimmt.	.46	3.27	.84	.01	3.21	.75
mdias3	Ich merke ziemlich schnell, wenn ein/eine Schüler/in Kummer hat.	.63	3.04	.77	.68	3.08	.31
mdias4	Ich spüre genau, wenn jemand richtig traurig ist.	.73	3.04	.71	.27	3.13	.63
Skala		$M =$	3.09		$M =$	3.08	
		$SD =$	0.50		$SD =$	0.51	
		$\alpha =$	.81		$\alpha =$	.62	
		$N =$	26		$N =$	24	

7.1.3.14 Diagnose im Leistungsbereich

Variablenname:	mdial
Quelle:	Baumert et al. (2008)
Frage/Instruktion:	Wie sicher sind Sie in Ihrer Schülerdiagnose?
Anzahl Items:	5
Wertelabels:	1 = trifft nicht zu 2 = trifft eher nicht zu 3 = trifft eher zu 4 = trifft zu

	Stichprobe 1	Prätest			Posttest		
		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$	r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
mdial1	Es fällt mir leicht festzustellen, ob ein/eine Schüler/in eine Aufgabe verstanden hat.	.61	3.08	.57	.22	3.04	-.35
mdial2	Ich merke sehr schnell, wenn jemand etwas nicht verstanden hat.	.68	3.23	.58	.05	3.00	-.21
mdial3	Ich weiß, bei welchen Aufgaben die einzelnen Schüler/innen Schwierigkeiten haben.	.40	2.85	.67	-.29	1.96	.33
mdial4	Ich kenne die Stärken und Schwächen der einzelnen Schüler/innen.	.43	3.12	.65	.17	2.96	-.43
Skala		$M =$	3.08		$M =$	2.82	
		$SD =$	0.45		$SD =$	0.28	
		$\alpha =$	.69		$\alpha =$	-.09	
		$N =$	26		$N =$	24	

7.1.4 Online-Fragebogen

7.1.4.1 Online-Skalen

7.1.4.1.1 Nützlichkeit

Variablenname:	bewt
Frage/Instruktion:	Inwieweit stimmen Sie folgenden Aussagen zu?
Anzahl Items:	6
Wertelabels:	1 = ++ 2 = + 3 = o 4 = - 5 = --

		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
bewt_prof	Die Schüler/-innen profitieren von der Durchführung der LSE.	.76	3.17	.92
bewt_paed	Von den LSE sind wichtige Impulse für die pädagogische Diskussion in meiner Schule ausgegangen.	.73	3.36	.92
bewt_qual	Die LSE leisten einen Beitrag zur langfristigen Qualitätsentwicklung an meiner Schule.	.84	3.22	.91
bewt_vergl	Die landesweite Durchführung der LSE nehme ich als Chance wahr, über den eigenen ‚Tellerrand‘ zu blicken.	.77	2.79	.92
bewt_gst	Die LSE nehme ich als Baustein in einem Gesamtkonzept zur Unterrichts- und Schulentwicklung wahr.	.84	2.96	.91
bewt_sinn	Ich halte die LSE insgesamt für eine sinnvolle Maßnahme.	.82	3.11	.91
Skala		$M =$	3.10	
		$SD =$	0.92	
		$\alpha =$	.93	
		$N =$	3878	

7.1.4.1.2 Schulentwicklung/Nutzung

Variablenname:	schentw
Frage/Instruktion:	Inwiefern haben Sie Maßnahmen/Schritte aus den Lernstandserhebungen abgeleitet?
Anzahl Items:	9
Wertelabels:	1 = systematisch/vollständig 2 = punktuell/teilweise 3 = gar nicht

		r_{it}	M	$\alpha_{\text{wenn Item entfernt}}$
lerns	Veränderung der Lernschwerpunkte für den aktuellen Unterricht	.60	2.14	.87
lehrin	Änderungen der Anordnung der Unterrichtsthemen im schuleigenen Stoffverteilungsplan/Fachcurriculum	.52	2.20	.88
verlk	Veränderung der Aufgabenformen in Leistungskontrollen	.58	2.13	.87
komp	Einsatz kompetenzorientierter Unterrichtsmethoden	.70	1.76	.86
met-did	methodisch-didaktische Überlegungen zum eigenen Unterricht	.72	1.88	.86
diff	Veränderung der Differenzierungsstruktur in meiner Klasse	.68	1.90	.87
neu	Neuanschaffung moderner Lehrwerke, die kompetenzorientiertes Lernen fördern	.60	1.91	.87
fortb	Besuch von Fortbildungsveranstaltungen	.65	1.96	.87
koor	stärkere Koordination der Arbeit in den Jahrgangsstufen	.64	1.94	.87
Skala		$M =$	2.42	
		$SD =$	0.85	
		$\alpha =$	.95	
		$N =$	3716	

7.1.4.2 Übersicht Antwortformate Online-Fragebogen

	Formulierung	Ausprägungen			
m8_umf	Der Umfang der LSE war	1 (zu kurz)	5 (zu lang)		
m8_niv	Die Aufgaben der LSE waren in ihrem Anspruchsniveau	1 (zu leicht)	5 (zu schwer)		
m8_zeit	Die zur Bearbeitung der LSE zu Verfügung gestellte Zeit war	1 (zu kurz)	5 (zu lang)		
m8_anst	Für die Schülerinnen und Schüler war die Teilnahme an der Lernstandserhebung sehr anstrengend.	++	+	0	--
m8_mot	Die Schülerinnen und Schüler waren bei der Bearbeitung der Lernstandserhebung sehr motiviert.	++	+	0	--
m8_nachv	Die Schülerinnen und Schüler konnten nachvollziehen, was in den einzelnen Aufgaben von ihnen verlangt wurde.	++	+	0	--
m8_unge	Die verwendeten Aufgabenformate waren für die Schülerinnen und Schüler sehr ungewohnt.	++	+	0	--
m8_aufgh	Aufgabenhefte	Schulnoten			
m8_korrek	Korrekturanweisungen	Schulnoten			
m8_hefta	Heft A genutzt?	O/NA (nein)	1 (ja)		
m8_heftb	Heft B genutzt?	O/NA (nein)	1 (ja)		
m8_heftc	Heft C genutzt?	O/NA (nein)	1 (ja)		
m8_korzeit	Wie lange haben Sie für die Korrektur der Lernstandserhebungen für eine Klasse benötigt (in Minuten)?	offen (in Minuten)			
m8_korang	Der zeitliche Aufwand war angemessen.	++	+	0	--
m8_eingzeit	Wie lange haben Sie für die Eingabe der Ergebnisse für eine Klasse benötigt (in Minuten)?	offen (in Minuten)			
m8_eingang	Der zeitliche Aufwand war angemessen.	++	+	0	--

	Formulierung	Ausprägungen			
tab_note	Tabellarische Übersicht	Schulnoten			
sof_note	Sofortrückmeldung	Schulnoten			
erg_note	Ergebnisbericht	Schulnoten			
schu_note	Schulbericht?	Schulnoten			
schue_note	Schülerbericht?	Schulnoten			
homp_note	Homepage	Schulnoten			
hotline_kontakt	Haben Sie die Hotline zu den LSE (per Mail oder Telefon) kontaktiert?	1 (ja)	2 (nein)		
hotline_hilfe	Wenn ja, konnten Ihnen weitergeholfen werden?	1 (ja)	2 (nein)		
mat_erh	Haben Sie die Materialien für Ihre Klasse(n) erhalten?	1 (ja)	2 (nein)		
mat_vollst	Waren die Materialien vollständig?	1 (ja)	2 (nein)		
mat_probl	Welche anderen Probleme mit dem Druck oder dem Versand der Materialien gab es?	offen			
vorb_info	Haben sie im Herbst des Vorjahres eine Informationsveranstaltung zu den LSE besucht?	1 (ja)	2 (nein)		
vorb_infil	Wenn ja, war der Besuch hilfreich für die Durchführung und Nutzung der LSE?	++	+	0	--
vorb_work	Haben Sie im Sommer des Vorjahres einen Workshop zu den LSE besucht?	1 (ja)	2 (nein)		
vorb_workhil	Wenn ja, war der Besuch hilfreich für die Nutzung der Ergebnisse der LSE für den Unterricht?	++	+	0	--
vorb_gut	Fühlen Sie sich insgesamt gut auf die LSE vorbereitet (z.B. durch Anschreiben, Informationsmaterial, Handreichungen, Homepage)?	++	+	0	--

	Formulierung	Ausprägungen				
inffi_elt	Ich habe den Informationsflyer für die Eltern ausgeteilt.	1 (ja)	2 (nein)			
inffi_sch	Ich habe den Informationsflyer für Schülerinnen und Schüler ausgeteilt.	1 (ja)	2 (nein)			
inffi_lg	Ich habe den Informationsflyer für Schülerinnen und Schüler mit meiner Lerngruppe besprochen.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_besp	Ich habe im Vorfeld die Aufgabenformate anhand älterer Lernstandsaufgaben besprochen.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_uebt	Ich habe im Vorfeld mit älteren Lernstandsaufgaben geübt.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_wied	Ich habe die Testdomänen nochmals im Unterricht wiederholt.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_didm	Ich habe mit den Didaktischen Materialien des vergangenen Jahres für die LSE geübt.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_verl	Ich habe spezielles Übungsmaterial der Verlage eingesetzt.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_demot	Um meine Schülerinnen und Schüler während der LSE nicht zu demotivieren, musste ich Hilfestellungen geben.	1 (systematisch, vollständig)	2 (punktuell, teilweise)			
vorb_mot	Ich habe großen Wert darauf gelegt, meine Schülerinnen und Schüler im Vorfeld zu motivieren.	++	+	0	-	--

	Formulierung	Ausprägungen			
nutz_diag	Die Lernstandserhebungen sind ein nützliches Instrument zur Lernstandsdiagnose.	++	+	0	--
nutz_unterricht	Die Lernstandserhebungen sind ein nützliches Instrument für die Unterrichtsentwicklung.	++	+	0	--
nutz_schule	Die Lernstandserhebungen sind ein nützliches Instrument für die Schulentwicklung.	++	+	0	--
umg_teil	Haben Sie persönlich an den LSE teilgenommen?	1 (ja)	2 (nein)		
teil_d8	Deutsch Jg. 8	0/NA (nein)	1 (ja)		
teil_m8	Mathe Jg. 8	0/NA (nein)	1 (ja)		
teil_e8	Englisch Jg. 6	0/NA (nein)	1 (ja)		
teil_f8	Französisch Jg. 8	0/NA (nein)	1 (ja)		
durch_sof	Durchsicht der Sofortrückmeldung	1 (systematisch, vollständig)		3 (gar nicht)	4 (keine Angabe möglich)
durch_erg	Durchsicht des Ergebnisberichts	1 (systematisch, vollständig)		3 (gar nicht)	4 (keine Angabe möglich)
durch_tab	Durchsicht der tabellarischen Auswertung	1 (systematisch, vollständig)		3 (gar nicht)	4 (keine Angabe möglich)
durch_schu	Durchsicht des Schulberichts	1 (systematisch, vollständig)		3 (gar nicht)	4 (keine Angabe möglich)

	Formulierung	Ausprägungen		
bsp_kl	Besprechung der Ergebnisse mit der gesamten Klasse	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
bsp_schue	Besprechung der Ergebnisse mit einzelnen Schülerin- nen/Schülern	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
bsp_elt	Besprechung der Ergebnisse mit den Eltern	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
bsp_team	Besprechung der Ergebnisse Klassenkonferenz/Team	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
bsp_fachs	Besprechung der Ergebnisse mit der Fachschaft	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
bsp_koll	Besprechung der Ergebnisse im gesamten Kollegium	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
bsp_schull	Besprechung der Ergebnisse mit der Schulleitung	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
vergl_erg	Vergleich der Ergebnisse (z.B. mit anderen Klassen, dem korr. Landesmittelwert, ...)	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)
did_mat	Kennen Sie das didaktische Material?	1 (ja)	2 (nein)	3 (gar nicht)
did_note	Mit welcher Schulnote bewerten Sie das didaktische Material?	Schulnoten		4 (keine Angabe möglich)

	Formulierung	Ausprägungen				
did_hilf	Die Zusatzinformationen zur Aufgabe helfen mir die Klassenergebnisse richtig einzuordnen.	++	+	0	-	--
did_anreg	Die Anregungen zur Weiterarbeit im Unterricht finde ich hilfreich für meine eigene Arbeit.	++	+	0	-	--
did_arb	Wie haben Sie mit dem didaktischen Material gearbeitet?	1 (systematisch, vollständig)				
vor_ueb	Ich habe die didaktischen Materialien aus dem Vorjahr als Übungsmaterial zur Vorbereitung auf den Lernstands verwendet.	1 (systematisch, vollständig)				
jetzt_ueb	Ich nutze die aktuellen didaktischen Materialien als Übungsmaterial nach dem Lernstand.	1 (systematisch, vollständig)				
did_sich	Ich habe die Didaktischen Materialien lediglich gesichtet, aber nicht weiter damit gearbeitet.	1 (systematisch, vollständig)				
did_besp	Ich habe den didaktischen Materialien für die Beschreibung der Ergebnisse mit der Klasse oder den Eltern zusätzliche Informationen zu den Lernstandsaufgaben entnommen.	1 (systematisch, vollständig)				
did_inf	Ich nutze die didaktischen Materialien zur Information über didaktische Konzepte und Entwicklungen.	1 (systematisch, vollständig)				

	Formulierung	Ausprägungen			
notw_aust	Austausch mit anderen Schulen	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)	4 (keine Angabe möglich)
notw_int	Integration der Auswertung in die Schulabläufe (z.B. Analyse der Ergebnisse im Rahmen des päd. Tages)	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)	4 (keine Angabe möglich)
notw_inf	mehr Informationen zur Weiterarbeit mit den Lernstandserhebungen im Unterricht	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)	4 (keine Angabe möglich)
notw_koop	verstärkte Kooperation in Jahrgangs-/Fachteams	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)	4 (keine Angabe möglich)
notw_hilf	Hilfe bei der Interpretation der Ergebnissrückmeldungen	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)	4 (keine Angabe möglich)
notw_plan	Unterstützung bei der Planung von Fördermaßnahmen	1 (systematisch, vollständig)	2 (punktuell, teilweise)	3 (gar nicht)	4 (keine Angabe möglich)

7.2 Leistungstest

7.2.1 Raterübereinstimmung

7.2.1.1 Testheft A (Hauptschule)

	N	%	Cohens K
Außenthermometer 1	22	100.00	1.00
Außenthermometer 2	22	95.45	0.65
Taxifahrt 1	22	90.91	-0.05
Taxifahrt 2	22	81.82	0.48
Autogas 1	22	100.00	1.00
Autogas 2	22	100.00	
Zahl gesucht	22	100.00	
Rechtecke 1	22	100.00	1.00
Rechtecke 2	22	100.00	1.00
Zug von Paderborn 1	22	90.91	0.46
Zug von Paderborn 2	22	86.36	0.71
Zug von Paderborn 3	22	90.91	0.80
Vier Spiegelungen	22	77.27	0.50
Superman	22	100.00	
Würfelnetz mit Buchstaben	22	86.36	0.64
Milchmenge	22	81.82	0.62
Winkel an Geraden	22	90.91	0.61
Größer-Kleiner	22	81.82	0.58
Handykauf 1	22	95.45	0.65
Handykauf 2	22	100.00	
Schultaschen 1	22	100.00	
Schultaschen 2	22	81.82	0.54
Fußballtabelle 1	22	95.45	0.00
Fußballtabelle 2	22	77.27	0.30
Fußballtabelle 3	22	95.45	0.86
Ebbe und Flut 1	22	95.45	0.00
Ebbe und Flut 2	22	95.45	0.65
Mineralwasser 1	22	90.91	0.69
Mineralwasser 2	22	100.00	1.00
Skatkarten 1	22	100.00	1.00
Skatkarten 2	22	90.91	0.80
Skatkarten 3	22	90.91	0.69
Würfel mit Muster	22	95.45	0.65

7.2.1.2 Testheft B (Realschule)

	N	%	Cohens K
Außenthermometer 1	108	98.15	0.87
Außenthermometer 2	108	99.07	0.94
Fehler in der Gleichung 1	108	93.52	0.83
Fehler in der Gleichung 2	108	96.30	0.93
Yardstick 1	108	99.07	0.98
Yardstick 2	108	100.00	1.00
Zahlenaussagen	107	100.00	
Spiegeleien 1	108	96.30	0.78
Spiegeleien 2	108	97.22	0.00
Ungewöhnlicher Mittelwert 1	108	94.44	0.88
Ungewöhnlicher Mittelwert 2	108	100.00	1.00
Ungewöhnlicher Mittelwert 3	108	98.15	-0.01
Zug von Paderborn 1	108	100.00	1.00
Zug von Paderborn 2	108	96.30	0.92
Zug von Paderborn 3	108	99.07	0.98
Vier Spiegelungen	108	99.07	0.96
Superman	108	90.74	0.79
Würfelnetz mit Buchstaben	108	96.30	0.91
Milchmenge	108	97.22	0.94
Winkel an Geraden	108	99.07	0.98
Größer-Kleiner	108	92.59	0.80
Handykauf 1	108	99.07	0.94
Handykauf 2	108	94.44	0.89
Schultaschen 1	108	99.07	0.98
Schultaschen 2	108	96.30	0.93
Fußballtabelle 1	108	96.30	0.92
Fußballtabelle 2	108	90.74	0.74
Fußballtabelle 3	108	98.15	0.91
Kreisdiagramme	108	100.00	1.00
Bälle ziehen 1	108	100.00	1.00
Bälle ziehen 2	108	99.07	0.97
Bälle ziehen 3	108	99.07	0.85
Durchschnittslinie 1	108	100.00	1.00
Durchschnittslinie 2	108	100.00	1.00
Handball 1	108	97.22	0.92
Handball 2	108	96.30	0.78

7.2.1.3 Testheft C (Gymnasium)

	N	%	Cohens K
Prozentrechnung 1	84	98.81	0.90
Prozentrechnung 2	84	100.00	1.00
Handball 1	84	98.81	0.95
Handball 2	84	88.10	0.76
Fehler in der Gleichung 1	84	94.05	0.81
Fehler in der Gleichung 2	84	97.62	0.94
Yardstick 1	84	98.81	0.97
Yardstick 2	84	97.62	0.95
Zahlenaussagen	84	94.05	0.82
Spiegeleien 1	84	97.62	0.94
Spiegeleien 2	84	96.43	0.80
Ungewöhnlicher Mittelwert 1	84	97.62	0.92
Ungewöhnlicher Mittelwert 2	84	96.43	0.93
Ungewöhnlicher Mittelwert 3	84	97.62	0.00
Buchstaben im Würfelnetz	84	96.43	0.92
Zwei Kreise 1	84	100.00	1.00
Zwei Kreise 2	84	95.24	0.31
Viereck finden 1	84	96.43	0.90
Viereck finden 2	84	97.62	0.94
Frankfurt-München	84	96.43	0.93
Quadratpuzzle 1	84	100.00	1.00
Quadratpuzzle 2	84	100.00	1.00
Weltrekordpizza	84	100.00	1.00
Zahlenschloss 1	84	88.10	0.49
Zahlenschloss 2	83	96.39	0.65
Zahlenstrahl 1	84	95.24	-0.02
Zahlenstrahl 2	84	91.67	0.32
Ergebnis kleiner als Null	84	88.10	0.74
Zahlensumme	84	100.00	1.00
Schülerzeitung	84	88.10	0.75
Kreisdiagramme	84	100.00	1.00
Bälle ziehen 1	84	96.43	0.93
Bälle ziehen 2	84	96.43	0.92
Bälle ziehen 3	84	96.43	0.92
Durchschnittslinie 1	84	100.00	1.00
Durchschnittslinie 2	84	100.00	1.00

7.2.2 Q₃-Statistiken des Leistungstests in Zusammenfassung

Tabelle 107: maximale Q₃-Werte je Items im Leistungstest

Item	Max. Q3	Max. Q3-Item ¹¹⁵	Item	Max. Q3	Max. Q3-Item ¹¹⁶
A1.1	0.13	A1.2	C7	0.07	C1.1
A1.2	0.13	A1.1	C8.1	0.35	C8.2
A2	0.15	A8.3	C8.2	0.35	C8.1
A3.1	0.18	A7.1	C8.3	0.26	C8.1
A3.2	0.32	A4.2	C9.1	0.59	C9.2
A4.1	0.46	A4.2	C9.2	0.59	C9.1
A4.2	0.46	A4.1	C10.1	0.07	A1.1
A5.1	0.07	A3.2	C10.2	0.08	A3.2
A5.2	0.10	A7.2	C11.1	0.41	C11.2
A5.3	0.14	A13.2	C11.2	0.41	C11.1
A6.1	0.08	A14.1	C12.1	0.04	C21
A6.2	0.11	A18	C12.2	0.06	C13
A7.1	0.18	A3.1	C13	0.07	A18
A7.2	0.15	A17	C14.1	0.31	C14.2
A8.1	0.30	A8.2	C14.2	0.31	C14.1
A8.2	0.30	A8.1	C15.1	0.30	C15.2
A8.3	0.21	A8.2	C15.2	0.30	C15.1
A9	0.11	A3.1	C15.3	0.13	C17.2
A10.1	0.31	A10.2	C1.1	0.17	C1.2
A10.2	0.32	A11.1	C1.2	0.17	C1.1
A11.1	0.32	A10.2	C2.1	0.04	C11.2
A11.2	0.15	A17	C2.2	0.04	C17.1
A12	0.13	A3.2	C3.1	0.27	C3.2
A13.1	0.21	A11.1	C3.2	0.27	C3.1
A13.2	0.14	A5.3	C4	0.05	C12.2
A14.1	0.40	A14.2	C5	0.05	C3.2
A14.2	0.40	A14.1	C6	0.07	C8.1
A14.3	0.25	A14.1	C16	0.04	C14.2
A15	0.17	A7.1	C17.1	0.04	C2.2
A16	0.15	A7.2	C17.2	0.13	C15.3
A17	0.15	A11.2	C18.1	0.11	C20.1
A18	0.11	A13.2	C18.2	0.07	C6
A19	0.10	A7.2	C19	0.07	C20.1
			C20.1	0.27	C20.2
			C20.2	0.27	C20.1
			C21	0.04	C12.1

¹¹⁵ Das Item, mit dem die höchste Korrelation in der Q₃-Statistik vorliegt.

7.2.3 Itemfit

Tabelle 108: Itemfit-Statistiken des Leistungstests (Teil 1)

	Outfit	Outfit t	Outfit p	Out- fit_pholm	Infit	Infit t	Infit p	In- fit_pholm
A1.1	1.65	2.14	0.03	1.00	1.08	0.60	0.55	1.00
A1.2	1.20	1.25	0.21	1.00	1.04	0.43	0.67	1.00
A2	1.08	1.46	0.14	1.00	1.05	1.48	0.14	1.00
A3.1	0.82	-0.42	0.68	1.00	0.99	0.01	0.99	1.00
A3.2	0.71	-2.96	0.00	0.18	0.88	-2.00	0.05	1.00
A4.1	0.94	-1.24	0.21	1.00	0.95	-1.68	0.09	1.00
A4.2	0.75	-3.22	0.00	0.08	0.86	-2.95	0.00	0.19
A5.1	0.94	-0.89	0.38	1.00	0.97	-0.79	0.43	1.00
A5.2	1.02	0.23	0.81	1.00	1.03	0.47	0.64	1.00
A5.3	0.98	-0.08	0.93	1.00	0.96	-0.33	0.74	1.00
A6.1	1.03	0.16	0.88	1.00	0.99	-0.01	0.99	1.00
A6.2	1.17	0.68	0.50	1.00	1.06	0.39	0.70	1.00
A7.1	1.67	1.48	0.14	1.00	1.11	0.51	0.61	1.00
A7.2	1.02	0.17	0.86	1.00	0.99	0.04	0.97	1.00
A8.1	1.10	1.02	0.31	1.00	1.07	1.11	0.27	1.00
A8.2	0.84	-0.64	0.52	1.00	0.94	-0.36	0.72	1.00
A8.3	0.79	-0.12	0.90	1.00	0.98	0.11	0.91	1.00
A9	1.58	1.24	0.21	1.00	1.07	0.35	0.73	1.00
A10.1	1.16	1.74	0.08	1.00	1.10	1.51	0.13	1.00
A10.2	0.99	0.04	0.97	1.00	1.01	0.13	0.90	1.00
A11.1	0.85	-0.09	0.93	1.00	0.96	0.04	0.97	1.00
A11.2	0.80	-0.27	0.79	1.00	0.96	-0.01	0.99	1.00
A12	0.81	-1.04	0.30	1.00	0.90	-0.88	0.38	1.00
A13.1	1.00	0.04	0.97	1.00	1.03	0.31	0.76	1.00
A13.2	1.21	0.58	0.56	1.00	1.04	0.24	0.81	1.00
A14.1	0.99	-0.16	0.88	1.00	1.00	0.04	0.97	1.00
A14.2	0.95	-0.75	0.46	1.00	1.01	0.18	0.86	1.00
A14.3	0.97	-0.34	0.73	1.00	1.02	0.38	0.70	1.00
A15	0.88	-0.61	0.54	1.00	0.99	-0.10	0.92	1.00
A16	1.05	0.73	0.47	1.00	1.04	1.06	0.29	1.00
A17	0.99	-0.30	0.76	1.00	1.00	-0.05	0.96	1.00
A18	1.11	2.10	0.04	1.00	1.09	2.64	0.01	0.48
A19	0.78	-2.05	0.04	1.00	0.90	-1.55	0.12	1.00
C7	1.15	1.78	0.07	1.00	1.03	0.72	0.47	1.00
C8.1	0.91	-1.70	0.09	1.00	0.92	-2.81	0.00	0.29
C8.2	0.87	-3.21	0.00	0.08	0.91	-3.74	0.00	0.01
C8.3	0.98	-0.15	0.88	1.00	1.01	0.23	0.82	1.00
C9.1	0.83	-3.98	0.00	0.00	0.89	-4.47	0.00	0.00
C9.2	0.88	-2.66	0.01	0.44	0.92	-3.01	0.00	0.16
C10.1	1.09	0.96	0.33	1.00	1.03	0.65	0.52	1.00
C10.2	0.89	-1.54	0.12	1.00	0.95	-1.32	0.19	1.00

Tabelle 109: Itemfit-Statistiken des Leistungstests (Teil 2)

	Outfit	Outfit_t	Outfit_p	Out- fit_pholm	Infit	Infit_t	Infit_p	In- fit_pholm
C11.1	1.12	2.46	0.01	0.76	1.08	3.01	0.00	0.16
C11.2	1.10	2.17	0.03	1.00	1.06	2.41	0.02	0.90
C12.1	1.19	4.28	0.00	0.00	1.14	5.42	0.00	0.00
C12.2	0.85	-1.73	0.08	1.00	0.93	-1.68	0.09	1.00
C13	0.64	-2.67	0.01	0.43	0.89	-1.58	0.12	1.00
C14.1	0.89	-0.79	0.43	1.00	0.99	-0.12	0.90	1.00
C14.2	0.55	-1.79	0.07	1.00	0.90	-0.78	0.44	1.00
C15.1	0.87	-3.04	0.00	0.14	0.90	-3.91	0.00	0.01
C15.2	0.78	-4.09	0.00	0.00	0.87	-4.39	0.00	0.00
C15.3	1.30	0.58	0.56	1.00	0.97	0.02	0.98	1.00
C1.1	0.84	-0.65	0.51	1.00	1.00	0.06	0.95	1.00
C1.2	1.07	1.03	0.30	1.00	1.03	0.78	0.44	1.00
C2.1	1.02	0.30	0.76	1.00	1.01	0.18	0.85	1.00
C2.2	1.04	0.27	0.79	1.00	1.03	0.28	0.78	1.00
C3.1	1.52	1.26	0.21	1.00	0.96	-0.13	0.90	1.00
C3.2	1.19	0.76	0.45	1.00	0.95	-0.36	0.72	1.00
C4	1.01	0.15	0.88	1.00	1.02	0.44	0.66	1.00
C5	1.51	4.26	0.00	0.00	1.18	3.12	0.00	0.11
C6	1.03	0.45	0.65	1.00	1.03	0.81	0.42	1.00
C16	1.26	3.97	0.00	0.00	1.09	2.50	0.01	0.70
C17.1	1.38	4.20	0.00	0.00	1.19	4.03	0.00	0.00
C17.2	0.87	-0.15	0.88	1.00	0.98	-0.03	0.98	1.00
C18.1	1.30	3.02	0.00	0.15	1.11	2.14	0.03	1.00
C18.2	1.03	0.38	0.70	1.00	1.04	0.93	0.35	1.00
C19	1.07	1.41	0.16	1.00	1.06	1.81	0.07	1.00
C20.1	1.15	0.56	0.58	1.00	1.08	0.54	0.59	1.00
C20.2	1.56	3.99	0.00	0.00	1.15	2.34	0.02	1.00
C21	0.88	-0.29	0.77	1.00	0.93	-0.36	0.72	1.00

7.3 Ergebnisergänzungen

7.3.1 Rangtransformierte, messwiederholte Varianzanalyse der Skalen zu Lernstandserhebungen (rmANOVA)

Tabelle 110: Deskriptive Beschreibung der Rangtransformation für Lernstandserhebungsskalen– Verfahren rmANOVA

Skala	Gruppe	MZP	N	M	SD	FLSD
Akzeptanz (H1 _{anpara} ¹¹⁶)	Kontrollgruppe	Prätest	9	11.67	6.4	4.83
	Kontrollgruppe	Posttest	9	10.67	7.52	4.83
	Projektgruppe	Prätest	16	15.62	8.55	4.83
	Projektgruppe	Posttest	16	16.44	7.67	4.83
Belastung (H3 _{nonpara})	Kontrollgruppe	Prätest	9	14.11	7.41	4.98
	Kontrollgruppe	Posttest	9	16.78	7.10	4.98
	Projektgruppe	Prätest	16	15.88	8.61	4.98
	Projektgruppe	Posttest	16	12.25	7.65	4.98
Förderung (H4 _{nonpara})	Kontrollgruppe	Prätest	9	12.56	9.03	5.43
	Kontrollgruppe	Posttest	9	10.89	7.98	5.43
	Projektgruppe	Prätest	15	15.20	7.71	5.43
	Projektgruppe	Posttest	15	16.47	7.31	5.43
Planung (H5 _{nonpara})	Kontrollgruppe	Prätest	9	13.67	6.73	5.75
	Kontrollgruppe	Posttest	9	12.00	7.42	5.75
	Projektgruppe	Prätest	15	13.07	7.81	5.75
	Projektgruppe	Posttest	15	15.67	8.19	5.75
Selektion (H6 _{nonpara})	Kontrollgruppe	Prätest	9	13.78	7.21	4.51
	Kontrollgruppe	Posttest	9	14.22	8.83	4.51
	Projektgruppe	Prätest	15	14.47	7.88	4.51
	Projektgruppe	Posttest	15	13.60	7.41	4.51

¹¹⁶ Der Index „nonpara“ zeigt an, dass es sich hierbei um eine nonparametrische Auswertungsergänzung der jeweiligen Hypothese handelt.

Anhang

Tabelle 111: Ergebnisse der rmANOVA für Lernstandserhebungsskalen

Skala	Effekt	dfn	dfd	Quadratsumme Effekt	Quadratsumme Fehler	F	p	p < .05	generalisiertes η^2
Akzeptanz (H1_{nonpara})	(Intercept)	1	23	10195.92	1977.47	118.59	>.01	*	0.79
	Gruppe	1	23	272.61	1977.47	3.17	.09		0.09
	Zeit	1	23	0.32	782.22	0.01	.92		0.00
	Gruppe x Zeit	1	23	9.46	782.22	0.28	.60		0.00
Belastung (H3_{nonpara})	(Intercept)	1	23	10599.68	1999.32	121.94	>.01	*	0.79
	Gruppe	1	23	22	1999.32	0.25	.62		0.01
	Zeit	1	23	23.12	831.87	0.64	.43		0.01
	Gruppe x Zeit	1	23	114	831.87	3.15	.09		0.04
Förderung (H4_{nonpara})	(Intercept)	1	22	9804.08	1837.78	117.36	>.01	*	0.78
	Gruppe	1	22	190.14	1837.78	2.28	.15		0.06
	Zeit	1	22	0.33	903.47	0.01	.93		0.00
	Gruppe x Zeit	1	22	24.2	903.47	0.59	.45		0.01
Planung (H5_{nonpara})	(Intercept)	1	22	9130.08	1580.47	127.09	<.01	*	0.78
	Gruppe	1	22	26.45	1580.47	0.37	.55		0.01
	Zeit	1	22	12	1013.8	0.26	.61		0.00
	Gruppe x Zeit	1	22	51.2	1013.8	1.11	.30		0.02
Selektion (H6_{nonpara})	(Intercept)	1	22	9436.02	2054.47	101.04	>.01	*	0.78
	Gruppe	1	22	0.01	2054.47	0	.99		0.00
	Zeit	1	22	1.69	623.98	0.06	.81		0.00
	Gruppe x Zeit	1	22	4.83	623.98	0.17	.68		0.00

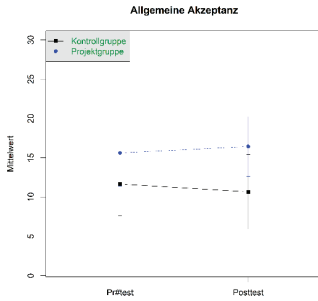


Abbildung 51: Rangmittelwerte *Allgemeine Akzeptanz* ($H1_{\text{nonpara}}$)

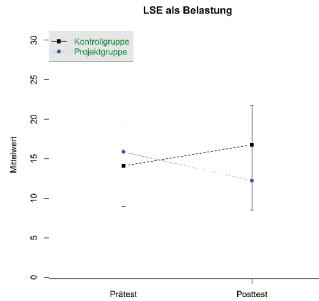


Abbildung 52: Rangmittelwerte *LSE als Belastung* ($H3_{\text{nonpara}}$)

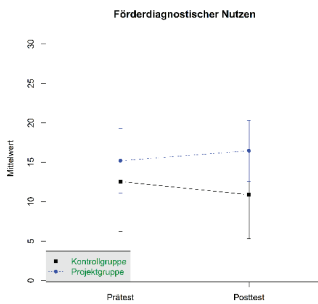


Abbildung 53: Rangmittelwerte *förderdiagnostischer Nutzen* ($H4_{\text{nonpara}}$)

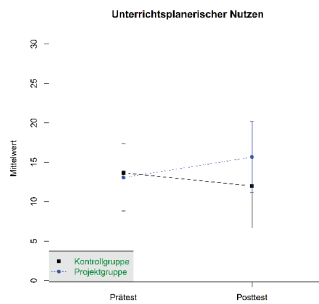


Abbildung 54: Rangmittelwerte *unterrichtsplanerischer Nutzen* ($H5_{\text{nonpara}}$)

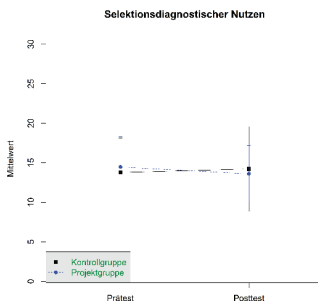


Abbildung 55: Rangmittelwerte *selektionsdiagnostischer Nutzen* ($H6_{\text{nonpara}}$)

7.3.2 Deskriptive Statistiken zur mANOVA für die Skalen der Unterrichtsqualität

Tabelle 112: Deskriptive Ergebnisse der Unterrichtsqualitätsskalen – Verfahren mANOVA

Skala	Gruppe	MZP	N	M	SD	FLSD
Begründungen und Erklärungen (H7)	Kontrollgruppe	Prätest	8	3.12	0.58	0.30
	Kontrollgruppe	Posttest	8	2.75	0.38	0.30
	Projektgruppe	Prätest	15	3.03	0.72	0.30
	Projektgruppe	Posttest	15	3.10	0.52	0.30
Variation Schwierigkeit (H8)	Kontrollgruppe	Prätest	8	2.77	0.57	0.26
	Kontrollgruppe	Posttest	8	2.73	0.58	0.26
	Projektgruppe	Prätest	15	2.74	0.50	0.26
	Projektgruppe	Posttest	15	2.76	0.51	0.26
kA Aufgaben (H9)	Kontrollgruppe	Prätest	10	2.74	0.54	0.18
	Kontrollgruppe	Posttest	10	2.65	0.49	0.18
	Projektgruppe	Prätest	16	2.77	0.51	0.18
	Projektgruppe	Posttest	16	2.76	0.51	0.18
kA Üben (H10)	Kontrollgruppe	Prätest	8	2.86	0.46	0.21
	Kontrollgruppe	Posttest	8	2.96	0.42	0.21
	Projektgruppe	Prätest	15	2.83	0.41	0.21
	Projektgruppe	Posttest	15	2.96	0.38	0.21
Repetitives Üben (H11a)	Kontrollgruppe	Prätest	8	2.85	0.75	0.18
	Kontrollgruppe	Posttest	8	2.95	0.54	0.18
	Projektgruppe	Prätest	15	2.15	0.49	0.18
	Projektgruppe	Posttest	15	2.17	0.31	0.18
Lernpartnerschaften (H18)	Kontrollgruppe	Prätest	8	2.99	0.63	0.21
	Kontrollgruppe	Posttest	8	2.65	0.55	0.21
	Projektgruppe	Prätest	15	2.92	0.52	0.21
	Projektgruppe	Posttest	15	3.00	0.53	0.21

Tabelle 113: Ergebnisse der rmANOVA für Unterrichtsqualitätsskalen

Skala	Effekt	dfn	dfd	Quadratsumme Effekt	Quadratsumme Fehler	F	p	p < .05	generalisiertes η^2
Begründungen und Erklärungen (H7_{nonpara})	(Intercept)	1	23	15948.98	3138.88	116.87	<.01	*	.81
	Gruppe	1	23	10.64	3138.88	0.08	.78		<.01
	Zeit	1	23	180.50	624.50	6.65	.02	*	.05
	Gruppe x Zeit	1	23	151.50	624.50	5.58	.03	*	.04
Variation Schwierigkeit (H8_{nonpara})	(Intercept)	1	23	16200.00	3939.56	94.58	<.01	*	.79
	Gruppe	1	23	57.44	3939.56	0.34	.57		.01
	Zeit	1	23	9.68	359.88	0.62	.44		<.01
	Gruppe x Zeit	1	23	1.44	359.88	0.09	.76		<.01
kA Aufgaben (H9_{nonpara})	(Intercept)	1	23	16200.00	3939.56	94.58	<.01	*	.79
	Gruppe	1	23	57.44	3939.56	0.34	.57		.01
	Zeit	1	23	9.68	359.88	0.62	.44		<.01
	Gruppe x Zeit	1	23	1.44	359.88	0.09	.76		<.01
kA Üben (H10_{nonpara})	(Intercept)	1	23	15382.58	2913.88	121.42	<.01	*	.82
	Gruppe	1	23	0.04	2913.88	0.00	.99		<.01
	Zeit	1	23	255.38	537.12	10.94	<.01	*	.07
	Gruppe x Zeit	1	23	0.00	537.12	0.00	.99		<.01
Repetitives Üben (H11_{nonpara})	(Intercept)	1	23	13612.50	2566.17	122.01	<.01	*	.82
	Gruppe	1	23	900.83	2566.17	8.07	.01	*	.24
	Zeit	1	23	106.58	336.50	7.28	.01	*	.04
	Gruppe x Zeit	1	23	6.42	336.50	0.44	.51		<.01
Lernpartnerschaften (H18_{nonpara})	(Intercept)	1	23	14552.18	3272.17	102.29	<.01	*	.80
	Gruppe	1	23	13.15	3272.17	0.09	.76		<.01
	Zeit	1	23	158.42	282.50	12.90	<.01	*	.04
	Gruppe x Zeit	1	23	122.58	282.50	9.98	<.01	*	.03

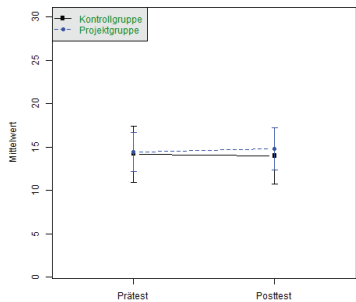


Abbildung 56: Rangmittelwerte *Insistieren auf Begründungen* (H7_{nonpara})

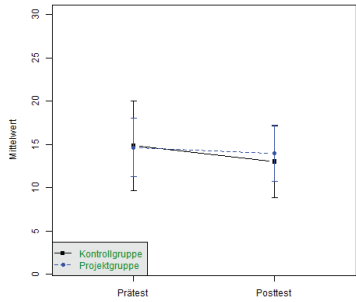


Abbildung 57: Rangmittelwerte *Variation Aufgabenschwierigkeit* (H8_{nonpara})

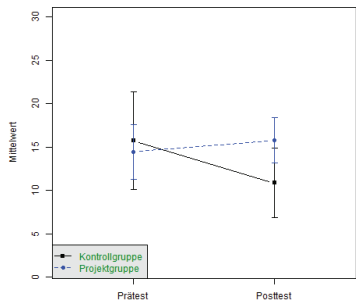


Abbildung 58: Rangmittelwerte *Kognitiv herausfordernde Aufgaben* (H9_{nonpara})

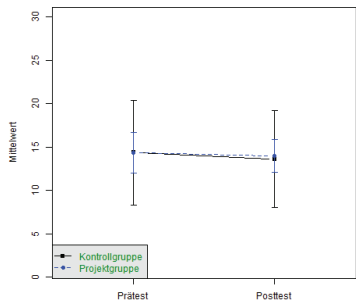


Abbildung 59: Rangmittelwerte *Kognitiv herausforderndes Üben* (H10_{nonpara})

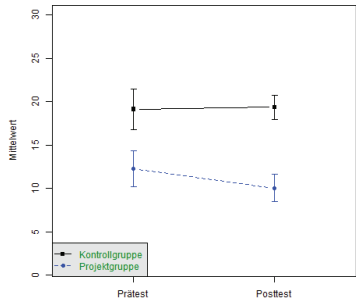


Abbildung 60: Rangmittelwerte *Repetitives, einschleifendes Üben* (H11_{nonpara})

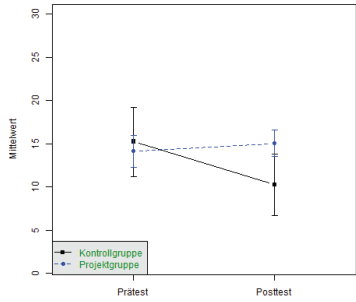


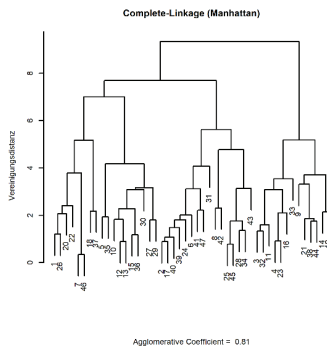
Abbildung 61: Rangmittelwerte *Lernpartnerschaften* (H12_{nonpara})

7.3.3 Deskriptive Statistiken zur rmANOVA für die Skalen der Unterrichtsqualität

Tabelle 114: Deskriptive Ergebnisse der Rangtransformation für Unterrichtsqualitätskalen– Verfahren rmANOVA

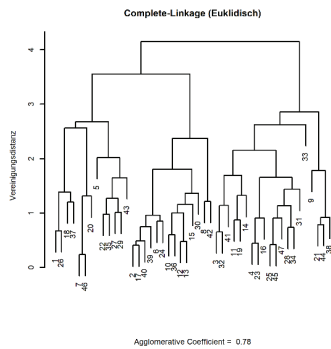
Skala	Gruppe	MZP	N	M	SD	FLSD
Schülerbeiträge (H12a)	Kontrollgruppe	Prätest	8	13.50	6.89	6.68
		Posttest	8	16.12	8.37	6.68
	Projektgruppe	Prätest	15	17.07	8.69	6.68
		Posttest	15	16.20	7.21	6.68
Erarbeitung LPS (H13a)	Kontrollgruppe	Prätest	8	20.50	6.28	4.38
		Posttest	8	18.38	5.42	4.38
	Projektgruppe	Prätest	15	12.60	8.21	4.38
		Posttest	15	9.93	5.81	4.38
Fürsorgliche Anleitung (H14a)	Kontrollgruppe	Prätest	8	17.38	6.76	3.38
		Posttest	8	14.62	4.81	3.38
	Projektgruppe	Prätest	15	13.33	8.10	3.38
		Posttest	15	13.73	7.69	3.38
Kognitive Selbstständigkeit (H15a)	Kontrollgruppe	Prätest	8	15.12	9.36	2.85
		Posttest	8	13.00	8.62	2.85
	Projektgruppe	Prätest	15	13.27	7.00	2.85
		Posttest	15	13.67	5.91	2.85
Soziale Konstruktion (H17)	Kontrollgruppe	Prätest	8	19.62	12.06	3.40
		Posttest	8	15.00	9.86	3.40
	Projektgruppe	Prätest	17	17.24	8.71	3.40
		Posttest	17	15.82	6.29	3.40
Persönliches Vertrauen (H19a)	Kontrollgruppe	Prätest	8	17.38	9.27	5.71
		Posttest	8	13.38	9.18	5.71
	Projektgruppe	Prätest	15	19.20	8.67	5.71
		Posttest	15	15.47	5.38	5.71
Soziale Diagnose (H20)	Kontrollgruppe	Prätest	8	15.62	9.07	4.11
		Posttest	8	10.38	7.35	4.11
	Projektgruppe	Prätest	15	15.80	7.24	4.11
		Posttest	15	16.13	5.94	4.11

7.3.4 Alternative Clusterlösungen



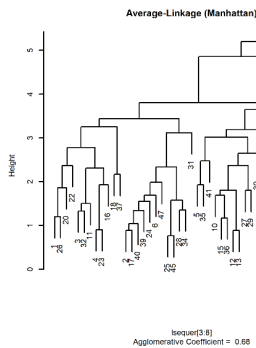
Agglomerationskoeffizient: .81

Abbildung 62: Dendrogramm Complete-Linkage (Manhattan-Metrik)



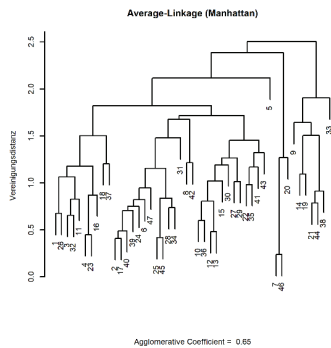
Agglomerationskoeffizient: .78

Abbildung 63: Dendrogramm Complete-Linkage (euklidische Metrik)



Agglomerationskoeffizient: .68

Abbildung 64: Dendrogramm Average-Linkage (Manhattan-Metrik)



Agglomerationskoeffizient: .65

Abbildung 65: Dendrogramm Average-Linkage (euklidische Metrik)

7.3.5 Online-Daten: Regression der Nützlichkeit auf die Nutzung (Einzelitems)

Tabelle 115: Regression der Nützlichkeitsdimensionen sowie der Nützlichkeits-Skala auf die Nutzung (Einzelitems)

	Lernschwerpunkte			Stoffverteilungsplan			Leistungskontrolle			komp. Methoden			met.-did. Überlegungen		
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²
Lernstandsdiagnose	0.08	<0.001	0.02	0.10	<0.001	0.04	0.09	<0.001	0.03	0.07	<0.001	0.02	0.07	<0.001	0.02
Unterrichtsentwicklung	0.10	<0.001	0.04	0.13	<0.001	0.06	0.12	<0.001	0.04	0.07	<0.001	0.02	0.08	<0.001	0.02
Schulentwicklung	0.12	<0.001	0.05	0.16	<0.001	0.08	0.13	<0.001	0.05	0.08	<0.001	0.02	0.11	<0.001	0.03
Nützlichkeit	0.14	<0.001	0.05	0.19	<0.001	0.09	0.16	<0.001	0.06	0.09	<0.001	0.02	0.10	<0.001	0.02
	Differenzierung			Neuanschaffung			Fortbildung			Koordination					
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²
Lernstandsdiagnose	0.05	<0.001	0.01	0.06	0.05	<0.001	0.01	0.06	0.05	<0.001	0.01	0.06			
Unterrichtsentwicklung	0.07	<0.001	0.01	0.07	0.07	<0.001	0.01	0.07	0.07	<0.001	0.01	0.07			
Schulentwicklung	0.09	<0.001	0.03	0.08	0.09	<0.001	0.03	0.08	0.09	<0.001	0.03	0.08			
Nützlichkeit	0.09	<0.001	0.02	0.08	0.09	<0.001	0.02	0.08	0.09	<0.001	0.02	0.08			

Tabelle 116: Regression der Nützlichkeitsdimensionen sowie der Nützlichkeits-Skala auf die Nutzung (Einzelitems; nur Mathematik-LPS)

	Lernschwerpunkte			Stoffverteilungsplan			Leistungskontrolle			komp. Methoden			met.-did. Überlegun- gen		
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²
Lernstandsdiagnose	0.08	0.55	-0.02	0.17	0.16	0.03	0.03	0.74	-0.03	0.28	0.02	0.11	0.25	0.03	0.10
Unterrichtsentwicklung	0.02	0.83	-0.03	0.21	0.05	0.08	0.01	0.92	-0.03	0.10	0.37	0.00	0.15	0.15	0.03
Schulentwicklung	-0.04	0.71	-0.03	0.18	0.13	0.04	-0.04	0.62	-0.02	0.09	0.49	-0.01	0.15	0.18	0.02
Nützlichkeit	-0.07	0.55	-0.02	0.24	0.05	0.08	0.03	0.76	-0.03	0.11	0.37	-0.01	0.17	0.13	0.04
	Differenzierung			Neuanschaffung			Fortbildung			Koordination					
	b	p	R ²	b	p	R ²	b	p	R ²	b	p	R ²			
Lernstandsdiagnose	0.14	0.25	0.01	0.07	0.60	-0.02	-0.04	0.76	-0.03	0.00	0.99	-0.03			
Unterrichtsentwicklung	0.01	0.93	-0.03	-0.03	0.81	-0.03	-0.12	0.30	0.00	-0.01	0.91	-0.03			
Schulentwicklung	-0.02	0.86	-0.03	-0.10	0.44	-0.01	-0.13	0.30	0.00	0.06	0.60	-0.02			
Nützlichkeit	0.05	0.69	-0.02	-0.06	0.63	-0.02	-0.08	0.51	-0.02	0.05	0.70	-0.02			

Die vorliegende Arbeit beschäftigt sich mit den deutschlandweit durchgeführten Mathematik-Vergleichsarbeiten *VerA-8*. Diese in Hessen Lernstandserhebungen genannten Arbeiten sollen als Orientierung verbindlicher Kompetenzerwartungen dienen, eine abschlussbezogene, objektive Leistungsstandsbestimmung und darauf aufbauend individuelle Förderung sowie gezielte Unterrichtsplanung ermöglichen. Damit zielen die Lernstandserhebungen u. a. auf die Unterrichtsentwicklung als einen Baustein der Schulentwicklung ab. Wie Befragungen zeigen, fällt die Akzeptanz der Vergleichsarbeiten durch Lehrkräfte jedoch indifferent aus, wobei Untersuchungen für Hessen bisher fehlen. Daher soll die vorliegende Untersuchung einen ersten Stand der Wahrnehmung der Lernstandserhebungen hessischer Lehrkräfte liefern sowie zur Suche nach Typen von Lehrkräften in Bezug auf deren Einstellungen zu den Lernstandserhebungen dienen. Weiterhin werden die Entwicklungsmöglichkeiten in den Einstellungen zu den Lernstandserhebungen im Rahmen einer Fortbildung (VELM-8; Verbesserung der Effektivität der Lernstandserhebungen Mathematik Klasse 8) eruiert, deren grundsätzliches Ziel es war, exemplarisch aufzuzeigen, wie Lehrkräfte die Lernstandserhebungen möglichst gewinnbringend für den eigenen Unterricht nutzen können.

ISBN 978-3-7376-0844-2



9 783737 608442 >