ZfDN (2019) 25:97–114
https://doi.org/10.1007/s40573-019-00094-7

ORIGINAL PAPER

Erfolgreicher Wissenserwerb im ersten Semester Physik

Analyse mithilfe eines Niveaumodells

David Woitkowski[1]

Eingegangen: 15. Oktober 2018 / Angenommen: 20. Mai 2019 / Online publiziert: 31. Mai 2019

Zusammenfassung

Der Erwerb von Fachwissen auf einem Niveau, welches auf die typischen komplexen und abstrakten Problemstellungen der Universitätsphysik angewandt werden kann, stellt eine zentrale Herausforderung in der Studieneingangsphase dar. Verschiedene Probleme dieser Phase, wie die hohen Abbruch-/Schwundquoten, können auf fachliche Schwierigkeiten in diesem Kontext zurückgeführt werden. Auch im späteren Studium wird immer wieder konzeptionell auf dem hier gelernten Fachwissen aufgebaut, so dass diese Phase als kritisch angesehen werden kann.

Die vorliegende Studie erhebt längsschnittlich das physikalische Fachwissen (Mechanik) bei 122 Physikstudierenden zu Beginn und Ende des ersten Studiensemesters. Die Studierenden können einem Niveaumodell zugeordnet werden, welches beschreibt, mit Aufgaben welcher Komplexität sie jeweils erfolgreich umgehen können.

Diese Niveaus stellen einerseits ein Bewertungskriterium für die Qualität des vorhandenen Fachwissens dar und können so ein an Anforderungsmerkmalen orientiertes Zielkriterium für erfolgreichen Fachwissenserwerb liefern. Andererseits können die Niveaus eine erste Einschätzung davon liefern, welche Ausprägung des Vorwissen prädiktiv für erfolgreichen Wissenserwerb ist.

Die Daten sprechen für einen wesentlichen Einfluss physikalischen und mathematischen Vorwissens auf die mittlere weitere Wissensentwicklung, insbesondere zeigen sich Hinweise auf eine additive Prädiktionskraft des schulischen Vorwissens für den Erwerb universitären Wissens bis Ende des ersten Semesters. Die Ergebnisse legen nahe, dass vor allem die Fähigkeit zum Umgang mit komplexen Problemstellungen prädiktiv für das universitäre Physiklernen ist – auch und gerade, wenn es im Rahmen der schulischen Physik erlernt wurde.

Schlüsselwörter Physikalisches Fachwissen · Wissenserwerb · Studieneingangsphase · Längsschnitt · Niveaumodell

Successful Knowledge Acquisition in First Semester Physics

Analysis Based on Competency Levels

Abstract

Acquiring usable levels of content knowledge which can be applied to complex and abstract problems is one of the major challenges when studying physics at universities. A lot of current issues in the introductory study phase today (like high dropouts) can be traced back to problems acquiring that knowledge. Due to the cumulative nature of physics knowledge the knowledge acquired in this phase is also critical for success in later studies.

✉ David Woitkowski
david.woitkowski@upb.de

[1] AG Didaktik der Physik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Deutschland

The study presented here uses a longitudinal analysis of 122 students' test results in physics content knowledge (mechanics) at the beginning and end of the first semester at university. Based on the test results participants are assigned competency levels according the complexity of tasks they can handle.
These levels on the one hand serve as a criterion for the quality of the acquired knowledge and thus give a task-characteristics based target criterion for successful learning of content knowledge. On the other hand the levels can give a first impression of what amount and quality of knowledge is predictive for successful acquisition of knowledge.
The data show a major impact of physics and mathematics foreknowledge. Especially first indications of an additive impact of school knowledge can be observed. The major predictor for learning physics at university can be identified as the ability to solve complex problems – even and especially if that is acquired within school physics.

Keywords Physics content knowledge · Knowledge acquisition · Introductory study phase · Longitudinal analysis · Competency levels

Einleitung

In der Physik an deutschen Hochschulen – wie auch in anderen MINT-Fächern – zeigt sich ein vielfältiges Problemspektrum, in dessen Kontext u. a. fachliche Schwierigkeiten eine zentrale Position einnehmen (Heublein et al. 2014). Dabei ist das erste Studienjahr besonders relevant, da sich hier Studienabbrüche und -wechsel stark häufen (Heublein et al. 2017), und gleichzeitig aufgrund der starken Kumulativität des Physikstudiums wesentliche Grundlagen für das weitere Studium gelegt werden (Schecker und Parchmann 2006). Empirisch zeigt sich, dass etwa ein Drittel der im Studium verbliebenen Studierenden auch nach einigen Semestern Studiendauer in der Mechanik, dem typischen Gegenstand des ersten Semesters, nicht über Fachwissen auf angemessenem Niveau verfügt (Woitkowski und Riese 2017). Neben Studienverbleib und -zufriedenheit gilt aber vielfach ein erfolgreicher Fachwissenserwerb (z. T. erhoben über Klausurnoten) als zentrales Merkmal von Studienerfolg (Albrecht 2011; Blüthmann et al. 2008; Freyer 2013; Fries 2002; Rindermann und Oubaid 1999; Sorge et al. 2016).

Buschhüter et al. (2016) berichten ein generelles Passungsproblem zwischen (vor allem mathematischen) Vorkenntnissen und Studienanforderungen. Für die Chemie konnte dies als zentraler Faktor für die von vielen Studierenden wahrgenommene Überforderung identifiziert werden (Schwedler 2017).

Mitunter wird z. B. von Lehrenden die Auffassung geäußert, dass das Physikstudium zwar auf mathematische, aber in der Regel nicht auf physikalische Vorkenntnisse zurückgreife (z. B. auch Agarwala, 2015; Shumba und Glass 1994). Im Kontrast dazu zeigen Buschhüter et al. (2017) eine starke Vorwissensabhängigkeit der Klausurnoten Ende des ersten Semesters.

Zur kriterialen Analyse des fachlichen Wissenserwerbs wäre ein objektiv an Anforderungsmerkmalen orientiertes Erfolgsmaß hilfreich (Kauertz 2008), statt eine Orientierung an der Stichprobe oder rein numerischen Cutoff-Scores (wie z. B. bei IQB 2013). Ähnlich stellt sich für das Vorwissen die Frage, ob hier lediglich ein „viel hilft viel" gilt, oder ob spezifischere Aussagen über die Menge oder Qualität des notwendigen Vorwissens getroffen werden können.

Im Folgenden wird auf Grundlage eines Komplexitätsmodells ein Niveaumodell erstellt, welches als Kriterium für die Beschreibung des Fachwissenserwerbs im ersten Fachsemester genutzt wird. Dazu werden Längsschnittdaten analysiert, mit denen das fachliche Vorwissen der Studierenden mit erfolgreichem und weniger erfolgreichem Fachwissenserwerb (im Sinne des Komplexitätskriteriums) charakterisiert und kontrastiert werden kann. Dasselbe Niveaumodell wird für eine kriteriale Analyse der Prädiktion des Wissenserwerbs durch das Vorwissens eingesetzt. Eine Analyse des gesamten Phänomens *Studienerfolg und -abbruch* in seinen vielfältigen Facetten wird damit jedoch nicht angestrebt.

Der Beitrag steht dabei in der Logik eines größeren Forschungsprogramms, in dem ausgehend von einem Kompetenzstrukturmodell (Woitkowski et al. 2011) ein Testinstrument entwickelt wurde (Woitkowski 2015). Mit den so erhobenen Daten konnte ein Kompetenzniveaumodell erstellt werden (Woitkowski und Riese 2017), mit dem ein Werkzeug zur Beobachtung von Kompetenzentwicklungsverläufen zur Verfügung steht. Die Ergebnisse können im weiteren Verlauf in die Formulierung eines Kompetenzentwicklungsmodells münden.

Theorie

Fachwissen als Kompetenzfacette

Kompetenz wird im Folgenden aufgefasst als 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 ver-

antwortungsvoll nutzen zu können“ (Weinert 2001, S. 27). Dieser übergreifende Kompetenzbegriff wird durch Angabe von Kompetenzmodellen jeweils inhaltlich konkretisiert. So umfasst z. B. das Modell der Kompetenz von Lehrkräften von Riese (2009, S. 26) das physikalische Fachwissen neben dem fachdidaktischen und pädagogischen Wissen und den motivationalen Orientierungen und Beliefs. Das Modell der Kompetenz von Physikern von Woitkowski (2017) enthält ebenfalls im Physikstudium zu erlernendes Fachwissen neben mathematischen und sonstigen wissenschaftlichen Fähigkeiten und Fertigkeiten sowie den physikbezogenen (motivationalen) Einstellungen und Beliefs.

Wie in vielen Studien in diesem Kontext (Kirschner 2013; Krauss et al. 2008; Riese 2009; Vogelsang et al. 2016; Walzer et al. 2013; Woitkowski und Borowski 2017), werden im Folgenden zwei Facetten des physikalischen Fachwissens unterschieden:

- Das *schulische Wissen* bezeichnet in Anlehnung an Krauss et al. (2008) dasjenige Wissen, welches ein durchschnittlicher Schüler am Ende der Sekundarstufe I erworben haben sollte. Zur Operationalisierung werden dann Items verwendet, die in Bezug auf ihren konzeptuell-begrifflichen Horizont auch in der Schule verwendet werden könnten.
- Demgegenüber ist das *universitäre Wissen* im Sinne der Konzeption von Riese (2009) vollständig von der Schule losgelöst. Es geht in den genutzten Begriffen und/oder im Mathematisierungsgrad (z. B. in der Nutzung von Differential- und Integralrechnung) über das für Schüler typischerweise leistbare hinaus. Entsprechende Testaufgaben können aufgrund des nicht ausreichenden begrifflichen oder mathematisch-methodischen Horizontes auch von sehr guten Schülern zumindest in der Mittelstufe in der Regel nicht gelöst werden.

Eine *vertiefte* Wissensfacette (vgl. Woitkowski und Borowski 2017) oder *Oberstufenwissen* wird hier zugunsten einer besseren Abgrenzbarkeit zwischen *schulischem* und *universitärem Wissen* nicht betrachtet. Diese beiden Facetten bilden gewissermaßen dasjenige Wissen ab, welches aus der Schule ins Studium mitgebracht werden sollte, und dasjenige, welches im Studium selbst erworben werden müsste.

Auf die Physik bezogene Beliefs spielen beim Erwerb dieses Wissens eine wesentliche Rolle, da Konstruktion von Wissen immer auf der Grundlage dessen geschieht, was Lernende vom Lerngegenstand wissen oder zu wissen glauben (Putnam und Borko 1997). Beliefs stellen in diesem Prozess eine Art Filter dar, da nur das effektiv gelernt wird, was nicht mit den Beliefs des Lerners in Konflikt steht (Blömeke 2004).

Eine etwas andere Rolle im Lernprozess nehmen dagegen motivationale Faktoren ein: Sie bestimmen eher inwieweit Lerner angebotene Lerngelegenheiten aktiv nutzen (Eccles und Wigfield 2002). Besonders im (im Vergleich mit der Schule) deutlich selbstgesteuerten Lernraum der Universität beeinflusst Motivation also, inwieweit Studierende z. B. Lehrveranstaltungen überhaupt besuchen, Aufgaben bearbeiten und anderweitig regelmäßig aktive Lernhandlungen ausführen (vgl. dazu auch Schulmeister 2015).

Komplexität als Anforderungsmerkmal

Neben der Untergliederung in Wissensfacetten werden Anforderungen in diesem Kontext typischerweise noch nach weiteren Merkmalen kategorisiert. Im Kontext der Konstruktion von Fachwissens- oder Kompetenztests in den Naturwissenschaftsdidaktiken hat sich dabei u. a. ein als *Komplexität* bezeichnetes Aufgabenmerkmal etabliert. Die zentrale Idee ist, dass es innerhalb jeder Wissensfacette Aufgabentypen gibt, bei denen zur Lösung nur einfache Wissenselemente benutzt (z. B. genannt oder wiedergegeben) werden müssen. Bei anderen Aufgabenstellungen müssen diese Elemente weiter verknüpft werden, um in angemessener Zeit zu einer Lösung zu kommen. Formaler kann man sagen, dass Lernende den Schritt von einer niedrigen zu einer höheren Komplexität schaffen, wenn es gelingt, Elemente einer niedrigeren Komplexität so zu kombinieren und zu transformieren, dass damit eine Anforderung bewältigt werden kann, die nur durch Aneinanderreihung von Elementen der niedrigeren Komplexität nicht bewältigbar wäre (Commons et al. 1998).

Dieses Konzept von Komplexität spiegelt die Auffassung von Wissen als ein propositionales Netzwerk wider (vgl. z. B. Schnotz 1994), dessen Qualität mit dem Verknüpfungsgrad dieses Netzwerkes steigt (Peuckert und Fischler 2000). Der Schritt von einer Komplexität zur nächst höheren entspräche dem Vorgang des *Chunking*, bei dem vorhandene Entitäten des Netzwerks zu größeren und komplexeren Bedeutungseinheiten zusammengefasst werden (Laird et al. 1986). Liegt in einem Wissensbereich bereits ein komplexverknüpftes Wissensnetzwerk vor, kann dieser Verknüpfungsgrad auch in einem nah angrenzenden Bereich vergleichsweise schnell aufgebaut werden, sofern die Regeln, nach denen Verknüpfungen sinnvoll hergestellt werden können, zwischen den Bereichen übertragbar sind (Dawson-Tunik 2006). Verglichen damit geschieht der Verknüpfungsaufbau ohne diese Übertragbarkeit deutlich langsamer (Armon und Dawson 1997).

Bei der Bestimmung der Komplexität von Anforderungen (d. h. Testitems) wird üblicherweise so vorgegangen, dass die in der Aufgabenstellung vorkommenden und die zur Lösung nötigen Begriffe und Konzepte auf ihren Verknüpfungsgrad hin analysiert werden (Bernholt 2010; Kauertz 2008). Es handelt sich somit um ein „objektives“ Aufgabenmerkmal, das nicht vom Lösenden oder dessen kon-

kreter Vorgehensweise bei der Lösung abhängt und niedrig-inferent erfassbar ist (Kauertz 2008, S. 22).

Werden nun in einem Testinstrument Items unterschiedlicher Komplexität verwendet, zeigt sich in verschiedenen Studien eine hohe Auswirkung auf die Itemschwierigkeit – die Komplexität kann als *schwierigkeitserzeugendes Aufgabenmerkmal* genutzt werden (Bernholt 2010; Kauertz 2008; Ohle et al. 2011; Woitkowski 2015). Dies bildet die Basis für Niveaumodelle, welche zur Analyse des Fachwissenserwerbs herangezogen werden können (vgl. Klieme et al. 2003, S. 85).

Im Folgenden wird das Komplexitätsmodell von (Bernholt 2010) mit den Komplexitätsausprägungen *Fakten, Prozessbeschreibungen, Lineare Kausalität* und *Multivariate Interdependenz* adaptiert. Dabei „bauen obere Stufen auf unteren Stufen auf, wobei die unteren Stufen durch die oberen Stufen organisiert werden. Jedes Element entsteht durch eine Verknüpfung und Koordination von Elementen der darunter liegenden Stufe." (Commons et al. 1998, Übersetzung Bernholt 2010, S. 22) Dabei unterscheidet sich das beschriebene *universitäre* und *schulische Wissen* zwar im Mathematisierungs- und Abstraktionsgrad, es können aber jeweils Anforderungen aller genannten Komplexitäten beschrieben werden.

Niveaumodelle

Klieme et al. (2003) empfehlen die kriteriale Interpretation von Testwerten anhand von Kompetenzniveaus. Das sind „Abschnitte auf kontinuierlichen Kompetenzskalen, die mit dem Ziel einer kriteriumsorientierten Beschreibung der erfassten Kompetenzen gebildet werden." (Hartig 2007, S. 86) Zur Niveaukonstruktion werden in der Literatur mehrere Verfahren diskutiert (Woitkowski und Riese 2017). Die Zuordnung geschieht in unserem Fall kriterial anhand des Aufgabenmerkmals Komplexität. Das heißt, dass die Probanden eines Niveaus Anforderungen einer Komplexität erfolgreich bewältigen können, Anforderungen der nächst höheren Komplexität aber nicht. Dabei erscheint es sinnvoll, für das *schulische* und *universitäre Wissen* getrennte Niveaumodelle zu erzeugen, so dass die in den verschiedenen Wissensbereichen belegten Niveaus miteinander in Beziehung gesetzt werden können.

Die Interpretation der Testdaten mit Hilfe dieses komplexitätsbasierten Niveaumodells ermöglicht dann zwei interpretative Zugänge zu den Wissensständen der Probanden, die allein auf Basis von numerischen Testwerten nicht möglich wären:

Erstens legt der Literaturbefund nahe, dass es länger dauert, Anforderungen höherer Komplexität zu erlernen, als auf eine in einer benachbarten Wissensfacette bereits beherrschte Komplexität aufzuschließen (Armon und Dawson 1997; Dawson-Tunik 2006).

Zweitens liefern die Niveaus eine kriteriale Einordnung des Wissensstandes – so kann z. B. vermutet werden, dass die typische Anfängervorlesung in der Universität komplexe und stark mathematisierte Anforderungen an die Studierenden stellt. Im Niveaumodell entspräche das einem hohen Niveau im *universitären Wissen*, also dem Umgang mit komplexen Problemstellungen in der u. a. über Abstraktion und Mathematisierung definierten Wissensfacette (Woitkowski 2015, S. 262). Somit kann dieses Niveau als normatives Lernziel im ersten Semester angenommen werden.

Wissenserwerb im Physikstudium

Bisherige Erkenntnisse über den universitären Wissenserwerb liegen vor allem aus längsschnittlich interpretierten Querschnittserhebungen vor, haben also das interpretative Problem, dass nicht dieselben Personen zu mehreren Zeitpunkten getestet werden. So können individuelle Entwicklungen und Kohorten- oder andere Gruppeneffekte nicht wirksam unterschieden werden. In diesen Studien zeigt sich eine mit der Studiendauer größer werdende Differenz zwischen fähigen und weniger fähigen Studierenden (Riese 2009).

Die Konstruktion komplexitätsbasierter Kompetenzniveaus wurde für die hier getesteten Wissensfacetten im Test mit Studierenden bereits erprobt (Woitkowski 2015, 2017). Auf dieser Basis konnte bereits ein Niveau als wünschenswertes Ziel im Physikstudium angegeben werden. Dieses wird jedoch auch nach mehreren Semestern Studiendauer von etwa einem Drittel der Probanden nicht erreicht (Woitkowski und Riese 2017). Hier wurde die Frage nach Determinanten dieser Entwicklung bisher nur in Querschnittsanalysen beantwortet. Ebenso ist die Geschwindigkeit der Entwicklung bzw. des Niveauaufstiegs auf der bisherigen Datengrundlage kaum zu beantworten.

Die Analyse längsschnittlichen Wissenserwerb mittels Komplexitätsniveaus wurde in der Hochschule noch nicht durchgeführt. Aufgrund der Niveaukonstruktion entlang von Itemschwierigkeiten ist auch hier ein Anstieg zu erwarten, der Ausmaß ist aber nicht klar. Auf der Basis der von Dawson-Tunik (2006) gefundenen Übertragbarkeit zwischen benachbarten Wissensfacetten kann vermutet werden, dass Probanden, die im *Schulwissen* ein höheres Niveau erreichen, auch im *universitären Wissen* eher in höhere Niveaus aufsteigen. Auf Basis der bisherigen Querschnittserhebungen könnte diese Hypothese noch nicht geprüft werden.

In der Vergangenheit wurden Niveaukonstrukte häufig herangezogen, um Bildungsziele oder Erfolgskriterien festzulegen (z. B. IQB 2013; Woitkowski und Riese 2017). Längsschnittliche Daten zum Zusammenhang der Zielerreichung mit Lernvoraussetzungen zu Studienbeginn liegen hier jedoch nicht vor. Hier kommen neben dem o. g. Vor-

wissen auch eine Reihe weiterer Merkmale als Prädiktor in Frage. Als erste Heuristik wurden hier Merkmale untersucht die auch im Kontext von Studienabbruch erhoben werden, da dieser häufig mit fachlichen Schwierigkeiten in Verbindung gebracht wird (z. B. Heublein et al. 2014): Motivation, Buoyancy, soziale und institutionelle Voraussetzungen, mathematische Kenntnisse (Albrecht 2011; Bosse und Trautwein 2014; Burger und Groß 2016; Buschhüter et al. 2016; Neumann et al. 2016; Sorge et al. 2016).

Forschungsfragen

Der Literaturbefund zeigt durchgehend eine Zunahme des Fachwissens über die Studiendauer (Riese 2009; Woitkowski 2015). Längsschnittdaten mit zwei Testzeitpunkten zu Beginn (TZP 1) und Ende (TZP 2) des ersten Semesters sollten diesen Befund reproduzieren können. Vor dem theoretischen Hintergrund sollten die Zuwächse im *universitären Wissen* höher als im *schulischen Wissen* ausfallen, da ersteres in höherem Maße Gegenstand universitärer Lehre ist.

F1: Zeigen sich zwischen den Testzeitpunkten Zuwächse im *schulischen* und *universitären Wissen*? In welcher Facette fällt der Zuwachs im Mittel höher aus?

Das Niveaumodell erlaubt nun einen Vergleich der zu Studienbeginn (TZP 1) erreichten Niveaus in den beiden Wissensfacetten. Für Studienanfänger (TZP 1) ist das häufige Erreichen hoher Niveaus im *schulischen*, nicht jedoch

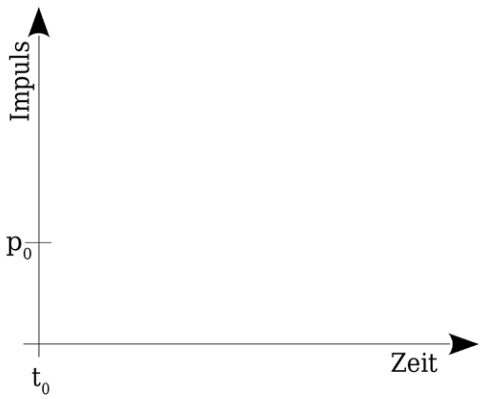
a

Nennen Sie eine Formel zur Berechnung der **Rotationsenergie** eines Körpers, wenn sein Trägheitsmoment I und die Winkelgeschwindigkeit ω bekannt sind.

b

An einem Körper, der **reibungsfrei** über eine Unterlage gleitet, greift eine **konstante Kraft in Bewegungsrichtung** an.

Zeichnen Sie rechts den **Impuls** als Funktion der Zeit ein, wenn der Impuls des Körpers zum Zeitpunkt $t = t_0$ gleich p_0 war.



c

Ein Rad mit dem Trägheitsmoment $I = 10\ kg \cdot m^2$ rotiert mit einer Winkelgeschwindigkeit von $\omega = 5\ rad/s$. Sie bremsen das Rad innerhalb von $\Delta t = 10\ s$ bis zum Stillstand ab.
Wie groß war das **mittlere bremsende Drehmoment**?

d

Sie konstruieren ein neuartiges Automobil, das die Energie, die beim Bremsen frei wird, in einem Schwungrad (Masse $m_S = 160\ kg$; Radius $r_S = 1\ m$) zwischenspeichert.

Sie stehen an einer roten Ampel, das Schwungrad rotiert mit $\omega_S = 5\ s^{-1}$.

Auf welche Geschwindigkeit können Sie mit der im Schwungrad gespeicherten Energie beschleunigen, wenn ihr Automobil die Gesamtmasse $m_A = 2 \cdot 10^3\ kg$ hat?

Bedenken Sie auch, dass Ihr Automobil vier Räder (jeweils: $m_R = 1\ kg$; $r_R = 0{,}5\ m$) besitzt, die in Rotation versetzt werden müssen. Rechnen Sie allerdings nur auf eine signifikante Stelle genau.

Verwenden Sie weiterhin für alle Räder $I = \frac{1}{2} \cdot m \cdot r^2$ und gehen Sie von einem Wirkungsgrad von $\eta = 1$ aus.

Abb. 1 **a** Beispielitem J5. Universitäres Wissen, (I) Fakten, Energie/Impuls, **b** Beispielitem D5. Universitäres Wissen, (II) Prozessbeschreibungen, Energie/Impuls, **c** Beispielitem E7. Universitäres Wissen, (III) Lineare Kausalität, Kraft, **d** Beispielitem C6. Universitäres Wissen, (IV) Multivariate Interdependenz, Energie/Impuls (Woitkowski 2015, S. 337, 340, 344, 361)

Tab. 1 Stichprobenüberblick mit Geschlechterverteilung, durchschnittliche Abiturnote, letzte schulische Mathematik- und Physiknote (jeweils M ± SD), aufgeschlüsselt nach Fach- und Lehramts-Studiengang sowie Dropout-Gruppe (d. h. Probanden, die nur am ersten Testzeitpunkt teilgenommen haben)

	N (Längsschnitt)	Anteil weiblich (%)	Abiturnote	Mathematiknote	Physiknote
Analysierte Stichprobe	122	27,8	1,90 ± 0,67	1,64 ± 0,84	1,49 ± 0,74
– davon Fach	99	25,3	1,83 ± 0,64	1,57 ± 0,82	1,42 ± 0,73
– davon Lehramt	23	39,1	2,20 ± 0,74	1,96 ± 0,84	1,78 ± 0,80
Dropout-Gruppe	(136)	19,8	2,29 ± 0,71	2,00 ± 0,96	1,79 ± 0,84
– davon Fach	(79)	19,0	2,12 ± 0,76	1,84 ± 1,00	1,57 ± 0,79
– davon Lehramt	(57)	21,1	2,53 ± 0,58	2,23 ± 0,85	2,06 ± 0,82

im *universitären Wissen* zu erwarten. Da es sich aber um zwei Wissensfacetten einer gemeinsamen Domäne handelt, könnte bei Anfängern auf hohen *schulischen* Niveaus am ehesten auch höhere *universitäre* Niveaus erwartet werden.

F2: Welche Niveaus werden zu TZP 1 in den beiden Wissensfacetten erreicht? Inwiefern korrespondiert das Erreichen hoher Niveaus zwischen den Facetten?

Der in F1 untersuchte Zuwachs kann mit Hilfe des Niveaumodells kriterial analysiert werden: Dabei interessiert vor allem, ab welchem *schulischen* Niveau im Vorwissen ein Effekt sichtbar wird.

F3: Welche Niveaus im *universitären Wissen* werden zu TZP 2 erreicht? Inwiefern prädizieren die zu TZP 1 erreichten Niveaus diese?

Als ein kriteriales Maß für erfolgreichen Fachwissenserwerb in der universitären Ausbildung kann das Erreichen des obersten *universitären* Niveaus festgelegt werden.

F4: Wie viele Studierende erreichen zu TZP 2 das oberste Niveau des *universitären Wissens* und wie unterscheiden sie sich von den Probanden, die dieses Niveau nicht erreichen?

Methoden

Stichprobe und Testzeitpunkte

Die Gelegenheitsstichprobe rekrutiert sich aus den Teilnehmerinnen und Teilnehmern in 7 Experimentalphysik-Anfängervorlesungen an 6 deutschen Universitäten in den Wintersemestern 2016/17 und 2017/18, die von Studierenden in Fach- und Gymnasial-Lehramtsstudiengängen belegt werden. Probanden, die nicht mehr im ersten Fachsemester waren, wurden aussortiert.

Gegenstand aller Lehrveranstaltungen war jeweils die klassische newtonsche Mechanik. Den Studierenden wurde das Forschungsprojekt in der ersten Vorlesung im ersten Semester kurz vorgestellt, bevor dann noch in der ersten Semesterwoche der erste Test stattfand. Der zweite Testzeitpunkt fand in der letzten Semesterwoche ebenfalls in der Lehrveranstaltung statt. Ein dritter Testzeitpunkt fand Ende des zweiten Semesters statt; Daten daraus werden hier jedoch aus Platzgründen nicht berichtet. Für eine Teilnahme an allen Tests des Längsschnitts wurden 50 € Probandengeld gezahlt. In die hier gezeigten Analysen gehen nur diejenigen Probanden ein, von denen Datensätze von den ersten beiden Testzeitpunkten vorliegen. Damit können im Folgenden $N = 122$ Probanden, davon 99 Fach- und 23 Lehramts-Studierende, analysiert werden (Tab. 1). Der Frauenanteil liegt bei 27,8 % (im Lehramt etwas höher), die mittlere Abiturnote bei 1,90 (SD = 0,67), die mittlere letzte Schulnote in Mathematik bei 1,64 (SD = 0,84) und in Physik bei 1,49 (SD = 0,74).

Vergleicht man die im Folgenden analysierten Gruppe mit vollständigem Datensatz mit der Dropout-Gruppe, also den Probanden, von denen nur Daten zum ersten Testzeitpunkt vorliegen, so zeigt sich insgesamt ein Dropout von 52,7 %. Die Dropout-Gruppe weicht in allen drei Noten von der hier analysierten Stichprobe signifikant nach unten ab (Abiturnote: $W = 10.564$; $p < 0{,}001$; Mathematiknote: $W = 9735{,}5$; $p = 0{,}003$; Physiknote: $W = 9250{,}5$; $p = 0{,}003$). Die hohe Dropout-Quote im Längsschnitt steht im Einklang mit den bekannten Studienabbruch- bzw. -wechselquoten in diesem Bereich (Heublein et al. 2014) und stellt eines der üblichen Probleme bei der Akquise längsschnittlicher Stichproben dar. Allerdings ist nicht feststellbar, ob (und in welchem Umfang) Probanden von TZP 1 zu TZP 2 zwar noch studierten aber aus anderen Gründen nicht am Test teilnahmen – die Studierenden wurden zum zweiten Test nicht persönlich, sondern als Gruppe in der jeweiligen Lehrveranstaltung angesprochen. In jedem Fall handelt es sich bei der hier analysierten Stichprobe aber um eine Positivauswahl in Bezug auf Merkmale wie Testmotivation, regelmäßiger Veranstaltungsteilnahme und möglicherweise auch Selbstkonzept.

Testinstrument

In der hier berichteten Studie kommen die Skalen zum *schulischen* und *universitären Fachwissen* von Woitkow-

Tab. 2 Auszug aus dem Entscheidungsbaum für die Item-Modell-Zuordnung (Woitkowski 2015, Anhang B)

Wissensfacette	Komplexität
1) Erfordert die Aufgabenlösung höhere Mathematik wie Vektorrechnung, Differential- oder Integralrechnung? Und handelt es sich bei ersterer nicht nur um die Darstellung von Vektoren als Pfeile oder Spaltenvektoren oder um deren Addition oder Multiplikation mit einem Skalar? **Ja → Universitäres Wissen;** **Nein → Nächste Frage** 2) Wäre es möglich, diese Aufgabe im Hinblick auf die in der Fragestellung vorkommenden physikalischen Begriffe ohne wesentliche Änderungen in einem Schulbuch zu verwenden? **Ja → Schulwissen; Nein → Universitäres Wissen**	1) Erfordert die Aufgabenlösung lediglich die Wiedergabe eines Merksatzes, einer Formel oder Definition? **Ja → Fakten; Nein → Nächste Frage** 2) Erfordert die Aufgabenlösung eine Rechnung oder verlangt der Aufgabentext explizit eine Begründung, Erklärung oder argumentative Erläuterung? **Ja → Weiter bei 4; Nein → Nächste Frage** 3) Erfordert die Aufgabenlösung lediglich die Beschreibung eines Prozesses durch Worte, Skizzen oder Diagramme? **Ja → Prozessbeschreibung; Nein → Nächste Frage** 4) Erfordert die Aufgabenlösung lediglich eine einzige lineare Begründung der Form „Weil x, darum y." Ohne dass weitere Faktoren oder Einschränkungen beachtet werden müssten? **Ja → Lineare Kausalität; Nein → Nächste Frage** Erfordert die Aufgabenlösung nur die Berücksichtigung eines wesentlichen Einflussfaktors? **Ja → Lineare Kausalität; Nein → Multivariate Interdependenz**

Insgesamt werden für die Wissensfacette 13 und für die Komplexität 18 Kriterien abgefragt

ski (2015) zum Einsatz. Die Items operationalisieren Inhalte der Mechanik, die üblicherweise Gegenstand des ersten Studiensemesters ist (vgl. KFP 2010). Beispielitems zeigt Abb. 1. Die Zuordnung der Items zu den Wissensfacetten und Komplexitäten, geschah im Rahmen einer Befragung eines Physikdidaktikers und eines Fachleiters als Experten mithilfe von Entscheidungsbäumen, wobei in strukturierter Reihenfolge einzelne Kriterien zur Einordnung abgefragt wurden. Dazu ist jeweils auf eine vom Experten angefertigte und in Bezug auf die nötigen Wissensbestände und Strukturen reflektierte Lösung als Basis nötig. Die Wissensfacetten-Kriterien beziehen sich auf die im Item verwendeten Begriffe, deren Mathematisierungs- und Abstraktionsgrad, die Komplexitäts-Kriterien aber auf die Struktur des zur Lösung notwendigen Vorgehens (Woitkowski 2015, Kap. 12). Aus den Rückmeldungen der beiden Experten wurde eine Konsens-Einordnung erarbeitet. Im Test sind in beiden Wissensfacetten Items auf allen Komplexitäten vorhanden, die Kriterien für die Komplexitätszuordnung sind für beide Wissensfacetten identisch. Die Merkmale Komplexität und Wissensfacetten sind also weitestmöglich orthogonal und zwischen den Facetten vergleichbar. Beispielhafte Kriterien zur Zuordnung zeigt Tab. 2.

Das mathematische Wissen wird mit 15 Items aus dem Studieneingangstest von Krause und Reiners-Logothetidou (1981) erhoben, welche die Bereiche Vektorrechnung, Geraden- und Ellipsengleichung, Quadratische Gleichungen, Funktionsgraphen und Ableitungen umfassen. Dies geht leicht über die mathematischen Kenntnisse hinaus, die zur Lösung der Fachwissens-Items nötig sind.

Weitere Entwicklungsprädiktoren wie Motivation, Einstellungen und Beliefs werden durch aus der Literatur übernommene Skalen abgedeckt: Belief- und Selbstkonzept-Skalen von Riese (2009) und Lamprecht (2011), Skalen zu Studienzufriedenheit, Kontextbedingungen, Lernschwierigkeiten und Studienklima nach Albrecht (2011) und Burger und Groß (2016), zur fachspezifischen Academic Buoyancy (Neumann et al. 2016) sowie zwei von Sundre (2007) für die im Physikstudium relevanten Übungs- und Klausursitationen adaptierte Skalen zu Effort und Importance.

Das Testinstrument ist für 60 min Testdauer ausgelegt. Die Fachwissens-Skalen folgen (wie in der ursprünglichen Veröffentlichung) einem *partially balanced incomplete Block design* (pBIBD; vgl. Kubinger et al. 2011), wobei jeder Proband 3 von 10 Item-Blöcken zur Bearbeitung vorgelegt bekommt. Die 3 Blöcke von TZP 1 und 2 sind vollständig disjunkt, so dass Erinnerungseffekte ausgeschlossen werden können. Jedes Testheft enthält im Rahmen dieses Designs zwischen 10 und 16 Items zum *schulischen Wissen* ($M = 12{,}6$; $SD = 2{,}0$) und zwischen 3 und 8 Items zum *universitären Wissen* ($M = 5{,}1$; $SD = 1{,}5$). Abgesehen von den Fachwissens-Items sind die Testhefte zu jedem Testzeitpunkt identisch.

Die Fachwissens- wie auch die Mathematik-Items sind teils offene Items, teils geschlossene (Single-Choice-)Items. Die weiteren Items sind (wie auch schon in früheren Testeinsätzen) als 4-stufige Likert-Skalen formuliert – die Skala N1 Academic Buoyancy jedoch 7-stufig[1].

Die Kodierung der Tests geschah durch geschulte Hilfskräfte mittels eines ausführlichen Kodiermanuals mit Er-

[1] Die Skalen A1–A3 Albrecht (2011) waren in der Originalveröffentlichung 6-stufig; die Skalen B1–B3 Burger und Groß (2016) 7-stufig; die Vorlage für die Skalen S1–S4 Sundre (2007) 5-stufig. Um eine Irritation der Probanden durch viele verschiedene Skalenlängen zu vermeiden, wurden diese aber auf ein einheitliches, bei Riese (2009) erprobtes 4-stufiges Format gebracht. Damit kann einerseits eine zügige Bearbeitung ermöglicht und andererseits eine Tendenz zur Mitte minimiert werden Busker (2014). Die 7-Stufigkeit der Skala N1 wurde mit Hinblick auf eventuelle spätere Datenzusammenführungen beibehalten.

Tab. 3 Überblick über die Fachwissens-Skalen sowie die Gesamtskala mit allen Items zum Vergleich. Die angegebenen Skalen-Kennwerte beziehen sich jeweils auf die bereinigten Skalen nach Item-Ausschluss. Im Folgenden wurden beide Testzeitpunkte gemeinsam analysiert, die Ergebnisse der nach Testzeitpunkt getrennten Analyse sind unten in der Tabelle aufgeführt

Skala	Anzahl Items	Varianz	EAP-Reliabilität	WLE-Reliabilität	AIC	BIC	χ^2-Test
Gemeinsame Analyse beider Testzeitpunkte (244 Fälle)							
Schulisches Wissen	31	1,622	0,771	0,614	4038	4238	$X^2 = 4{,}519$
Universitäres Wissen	15	2,219	0,727	0,272			df = 2
Gesamtskala zum Vergleich	46	1.652	0.789	0,760	4039	4232	p = 0,09
Nur Testzeitpunkt 1 (122 Fälle)							
Schulisches Wissen	31	1,573	0,762	0,621	1958	2118	$X^2 = 0{,}429$
Universitäres Wissen	15	1,892	0,731	0,062			df = 2
Gesamtskala zum Vergleich	46	1,576	0,768	0,751	1955	2109	p = 0,81
Nur Testzeitpunkt 2 (122 Fälle)							
Schulisches Wissen	31	1,819	0,778	0,615	2056	2215	$X^2 = 0{,}952$
Universitäres Wissen	15	1,719	0,743	0,309			df = 2
Gesamtskala zum Vergleich	46	1,702	0,790	0,758	2053	2207	p = 0,62

wartungshorizont zu allen Wissens-Items. Die Qualität der Kodierung wurde laufend durch Doppelkodierung von ca. 10 % der Testhefte überprüft und ggf. korrigiert. Cohens $\kappa = 0{,}874$ liegt im sehr guten Bereich (Bortz und Döring 2006, S. 277).

Analysen

Für dargestellten statistischen Analysen wurden lediglich die $N = 122$ Probanden herangezogen, für die ein vollständiger Datensatz zu beiden Testzeitpunkten vorliegt. Für die Skalenbildung und Rasch-Analyse wurde die Technik der virtuellen Probanden genutzt, es wurde also jeder Proband zu jedem Zeitpunkt als einzelner Fall im Datensatz abgebildet (Hartig und Kühnbach 2006; König et al. 2018; Plöger et al. 2016; Seifert und Schaper 2012). Insgesamt wurden die Skalen also mit 244 Datensätzen gebildet.

Die Fachwissens-Skalen werden mit dem dichotomen Rasch-Modell mit dem R-Paket TAM (Robitzsch et al. 2017) analysiert, wobei Items mit einem Infit von MNSQ > 1,25 oder T > 1,96 von der weiteren Verwendung ausgeschlossen wurden (vgl. Adams und Wu 2007); ebenso Items mit Item-DIF von mehr als 0,638 Logits zwischen den beiden Testzeitpunkten, was einem *großen* DIF entspräche (vgl. Wilson 2005, S. 167). Die Rasch-Analyse wurde einerseits für die nach Wissensfacette in zwei Skalen getrennten Items durchgeführt und andererseits für eine gemeinsame Skala mit allen Items, um die Trennbarkeit der Wissensfacetten zu prüfen. Zur weiteren Analyse der Probanden-Fähigkeiten werden die Personenparameter (WLE-Schätzer) als Testscores verwendet. Dieses Verfahren hat insgesamt den Vorteil, dass die Itemparameter zwischen den Testzeitpunkten nicht variieren (was für eine Vergleichbarkeit der darauf aufbauenden Niveaukonstruktion notwendig ist) und dass die Nutzung von WLEs als Scores reliablere Ergebnisse liefert (Hartig und Kühnbach 2006). Die Verwendung von plausible Values (PV) statt WLE-Schätzer würde außerdem die Aussage über einzelne Individuen erschweren. Der Nachteil, dass Veränderungen der Skalenzusammensetzung oder der zugrunde liegenden Kompetenzstruktur so nicht abgebildet werden können, wird hier zugunsten einer einfacheren Interpretierbarkeit in Bezug zur Fragestellung hingenommen.

Das Niveaumodell wird aus den Item-Parametern nach dem bei Woitkowski (2015) erprobten Verfahren konstruiert. Dabei wird zuvor mittels linearer Regression überprüft, ob die Itemschwierigkeiten gut mit der Komplexität prädiziert werden können (in anderen Studien zeigt sich hier ggf. auch ein Einfluss anderer Itemmerkmale; z. B. Kauertz und Fischer 2006). Alle weitere Skalen werden mit den Mitteln der klassischen Testtheorie ausgewertet. Als Maß für die Reliabilität wird Cronbach's α ermittelt und bei $\alpha < 0{,}6$ die Skala nicht weiter verwendet. Nach der Rasch-Analyse und Skalenbildung wird dann die zwischen TZP 1 und 2 zusammengehörigen Fälle im Datensatz identifiziert und zusammengeführt.

Zu Forschungsfrage F1 werden Unterschiede in den Testzeitpunkten in den jeweiligen Scores berechnet. Für F2 werden Niveau-Belegungen ausgezählt. Bei F3 wird zunächst deskriptiv berichtet, wie viele Studierenden zwischen TZP 1 und 2 von welchem Niveau auf welches Niveau wechseln. Zur Prüfung der Prädiktion der zu TZP 1 belegten Niveaus werden dann die Scores im universitären Wissen zu TZP 2 zwischen den Gruppen verglichen und das Ergebnis mit einer ANOVA abgesichert. Die Interaktion zwischen schulischem und universitärem Wissen wird über einen Vergleich verschiedener ANOVAs und linearer Regressionsmodelle aufgeklärt. Für Frage F4 werden Gruppenunterschiede zwischen den Probanden berichtet, die das Zielniveau zu TZP 2 erreichen, und denen, die es nicht erreichen.

Tab. 4 Überblick über die Skalen zu Einstellungen und Beliefs (jeweils 4-stufige Likert-Skalen) mit Cronbach's α für gesamte Stichprobe und Anzahl der Items. (Quellen: R1–R4: Riese (2009); N1: Neumann, Sorge, Jeschke, Heinze und Neumann (2016); A1–A3: Albrecht (2011); B1–B3: Burger und Groß (2016); S3–S4: adaptiert nach Sundre (2007))

Nr	Skala	Beispielitem	α	N_{Items}
R1	Experimentierbezogener Enthusiasmus	Experimentieren war für mich die interessanteste Tätigkeit in der Schule	0,614	4
R2	Ontologie/Natur des Wissens	Physik beschreibt, wie die Natur wirklich ist. *(inv)*	0,704	4
R3	Leistungsmotivation	Ich mag Situationen, in denen ich feststellen kann, wie gut ich bin	0,812	5
R4	Leistungsängstlichkeit	Wenn ich ein Problem nicht sofort verstehe, werde ich ängstlich	0,838	5
N1	Academic Buoyancy (7-stufige Likert-Skala)	Selbst, wenn ich bei einer schwierigen Physikaufgabe nach mehreren Anläufen keine Lösungsidee habe, versuche ich es immer wieder	0,859	11
A1	Lernschwierigkeiten	Ich lasse mich oft durch andere Dinge vom Lernen abbringen	0,805	10
A2	Studienklima	Es herrscht keine angenehme Arbeitsatmosphäre. *(inv)*	0,710	8
A3	Studienzufriedenheit	Mein Interesse am Studienfach ist im Verlauf meines Studiums weitgehend verloren gegangen. *(inv)*	0,784	5
B1	Leistungszufriedenheit	Meine Leistungserwartungen und -ansprüche haben sich im Studium voll erfüllt	0,732	3
B2	Soziale Integration	Mir ist es während meines bisherigen Studiums gut gelungen, Kontakte zu anderen Studierenden aufzubauen	0,900	3
B3	Prozedurale Gerechtigkeit	Meine Dozenten sind bei der Notenvergabe unvoreingenommen	0,721	6
S3	Importance (Übungszettel)	Es ist mir wichtig, bei den Übungszetteln gut abzuschneiden	0,696	5
S4	Effort (Übungszettel)	Ich versuche, auf jeden Fall alle Aufgaben auf dem Übungszettel zu lösen	0,812	5

Tab. 5 Lineare Regression der Itemparameter auf die Item-Komplexität. Jeweils einmal für jede Wissensfacette. Angegeben sich jeweils nicht-standardisierte Regressionskoeffizienten mit Standardfehler und p-Werte

Unabhängige Variable	Schulisches Wissen b ± SE/Logits	p	Universitäres Wissen b ± SE/Logits	p
(Intercept)	–2,13 ± 0,34	<0,001	–1,00 ± 0,54	0,086
Komplexität II	1,07 ± 0,42	0,016	1,80 ± 0,76	0,035
Komplexität III	2,03 ± 0,45	<0,001	2,68 ± 0,70	0,002
Komplexität IV	3,28 ± 0,65	<0,001	3,12 ± 0,82	0,002
F-Statistik	$F(3,33) = 11,36$, $p < 0,001$		$F(3,13) = 6,44$, $p = 0,007$	
R^2	0,46		0,50	

Für die Gruppenunterschiede wird jeweils der zweiseitige Wilcox-Mann-Whitney-Test genutzt. Dieser ist im Vergleich mit dem gängigen t-Test robuster in Bezug auf Stichprobengröße und Normalverteilung der Daten; das Signifikanzniveau p kann aber analog mit *<0,05; **<0,01; ***<0,001 angegeben und interpretiert werden (Hollander und Wolfe 1973). Bei mehrfaktoriellen Unterschieden wird zusätzlich eine ANOVA gerechnet. Als Effektstärkemaß wird Cohen's d angegeben. Dabei markiert d > 0,2 kleine, d > 0,5 mittlere und d > 0,8 große Effekte (Tiemann und Körbs 2014, S. 291).

Skalenkennwerte

Die Testitems wurden nach Wissensfacetten zu zwei Skalen zusammengeführt (Tab. 3). Im *schulischen Wissen* wurden 2 Items aufgrund zu geringen Infits ausgeschlossen. Die WLE-Reliabilität ist im *schulischen Wissen* akzeptabel, im universitären Wissen jedoch schwach. Dies kann zum Teil auf die kurze *universitäre* Skala zurückgeführt werden, die durch das rotierende Testheftdesign zusätzlich verkürzt wird (Adams 2005). Zum Teil scheint es sich aber auch um einen Prä-Test-Effekt zu handeln, wie die zum Vergleich nach Testzeitpunkten getrennt durchgeführte Analyse zeigt, bei der die Reliabilität zu TZP 2 etwas höher liegt. Im Falle eines Prä-Tests kann die WLE-Reliabilität jedoch i. d. R. nicht interpretiert werden (Rost 2004, S. 382). Die in diesem Fall besser interpretierbare EAP-Reliabilität ist in allen Fällen akzeptabel.

Unabhängig von den konkreten Ursachen bedeutet die geringe WLE-Reliabilität anschaulich einen relativ hohe Messunsicherheit der einzelnen Probandenfähigkeiten. Die auf dieser Basis durchgeführte Niveauzuordnung wird damit besonders an de Niveaugrenzen unsicherer, als sie bei angemessener Reliabilität wäre. Die Messunsicherheit führt außerdem zu einer geringer angegebenen Signifikanz von Korrelationen der Probandenfähigkeiten mit anderen Größen (Adams 2005), Effekte werden im Folgenden also möglicherweise in ihrer Bedeutsamkeit unterschätzt. Die Nutzung von plausible Values (PV) würde hier zwar Abhilfe

Tab. 6 Mittlere Itemparameter (in Logits), die im Rahmen der Niveaukonstruktion als untere Niveaugrenzen herangezogen werden. Im *schulischen Wissen* umfasst das mittlere Niveau die Komplexitäten (II) und (III) im *universitären Wissen* wurden jeweils zwei Komplexitäten zu einem Niveau zusammengezogen

	Schulisches Wissen	Universitäres Wissen
(I) Fakten	–2.13	–0,11
(II) Prozessbeschreibungen	–0,65	
(III) Lineare Kausalität		1,82
(IV) Multivariate Interdependenz	1,15	

schaffen, erschwert aber die Niveauzordnung (s. oben), so dass hier mit WLE-Schätzern weitergearbeitet wird.

Zur Überprüfung der Passung eines Modells mit zwei Skalen wurde außerdem zum Vergleich eine Gesamtskala mit allen Items erstellt. Fit-Indizes und Modellvergleiche zeigt Tab. 3. Der AIC spricht knapp für getrennte Skalen, der BIC eher für eine gemeinsame Skala. Ein χ^2-Test ist knapp nicht signifikant. Dieses uneinheitliche Bild lässt somit prinzipiell beide Modelle zu. Die nach Testzeitpunkten getrennte Analyse spricht jeweils stärker für eine gemeinsame Skala. Dies lässt sich so interpretieren, dass die beiden Skalen zu jedem Testzeitpunkt hoch miteinander korrelieren, sich dieser Zusammenhang zwischen den TZP aber verschiebt, bei der gemeinsamen Analyse beider Testzeitpunkte also geringer ausfällt. Dies spricht für eine differentielle Entwicklung der Wissensfacetten zwischen den Testzeitpunkten und somit für eine getrennte Analyse. Da diese auch theoretisch abgrenzend beschrieben sind und es im Erkenntnisinteresse liegt, denjenigen Wissensbestand, der aus der Schule mitgebracht werden sollte, von demjenigen, der in der Universität erworben werden sollte, zu trennen, wird im Folgenden mit dem nach Facetten getrennten Modell weiter gerechnet.

Zur Überprüfung der Homogenität zwischen den Testzeitpunkten stellt Tab. 3 außerdem Varianz und Reliabilität für jeden Testzeitpunkt dar. Diese befinden sich jeweils in derselben Größenordnung für eine gemeinsame Skalierung der Testzeitpunkte. Die Item-Parameter korrelieren höchst signifikant zwischen den Testzeitpunkten (Schulwissen: $r(35) = 0{,}833$; $p < 0{,}001$; Universitäres Wissen: $r(15) = 0{,}820$; $p < 0{,}001$). Die gemeinsame Skalierung mit der Methode der virtuellen Fälle kann somit durchgeführt werden (vgl. bei Seifert und Schaper 2012).

Die so entstandenen Skalen korrelieren latent mäßig hoch miteinander ($r_{lat} = 0{,}884$, man beachte, dass latente Korrelationen zahlenmäßig deutlich höher ausfallen als manifeste; Wu et al. 1998).

Die Skala zur Mathematik wurde klassisch ausgewertet. Sie ist mit Cronbach's $\alpha = 0{,}76$ als hinreichend reliabel anzusehen; im Mittel wurden 73 % (SD = 19 %) der Items korrekt gelöst. Die Skalen zu Beliefs, Einstellungen, Motivation und Wahrnehmung des eigenen Studiums stellt Tab. 4 dar. Die beiden Skalen zu Importance und Effort mit Bezug auf die Klausur zeigten auch nach einer Bereinigung ein Cronbach's $\alpha < 0{,}6$ für die gesamte Stichprobe und wurden von der Analyse ausgeschlossen. Die Importance- und Effort-Skalen mit Bezug zum Übungszettel können aber interpretiert werden.

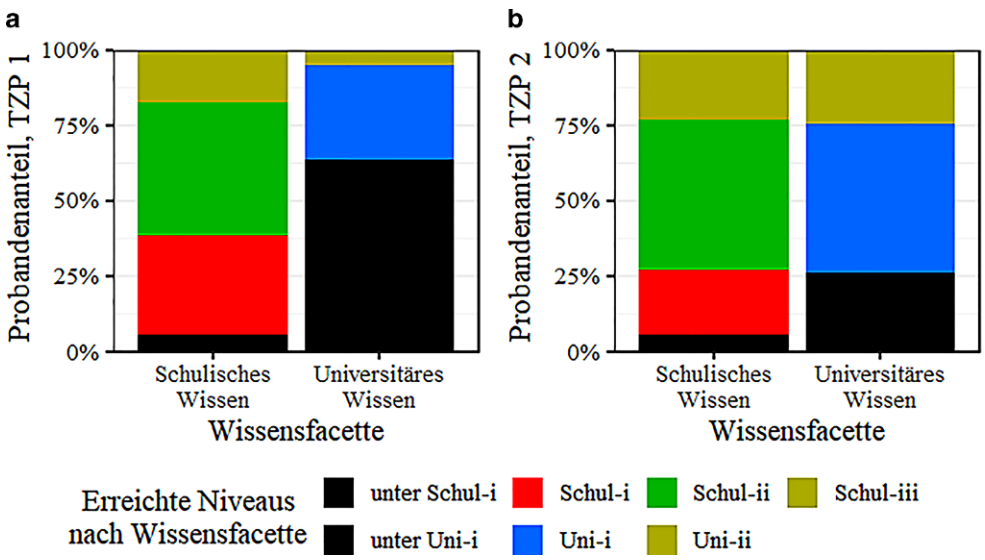


Abb. 2 Häufigkeit der erreichten Niveaus zu TZP 1 (**a**) und TZP 2 (**b**)

Tab. 7 Zuwächse der *schulischen* und *universitären* Testscores (Mittlere Personenparameter, jeweils M ± SD) zwischen den Testzeitpunkten sowie der WMW-Test für abhängige Stichproben und Cohen's d

Skala	Mittlerer Score TZP 1	Mittlerer Score TZP 2	Zuwachs	WMW-Test	Effektstärke d
Schulisch	–0,23 ± 1,40	0,17 ± 1,44	0,39 ± 1,35	V = 2559	0.28** (klein)
Universitär	–0,48 ± 1,42	0,78 ± 1,58	1,26 ± 1,70	V = 1164	0.84*** (groß)

Niveaukonstruktion

Auf Basis der Testwerte kann mit dem folgenden Verfahren (Hartig 2007; Schaper et al. 2008) für jede Wissensfacette ein Niveaumodell erstellt werden (Woitkowski und Riese 2017; ausführlich bei Woitkowski 2015, Kap. 17):

1. Die Itemparamter (d. h. die Itemschwierigkeit) aller Items der jeweiligen Wissensfacette werden bestimmt.
2. Für jede Komplexität wird der Mittelwert aller Itemparameter der Items dieser Komplexität bestimmt.
3. Es wird geprüft, (a) ob sich die Itemparameter zwischen den Komplexitäten signifikant voneinander unterscheiden und (b) ob sich die mittleren Itemparameter absolut um mindestens 1,1 Logits unterscheiden – dies ist gleichbedeutend mit einem Unterschied von 25 % in der Lösungswahrscheinlichkeit. Trifft ein Kriterium nicht zu, müssen Itemgruppen so zusammengefasst werden, dass die Gruppen mehrere benachbarte Komplexitäten umfassen, bis die beiden Bedingungen erfüllt sind (Hartig 2007). Dies sichert voneinander abgrenzbar interpretierbare Niveaus.
4. Als Niveaugrenzen werden nun die mittleren Itemparameter dieser Gruppen angesetzt (Schaper et al. 2008).
5. Die Fähigkeiten der Probanden werden anhand der Personenparameter den Niveaus zugeordnet. Fähigkeiten unterhalb der unteren Grenze liegen auf *Niveau unter i*, Fähigkeiten zwischen der unteren und zweit-unteren Grenze liegen auf *Niveau i* usw. Das Niveaumodell kann nun so interpretiert werden, dass Probanden mit Fähigkeiten eines Niveaus das typische Item der darunterliegenden Gruppe hinreichend wahrscheinlich (d. h. mit mindestens 50 % Wahrscheinlichkeit) lösen können, das typische Item der nächsten Gruppe aber nicht (ausführliche Diskussion der Details der Niveauinterpretation bei Woitkowski und Riese 2017).

Das Verfahren basiert auf der Annahme, dass die Itemkomplexität das vorrangige schwierigkeitserzeugende Aufgabenmerkmal darstellt. Zur Prüfung wurde für jede der beiden Skalen die Itemschwierigkeit in einem linearen Regressionsmodell auf die Itemkomplexität zurückgeführt. Die Regressionskoeffizienten zeigt Tab. 5. Wie erwartet steigt die Itemschwierigkeit mit der Komplexität.

Nun wird die Niveaukonstruktion für beide Wissensfacetten nach dem o. g. Verfahren durchgeführt. Wie Tab. 5 zeig, liegen im *schulischen Wissen* die Itemgruppen der Komplexitäten (II) und (III) im Mittel nur 2,03 Logits – 1,07 Logits = 0,96 Logits auseinander (statt der geforderten 1,1 Logits), so dass diese Itemgruppen zu einem Niveau zusammengelegt wurden. Hier wurden also nur drei Niveaus definiert. Im universitären Wissen tritt dasselbe Problem noch einmal auf. Hier wurden die Komplexitäten (I) und (II) sowie (III) und (IV) zusammengelegt, so

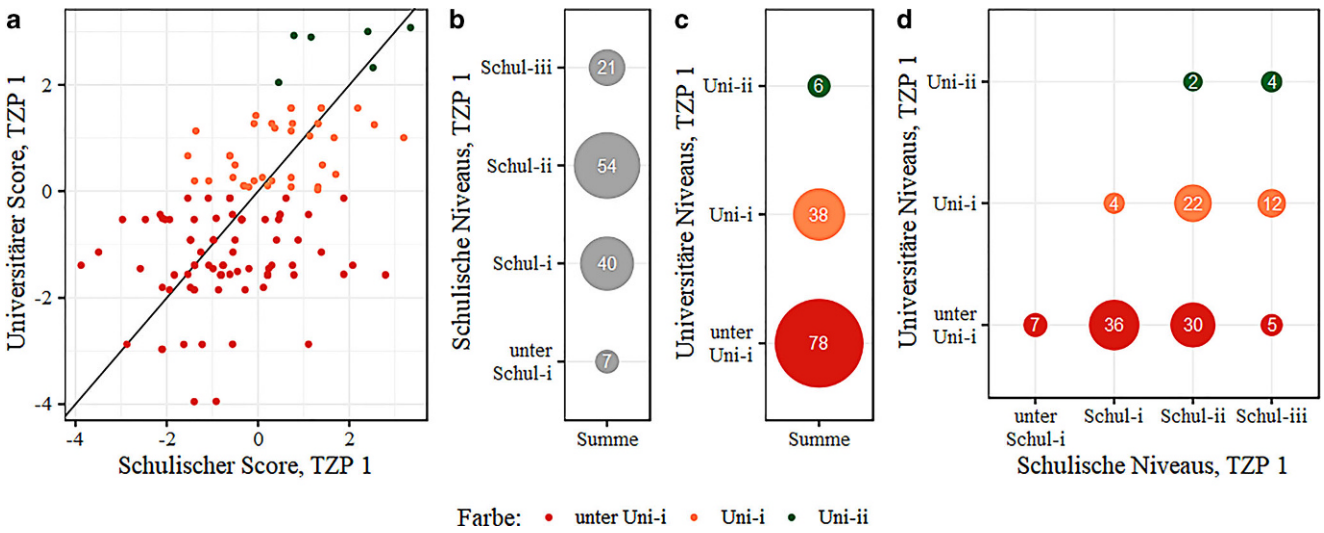


Abb. 3 Zusammenhang in der Verteilung der erreichten **a** Scores (WLE) bzw. **b–d** Niveaus im *schulischen* und *universitären Wissen*

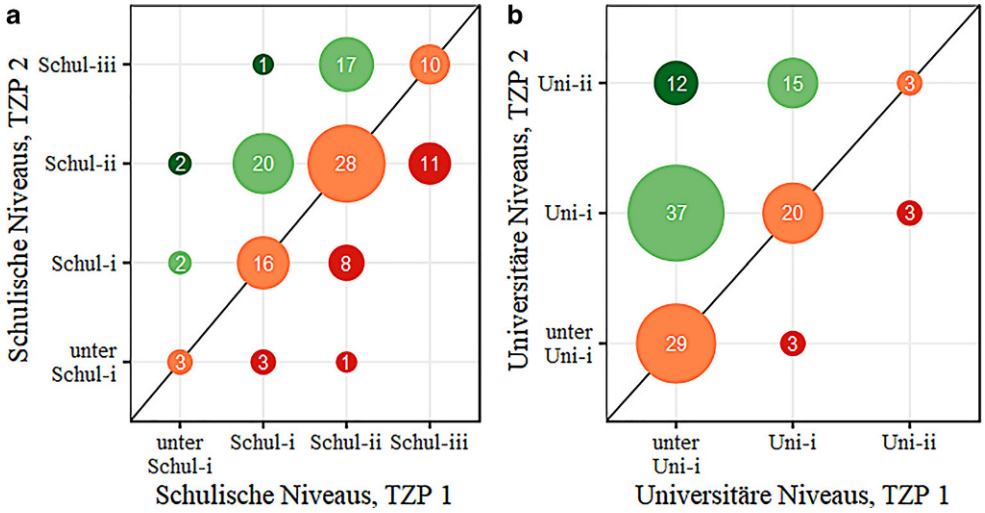


Abb. 4 Anstieg der erreichten Niveaus zwischen den beiden Testzeitpunkten. Personen oberhalb der Diagonale (*grün*) erreichen zu TZP 2 ein höheres Niveau als zu TZP 1. Personen darunter (*rot*) ein niedrigeres. Probanden auf der Diagonale (*orange*) zeigen keine Veränderung im erreichten Niveau

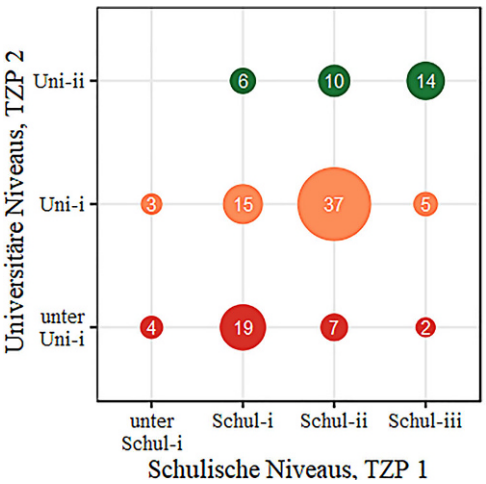


Abb. 5 Erreichtes *universitäres* Niveau zu TZP 2 in Abhängigkeit des *schulischen* Niveaus zu TZP 1

dass hier sogar nur zwei Niveaus definiert sind.[2] Diese Zusammenfassung ist kein Effekt der hier betrachteten Stichprobe sondern zeigt sich auch bei einer von dieser disjunkten Stichprobe (Woitkowski 2015, Kap. 17). Die so festgelegten Niveaugrenzen zeigt Tab. 6. Eine nach Testzeitpunkten getrennte Niveaubelegung zeigt Abb. 2.

In Analogie zu den vorhandenen Niveaus kann das mittlere Niveau im schulischen Wissen nun so interpretiert werden, dass die Studierenden hier Anforderungen, die die Nutzung von einfachen Zusammenhängen erfordern, hinreichend gut beherrschen. Eine Reanalyse der Items der universitären Skala liefert zudem folgende Interpretationen für die Niveaus im universitären Wissen: Das untere Niveau bezieht sich auf Anforderungen, bei denen physikalische Fakten oder einfache Zusammenhänge zur Lösung genutzt werden, die zwar stark mathematisiert formuliert werden, die aber keine Rechnung zur Begründung erfordern. Das obere Niveau hingegen umfasst Anforderungen, bei denen ein- oder mehrschrittige Zusammenhänge zur Bewältigung genutzt werden, bei denen durchgängig auch eine Rechnung erforderlich ist (vgl. Woitkowski 2015, Kap. 17).

Ergebnisse

F1: Zuwächse im ersten Semester

Als Zuwachs im ersten Semester wird in Tab. 7 jeweils die Differenz zwischen den Testscores (WLE) zu TZP 1 und 2 angegeben. Dabei zeigt sich, dass der Score im *schulischen Wissen* mit geringer Effektstärke und Signifikanz ansteigt, im *universitären Wissen* zeigt sich dagegen ein höchst-signifikanter Anstieg mit großer Effektstärke. In beiden Skalen ist die Varianz des Zuwachses vergleichsweise groß.

Dieser Befund ist konsistent mit der Annahme, dass das im Test als *schulisch* erfasste Wissen in geringerem Maße Gegenstand universitärer Lehrveranstaltungen sein sollte als die *universitäre* Wissensfacetten. Eine ANOVA mit Messwiederholung (Innersubjektfaktoren Testzeitpunkt und Wissensfacette) zeigt einen hochsignifikanten Interaktionseffekt ($F(1, 121) = 26{,}89$; $p < 0{,}001$), der Zuwachs unterscheidet sich also signifikant zwischen *schulischem* und *universitären Wissen*. Dieser Unterschied spricht noch einmal empirisch für die Skalentrennung zwischen *schulischem* und *universitärem Wissen*; die Annahme, die *schulischen* Items wären innerhalb einer gemeinsamen Skala ins-

[2] Es wäre im universitären Wissen im Prinzip jeweils möglich, andere Itemgruppen zusammenzufassen, so dass man andere Niveaus erhielte. Neben der Diskussion bei Woitkowski (2015, Kap. 17) über Vor- und Nachteile, sei hier vor allem darauf hingewiesen, dass bei der gewählten Niveaubildung ein oberes Niveau entsteht, das als Zielniveau für die universitäre Ausbildung festgelegt werden kann, unabhängig davon ob man die Komplexität (III) Lineare Kausalität oder (IV) Multivariate Interdependenz als angemesseneres Ziel ansieht.

Tab. 8 Drei mögliche Varianzanalyse-Modelle (ein-/mehrfaktorielle ANOVA). Der universitäre Score zu TZP 2 wird durch Niveaus zu TZP 1 erklärt. Angegeben sind jeweils Faktoren und F-Statistik des ANOVA-Modells

	AM1	AM2	AM3	AM4
Unabhängige Variable	Universitärer Score TZP 2			
Faktoren (TZP 1)	Niveau Schul***	Niveau Uni***	Niveau Schul** Niveau Uni***	Niveau Schul** Niveau Uni*** Interaktionseffekt
F-Statistik (Modell)	F(3,118) = 9,919, $p < 0{,}001$	F(2, 119) = 12,07, $p < 0{,}001$	F(3, 116) = 4,05, $p = 0{,}008$ F(2, 116) = 12,99, $p < 0{,}001$	F(3, 113) = 0,331, $p = 0{,}009$ F(2, 113) = 12,79, $p < 0{,}001$ F(3, 113) = 0,331, $p = 0{,}8$

Tab. 9 Lineare Regression des universitären Scores zu TZP 2 mit den Niveaus zu TZP 1

		LM1		LM2		LM3	
	N_{Niveau}	b ± SE/Logits	*p*	b ± SE/Logits	*p*	b ± SE/Logits	*p*
Intercept	–	–0,48 ± 0,54	0,38	0,30 ± 0,16	0,07	–0,48 ± 0,53	0,36
Schul-i	40	0,58 ± 0,58	0,33	–	–	0,50 ± 0,57	0,39
Schul-ii	54	1,48 ± 0,57	0,011	–	–	1,12 ± 0,58	0,05
Schul-iii	21	2,42 ± 0,62	<0,001	–	–	1,80 ± 0,65	0,006
Uni-i/Uni-ii	38 + 6	–	–	1,34 ± 0,27	<0,001	0,81 ± 0,31	0,010
F-Statistik	–	F(3,118) = 9,919, $p < 0{,}001$		F(1,120) = 24,33, $p < 0{,}001$		F(4,117) = 9,55	
R^2	–	0,18		0,16		0,22	

gesamt einfacher als die *universitären*, würde diesen Befund nicht erklären können.

F2: Zu Beginn erreichte Niveaus nach Wissensfacette

Zunächst kann der Wissensstand der Probanden zu Beginn des Studiums (TZP 1) analysiert werden. Abb. 3b, c zeigt, dass die Probanden zu Studienbeginn im *schulischen Wissen* vor allem mittlere bis obere Niveaus belegen, im *universitären Wissen* aber lediglich ein Drittel (36 %) überhaupt über das Niveau *unter Uni-i* hinauskommt. Auf dieser ersten Analyseebene kann also gesagt werden, dass ein wesentlicher Teil der Studienanfänger vor allem mit *schulischem Wissen* ins Studium startet.

Setzt man nun die beiden Wissensfacetten in Beziehung, zeigt Abb. 3a zunächst einen Plot der Scores (Probanden-Parameter) im *universitären* über dem *schulischen Wissen*. Diese beiden korrelieren signifikant miteinander (Pearson-Korrelation der WLE-Scores zu TZP 1: r = 0,55, $p < 0{,}001$). Abb. 3d zeigt denselben Zusammenhang auf Ebene der Niveaus. Hier lässt der Zusammenhang durch die Betrachtung der Niveaus dahingehend präzisieren, dass das Belegen eines hohen *schulischen* Niveaus notwendige, keinesfalls jedoch hinreichende Bedingung für das Belegen eines hohen *universitären* Niveaus ist. Demgegenüber kommen hohe *schulische* Niveaus in Kombination mit niedrigen *universitären* Niveaus durchaus vor.

F3: Universitäre Niveaus zu Semesterende

Die Betrachtung durch die Brille des Niveaumodells liefert gegenüber Forschungsfrage F1 die zusätzliche Information, ob die Probanden zu TZP 2 substantiell schwierigere Problemstellungen lösen können als zu TZP 1, was hier durch die Verortung auf einem höheren Niveau operationalisiert wird. Abb. 4 zeigt zunächst für das *schulische* und *universitäre Wissen* den Anstiegt der von den Probanden erreichten Niveaus zwischen den beiden Testzeitpunkten.

Im *schulischen Wissen* zeigt sich eine schwache Tendenz hin zu höheren Niveaus mit viel Varianz (Niveauanstiege bei 34 %, -abstiege bei 23 % der Probanden). Dagegen zeigen sich im *universitären Wissen* bei etwa der Hälfte (52 %) der Probanden Niveauanstiege. Das ist konsistent mit den Zuwächsen in den Scores in F1, wo der Anstieg im *schulischen Wissen* ebenfalls deutlich geringer ausfällt. Es ist außerdem konsistent mit der Annahme, dass im Studium vor allem dasjenige Wissen thematisiert wird, was hier durch die Skala zum *universitären Wissen* erhoben wird. Im *schulischen Wissen* sind also durchaus Vergessenseffekte oder auch Missverständnisse durch die neue und vielleicht ungewohnte Darstellung in der Universität als Gründe für Niveauabfälle denkbar. Dort, wo im *schulischen Wissen* ein Niveauanstieg sichtbar ist, könnte umgekehrt der (an sich wünschenswerte) Fall angenommen werden, dass die universitären Lehrveranstaltungen zu einem besseren Verständnis des *schulischen Wissens* geführt hat. Ohne eine

Tab. 10 Überblick über Gruppenunterschiede zwischen der Uni-ii-Gruppe und den restlichen Probanden. Angegeben sind die Scores der Skalen (M ± SD) sowie WMW-Test-Statistik für unabhängige Stichprobe und Cohen's d als Effektstärke

Skala	Erreichen Uni-ii	Erreichen Uni-ii nicht	WMW-Test	Effektstärke d
Mathematik-Score, TZP 1	0,83 ± 0,11	0,64 ± 0,20	W = 556	1,05*** (groß)
Mathematik-Score, TZP 2	0,86 ± 0,10	0,74 ± 0,19	W = 803	0,72*** (mittel)
Mathematik-Zuwachs TZP 1 → TZP 2	0,03 ± 0,10	0,10 ± 0,14	W = 1746	0,51* (mittel)
R2: Ontologie/Natur des Wissens TZP 1	2,74 ± 0,70	2,45 ± 0,62	W = 943	0,45 (n. s.)
R2: Ontologie/Natur des Wissens TZP 2	2,88 ± 0,67	2,57 ± 0,68	W = 1021	0,47* (klein)
R3: Leistungsmotivation TZP 1	3,53 ± 0,44	3,37 ± 0,49	W = 897	0,33 (n. s.)
R3: Leistungsmotivation TZP 2	3,52 ± 0,42	3,30 ± 0,46	W = 1005	0,44* (klein)

Beim Mathematik-Score ist der Anteil gelöster Items angegeben
Die Belief-Skalen sind 4-stufige Likert-Skalen
Für Beispielitems siehe Tab. 4. Skalen ohne signifikante Unterschiede wurden unterdrückt

differenzierte qualitative Analyse ist hier aber kaum zu klären, warum bei manchen Probanden der eine, bei anderen Probanden der andere Effekt auftritt.

Aufgrund des Literaturbefundes ist nun eine Prädiktion des *universitären Wissens* zu TZP 2 durch das zu TZP 1 vorhandene *schulische* Wissen zu erwarten – und zwar über den allgemeinen Effekt der Korrelation dieser beiden Wissensfactten hinaus (vgl. z. B. Dawson-Tunik 2006). Dies kann überprüft werden, indem die Probanden nach dem zu TZP 1 belegten *schulischen* Niveau in Gruppen getrennt werden und auf Unterschiede im *universitären* Wissensscore zu TZP 2 überprüft wird. Ein Einfluss auf die universitären *Niveaus* zu TZP 2 (z. B. im Rahmen einer multinomialen logistischen Regression) ist aufgrund der z. T. geringen Besetzungszahlen hier leider nur wenig aussagekräftig.

Die Darstellung in Abb. 5 suggeriert zunächst den genannten Zusammenhang. Rechnerisch zeigt eine einfaktorielle ANOVA hier einen hochsignifikanten Effekt ($F(3,118) = 9{,}919$; $p < 0{,}001$) des *schulischen* Niveaus zu TZP 1 auf den *universitären* Score zu TZP 2.

Da die Scores im *schulischen* und *universitären Wissen* jeweils miteinander korrelieren, könnte es sich hier auch um einen indirekten „Matthäus-Effekt" handeln, derart dass ein hohes *schulisches* Niveau zu TZP 1 qua Korrelation mit einem hohen *universitären* Niveau zu TZP 1 korrespondiert und dies direkt zu einem höheren *universitären* Niveau zu TZP 2 führt, ohne dass das *schulische Wissen* einen direkten Einfluss hätte.

Um dies zu überprüfen, werden vier ANOVA-Modelle miteinander verglichen (Tab. 8), in denen der *universitäre* Score zu TZP 2 durch verschiedene Kombinationen von Niveaus zu TZP 1 aufgeklärt wird. Tatsächlich zeigt sich sowohl in AM2 (*universitäre* Niveaus) als auch in AM1 (*schulische* Niveaus) ein signifikanter Effekt. Nimmt man beide Niveaus als Faktoren zusammen (AM3), sind die Effekte beider Niveaus signifikant. Einen signifikanten Interaktionsterm findet man hingegen nicht (AM4), der Einfluss einer Wissenfacette wird also nicht durch eine hohe oder niedrige Ausprägung der anderen Facette verstärkt oder gemindert. Insgesamt lässt sich somit der „Matthäus-Effekt" soweit spezifizieren, dass über Vorkenntnisse in der *universitären Wissensfacette* hinaus *schulisches* Vorwissen einen hochsignifikanten Mehrwert bietet.

Um die prädiktiven Niveaus konkret zu bestimmen, kann nun ein lineares Regressionsmodell mit Niveaus als Prädiktoren genutzt werden (Tab. 9). Für das *schulische Wissen* zeigt sich hier jeweils, dass ein Erreichen des Niveaus *Schul-iii* (und in LM1 schwächer signifikant *Schul-ii*) einen signifikanten Vorteil für den *universitären* Wissenserwerb bietet. Beim *universitären Wissen* ist zu TZP 1 das Niveau *Uni-ii* zu gering besetzt, um als eigenständiger Prädiktor eingesetzt zu werden. Ein Modell mit dem Prädiktor *Uni-i oder Uni-ii* zeigt aber einen signifikanten Effekt. Wie schon bei den Varianzanalysen in Tab. 9 zeigt sich hier in LM3 der additive Effekt des *schulischen* und des *universitären Wissens*, die beide signifikante Beiträge liefern. Vergleicht man die Passung von LM3 mit einem linearen Regressionsmodell, dass als Prädiktoren einfach den *schulischen* und *universitären* Score zu TZP 1 enthält, ergibt sich kein signifikanter Unterschied in der Modellpassung zu LM3 ($F(3,116) = 0{,}9885$; $p = 0{,}4$). Durch die Diskretisierung von Scores in Niveaus entsteht also kein relevanter Verlust der Prädiktionskraft oder der Modellpassung.

Zusammengefasst erscheint das *schulische Wissen* zu TZP 1 also inkrementell prädiktiv für das Erreichen eines höheren *universitären* Niveaus zu TZP 2. Dabei unterscheiden sich Probanden, die auf oder unter dem Niveau *Schul-i* in ihr Studium starten deutlich von der Gruppe auf höheren *schulischen* Niveaus, der Unterschied liegt also in der Fähigkeit zum Umgang mit Anforderungen zu deren Lösung die Nennung von Fakten nicht ausreicht, sondern (ein- oder) mehrschrittige Zusammenhänge hergestellt werden müssen.

F4: Studierende, die das höchste universitäre Niveau erreichen

Zu TZP 2 erreichen nur 30 der 122 Probanden das Niveau *Uni-ii*. Diese Uni-ii-Gruppe wird nun bezüglich der erhobenen Begleitvariablen mit der Rest-Gruppe verglichen.

In der Uni-ii-Gruppe sind 13,3 % der Probanden weiblich, in der Rest-Gruppe 32,6 %. Der Unterschied ist im WMW-Test knapp signifikant ($W = 1114$; $p = 0{,}042$). Die beiden Gruppen unterscheiden sich nur in wenigen Merkmalen (Tab. 10; Merkmale ohne signifikante Gruppenunterschiede wurden unterdrückt). Zunächst ist hier der Mathematik-Score zu nennen, der sich zu beiden Testzeitpunkten deutlich zugunsten der Uni-ii-Gruppe unterscheidet. Allerdings ist der Zuwachs bei der Rest-Gruppe signifikant größer. Dabei kann ein Deckeneffekt jedoch nicht ausgeschlossen werden.

Weitere Unterschiede zeigen nur zwei Begleitskalen. Die Uni-ii-Gruppe weist zu TZP 2 angemessenere Vorstellungen bezüglich der Natur der Naturwissenschaften auf (Skala R2). Zu TZP 2 zeigt die Rest-Gruppe eine niedrigere allgemeine Leistungsmotivation (R3). Zu TZP 1 sind die Unterschiede jeweils nicht signifikant, sie bilden sich also erst im Laufe des Semesters aus.

Zusammenfassung und Diskussion

Die Qualität von Wissensstrukturen wird im vorliegenden Beitrag mit Hilfe von nach *hierarchischer Komplexität* (Bernholt 2010) gestaffelten Items erfasst, die den Verknüpfungsgrad des physikalischen Fachwissen operationalisieren (Peuckert und Fischler 2000). Dazu werden die Fähigkeiten der Probanden im Rahmen eines Niveaumodells für die Wissensfacetten des schulischen und universitären Wissens Niveaus zugeordnet, die die von ihnen jeweils bewältigbare Komplexität wiedergeben.

Die Zuordnung von Items zu Komplexitäten einerseits und zu Wissensfacetten andererseits basiert auf disjunkten Kriterien, die Itemmerkmale sind so weit wie möglich orthogonal zueinander. Für die Zuordnung zu Wissensfacetten kann das unterschiedliche Ansteigen der Scores, bei der Zuordnung zu Komplexitäten die Prädiktion der Itemschwierigkeit als Hinweis auf die Validität der Zuordnung verstanden werden. Aussagen über die zeitliche Stabilität dieser Zuordnung liegen hier nicht im Fokus. Daher wurde ein Vergleich der hier theoretisch begründeten Skalenzusammensetzung mit einer explorativen Skalenbildung nicht durchgeführt.

Bei der Rasch-Analyse der Daten von 122 Probanden zu zwei Testzeitpunkten zu Beginn und Ende des ersten Semesters wurde das Verfahren der virtuellen Personen gewählt, bei dem die Item-Parameter zwischen den Testzeitpunkten konstant gehalten werden, so dass das Niveaumodell zwischen den Testzeitpunkten übertragbar ist.

Bei der Interpretation muss klargestellt werden, dass es sich zwar um einen echten Längsschnitt, dennoch aber nur um zwei Messzeitpunkte handelt. Das rotierende Testheftdesign verhindert zwar Erinnerungseffekte so weit wie möglich, Effekte der Tagesform oder der Konzentration auf einen (möglicherweise wenig relevant eingeschätzten) Test können aber nicht ausgeschlossen werden. Weiterhin bezieht sich die Niveauinterpretation immer nur auf *typische Items* einer Itemgruppe (vgl. Woitkowski und Riese 2017) und Mittelwerte bilden immer nur die *typischen Probanden* einer Gruppe ab. Weiterhin weist die Skala zum *universitären Wissen* eine sehr geringe Reliabilität auf, was die Niveauzuordnung zusätzlich unsicherer macht und insgesamt eine höhere Messunsicherheit bedeutet.

In der Analyse zeigt sich eine mittlere längsschnittliche Wissenszunahme mit kleinem Effekt im *schulischen* und mit großem Effekt im *universitären Wissen* (das ist konsistent mit den Befunden von Buschhüter et al. 2017). Da es sich in Bezug auf demographische Merkmale um eine Positivauswahl handelt, muss jedoch angenommen werden, dass es eine wesentliche Teilpopulation mit wahrscheinlich geringerem Wissenszuwachs gibt, von der hier keine Daten vorliegt. Dieses Problem zeigt sich in einem Vergleich der Schulnoten, die in anderen Studien regelmäßig hoch mit den fachlichen Testscores korrelieren (Woitkowski 2015).

Zu Studienbeginn erreichen die Probanden im *schulischen Wissen* typischerweise mittlere und im *universitären Wissen* häufiger nur das unterste Niveau. Der bis zum Semesterende beobachtete Niveauanstieg im *universitären Wissen* kann dabei nicht allein mit einem höheren *universitären* Vorwissen erklärt werden. Vielmehr ist das *schulische* Vorwissen und hier vor allem das Erreichen der oberen Niveaus inkrementell prädiktiv für ein hohes *universitäres Wissen* zu Semesterende. Eine Analyse auf Einzelniveauebene ist hier aufgrund der geringen Probandenzahlen nur begrenzt möglich. Dennoch erlauben die Analysen die Interpretation, dass vor allem die Fähigkeit zum Umgang mit komplexen, d. h. aus verschiedenen Einzelschritten und Verknüpfungen zusammengesetzten Begründungslinien, welche die oberen Niveaus charakterisieren, einen relevanten Prädiktor für den universitären Wissenserwerb darstellt. Liegt diese Fähigkeit für das Schulwissen vor, so ist sie leichter auf die universitäre Physik übertragbar (Dawson-Tunik 2006).

Formuliert man das Erreichen des obersten *universitären* Niveaus als Zielkriterium, zeigen die (relativ wenigen) Probanden, die es erreichen, bereits zu Studienbeginn höhere Mathematik-Scores (vgl. auch Buschhüter et al. 2016). Probanden mit in diesem Sinne gelingendem Fachwissenserwerb zeigen am Semesterende reflektiertere epistemologische Vorstellungen, die restlichen Probanden zeigen zu

Semesterende eine geringere Leistungsmotivation. Beide Unterschiede treten zu Semesterbeginn nicht auf. Andere Merkmale wie Academic Buoyancy, Studienklima und -zufriedenheit sowie soziale Integration und wahrgenommene prozedurale Gerechtigkeit unterscheiden sich zwischen den beiden Probandengruppen nicht. Da diese Merkmale häufig mit Studienabbruch in Verbindung gebracht werden (Burger und Groß 2016; Heublein et al. 2014), kann aber auch hier die Positivauswahl eine Rolle spielen.

Insgesamt liefert die Betrachtung von Niveaus hier wesentlichen Mehrwert. Einerseits, weil ein Zielkriterium für das *universitäre Wissen* angegeben werden kann, wobei der zu erreichende Score durch Anforderungscharakteristika statt durch eine letztlich willkürliche numerische Grenze festgelegt ist (Klieme et al. 2003). Andererseits, ermöglicht die Niveaubetrachtung die Angabe, ab welchem Niveau das Vorwissen der Studierenden tatsächlich einen positiven Effekt auf den Wissensstand zu Semesterende hat. Der Informationsverlust beim Übergang von intevallskalierten Testwerten zu ordinalen Niveaus erscheint nicht unerheblich, aber angesichts der interpretativen Vorteile gerechtfertigt.

Mit Blick auf die universitäre Lehrpraxis in der Physik wird von Lehrenden gelegentlich die Auffassung geäußert, dass man hier „von null“ anfange und deshalb schulische Vorkenntnisse für das Physiklernen in der Universität wenig bedeutsam seien. Die hier vorgestellten Ergebnisse legen im Gegenteil nahe, dass gerade die Fähigkeit zum Umgang mit komplexen Problemstellungen prädiktiv für das universitäre Physiklernen ist – auch und gerade, wenn es im Rahmen der schulischen Physik erlernt wurde. Dieser Befund ist auch konsistent mit einer Untersuchung der Vorwissensabhängigkeit von Klausurnoten im Physikstudium (Buschhüter et al. 2017).

Um dieses Vorwissen und vor allem die Fähigkeit im Umgang mit komplexen Problemstellungen zu Studienbeginn zu stärken, wären mehrere Möglichkeiten denkbar. Weit verbreitet, aber in der aktuellen Form nicht sehr wirksam, sind Vorkurse (Buschhüter et al. 2016). Statt der häufigen Form der mathematikzentrierten, könnten physik- und problemlösezentrierte Vorkurse entwickelt werden. Auch in Bezug auf die Schule wäre hier zu fragen, wie dort der Umgang mit komplexen Problemstellungen effektiver als bisher vorbereitet werden kann. Eingangstest erscheinen vor dem Hintergrund der ohnehin als nicht sehr attraktiv wahrgenommenen Studienwahl Physik wohl nicht wünschenswert (s. dazu z. B. auch Merzyn 2010).

Funding Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) – WO 2181/2-1.

Literatur

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*(2–3), 162–172.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients Multinomial Logit model: a generalized form of the Rasch model. In M. von Davier & C. H. Carstensen (Hrsg.), *Multivariate and mixture distribution Rasch models. Extensions and applications* (S. 57–75). New York: Springer. Statistics for Social and Behavioral Sciences.

Agarwala, A. (2015). Physikstudium: Ganz schön Verrechnet. https://www.zeit.de/2015/23/physikstudium-mathematik-hochschule-ranking (Erstellt: 3. Juni 2015). ZEIT Campus.

Albrecht, A. (2011). Längsschnittstudie zur Identifikation von Risikofaktoren für einen erfolgreichen Studieneinstieg in das Fach Physik. http://www.diss.fu-berlin.de/diss/servlets/MCRFileNodeServlet/FUDISS_derivate_000000010456/Dissertation_Druckversion_Andre_Albrecht_UB.pdf. Zugegriffen: 29. Jan. 2014. Dissertation, Freie Universität Berlin. Berlin.

Armon, C., & Dawson, T. L. (1997). Developmental Trajectories in moral reasoning across the life span. *Journal of Moral Education*, *26*(4), 433–453.

Bernholt, S. (2010). *Kompetenzmodellierung in der Chemie – Theoretische und empirische Reflexion am Beispiel des Modells hierarchischer Komplexität*. Berlin: Logos.

Blömeke, S. (2004). Empirische Befunde zur Wirksamkeit der Lehrerbildung. In S. Blömeke, P. Reinhold, G. Tulodziecki & J. Wildt (Hrsg.), *Handbuch Lehrerbildung* (S. 59–91). Bad Heilbrunn: Klinkhardt.

Blüthmann, I., Lepa, S., & Thiel, F. (2008). Studienabbruch und -wechsel in den neuen Bachelorstudiengängen: Untersuchung und Analyse von Abbruchquoten. *Zeitschrift für Erziehungswissenschaft*, *11*(3), 406–429.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation: Für Human- und Sozialwissenschaftler* (4. Aufl.). Berlin, Heidelberg: Springer.

Bosse, E., & Trautwein, C. (2014). Individuelle und institutionelle Herausforderungen der Studieneingangsphase. *Zeitschrift für Hochschulentwicklung*, *9*(5), 41–62.

Burger, R., & Groß, M. (2016). Gerechtigkeit und Studienabbruch. Die Rolle der wahrgenommenen Fairness von Benotungsverfahren bei der Entstehung von Abbruchsintentionen. *Zeitschrift für Erziehungswissenschaft*, *19*(3), 625–647.

Buschhüter, D., Spoden, C., & Borowski, A. (2016). Mathematische Kenntnisse und Fähigkeiten von Physikstudierenden zu Studienbeginn. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*(1), 61–75.

Buschhüter, D., Spoden, C., & Borowski, A. (2017). Studienerfolg im Physikstudium: Inkrementelle Validität physikalischen Fachwissens und physikalischer Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*(1), 127–141. https://doi.org/10.1007/s40573-017-0062-7.

Busker, M. (2014). Entwicklung eines Fragebogens zur Untersuchungen des Fachinteresses. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 269–281). Berlin: Springer.

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, *18*, 237–278.

Dawson-Tunik, T. L. (2006). Stage-like patterns in the development of conceptions of energy. In X. Liu & W. J. Boone (Hrsg.), *Applica-*

tions of Rasch measurement in science education (S. 111–136). Maple Grove: JAM.
Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values and goals. *Annual Review of Psychology*, *53*(1), 109–132.
Freyer, K. (2013). *Zum Einfluss von Studieneingangsvoraussetzungen auf den Studienerfolg Erstsemesterstudierender im Fach Chemie*. Berlin: Logos.
Fries, M. (2002). Abitur und Studienerfolg: Welchen „Wert" hat das Abitur für ein erfolgreiches Studium? *Beiträge zur Hochschulforschung*, *24*(1), 30–51.
Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung* (S. 83–99). Weinheim: Beltz.
Hartig, J., & Kühnbach, O. (2006). Schätzung von Veränderung mit „plausible values" in mehrdimensionalen Rasch-Modellen. In A. Ittel & H. Merkens (Hrsg.), *Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft* (S. 27–44). Wiesbaden: VS.
Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J., et al. (2017). *Studienerwartungen und Studienwirklichkeit, Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen*. Hannover: DZHW.
Heublein, U., Richter, J., Schmelzer, R., & Sommer, D. (2014). *Die Entwicklung der Studienabbruchquoten an den deutschen Hochschulen: Statistische Berechnungen auf der Basis des Absolventenjahrgangs 2012*. Hannover: DZHW.
Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York: Wiley.
Institut zur Qualitätsentwicklung im Bildungswesen (IQB) (2013). Kompetenzstufenmodelle zu den Bildungsstandards im Fach Physik für den Mittleren Schulabschluss: Kompetenzbereiche „Fachwissen" und „Erkenntnisgewinnung". http://www.iqb.hu-berlin.de/bista/ksm/KSM_Physik.pdf. Zugegriffen: 29. Okt. 2013.
Kauertz, A. (2008). *Schwierigkeitserzeugende Merkmale physikalischer Leistungstestaufgaben*. Berlin: Logos.
Kauertz, A., & Fischer, H. E. (2006). Assessing students' level of knowledge and Analysing the reasons for learning difficulties in physics by Rasch analysis. In X. Liu & W. J. Boone (Hrsg.), *Applications of Rasch measurement in science education* (S. 121–246). Maple Grove: JAM.
Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Berlin: Logos.
Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2003). *Zur Entwicklung nationaler Bildungsstandards: Eine Expertise. Stand Juni 2003*. Bonn: BMBF.
Konferenz der Fachbereiche Physik (KFP) (2010). *Zur Konzeption von Bachelor- und Master-Studiengängen in der Physik: Handreichung*. Berlin: KFP.
König, J., Darge, K., Klemenz, S., & Seifert, A. (2018). Pädagogisches Wissen von Lehramts-studierenden im Praxissemester: Ziel schulpraktischen Lernens? In J. König, M. Rothland & N. Schaper (Hrsg.), *Learning to Practice, Learning to Reflect? Ergebnisse aus der Längsschnittstudie LtP zur Nutzung und Wirkung des Praxissemesters in der Lehrerbildung* (S. 287–323). Wiesbaden: Springer VS.
Krause, F., & Reiners-Logothetidou, A. (1981). *Kenntnisse und Fähigkeiten naturwissenschaftlich orientierter Studienanfänger in Physik und Mathematik: Die Ergebnisse des bundesweiten Studieneingangstests Physik 1978*. Bonn: Universität Bonn.
Krauss, S., Neubrand, J., 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 Mathematikdidaktik*, *29*(3/4), 223–258.
Kubinger, K. D., Hohensinn, C., Hofer, S., Khorramdel, L., Frebort, M., Holocher-Ertl, S., et al. (2011). Designing the test booklets for Rasch model calibration in a large-scale assessment with reference to numerous moderator variables and several ability dimensions. *Educational Research and Evaluation*, *17*(6), 483–495.
Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in soar: the anatomy of general learning mechanisms. *Machine Learning*, *1*, 11–46.
Lamprecht, J. (2011). *Ausbildungswege und Komponenten professioneller Handlungskompetenz: Vergleich von Quereinsteigern mit Lehramtsabsolventen für Gymnasien im Fach Physik*. Berlin: Logos.
Merzyn, G. (2010). Kurswahlen in der gymnasialen Oberstufe. Leistungskurs Physik, Chemie, Mathematik. In *PhyDid B – Didaktik der Physik – Beiträge zur DPG-Frühjahrstagung Hannover*.
Neumann, I., Sorge, S., Jeschke, C., Heinze, A., & Neumann, K. (2016). Zur Academic Buoyancy von Physikstudierenden. In C. Maurer (Hrsg.), *Authentizität und Lernen – das Fach in der Fachdidaktik* (S. 86–88). Regensburg: Universität Regensburg.
Ohle, A., Fischer, H. E., & Kauertz, A. (2011). Der Einfluss des physikalischen Fachwissens von Primarstufenlehrkräften auf Unterrichtsgestaltung und Schülerleistung. *Zeitschrift für Didaktik der Naturwissenschaften*, *17*, 357–389.
Peuckert, J., & Fischler, H. (2000). Concept Maps als Diagnose- und Auswertungsinstrument in einer Studie zur Stabilität und Ausprägung von Schülervorstellungen. In H. Fischler & J. Peuckert (Hrsg.), *Concept Mapping in fachdidaktischen Forschungsprojekten der Physik und Chemie* (S. 91–116). Berlin: Logos.
Plöger, W., Scholl, D., & Seifert, A. (2016). „Und sie bewegt sich doch!" – Wie spezifische Lerngelegenheiten die bildungswissenschaftlichen Kompetenzen von Lehramtsstudierenden fördern können. *Zeitschrift für Pädagogik*, *62*(1), 109–130.
Putnam, R. T., & Borko, H. (1997). Teacher learning: implications of new views of cognition. In B. J. Biddle, T. L. Good & I. F. Goodson (Hrsg.), *International handbook of teachers and teaching* (S. 1223–1296). Dordrecht: Springer.
Riese, J. (2009). *Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften*. Berlin: Logos.
Rindermann, H., & Oubaid, V. (1999). Auswahl von Studienanfängern durch Universitäten: Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *20*(3), 172–191.
Robitzsch, A., Kiefer, T., & Wu, M. L. (2017). TAM: test analysis modules. https://CRAN.R-project.org/package=TAM
Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion* (2. Aufl.). Bern: Huber.
Schaper, N., Ulbricht, T., & Hochholdinger, S. (2008). Zusammenhang von Anforderungsmerkmalen und Schwierigkeitsparametern der MT21-Items. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare. Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung* (S. 451–480). Münster: Waxmann.
Schecker, H., & Parchmann, I. (2006). Modellierung naturwissenschaftlicher Kompetenz. *Zeitschrift für Didaktik der Naturwissenschaften*, *12*, 45–66.
Schnotz, W. (1994). *Aufbau von Wissensstrukturen: Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten*. Weinheim: Beltz.
Schulmeister, R. (2015). Abwesenheit von Lehrveranstaltungen: Ein nur scheinbar triviales Problem. http://www.rolf.schulmeister.com/pdfs/Abwesenheit.pdf. Zugegriffen: 17. Dez. 2015.
Schwedler, S. (2017). Was überfordert Chemiestudierende zu Studienbeginn? *Zeitschrift für Didaktik der Naturwissenschaften*, *23*(1), 165–179.
Seifert, A., & Schaper, N. (2012). Die Entwicklung von bildungswissenschaftlichem Wissen: Theoretischer Rahmen, Testinstrument, Skalierung und Ergebnisse. In J. König & A. Seifert (Hrsg.), *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erzie-*

hungswissenschaftlichen Lehrerausbildung (S. 183–214). Münster: Waxmann.

Shumba, O., & Glass, L. W. (1994). Perceptions of coordinators of college freshman chemistry regarding selected goals and outcomes of high school chemistry. *Journal of Research in Science Teaching*, *31*, 381–392.

Sorge, S., Petersen, S., & Neumann, K. (2016). Die Bedeutung der Studierfähigkeit für den Studienerfolg im 1. Semester in Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, *22*(1), 165–180.

Sundre, D. L. (2007). *The student opinion scale (SOS), A measure of examinee motivation: test manual*. Harrisonburg: Center for Assessment and Research Studies, James Madison University.

Tiemann, R., & Körbs, C. (2014). Die Fragebogenmethode, ein Klassiker der empirischen didaktischen Forschung. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 283–295). Berlin: Springer.

Vogelsang, C., Borowski, A., Fischer, H. E., Kulgemeyer, C., Reinhold, P., & Riese, J. (2016). ProfiLe-*P* + – Professionskompetenz im Lehramtsstudium Physik. In O. Zlatkin-Troitschanskaia, H. A. Pant, C. Lautenbach & M. Toepper (Hrsg.), *Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen (KoKoHs): Übersicht der Forschungsprojekte*. KoKoHs Working Papers, (Bd. 10, S. 39–43). Mainz, Berlin: Johannes-Gutenberg-Universität, Humboldt-Universität.

Walzer, M., Fischer, H. E., & Borowski, A. (2013). Fachwissen im Studium zum Lehramt der Physik. In S. Bernholt (Hrsg.), *Inquiry-based Learning – Forschendes Lernen*. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung, Hannover, 2012. (S. 530–532). Kiel: IPN.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessungen in Schulen* (2. Aufl. S. 17–32). Weinheim: Beltz.

Wilson, M. (2005). *Constructing measures. An item response modelling approach*. Mahwah: Lawrence Erlbaum.

Woitkowski, D. (2015). *Fachliches Wissen Physik in der Hochschulausbildung: Konzeptionalisierung, Messung, Niveaubildung*. Berlin: Logos.

Woitkowski, D. (2017). Studieneingangsprofile in Fach- und Lehramts-Studiengängen Physik: Eine kontrastierende Analyse auf Basis eines Kompetenzstrukturmodells für Fach-Physiker. *Physik und Didaktik in Schule und Hochschule*, *16*(1), 43–56.

Woitkowski, D., & Borowski, A. (2017). Fachwissen im Lehramtsstudium Physik. In H. Fischler & E. Sumfleth (Hrsg.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik* (S. 57–74). Berlin: Logos.

Woitkowski, D., & Riese, J. (2017). Kriterienorientierte Konstruktion eines Kompetenzniveaumodells im physikalischen Fachwissen. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*(1), 39–52.

Woitkowski, D., Riese, J., & Reinhold, P. (2011). Modellierung fachwissenschaftlicher Kompetenz angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *17*, 289–313.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). *ACER ConQuest: Generalised item reponse modelling software manual*. Melbourne: ACER Press.