ZfDN (2018) 24:131–150
https://doi.org/10.1007/s40573-018-0079-6

ORIGINAL PAPER

Formulierung eines evidenzbasierten Validitätsarguments am Beispiel der Erfassung physikdidaktischer Selbstwirksamkeitserwartungen mit einem neu entwickelten Instrument

Claudia Meinhardt[1] · **Thorid Rabe**[1] · **Olaf Krey**[1]

Eingegangen: 12. September 2017 / Angenommen: 5. Juli 2018 / Online publiziert: 12. Juli 2018

Zusammenfassung

Der vorliegende Artikel fokussiert die Entwicklung eines Validitätsarguments für die Erfassung physikdidaktischer Selbstwirksamkeitserwartungen (SWE) in den Handlungsfeldern „Experimentieren“, „Umgang mit Schülervorstellungen“, „Elementarisieren“ sowie „Umgang mit Aufgaben“ jeweils in den Dimensionen „Planung“ und „Durchführung“. Das neu entwickelte Instrument ist für den Einsatz bei Physiklehramtsstudierenden, PhysikreferendarInnen und Physiklehrkräften geeignet. In einem ersten Schritt wird das zugrundeliegende Konstrukt definiert und dann die intendierte Testwertinterpretation offengelegt sowie die ihr zugrundeliegenden wesentlichen Annahmen expliziert. Der Frage nach der Gültigkeit dieser Annahmen wird im Rahmen von vier Pilotstudien nachgegangen. Neben einer ersten quantitativen Studie (Pilotstudie 1), die die Skalen bereits als qualitativ hochwertig kennzeichnet, werden Interviews unter Einsatz der Methode des Lauten Denkens geführt (Pilotstudie 2) sowie Experten zu den Operationalisierungen der Handlungsfelder und Dimensionen befragt (Piotstudie 3). Diese Pilotstudien 2 und 3 bilden den Kern des Validierungsprozesses und helfen die Skalen erheblich zu verbessern. Schließlich wird das eingesetzte Antwortformat mithilfe einer vierten Pilotstudie optimiert, um dann im Rahmen der Hauptstudie ($N = 931$) die psychometrischen Eigenschaften der Skalen zu überprüfen. Es zeigt sich, dass Erkenntnisse über die Gültigkeit der Annahmen zu gewinnen sind, die die intendierte Testwertinterpretation im Sinne eines Validitätsarguments stützen. Der Beitrag versteht sich als Diskussionsanlass bezüglich der Gestaltung von Validierungsmaßnahmen und deren Implikationen für den Forschungsprozess.

Schlüsselwörter Validitätsargument · Selbstwirksamkeitserwartungen · Instrumentenentwicklung · Lehrerprofessionalisierung

✉ Thorid Rabe
thorid.rabe@physik.uni-halle.de

[1] Institut für Physik, Arbeitsbereich Didaktik der Physik, Martin-Luther-Universität Halle-Wittenberg, Hoher Weg 8, 06120 Halle/Saale, Deutschland

Establishing an Evidence-Based Validity Argument for Assessing Self-Efficacy Beliefs for Teaching Physics with a Newly Developed Instrument

Abstract

This article describes the development of a validity argument for assessing (future and trainee) physics teachers' self-efficacy beliefs for teaching physics, specifically in four areas of action ("experimenting", "dealing with students' conceptions", "analysing and preparing physics content" as well as "dealing with tasks"). Each field of action is subdivided into the dimensions "planning" and "conducting". To generate possible evidence for the validity argument, we first define the construct under consideration as well as the intended test score interpretation and identify underlying assumptions of this test score interpretation explicitly. We implemented four pilot studies to check whether these assumptions apply. In a quantitative study (pilot study 1) the scales were already identified as promising and meeting the necessary statistical criteria. Interviews using the think aloud method were conducted with physics teacher students, physics trainee teachers as well as in service physics teachers (pilot study 2) and experts were asked to judge the appropriateness of our operationalisations of the fields of action and dimensions (pilot study 3). These qualitative studies are at the core of our validity argument and also helped considerably to improve the items and scales. In pilot study 4 we optimized the used response format, before conducting a major study ($n = 931$) to ensure our scales meet psychometric standards. We found support for all of our assumptions and therefore can present an evidence-based validity argument for our intended test score interpretations. The paper aims to generate discussion about validity concepts and standards and their implications for future research.

Keywords Validity argument · Self-efficacy beliefs · Development of an instrument · Teachers' professional development

Einleitung

Die empirische Wende in der naturwissenschaftsdidaktischen Forschung vollzog sich spätestens seit dem TIMSS- und PISA-Schock, der zu einer Intensivierung der Forschungsanstrengungen, nicht zuletzt dank erheblicher Investitionsmittel des Bundes, führte. Wegen der inhaltlichen Nähe zur Lehr-Lern-Psychologie lag eine Orientierung an deren Konstrukten und Forschungsmethoden nahe. In der Folge hat sich die quantitative Lehr-Lernforschung als eine Hauptströmung naturwissenschaftsdidaktischer Forschung etabliert, die ihre Gültigkeit aus der methodisch kontrollierten Anwendung von etablierten oder neu entwickelten Erhebungsinstrumenten ableitet, mit denen Daten generiert und dann in der Regel statistisch ausgewertet werden. Forschungsinstrumenten und deren Entwicklung, ihrer Qualität und reflektierten Nutzung in Forschungsprozessen kommt dabei eine zentrale, die Qualität der Forschungsergebnisse beeinflussende Bedeutung zu. In jüngerer Vergangenheit wird nach dem wenig erfolgreichen Versuch, die Befunde von 100 in qualitativ hochwertigen Journals veröffentlichten psychologischen Studien zu replizieren (Open Science Collaboration 2015), unter dem Stichwort „replication crisis" oder „replicability crisis" ein Problem psychologischer – das meint hier quantitativ empirischer – Forschung diskutiert. Vor diesem allgemeinen Hintergrund wird die Frage nach den Qualitätsanforderungen an die Entwicklung von Testinstrumenten zu einer hochaktuellen. Auch in der deutschsprachigen Naturwissenschaftsdidaktik stellt sich das Problem angesichts der dokumentierten Interpretationsprobleme von Daten, die mit neu entwickelten Instrumenten erhoben wurden, derzeit als bedeutsam dar (vgl. u. a. Cauet et al. 2015; Reinhold et al. 2017; Vogelsang und Cauet 2017).

Eine Diskussion über den notwendigen und als angemessen erachteten Aufwand bei der Entwicklung neuer oder bei der Adaptation bereits vorliegender Instrumente wird bisher vorranging forschungsgruppenintern geführt. Das Ziel des vorliegenden Beitrages ist es, diese Diskussion entsprechend der skizzierten Relevanz aus ihrem Nischendasein zu heben und in die Forschungsgemeinschaft zu tragen. Dazu wird nachfolgend eine Instrumentenentwicklung und -validierung vorgestellt, die – sofern sie als überzeugend und gewinnbringend für die naturwissenschaftsdidaktische Forschungsgemeinschaft eingestuft wird – als Referenz für folgende Validierungsanliegen dienen kann. In jedem Fall – und das ist weitaus wichtiger – kann sie aber als konkreter kritisierbarer Vorschlag fungieren. Systematisch wird der argumentbasierte Validierungsansatz (vgl. Kane 2001, 2013; AERA 2014) aufgegriffen, auf den auch in der Physikdidaktik gelegentlich rekurriert (vgl. z. B. Vogelsang 2014; Gramzow 2015), der jedoch nur selten so umfassend wie z. B. von Dickmann (2016) umgesetzt wird. Kanes Validitätsverständnis geht auf die wegweisenden Vorarbeiten Messicks (1995) zurück, die ebenfalls Eingang in die Fachdidaktik gefunden haben (vgl. u. a. Leuders 2014; Hadenfeldt und Neumann 2012). Aufbauend auf dem argumentbasierten Ansatz wird im Folgenden ein Validitätsargument auf Grundlage einer Vielzahl von Teilstudien konstruiert und zur Diskussion gestellt. Konkret wurde im Rahmen eines von der DFG geförderten Projekts ein Erhebungsinstrument entwickelt, das es ermöglichen soll, Selbstwirksam-

keitserwartungen (SWE) auf dem Spezifitätsniveau physikdidaktischer Handlungsfelder (Experimentieren, Elementarisieren, Umgang mit Schülervorstellungen, Umgang mit Aufgaben) in den Dimensionen der Planung und Durchführung von Physikunterricht bei Studierenden, Referendaren und Lehrkräften mit dem Unterrichtsfach Physik zu erfassen.

Die Sinnhaftigkeit der Neuentwicklung eines solchen Instrumentes sowie der damit verbundene Aufwand des nachfolgend im Fokus stehenden Validierungsprozesses sollen zunächst anhand einiger ausgewählter Argumente legitimiert werden (vgl. auch Rabe et al. 2012). Im Rahmen der sozial-kognitiven Theorie nach Bandura (1997) gelten SWE als entscheidend dafür, ob Handlungen aufgenommen werden und mit welcher Ausdauer bzw. mit welcher Anstrengung sie ggf. verfolgt werden. Damit haben sie zumindest indirekt auch Einfluss auf den Handlungserfolg (Bandura 1997). Auf der Ebene des Lehrerhandelns werden Einflüsse von Lehrer-SWE auf die Zielorientierung, die Qualität der Unterrichtsplanung und die Qualität des Lehrerhandelns im Unterricht postuliert sowie Zusammenhänge zu Jobzufriedenheit und Resilienz gegenüber Burnout hergestellt. Vermittelt über das Lehrerhandeln ist auch davon auszugehen, dass sich SWE auf Schülervariablen wie Motivation und Leistung auswirken (vgl. Woolfolk Hoy und Davis 2006).

Auch wenn heute als Konsens gilt, dass Beliefs (zu denen auch SWE zu rechnen sind) und nicht etwa nur vorhandenes Wissen einen entscheidenden Einfluss auf die Aufnahme konkreter Handlungen, z. B. die Art und Weise der Planung und Durchführung von (Physik-)Unterricht haben (Wallace 2014, S. 17), so sind dennoch viele der theoretisch abgeleiteten Vermutungen – nicht zuletzt aufgrund als defizitär einzuschätzender Instrumente – empirisch unzureichend belegt. Ein Instrument zur Erfassung von SWE in entsprechenden Domänen ist daher von großer Relevanz. Zwar existiert mit dem Science Teaching Belief Instrument, kurz STEBI (Enochs und Riggs 1990; Riggs und Enochs 1990), ein weit verbreitetes Instrument, welches in der überwiegenden Mehrzahl der Forschungsarbeiten Verwendung findet (vgl. Klassen et al. 2011) und auch als Standard propagiert wird (Shroyer et al. 2014). Das Instrument beruht jedoch in großen Teilen auf der Teacher Efficacy Scale, kurz TES (Gibson und Dembo 1984), die häufig aufgrund fehlender theoretischer Präzision z. B. hinsichtlich der Operationalisierung abgebildeter Konstrukte sowie messtheoretischer Probleme in Frage gestellt wird (Guskey 1988; Kushner 1993; Guskey und Passaro 1994; Coladarci und Fink 1995; Tschannen-Moran et al. 1998; Henson 2001; Brouwers und Tomic 2003; Denzine et al. 2005; Dellinger et al. 2008). Roberts und Henson (2000, S. 5) fassen zusammen: „TES has come under an increasing amount of fire." Details zu der Kritik an existierenden Erhebungsinstrumenten können hier nicht referiert werden, dazu sei auf Meinhardt (2018) verwiesen. In jedem Falle sprechen sich einige Forschende dafür aus, auf die Anpassung oder Übersetzung bereits existierender SWE-Instrumente für die Domäne des Unterrichtens naturwissenschaftlicher Fächer zu verzichten und stattdessen ein Instrument in Gänze neu zu entwickeln (vgl. u. a. Cakiroglu et al. 2012, S. 458). Dies bringen auch Pruski et al. (2013, S. 1151) zum Ausdruck: „Rather than attempting to revamp a scale, it might be better to go back to the ‚master' and begin again." Sie geben weiter zu bedenken: „By trying to improve or revamp each others' scales, we, the research community, may have created a type of ‚in-breeding' that clouds better thinking about efficacy item construction" (ebd.). Im angesprochenen Beispiel der Anpassung der TES für den naturwissenschaftlichen Kontext (STEBI) ist die Problematik der Validität abgeleiteter Interpretationen besonders virulent, da beide Instrumente bereits seit Jahrzehnten eingesetzt werden und ganze Forschungstraditionen auf diesen nun in Frage stehenden Instrumenten aufbauen.

Dass zu Fragen nach Umfang und Anspruch von Validierungsstudien eine Debatte zu führen ist, zeigt sich auch, wenn Projekte wie das hier vorzustellende mit dem Argument konfrontiert werden, dass ein „quick-and-dirty"-Verfahren für eine ausreichende Validierung genüge und zu zeigen sei, dass sich der mit zeitlichen und finanziellen Kosten verbundene Mehraufwand auszahle. Einer solchen Forderung liegt mindestens implizit die Annahme zugrunde, dass ein Abweichen von der seit langer Zeit etablierten (und sich erst in den letzten Jahren langsam wandelnden) Validierungspraxis begründungsbedürftig sei. Argumentativ unterlegt werden muss jedoch aus Sicht der Autoren vielmehr das Abweichen von wissenschaftlich ausgehandelten und kommunizierten Standards (vgl. AERA 2014) – auch und gerade, wenn diese die gängige Praxis in Frage stellen. In der darzustellenden exemplarischen Validierungsstudie soll jedenfalls gezeigt werden, dass jeder einzelne Validierungsschritt im Rahmen dieser Studie zu einer Qualitätssteigerung des Erhebungsinstruments beigetragen hat. (Anmerkung: Ob für einen spezifischen Einsatzzweck die vorliegenden Skalen „besser" (in welchem Sinne auch immer) als auf anderem Wege entwickelte Skalen sind, wäre eine interessante Anschlussfrage, die dieser Artikel ausdrücklich nicht adressiert. Zum einen, weil kein theoretisch fundiertes Instrument auf vergleichbarem Spezifitätsniveau existiert; zum anderen, weil die vorliegende Studie sich auf den deskriptiven Teil eines Validitätsargumentes beschränkt, wie weiter unten ausgeführt wird).

Dazu werden zunächst die zugrundeliegende Auffassung von Validität erläutert sowie die Basis des Validierungsunterfangens, d. h. Aspekte der theoretischen Fundierung in aller Kürze vorgestellt. In diesem Zusammenhang wird das zentrale Konstrukt der (physikdidaktischen) SWE auch in Abgrenzung zu benachbarten Konstrukten definiert; kon-

struktkonstituierende Merkmale werden abgeleitet. Darauf aufbauend wird die intendierte Testwertinterpretation dargelegt und dann schrittweise auf Basis des (bisherigen) Validierungsprozesses mit Evidenzen für ihre Belastbarkeit verknüpft. Die einzelnen durchgeführten Studien werden dazu so ausführlich vorgestellt und die Ergebnisse so umfänglich referiert, wie es nötig erscheint, um die abschließende Ableitung eines Validitätsarguments nachvollziehbar und den in Kauf genommenen Aufwand beurteilbar zu machen. Für jede dargestellte Studie werden einzelne, der intendierten Testwertinterpretation zugrunde liegende Annahmen auf ihre Gültigkeit geprüft. Dabei werden die Ausführungen exemplarisch sowohl bei der Darstellung der qualitativen als auch der quantitativen Teilstudie auf die beiden neu konstruierten Skalen zum Handlungsfeld Experimentieren beschränkt. Abschließend wird das Vorgehen noch einmal diskutiert; insbesondere werden die Grenzen der vorliegenden Validierung einer Testwertinterpretation aufgezeigt.

Validität als Argument

Gegenüber früheren Auffassungen von Validität als Eigenschaft eines Testes, die einmal nachgewiesen, nahezu uneingeschränkte Gültigkeit besaß (Moosbrugger und Kelava 2008), hat sich gegenwärtig die Konzeption von Validierung als fortdauernden Prozess, der sich auf die Interpretation und den Umgang mit Testwerten bezieht, weitgehend durchgesetzt: „Validity is not a property of the test. Rather, it is a property of the proposed interpretations and uses of the test scores. Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (...)“ (Kane 2013, S. 3). Validiert wird demnach nicht ein Instrument an sich, vielmehr werden die Annahmen, die bei der Interpretation und beim Gebrauch eines Erhebungsinstruments getroffen werden, auf ihre Belastbarkeit geprüft: „Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of the tests“ (AERA 2014, S. 11). Auch die *Standards for Educational and Psychological Testing* der AERA legen demnach eine argumentbasierte Auffassung von Validität nahe. Mögliche Testwertinterpretationen und der vorgesehene Umgang mit den Testwerten müssen möglichst klar und genau beschrieben werden, damit dann Hinweise bzw. Evidenzen gesammelt und zusammengefügt werden können, die die Angemessenheit des intendierten Umgangs mit den Testwerten stützen oder in Frage stellen (vgl. Kane 2001). Herangezogen werden dazu theoretische Argumente und Plausibilitäten ebenso wie empirische Belege. Erst zusammen mit der Abwägung von Einschränkungen und Grenzen kann ein übergreifendes – aber vorläufig bleibendes – Validitätsargument formuliert werden. An Qualität gewinnt dieses Argument, wenn auch konkurrierende Testwertinterpretationen (beispielsweise, dass nicht SWE, sondern z. B. Handlungsergebniserwartungen erfasst werden) ernsthaft auf ihre mögliche Gültigkeit hin untersucht werden. Bedeutsam ist, dass sich ein Standardverfahren für Validierungsprozesse aus dem argumentativen Ansatz *nicht* ableiten lässt. Welche Schritte sinnvoll und notwendig erscheinen, ergibt sich aus der jeweiligen spezifischen intendierten Testwertinterpretation. Konsequenzen der Testnutzung sollten schon im Validierungsprozess mitgedacht werden, sofern schon bestimmte Einsatzzwecke festgelegt oder intendiert sind, wobei hier auch den (zukünftigen) Testnutzern Verantwortung bei der Aufstellung des Validitätsarguments zugeschrieben wird: „Validation is the joint responsibility of the test developer and the test user“ (AERA 2014, S. 13). Bei jeder Anwendung eines Instruments ist neu zu prüfen, ob das bisherige Validitätsargument ausreicht und seine Gültigkeit behält.

Kane (2001, S. 337) legt das Fundament für diese Sichtweise, indem er sowohl einen deskriptiven als auch einen präskriptiven Teil des Validitätsarguments postuliert. Während ersterer im Sinne einer „Interpretationskomponente“ auf die intendierte Testwertinterpretation rekurriert, bezieht sich der zweite Teil im Sinne einer „Gebrauchskomponente“ auf die Entscheidungen, die auf Grundlage der Testwerte und ihrer Interpretation getroffen werden. Wollte man z. B. Selbstwirksamkeitserwartungen zu Beginn des Studiums als Prädiktor für Studienerfolg nutzen, so wäre sicherzustellen, dass die erhobenen Messwerte tatsächlich zuverlässig Auskunft über Selbstwirksamkeitserwartungen der betreffenden Person geben (deskriptiver Teil). Die Validität der ggf. aus den Testwerten abzuleitenden Schlüsse, z. B. die Zulassung oder Ablehnung für einen Studiengang sind zu analytischen Zwecken davon zu trennen (präskriptiver Teil).

Den notwendigen Ausgangspunkt für die Aufstellung eines Validitätsarguments bildet „ein theoretisches Konstrukt oder eine operationale Merkmalsdefinition, auf denen der Test beruht. Oder vereinfacht die Frage: Was und wozu soll überhaupt gemessen werden?“ (Schmiemann und Lücken 2014, S. 108). Insofern ist auch für den nachfolgend dargestellten Validierungsprozess eine theoretisch verankerte Arbeitsdefinition auf Basis des aktuellen Forschungsstands ein wichtiger erster Arbeitsschritt, aus dem abgeleitet wird, wie das Instrument in seinen Grundzügen angelegt sein und welche Testwertinterpretation der Validierung zugrunde liegen sollen. Die qualitativen und quantitativen Studien dienen dann vor allem dazu, den deskriptiven Teil eines Validitätsarguments vorzubereiten, tragen aber auch zu weiteren definitorischen Ausschärfungen bei. Eine wie auch immer geartete spezifische Testnutzung sowie die damit einhergehende Präzisierung der Testwertinterpretation, also der präskriptive Anteil des Validitätsarguments wird im bisherigen Validierungsprozess aus pragmatischen Gründen wie

fehlender Ressourcen weitgehend ausgeklammert. Damit liegt der Schwerpunkt der darzustellenden Validierungsbemühungen auf der Absicherung notwendiger Voraussetzungen für eine ganze Klasse möglicher Testwertinterpretationen. Die Überprüfung der Validität der aus den Testwerten zu ziehenden Schlüsse, die mit einer konkreten Testanwendung einhergeht (Stichwort Prädiktivität), ist Teil des ohnehin immer unabgeschlossenen, fortlaufenden Validierungsprozesses, für den eben auch die Anwender des Testes Verantwortung tragen.

Als grundsätzliche Problematik im Zusammenhang mit Validierungsprozessen – gerade wenn es sich um neu entwickelte Instrumente handelt – erweist sich, dass eher solche Evidenzen wahrgenommen bzw. evoziert werden, die für die Stützung der eigenen Testwertinterpretation sprechen („confirmationist bias“, vgl. Kane 2001). Neben der Prüfung von konkurrierenden Testwertinterpretationen wurde in dem hier vorgestellten Projekt eine Qualitätssicherung dadurch angestrebt, dass alle Validierungsschritte, -ergebnisse und -implikationen intensiv und fortlaufend im dreiköpfigen Projektteam (die AutorInnen des Artikels als Lehrende und Forschende in der Physikdidaktik) diskutiert wurden. Die folgenden Darstellungen sollen es ermöglichen, das bis dato formulierte Validitätsargument kritisch zu hinterfragen und weitere notwendige Validierungsanstrengungen zu identifizieren. Wir erhoffen uns, durch dieses konkrete Beispiel einen Beitrag zur allgemeinen Diskussion um Validierungsbemühungen in der naturwissenschaftsdidaktischen Forschungsgemeinschaft zu leisten.

Physikdidaktische Selbstwirksamkeitserwartungen

Da das Konstrukt der (physikdidaktischen) SWE und seine Relevanz im Kontext Lehrerprofessionalisierung bereits an anderer Stelle ausführlich dargestellt wurde (vgl. Rabe et al. 2012), wird hier nur zusammenfassend und als Ausgangspunkt für die Ableitung sinnvoller Validierungsmaßnahmen die im Rahmen des Projektes erarbeitete Konstruktdefinition vorgestellt. Diese wurde im Laufe des Projektes entwickelt und ausgeschärft, diente aber in der jeweiligen Version im Validierungsprozess als wichtiger Referenzpunkt. Einerseits, weil eine scharfe Beschreibung des Konstruktes zur Prüfung der Güte der Operationalisierungen herangezogen werden kann, andererseits, weil so ein Orientierungspunkt vorliegt, auf den Erkenntnisse aus den Validierungsschritten rückbezogen werden können, beispielsweise um einzelne Konstruktfacetten im Zuge der vertieften Theoriearbeit zu konkretisieren. Insofern dient eine Arbeitsdefinition dazu, Theoriearbeit und konkrete Operationalisierungsbemühungen immer wieder miteinander abzugleichen.

Das Konstrukt der Selbstwirksamkeitserwartungen geht auf die sozial-kognitive Theorie Banduras zurück, der die folgende Definition gibt: „Perceived self-efficacy refers to beliefs in one's capabilities to organize and execute the courses of action required to produce given attainments“ (Bandura 1997, S. 3). Hier wird also das Vertrauen in die eigene Handlung in den Mittelpunkt gerückt. Dies unterscheidet Selbstwirksamkeitserwartungen einerseits von Handlungsergebniserwartungen, bei denen die Konsequenzen/Folgen eigener Handlungen antizipiert werden, und andererseits von Kontrollüberzeugungen, bei denen die prinzipielle Erreichbarkeit eines Ziels durch die (wie auch immer) handelnde Person von zentraler Bedeutung ist (vgl. Skinner et al. 1988).

Die folgenden zentralen Eigenschaften des Konstrukts sind offensichtlich mit der oben angegebenen Definition kompatibel und bilden den Kern der Beschreibung des Konstrukts nach Bandura (1997) ab. Sie sind implizit auch in Banduras „Guide for Constructing Self-Efficacy Scales“ (Bandura 2006) enthalten:

- *Selbstreferentialität/Subjektivität*: Es handelt sich um eine *Selbst*einschätzung *eigener* Fähigkeiten.
- *Handlungsbezug*: Die Selbsteinschätzung bezieht sich auf (konkrete) Handlungen.
- *Schwierigkeitsbezug*: Die Selbsteinschätzung ist nur dann sinnvoll, wenn keine „trivialen“ Handlungen im Vordergrund stehen, sodass „[p]erceived efficacy should be measured against levels of task demands that represent graditions of challenges or impediments to successful performance“ (Bandura 2006, S. 311).
- *Situationsspezifität*: Die Selbsteinschätzung ist abhängig von der jeweiligen Situation/von dem Kontext, auf den Bezug genommen wird.
- *Domänenspezifität*: Die Selbsteinschätzung ist abhängig von der Domäne.

Aus den letzten zwei Aspekten ergibt sich, dass SWE auf sehr unterschiedlichen Spezifitätsniveaus konzeptualisiert werden können, angefangen bei allgemeinen SWE über beispielsweise Lehrer-SWE bis hin zu SWE, die sich auf eine spezifische Domäne und eine spezifische Handlungsdimension beziehen. Auf letzterer Spezifitätsebene befinden sich die SWE in physikdidaktischen Handlungsfeldern (Experimentieren, Umgang mit Schülervorstellungen, Elementarisieren, Umgang mit Aufgaben) in den Dimensionen Planung und Durchführung von Physikunterricht, die im Fokus des Projekts stehen. Damit wird bewusst auf eine Spezifizierung der Inhaltsgebiete (z. B. Mechanik oder Optik) verzichtet. Diese Entscheidung ist letztlich willkürlich, da Fragen nach dem optimalen Spezifitätsniveau immer noch Gegenstand der Forschung sind (vgl. u. a. Woolfolk Hoy et al. 2009, S. 631). Dem Projektteam lag das etwas allgemeinere Spezifitätsniveau näher, weil die Aussicht viel-

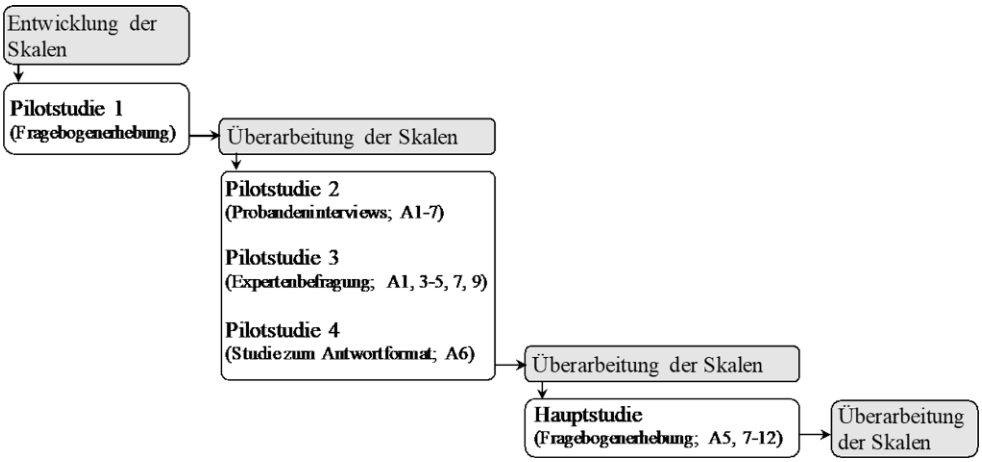

Abb. 1 Teilstudien und Beiträge zur Prüfung von Annahmen, die der intendierten Testwertinterpretation zugrunde liegen

versprechend schien, das Instrument gemeinsam mit (fachlichen, pädagogischen oder fachdidaktischen) Wissenstests einsetzen zu können, die in der Regel auch mehr als einen Themenbereich umfassen. Die Tests wären dann auf einem ähnlichen Spezifitätsniveau angesiedelt, was deren Prädiktivität vermutlich erhöhen würde. Man hätte aber ebenso, vielleicht sogar zuverlässiger, eine Mechanik- oder Optikversion der Skalen zur Erfassung physikdidaktischer SWE entwickeln können. Die Frage nach dem jeweils angemessenen Spezifitätsniveau ist erst dann sinnvoll zu beantworten, wenn ein konkretes Erkenntnisinteresse formuliert ist. Als Bezugspunkt für die Entwicklung eines Instruments zur Erfassung der physikdidaktischen SWE und für die Ableitung von Testwertinterpretationen dient die folgende **Arbeitsdefinition**:

Lehrer-Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern beziehen sich auf die subjektive Selbsteinschätzung, komplexe Handlungen der Planung und Durchführung von Physikunterricht ausführen zu können, auch wenn mit Widerständen umgegangen werden muss.

Validierungsschritte zur Entwicklung eines Validierungsarguments zu physikdidaktischen SWE

Der im folgenden beschriebene Validierungsprozess wird um der Übersichtlichkeit willen als linearer Ablauf dargestellt. Realiter fanden einige Schritte zeitlich parallel statt bzw. wurden die Teilstudien wiederholt zur Ausschärfung bzw. Anpassung von Arbeitsdefinition, Konstruktionsvorschriften und Annahmen zur Testwertinterpretation herangezogen. Der Ablauf ist vereinfacht in Abb. 1 veranschaulicht.

Der erste Schritt im Validierungsprozess bestand, wie bereits erwähnt, in einer theoretischen Klärung des avisierten Konstrukts physikdidaktischer SWE, der Formulierung der Arbeitsdefinition und schließlich in der Ableitung von Konstruktionsvorschriften. Diese sind in Tab. 1 vorausgreifend in zwei Versionen dargestellt, da die vor der ersten Pilotierung aufgestellten Regeln aufgrund der weiteren Pilotierungen überarbeitet wurden.

Entlang dieser Regeln wurden insgesamt acht Skalen für die vier Handlungsfelder Experimentieren, Umgang mit Schülervorstellungen, Elementarisieren, Umgang mit Aufgaben und jeweils in den zwei Dimensionen Planung und Durchführung von Physikunterricht entworfen. Der Validierungsprozess wird im Folgenden exemplarisch anhand der Skalen für das Handlungsfeld Experimentieren dargestellt.

Dem Prozess liegt folgende **intendierte Testwertinterpretation** zugrunde:

> Die Skalenwerte, die auf den neu entwickelten Skalen vergeben werden, entsprechen den Lehrer-SWE in dem jeweiligen physikdidaktischen Handlungsfeld (hier: Experimentieren) und in der jeweiligen Dimension (Planung versus Durchführung) der befragten Personen. Diese können den Gruppen der Physik-Lehramtsstudierenden, der Physik-Lehramtsanwärter/innen und der Physik-Lehrpersonen angehören.

Die genannte Testwertinterpretation umfasst damit entsprechend den Begrifflichkeiten nach Kane (2001, S. 337) lediglich beschreibende Elemente („descriptive part"), ist also auf keinen spezifischen Nutzen zugeschnitten. Weitergehende Schlüsse (z. B. bzgl. der Eignung oder Erfolgsaussicht befragter Person) auf Grundlage der erhobenen Daten sind damit derzeit nicht legitimiert. Die Testwertinterpretation beruht u. a. auf den folgenden Annahmen, die im Rahmen des Validitätsarguments auf Plausibilität bzw. auf empirische Haltbarkeit geprüft werden müssen. Auf Itemebene ist sicherzustellen, dass …

Tab. 1 Ableitung der Itemkonstruktionsregeln in ursprünglicher und in überarbeiteter Fassung

Konstruktaspekt	Konstruktionsregel vor Pilot 1	Konstruktionsregel nach Pilot 2 und 3
Subjektivität/ Selbstreferentialität	Formulierung in der 1. Person Singular	Formulierung in der 1. Person Singular, jedes Item beginnt mit „Ich kann …“
Handlungsorientierung	Verwendung von Phrasen wie „Ich kann …“, „Ich bin in der Lage“ gefolgt von komplexen Handlungen	Verwendung ausschließlich der Phrase „Ich kann …“, gefolgt von komplexen Handlungen
Domänenspezifität	Handlungen müssen spezifisch für ein Handlungsfeld und eine der Dimensionen Planung vs. Durchführung sein	Handlungen müssen spezifisch für ein Handlungsfeld und eine der Dimensionen Planung vs. Durchführung sein
Schwierigkeitsbezug	Formulierung passender Handlungsbarrieren mit Phrasen wie „auch wenn“, „obwohl“ Zuspitzung der Schwierigkeit durch Formulierungen wie „immer“ oder „in jedem Fall“	Formulierung passender Handlungsbarrieren ausschließlich mit „auch wenn“, keine Zuspitzungen

A1 … die entwickelten Items Selbstwirksamkeitserwartungen im Sinne der zugrundeliegenden Definition formulieren.
A2 … die Items bei den befragten Personen kognitive Prozesse evozieren, die sich auf Selbstwirksamkeitserwartungen beziehen.
A3 … die Items von den befragten Personen inhaltlich in der intendierten Weise verstanden werden.
A4 … Probanden die Items als authentisch (hinsichtlich der Handlungsfelder und Dimensionen) wahrnehmen.
A5 … befragte Personen aus den drei vorgesehenen Befragungsgruppen (Studierende, Anwärter, Lehrpersonen) die Items ähnlich verstehen.
A6 … die befragten Personen die Möglichkeit haben, sich hinsichtlich der in den Items beschriebenen SWE differenziert einzuschätzen.

Auf Skalenebene muss verlangt werden, dass …

A7 … die Konstrukte (das heißt physikdidaktische SWE in einem Handlungsfeld und in einer Dimension) treffend und repräsentativ abgebildet werden.
A8 … das jeweilige Konstrukt mittels einer Skala zuverlässig abgebildet wird.
A9 … die mittels der Skalen dargestellten Konstrukte ausreichend gegeneinander abgrenzbar sind.
A10 … die mittels der Skalen dargestellten Konstrukte auch gegen „benachbarte“ Konstrukte abgrenzbar sind.
A11 … die mittels der Skalen dargestellten Konstrukte von den Personen aus allen vorgesehenen Befragungsgruppen ähnlich wahrgenommen werden.

Vorausgesetzt, dass sich die genannten elf Annahmen der Testwertinterpretation als belastbar erweisen, kann auch die Annahme geprüft werden, dass …

A12 … Mittelwertunterschiede auf den Skalen zwischen den verschiedenen Befragungsgruppen theoriekonform plausibel gemacht werden können.

Auf Grundlage der dargestellten Annahmen können gezielte Prüfmaßnahmen abgeleitet werden, die dann in einzelnen Studien jeweils im Fokus stehen (vgl. Abb. 1). Beispielsweise könnten Experten bzw. Probanden nach der Authentizität der konstruierten Items befragt und damit Annahme 4 adressiert werden. Erkennbar ist, dass eine Studie mehrere Annahmen parallel aufgreift, eine Annahme aber auch mehreren Prüfverfahren unterworfen werden kann. So steht Annahme 1 nicht nur in den Pilotstudien 2 und 3 auf dem Prüfstand, sondern muss auch einer theoretischen Plausibilitätsprüfung genügen. Um eine theoriegeleitete und letztlich wissenschaftliche Überprüfung der einzelnen Annahmen zu gewährleisten, sollten stichhaltige und literaturgestützte Hypothesen erarbeitet werden. So sind auf Grundlage des bisherigen Forschungsstands z. B. mit steigender Berufserfahrung auch steigende SWE zu erwarten, die durch Mittelwertvergleiche (A12) detektiert werden könnten. Dezidiertere Hypothesen sind jedoch aufgrund eines begrenzten und teils widersprüchlichen Forschungsstands bzw. aufgrund fehlender domänenspezifischer Studien nicht ableitbar. In Bezug auf Annahme 10 folgt z. B. aus Überlegungen hinsichtlich der Hierarchie des Konstrukts, dass globalere Messungen der Selbstwirksamkeitserwartungen (z. B. Lehrer-SWE) eher gering mit spezifischen, domänenbezogenen SWE korrelieren sollten (Bandura 1997, S. 42).

Über die oben genannten Annahmen hinaus ist es auch sinnvoll, möglichen konkurrierenden Testwertinterpretationen nachzugehen, auch um dem bereits erwähnten „confirmationist bias“ entgegenzuwirken. Aufgrund der theoretischen Nähe sollen die konkurrierenden Interpretationen geprüft werden, dass es sich bei dem mit dem entwickelten Instrument erhobenen Konstrukt um das Selbstkonzept, Handlungsergebniserwartungen oder Kontrollüberzeugungen bezüglich physikdidaktischer Handlungsfelder handelt.

Das bisherige Vorgehen zeichnet sich durch Transparenz bezüglich Ausgangspunkt (Theoriefundament und Testwertinterpretation) und Zweck (Prüfung relevanter Annahmen) der angedachten Studien aus. Durch den Vierschritt

Tab. 2 Skalenversion vor und nach dem Pilot 1; dargestellt ist die ursprüngliche Version der Items plus die aufgrund der Pilotstudie 1 vorgenommenen Veränderungen an einzelnen Items (kursiv); in dieser veränderten Fassung gehen die Skalen in die Pilotstudien 2 und 3 ein; Antwortformat: vierstufig mit „stimmt nicht", „stimmt kaum", „stimmt eher", „stimmt genau"

Items der Skala Experimentieren/Planung	
exp1	Es bereitet mir keine Probleme, zu meinem Unterricht passende Experimente vorzubereiten, auch wenn die Physiksammlung schlecht ausgestattet ist.
exp2	Ich bin stets in der Lage, in der Unterrichtsvorbereitung lernförderliche Experimente auszuwählen, auch wenn ich das Themengebiet, das erste Mal unterrichte. *verändert zu: Ich bin stets in der Lage, für eine Unterrichtssequenz ein motivierendes Einstiegsexperiment zu planen, auch wenn ich die Sequenz das erste Mal unterrichte.*
exp3	Ich kann jedes Experiment auch für andere Phasen des Unterrichts einplanen, obwohl ich es bisher nur zur Erarbeitung verwendet habe.
exp4	Auch für eine sehr heterogene Klasse kann ich Schülerexperimente so planen, dass ich sie zur Differenzierung nutzen kann.
exp5	Auch Zeitdruck während der Unterrichtsvorbereitung hindert mich nicht daran, ein zu meinen Unterrichtszielen passendes Experiment zu entwickeln.
exp6	Es gelingt mir immer, zu einem Thema Schülerexperimente leistungsdifferenziert vorzubereiten, auch wenn es in der Literatur dazu keine Vorschläge gibt.
exp7	Ich kann jedes Experiment als Demonstrationsexperiment planen, auch wenn ich das betreffende Experiment bisher ausschließlich als Schülerexperiment eingesetzt habe. *verändert zu: Ich kann mich für jedes Experiment begründet entscheiden, ob ich es als Schüler- oder Demonstrationsexperiment plane, auch wenn ich noch keine Erfahrung mit dem Experiment habe.*
Items der Skala Experimentieren/Durchführung	
exd1	Immer, wenn der Unterrichtsverlauf es sinnvoll erscheinen lässt, gelingt es mir, ein Experiment angemessen zu variieren, auch wenn ich das vorher nicht geplant habe.
exd2	Ich kann ein Demonstrationsexperiment für Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine sehr komplexe Versuchsanordnung handelt.
exd3	Auch unvorbereitet schaffe ich es stets, anhand eines Experiments über naturwissenschaftliches Arbeiten mit den Schülerinnen und Schülern zu reflektieren, auch wenn ich mich nicht explizit darauf vorbereitet habe.
exd4	Während Schülerexperimentierphasen kann ich die Schülerinnen und Schüler individuell zielführend unterstützen, auch wenn ich nicht alle Schwierigkeiten vorhergesehen habe. *verändert zu: Auch auf unvorhergesehene Schwierigkeiten von Schülerinnen und Schülern beim Experimentieren kann ich so reagieren, dass sie selbständig weiterarbeiten können.*
exd5	Ich bin in jedem Fall in der Lage, auch längere Demonstrationsexperimente so zu gestalten, dass die Schülerinnen und Schüler bei der Sache bleiben. *verändert zu: Ich bin in jedem Fall in der Lage, ein Demonstrationsexperiment so durchzuführen, dass die Schülerinnen und Schüler bei der Sache bleiben, auch wenn die Durchführung viel Zeit in Anspruch nimmt.*
exd6	Ich weiß, dass ich beim Experimentieren in einer Vertretungsstunde auf Unvorhergesehenes reagieren kann, auch wenn ich das Experiment längere Zeit nicht eingesetzt habe.
exd7	Auch unter Zeitdruck bin ich fast immer in der Lage, ein Experiment zum Laufen zu bringen, wenn es im Unterricht nicht auf Anhieb funktioniert.

aus Testwertinterpretation, zugrundeliegenden Annahmen, abgeleiteten Maßnahmen und literaturbasierten Hypothesen wird der Validierungsprozess in viele Teilschritte zerlegt und dadurch nachvollziehbar bzw. hinsichtlich der Sinnhaftigkeit der jeweiligen Entschlüsse und Aktionen beurteilbar. Ergebnisse der einzelnen Studien können direkt in die Argumentation für oder gegen die Gültigkeit der intendierten Testwertinterpretation eingeflochten werden. Grenzen der Testwertinterpretation werden durch eine Charakterisierung selbiger offengelegt, sodass für potentielle Anwender des Instrumentes expliziert ist, welche Art von Aussagen aus den Daten abgeleitet werden können und welche eher unterbleiben sollten oder zusätzliche Validierungsanstrengungen einfordern.

Nachfolgend werden die einzelnen Studien sowie die erzielten Ergebnisse mit Bezug zu den dargelegten Annahmen berichtet und Überarbeitungsschritte transparent gemacht, sodass im Anschluss daran das vorläufige Validitätsargument nachvollzogen werden kann.

Pilotstudie 1: Fragebogenerhebung bei Lehramtsstudierenden

Details zur ersten Pilotstudie – einer ersten quantitativ angelegten Fragebogenstudie mit Physiklehramtsstudierenden – sind bereits an anderer Stelle ausführlich dargestellt worden (Rabe et al. 2012). Die Studie wird deshalb hier nur angerissen. Sie diente primär dazu, die prinzipielle Realisierbarkeit eines Instruments zur Erhebung von physikdidaktischen SWE zu erproben. Die Skalen zum Experimentieren, die in diese erste Pilotstudie eingingen und somit

den Ausgangpunkt des gesamten Validierungsprojekts darstellen, sind in Tab. 2 aufgelistet. Es sollte erkennbar sein, dass die Items den ursprünglichen und nicht den überarbeiteten Konstruktionsregeln in Tab. 1 genügen.

Durch Kursivsetzung ist in Tab. 2 kenntlich gemacht, welche Items aufgrund dieser ersten Studie bereits überarbeitet wurden. Anlass zur Veränderung der Items waren Ergebnisse konfirmatorischer Faktorenanalysen gepaart mit inhaltlichen Erwägungen. Ein direkter Bezug zu den Annahmen der Testwertinterpretation wird nicht hergestellt, da diese ebenfalls erst im Anschluss an Pilotstudie 1 ausgeschärft wurde.

Pilotstudie 2: Interviews mit Probanden der Befragungsgruppen

Eine zweite Pilotstudie liefert Hinweise zu den Annahmen 1 bis 7 der Testwertinterpretation. Konzipiert wurde dieser Validierungsschritt als Interviewstudie mit Vertretern der drei vorgesehenen Befragungsgruppen. Insgesamt 20 Personen, davon je vier Physiklehramtsstudierende mit gar keiner, wenig (maximal Schulpraktische Übungen absolviert, keine Nachhilfeerfahrung) oder viel (mindestens Praxissemester absolviert) Praxiserfahrung, Physikreferendare und Physiklehrkräfte konnten für die Studie gewonnen werden. Mit jedem Interviewpartner wurde ein Interview zu vier Skalen aus zwei Handlungsfeldern (beispielweise zum Experimentieren in den Dimensionen Planung und Durchführung in der überarbeiteten Version aus Tab. 2 und zu einem weiteren Handlungsfeld) geführt, das Elemente des lauten Denkens (vgl. Sandmann 2014) umfasste, womit eine möglichst große Nähe zu den kognitiven Prozessen der Probanden ermöglicht werden sollte. Der Interviewleitfaden sah vor, dass die Interviewpartner jedes Item zunächst laut vorlesen, es im Anschluss mit eigenen Worten wiedergeben bzw. paraphrasieren, um sich dann hinsichtlich des Items selbst einzuschätzen und für die Einschätzung eine Begründung zu äußern. Im Anschluss an einen so gearteten Durchgang durch die Items einer Skala wurden durch den Interviewer weitere Fragen u. a. zu besonders schwierigen Items, zur wahrgenommenen Authentizität und Relevanz der Items und zur Eignung des Antwortformates gestellt. Die Interviews dauerten im Mittel 60 min und wurden nur in Teilen transkribiert. Anhand von Audiomitschnitten wurde im kriterial angelegten Auswertungsprozess zu jedem Item das Verständnis des Items durch den Interviewpartner, die Passung der kognitiven Prozesse zum intendierten Konstrukt SWE, die durch die Interviewpartner wahrgenommene Passung des Antwortformats und die Einschätzung der Authentizität in tabellarischer Form dokumentiert.

Schlussfolgerungen aus der Pilotstudie 2 für die Revision der Items und Skalen werden zusammen mit den Hinweisen aus Pilotstudie 3 dargestellt.

Pilotstudie 3: Expertenbefragung

Die Annahmen 1, 3, 5, 7 und 9 werden unter anderem in einer dritten Pilotierung hinsichtlich ihrer Gültigkeit geprüft. Es handelte sich um eine Befragung von Experten der Physikdidaktik aus dem universitären Bereich, die schriftlich über Fragebögen befragt wurden.

Von insgesamt 25 in einer persönlichen Email angefragten physikdidaktischen Expert/innen (in der Regel Professorinnen und Professoren sowie einige Postdocs) aus dem Bereich der Physikdidaktik haben 17 an der Befragung teilgenommen.

Mittels schriftlicher Fragebögen wurden jedem der Physikdidaktik-Experten die Items von vier ausgewählten Skalen aus zwei Handlungsfeldern zur Beurteilung vorgelegt. Zu den folgenden Aspekten wurden Rückmeldungen erbeten:

- Bewertung der Passung der Items zu dem jeweiligen Handlungsfeld,
- Bewertung der Relevanz der Items innerhalb des Handlungsfeldes,
- Einschätzung der Items hinsichtlich ihrer Schwierigkeit,
- Einschätzung der Authentizität der Handlungsbarrieren/Schwierigkeiten.

Außerdem konnten die Experten die Auswahl der Handlungsfelder kommentieren und weitere Handlungsfelder angeben, für die aus ihrer Sicht eine Skalenentwicklung sinnvoll wäre. Abschließend wurden die Experten nach der Sinnhaftigkeit der Trennung der Dimensionen Planung und Durchführung gefragt.

Im Ergebnis liegen zu jedem Handlungsfeld im Mittel acht Experteneinschätzungen vor. Die Auswertung fand wiederum für jedes einzelne Item statt, so dass die tabellarische Zusammenfassung aus Pilotstudie 2 um die Experteneinschätzungen ergänzt wurde.

Im Ergebnis lag für jedes einzelne Item eine umfangreiche Tabelle mit den Ergebnissen aus Probandeninterviews und Expertenbefragung vor, die als Grundlage für eine im Team erarbeitete Itemrevision genutzt wurde.

Auswertung der Pilotstudien 2 und 3 hinsichtlich der Testwertinterpretation

Zwar können Physikdidaktiker nicht als Experten für das psychologische Konstrukt der SWE im engeren Sinne betrachtet werden, jedoch wurde ihnen die den Items zugrun-

deliegende Arbeitsdefinition zur Verfügung gestellt, so dass zumindest das Passungsverhältnis beurteilt werden konnte. Kritik an diesem Passungsverhältnis wurde nicht geäußert, was als Indiz für die Gültigkeit von Annahme 1 gesehen werden kann.

Von den Interviewpartnern werden die meisten Items gut verstanden, es lassen sich aber auch Verstehenshürden wie Fachbegriffe („Schülervorstellungen“, „Lernaufgaben“), Verneinungen oder doppelte Hürden/Zuspitzungen identifizieren (Bezug zu A3). Aus den Interviews ergeben sich keine Hinweise auf Unterschiede in der Verständlichkeit bzw. in der inhaltlichen Auslegung der Items für die Befragungsgruppen der Studierenden, Referendare und Lehrkräfte.

Die Itemschwierigkeiten werden in den drei Befragungsgruppen ähnlich eingeschätzt (Bezug zu A5). Die Schwierigkeit der Items scheint hinreichend hoch zu sein, da ihnen nur in wenigen Fällen vollständig zugestimmt wurde. Die Expertenmeinungen hinsichtlich der Itemschwierigkeiten gehen auseinander, was unter anderem darin begründet sein kann, dass keine Befragungsgruppe als Referenz angegeben wurde. Allerdings können dennoch vorläufige Hypothesen über besonders leichte und besonders schwierige Items abgeleitet werden, die in nachfolgender Hauptstudie als Bezugspunkt dienen.

Hinsichtlich der intendierten kognitiven Prozesse (Bezug zu A2) zeigt sich, dass zu dem Spezifitätsniveau passende Situationen assoziiert werden und die Probanden selbstreferentiell auf die Items reagieren. Sie beziehen sich auf (anspruchsvolle) Handlungen, die zu den Handlungsfeldern passen, allerdings wird nicht immer die Schwierigkeit der Items bzw. die Hürde wahrgenommen. Von Expertenseite wurde der Hinweis gegeben, dass eine kognitive Entlastung beim Leseprozess möglich wäre, indem einheitliche Satzkonstruktionen für die Items verwendet werden.

Eine Authentizität der Items (Bezug zu A4) wird in solchen Fällen nicht konstatiert, in denen die Handlungsbarrieren als unpassend empfunden werden. In der Regel werden die beschriebenen Handlungen aber als relevant und zutreffend akzeptiert.

Hinsichtlich des vierstufigen Antwortformats ergibt sich kein einheitliches Bild (Bezug zu A6), da sich einige Interviewpartner eine stärkere Differenzierung wünschen, andere aber ein vierstufiges Antwortformat als intuitiv wahrnehmen. Dieser Befund gab Anlass für eine weitere Pilotstudie (s. unten Pilotstudie 4).

In der Regel werden die Items von den Experten als repräsentativ für die Handlungsfelder eingeschätzt (Bezug zu A7). Aus den Hinweisen auf weitere relevante Handlungsfelder konnte einerseits eine weitere Bestätigung der bisherigen Auswahl gewonnen werden, da mehrfach Experimentieren, Schülervorstellungen, Aufgaben und Elementarisieren von Experten genannt wurden, die das betreffende Handlungsfeld nicht vorliegen hatten. Als weitere Handlungsfelder werden beispielsweise Modelle, Mathematisierung und Erklären im Physikunterricht genannt.

Hinsichtlich der Trennschärfe der Konstrukte (vgl. A9) ergibt sich ein differenziertes Bild. Der überwiegende Anteil der Experten (17 von 18) halten eine Konstruktion von Skalen in den beiden Dimensionen Planung und Durchführung von Physikunterricht für relevant, aber es wird darauf verwiesen, dass sich die Trennbarkeit empirisch erweisen muss. Bei den Interviewpartnern zeigt sich die Tendenz, auch bei Planungsitems Situationen der Unterrichtsdurchführung zu assoziieren. Dazu kommt es insbesondere bei Items, in denen die Zugehörigkeit zu einer Dimension nicht expliziert wird (z. B. durch die Verwendung von „im Planungsprozess“ [Dimension Planung] oder „während des Unterrichts“ [Dimension Durchführung]). Items, die eine solche „Markierung“ bereits deutlich enthalten, sind hingegen unauffällig, so dass das Problem durch sprachliche Hervorhebung zu beheben ist.

Skalenrevision

Auf der Grundlage der Ergebnisse der Pilotstudien 2 und 3 fand eine umfangreiche Revision der Skalen statt. Dazu wurde in Sitzungen des Projektteams zunächst jedes einzelne Item überarbeitet, um dann auf Skalenebene durch Ergänzung einzelner Items oder Ausschluss von sehr ähnlichen Items die angemessene Darstellung des Handlungsfeldes abzusichern.

Folgende übergreifende Leitlinien (vgl. Konstruktionsregeln nach Pilot 2 und 3 in Tab. 1) wurden dabei berücksichtigt:

- Einführung einer einheitlichen Satzstruktur,
- Entfernung aller Verneinungen und extremen Zuspitzungen („immer“, „nie“),
- Ergänzung eines Instruktionstextes zu den Skalen bzgl. Schülervorstellungen,
- Beseitigung doppelter Hürden,
- sprachliche Hervorhebung der Planungsdimension bei den zugehörigen Items.

Für die Ergänzung von Items konnte auf die Empfehlungen der Experten zurückgegriffen werden.

Am Beispiel von drei Items soll der Überarbeitungsprozess – der entscheidend zur Qualität der Skalen beigetragen hat – transparent gemacht werden. Für das bisherige Item exp1 (vgl. Tab. 2) wiesen die Experten auf die problematische Negation („es bereitet mir keine Probleme“) hin, auch in den Interviews wurden dazu Schwierigkeiten sichtbar. Außerdem wurde von den Interviewpartnern zum Teil die Durchführungsdimension statt einer Planungssituation assoziiert. Die neue Version des Items, in der auch die über-

Tab. 3 Überarbeitete Skalenversion nach Pilotstudien 2 und 3; die Items wurden hier neu durchnummeriert, die kursiven Items wurden nach der Hauptstudie entfernt oder umformuliert; steil gekennzeichnet ist die Skala, wie sie zur Verwendung empfohlen wird; Antwortformat sechsstufig mit ausschließlicher Benennung der Extrema durch „stimmt nicht“ und „stimmt genau“

Items der Skala Experimentieren/Planung	
exp1	Ich kann in meiner Unterrichtsplanung zu den Lernzielen passende Experimente aufbauen, auch wenn die Physiksammlung schlecht ausgestattet ist.
exp2	Ich kann bei meiner Unterrichtsplanung ein Experiment gegebenenfalls so variieren, dass ich es in einer Übungsphase einsetzen kann, auch wenn ich es bisher nur als Einstiegsexperiment genutzt habe.
exp3	Ich kann Schülerexperimente so zusammenstellen, dass die praktischen Fähigkeiten meiner Schülerinnen und Schüler auf verschiedenen Niveaus gefördert werden, auch wenn ich bei der Planung unter Zeitdruck stehe.
exp4	*Ich kann zu dem Kontext einer Unterrichtsreihe ein Experiment entwickeln, auch wenn es zu diesem Kontext keine fertigen Experimentieranweisungen gibt.*
exp5	*Ich kann verschiedene Varianten eines Experimentes planen, mit denen sich Physikunterricht leistungsdifferenziert gestalten lässt, auch wenn es dazu keine Unterrichtsvorschläge gibt.*
exp6	Ich kann ein Experiment planen, das meine Schülerinnen und Schüler begeistert, auch wenn sie sich sonst wenig für Physik interessieren.
exp7	Ich kann für ein physikalisches Experiment begründet entscheiden, ob es didaktisch sinnvoller ist, es als Demonstrations- oder Schülerexperiment einzuplanen, auch wenn ich das Experiment noch nicht eingesetzt habe.
exp8	Ich kann in meiner Unterrichtsvorbereitung ein Experiment planen, welches meine Schülerinnen und Schüler dazu anregt, physikalische Fragestellungen zu entwickeln, auch wenn ich dieses Experiment neu entwickeln muss.
exp9	Ich kann mehrere Experimente so zusammenstellen, dass bei der Auswertung unterschiedliche Möglichkeiten des Umgangs mit Messdaten deutlich werden, auch wenn ich diesbezüglich keine Unterrichtsvorschläge kenne.
exp10	Ich kann bei der Unterrichtsplanung didaktisch begründet entscheiden, ob ein Experiment mit Hilfe von schultypischen Experimentiersets oder mit Alltagsgegenständen durchgeführt werden soll, auch wenn ich die Lerngruppe noch nicht lange kenne.
Items der Skala Experimentieren/Durchführung	
exd1	Ich kann physikalische Experimente an interessante Impulse meiner Schülerinnen und Schülern anpassen, auch wenn ich das vorher nicht geplant hatte.
exd2	Ich kann ein Demonstrationsexperiment für meine Schülerinnen und Schüler nachvollziehbar durchführen, auch wenn es sich um eine komplexe Versuchsanordnung handelt.
exd3	Ich kann beim Experimentieren spontan mit den Schülerinnen und Schülern über das Wechselspiel von Theorie und Experiment reflektieren, auch wenn ich den Anlass nicht vorgesehen hatte.
exd4	Ich kann auf unvorhergesehene Verständnisschwierigkeiten meiner Schülerinnen und Schülern beim Experimentieren so reagieren, dass sie selbstständig weiterarbeiten können, auch ohne einfach einen Lösungsweg vorzugeben.
exd5	Ich kann ein Experiment, das im Physikunterricht nicht auf Anhieb funktioniert, zum Laufen bringen, auch wenn ich unter Zeitdruck stehe.
exd6	Ich kann unerwartete Messwerte aus einem Demonstrationsexperiment spontan als Lernanlass für meine Schülerinnen und Schüler nutzen, auch ohne „unpassende“ Werte einfach zu übergehen.
exd7	*Ich kann spontan ein passendes Experiment einsetzen, um auf weiterführende physikalische Fragestellungen meiner Schülerinnen und Schüler zu reagieren, auch wenn ich das nicht vorgesehen hatte.*
exd8	*Ich kann Experiment so inszenieren, dass meine Schülerinnen und Schüler motiviert sind, eigene physikalische Fragestellungen zu entwickeln, auch wenn es sich um eine unbeliebte Randstunde handelt.* Ich kann ein Experiment so inszenieren, dass meine Schülerinnen und Schüler motiviert sind mitzuarbeiten, auch wenn es sich um eine unbeliebte Randstunde handelt.
exd9	Ich kann meine Schülerinnen und Schüler bei der Planung ihres experimentellen Vorgehens unterstützen, auch wenn sie im Physikunterricht Ihren eigenen Fragestellungen nachgehen.

greifenden Überarbeitungsleitlinien erkennbar sein sollten, ist in Tab. 3 (neu: exp1) zu finden.

Am Item exp2 (vgl. Tab. 2) wurde von den Experten bemängelt, es suggeriere, dass es immer sinnvoll sei, Experimente unter einer Motivationsperspektive zu planen, so dass ein starres Bild von Unterrichtsplanung entstehe. Die Zuspitzung „stets“ trägt zu dieser Einschätzung vermutlich bei und sorgt außerdem für eine hohe Schwierigkeit. Ähnliche Wahrnehmungen werden auch bei den Interviewpartnern sichtbar. Hinzu kommt, dass statt der intendierten „Unterrichtssequenz“ eher einzelne Unterrichtsstunden assoziiert und eher reine Showexperimente mit der Motivationsfunktion verbunden wurden. Die Hürde („auch wenn ich die Sequenz das erste Mal unterrichte“) wurde in diesem Zusammenhang als wenig relevant wahrgenommen. Entsprechend grundlegend fällt die Revision des Items (neu: exp6) aus, das ebenfalls Tab. 3 zu entnehmen ist.

Bei Item exd3 (vgl. Tab. 2) wurde die Barriere, die ungünstigerweise in dem Item doppelt benannt wurde („unvorbereitet“), sowohl von Experten als auch von Interviewpartnern als nicht authentisch kritisiert. Außerdem war einigen Interviewpartner/innen nicht klar, was mit „naturwissen-

schaftlichem Arbeiten“ gemeint sei. In der Überarbeitung (vgl. Tab. 3, neu: exd3) wurde versucht, diese Probleme auszuräumen, indem der spontane Anlass für die Handlung stärker herausgearbeitet wurde und die ursprüngliche Reflexion über „naturwissenschaftliches Arbeiten“ als Reflexion über „das Wechselspiel von Theorie und Experiment“ spezifiziert wurde.

Im Hinblick auf die Annahmen, die der intendierten Testwertinterpretation zugrunde liegen (hier Annahme 1 bis 7 und 9), wurde auf Basis der Pilotstudien 2 und 3 geschlussfolgert, dass diese aufrechterhalten werden können, sofern die Items so überarbeitet werden, dass wahrgenommene Probleme ausgeräumt werden. Es wurden solche kognitiven Prozesse beobachtet, die für die Erfassung von physikdidaktischen SWE intendiert waren und es gab keine Hinweise darauf, dass die Arbeitsdefinition grundsätzlich nicht angemessen umgesetzt wurde. Die Verständlichkeit der Items war entweder gegeben oder sollte durch gezielte Überarbeitung ermöglicht werden können. Die ausgewählten Handlungen und Hürden erschienen überwiegend als relevant und authentisch, konkrete Hinweise für die Herstellung von Relevanz konnten gewonnen werden. Die Trennschärfe der Skalen hinsichtlich der Dimensionen Planung und Durchführung wurde in der Itemrevision verstärkt, muss aber statistisch weiter geprüft werden.

Pilotstudie 4: Fragebogen zum Antwortformat

Neben der Vielzahl der ermutigenden Ergebnisse aus den Pilotstudien 2 und 3 deuten sich in den Probandeninterviews jedoch Schwierigkeiten mit dem genutzten vierstufigen Antwortformat an, sodass in einer zusätzlich durchgeführten Fragebogenerhebung, die Annahme 6 zur Eignung des Antwortformats für eine differenzierte Einschätzung aufgegriffen wird. Das vierstufige Antwortformat wurde zunächst gewählt, da dieses für etablierte Skalen zur allgemeinen SWE (Schwarzer und Jerusalem 1999) oder zur Lehrer-SWE (Schmitz und Schwarzer 2000) genutzt wurde und davon auszugehen war, dass die vorhandene Verwandtschaft der Konstrukte dies rechtfertige

Physiklehramtsstudierende ($N = 33$) wurden schriftlich befragt, welches Antwortformat ihnen geeignet erscheint, um die eigenen physikdidaktischen SWE differenziert angeben zu können.

Angeboten wurde eine zufällige Auswahl von 8 Items aus allen Skalen mit jeweils der Möglichkeit, sich auf einer 4-, 6-, 8- oder 10-stufigen Skala einzuschätzen. Ungeradzahlige Stufenzahlen wurden vermieden, um einer Tendenz zur Mitte vorzubeugen, auch wenn dies bei unipolaren Skalen eher weniger ein Problem darstellt. Für ein Unbehagen mit der geraden Anzahl und dem damit verbundenen Verzicht auf eine mittlere Antwortstufe ließen sich in den geführten Interviews keine Anzeichen finden. Die in Pilotstudie 4 mit dem Fragebogen Befragten wurden außerdem explizit nach der präferierten Stufung gefragt. Von den Studierenden zogen über 50 % eine sechsstufige Skala anderen Abstufungen vor, mit der Tendenz, dass die Zustimmung zu dieser 6-stufigen Skala bei Studierenden der höheren Semester noch höher lag. Das Ergebnis stimmt mit einer auf Miller (1956) zurückgehenden Empfehlung überein, 7 ± 2 Antwortstufen zu verwenden, was er mit der Leistungsfähigkeit unseres Kurzzeitgedächtnisses begründet. Diese Empfehlung wird mit dem Verweis auf neuere methodische Untersuchungen auch heute noch ausgesprochen (Franzen 2014, S. 705). Auch die Entscheidung für eine unipolare Antwortskala folgt den üblichen Empfehlungen, die in diesem Fall mit dem Bemühen um Einfachheit begründet wird (Franzen 2014, S. 707) und die sich auch empirisch stützen lässt (Schaeffer und Presser 2003). Die Ergebnisse der Pilotstudien 2, 3 und 4 deuten darauf hin, dass sich ein sechsstufiges Antwortformat der Likertskala als günstiger erweist als das bisher verwendete vierstufige Format. Dieses wurde entsprechend in der nachfolgenden Hauptstudie umgesetzt. An dieser Stelle sei vorwegnehmend berichtet, dass sich die Angemessenheit dieser Entscheidung auch darin zeigt, dass anders als in Pilotstudie 1, bei der eine vierstufige Skala verwendet wurde, Boden- und Deckeneffekte in der Hauptstudie weitgehend vermieden werden konnten. Auch die Untersuchungen der Wahrscheinlichkeitsverteilungen über die einzelnen Antwortstufen und der Rasch-Andrich-Tresholds stützen diese Entscheidung nachträglich.

Der Wert dieser Studie liegt damit insbesondere darin, den theoretisch empfohlenen Bereich der angebotenen Antwortstufen weiter einzugrenzen und die Wahrscheinlichkeit dafür zu erhöhen, dass die gewählte Anzahl der Stufen eine differenzierte Selbsteinschätzung der Probanden ermöglicht. Dies ist hier (wie die Ergebnisse der Hauptstudie zeigen werden) gelungen und zwar mit einem theoriekonformen, aber von etablierten Skalen abweichendem und insofern keineswegs trivialem Ergebnis. Dass ein solches Vorgehen in jedem Fall bei einer Skalenentwicklung notwendig ist, wollen wir nicht behaupten, wohl aber, dass es sich für dieses Projekt als hilfreich erwiesen hat, den Anzeichen für eine fehlende Passung des von etablierten Skalen übernommenen Antwortformats nachzugehen.

Hauptstudie: Fragebogenerhebung für alle Befragungsgruppen

Ziel der quantitativ angelegten Hauptstudie war, insbesondere die Annahmen 5 sowie 7 bis 11 der Testwertinterpretation auf den Prüfstand zu stellen.

Tab. 4 Deskription der Stichprobe

	Studierende	Referendare	Lehrkräfte	∑
Format				
Online	34	39	168	241
Papier & Bleistift	491	129	0	690
∑	525	238	168	931
Geschlecht				
Männlich	335	167	104	606
Weiblich	188	71	63	322
o. A.	2	0		3
∑	525	238	168	931
Studiengang				
Sekundar	140	39	36	215
Gymnasial	385	128	70	583
Quereinstieg	0	71	29	100
DDR-Diplom	0	0	33	33
∑	525	238	168	931
Schulform				
Sekundarschule	k. A.	k. A.	68	
Gymnasium	k. A.	k. A.	100	
∑			168	
Praxiserfahrung				
Keine	120	k. A.	k. A.	
Institutionelle	385	k. A.	k. A.	
Anderweitige	20	k. A.	k. A.	
∑	525			

Verwendet wurden dafür die grundlegend überarbeiteten Items bzw. Skalen, wie sie exemplarisch in Tab. 3 dargestellt sind. Zusätzlich wurden über den Fragebogen demographische Informationen erfragt und weitere Skalen (u. a. zum physikalischen Selbstkonzept, zu allgemeinen SWE und zu Lehrer-SWE) zur Bearbeitung vorgelegt, die dazu dienen sollten, strukturelle wie externale Facetten der Testwertinterpretation zu prüfen.

Befragt wurden Personen aus den drei Befragungsgruppen Physiklehramtsstudierende, Physikreferendare und Physiklehrkräfte, wobei sowohl paper-and-pencil Tests als auch online-Fragebögen eingesetzt wurden. Die resultierende Stichprobe umfasst nach Bereinigung der Daten $N = 931$ Personen. Details zur Stichprobe können Tab. 4 entnommen werden.

Die befragten Lehrkräfte waren im Mittel 45 Jahre alt (26–68 Jahre) und konnten auf eine durchschnittliche Berufserfahrung von 16 Jahren (0–40 Jahre) zurückblicken. Je ca. 30 % der befragten Lehrpersonen gaben an, in Brandenburg oder Rheinland-Pfalz zu unterrichten. In Berlin unterrichteten 20 %, in Hamburg 10 % der befragten Physik-Lehrkräfte. Am häufigsten vertraten die Physik-Lehrpersonen die Fachkombination Physik und Mathematik (87 %).

Im Durchschnitt waren die befragten Referendare 30 Jahre alt (23–58 Jahre) und absolvierten ihr Referendariat hauptsächlich in Berlin (32 %), Niedersachsen (27 %), Baden-Württemberg (17 %) oder Hamburg (13 %) an einem Gymnasium (53 %). Am häufigsten wurde das Fach Physik mit Mathematik kombiniert (66 %). 20 % der befragten Referendare gab an, zusätzlich ein drittes Fach studiert zu haben.

Die Studierenden waren zwischen 18 und 46 Jahre alt (Durchschnitt: 24 Jahre). 39 % der Lehramtsstudierenden waren in einem Bachelorstudiengang immatrikuliert, 18 % in einem Masterstudiengang und 42 % studierten, mit dem Ziel eines ersten Staatsexamens. Die meisten Studierenden absolvierten das 3. (21 %), 5. (31 %) oder 7. Semester (21 %). Bis auf vier Studierende, die an einer Pädagogischen Hochschule studierten, waren alle an einer Universität eingeschrieben. Eine Mehrheit von 72 % strebte den Abschluss eines Gymnasiallehramtsstudiums an, wobei Physik am häufigsten mit dem Fach Mathematik kombiniert wurde (67 %). Der auswertbare Datensatz enthält Antworten von Studierenden aller Bundesländer. Die größte Gruppe wurde am Standort Gießen befragt (ca. 9 %). Insgesamt 20 % der Befragten studierten an einer Berliner Universität.

Die statistische Auswertung der durch diese Stichprobe generierten Daten sah sowohl eine Analyse im Rahmen der klassischen Testtheorie als auch innerhalb des probabilistischen Paradigmas vor, wobei an dieser Stelle nur an den Stellen auf die Ergebnisse der Rasch-Analysen (Rating-Scale-Modell) verwiesen wird, an denen diese Ergebnis-

Tab. 5 Modellfit für die spezifizierten und revidierten Messmodelle zum Handlungsfeld Experimentieren: Ergebnisse der konfirmatorischen Faktorenanalyse

	Modell	Entf. Items	χ^2	df	χ^2/df	*p*	CFI	TLI	RMSEA	SRMR
Studierende	SWE-Ex-P spez	–	98,34	35	2,81	0,000	0,954	0,941	0,059	0,038
	SWE-Ex-P rev	exp4/5	49,18	20	2,46	0,000	0,971	0,959	0,053	0,032
	SWE-Ex-D spez	–	83,98	27	3,11	0,000	0,950	0,933	0,063	0,036
	SWE-Ex-D rev	exd7	59,23	20	2,96	0,000	0,959	0,942	0,061	0,035
Referendare	SWE-Ex-P spez	–	77,71	35	2,22	0,000	0,920	0,897	0,072	0,052
	SWE-Ex-P rev	exp4/5	21,61	20	1,08	0,362	0,995	0,993	0,018	0,034
	SWE-Ex-D spez	–	61,56	35	1,76	0,000	0,935	0,913	0,073	0,047
	SWE-Ex-D rev	exd7	29,07	20	1,45	0,086	0,979	0,970	0,044	0,036
Lehrkräfte	SWE-Ex-P spez	–	71,04	35	2,03	0,000	0,938	0,920	0,078	0,047
	SWE-Ex-P rev	exp4/5	34,90	20	1,75	0,009	0,954	0,935	0,073	0,042
	SWE-Ex-D spez	–	57,44	27	2,13	0,007	0,939	0,919	0,082	0,045
	SWE-Ex-D rev	exd7	40,89	20	2,05	0,004	0,948	0,927	0,079	0,044

CFI Comparative Fit Index, *TLI* Tucker-Lewis-Index, *RMSEA* Root Mean Square Error of Approximation, *SRMR* Standardized Root Mean Residual

se andere oder ergänzende Interpretationen zulassen. Alle Details zu den Auswertungsschritten (verwendete Verfahren mit den gewählten Kennwerten für Gütekriterien) und zur verwendeten Software sind in einem Skalenreport ausführlich dokumentiert und über die Plattform pedocs frei zugänglich (Meinhardt et al. 2016).

Im Folgenden werden die Auswertungsergebnisse bezüglich der statistisch zu überprüfenden Annahmen zur intendierten Testwertinterpretation nur in Auswahl und unter weitgehendem Verzicht auf statistische Details dargestellt. Für Details beispielsweise zum Umgang mit fehlenden Werten oder zu den Verteilungen auf den Items sei erneut auf die Skalendokumentation verwiesen.

Annahme 7: In Konfirmatorischen Faktorenanalysen (KFA) für die einzelnen Befragungsgruppen zeigt sich (unter Berücksichtigung, dass die Werte in der Regel nicht normalverteilt sind), dass der Modellfit für die spezifizierten Modelle, die der Skalenversion aus Tab. 3 entsprechen, akzeptabel ist. Nach einer Revision der Skalen durch die Entfernung von Items werden die Modellfits gut bis sehr gut (vgl. Tab. 5). Vor der endgültigen Entfernung von problematischen Items wird neben der statistischen Analyse auch eine inhaltliche Überprüfung vorgenommen. Es wird überprüft, ob inhaltliche Probleme erkennbar sind bzw. ob das Handlungsfeld inhaltlich auch ohne das Item ausreichend gut dargestellt wird. Exemplarisch soll dies an Item exp4 (vgl. Tab. 3) erläutert werden. Statistisch gesehen weist exp4 insbesondere in der Gruppe der Referendare in der Summe die höchsten Modifikationsindizes aller Items der Skala auf. Inhaltlich betrachtet liegt es wahrscheinlich eher am Rande des Konstruktes, was sich in den Interviews bereits angedeutet hatte. Lehrkräfte sehen es i. d. R. nicht als ihre Kernaufgabe an, selbst Experimente oder Aufgabenstellungen zu entwickeln, sondern nutzen vorhandene Experimentieranleitungen bzw. variieren diese. Diese Erwägungen führen schließlich zum Ausschluss des Items von der Skala.

Annahme 8: Die Reliabilitätsmaße weisen darauf hin, dass die Konstrukte mit den Skalen zuverlässig abgebildet werden. Es ergeben sich niedrige mittlere durchschnittliche Inter-Itemkorrelationen (dIK für die Skala Experimentieren/Planung:0,37 $\leq$ dIK $\leq$0,43; für Experimentieren/Durchführung:0,38 $\leq$ dIK $\leq$45; jeweils nach Revision der Skalen) und Itemtrennschärfen im mittleren Bereich.

Die Skalenreliabilitäten, hier durch Cronbach's α und die im Rahmen des konfirmatorischen Ansatzes geschätzten Faktorreliabilitäten bzw. durchschnittlich extrahierten Varianzen gekennzeichnet, liegen in einem sehr guten Bereich und verringern sich für jede Kohorte mit der Revision der Skalen nur leicht (vgl. Tab. 6).

Annahme 5 und Annahme 11: Beide Annahmen betreffen die Frage der Messinvarianz. Zunächst wird ein Vergleich der auf Grundlage der Pilotierungen antizipierten Itemschwierigkeiten mit der statistischen Schwierigkeit (auf Grundlage der Item-Mittelwerte) vorgenommen. Es zeigt sich eine weitgehende und kohortenübergreifende Übereinstimmung, wobei zusätzliche schwierige Items identifiziert werden können.

Desweiteren werden für die revidierten Modelle aus der KFA Mehrgruppenvergleiche (sowohl im Rahmen der KFA als auch als DIF-Analyse) zur Prüfung der Messinvarianz durchgeführt. Die Daten weisen bei allen Skalen auf mindestens partielle skalare Messinvarianz hin, wobei unterschiedlich viele Intercepts frei zu schätzen sind (vgl. Tab. 7). Für den Vergleich von Mittelwerten, wie er in Anwendungssituationen häufig erforderlich wird, ist die (partielle) Invarianz von Faktorladungen und Intercepts, also das Vorliegen partieller skalarer Invarianz hinreichend (Brown

Tab. 6 Skalenreliabiltiäten nach der Revision

	Studierende			Referendare			Lehrkräfte		
	α_C	FR	DEV	α_C	FR	DEV	α_C	FR	DEV
Ex-P	0,84	0,84	0,40	0,82	0,82	0,36	0,86	0,86	0,44
Ex-D	0,83	0,83	0,39	0,83	0,83	0,39	0,86	0,87	0,45

α_C Cronbach's Alpha, *FR* Faktorreliabilität, *DEV* Durchschnittlich Extrahierte Varianz

Tab. 7 Ergebnisse der Mehrgruppenvergleiche

Modell	Gruppen	Art der Messinvarianz; [frei zu schätzende Intercepts]
Ex-P	S, R, L	Partielle skalare Messinvarianz; [exp10], [exp2], [exp7], [exp6]
Ex-D	S, R, L	Partielle skalare Messinvarianz; [exd8], [exd9]

S Studierende, *R* Referendare, *L* Lehrkräfte

2006, 290, 300), so dass die Ergebnisse insgesamt darauf hindeuten, dass die Items in den Befragungsgruppen überwiegend ähnlich verstanden werden und Mittelwertvergleiche auf dieser Basis problemlos möglich sind. Angemerkt sei, dass wir uns hier einem recht hohen Standard stellen und dass die Verletzung der partiellen skalaren Invarianzbedingung nicht automatisch zum Verwerfen der Durchführbarkeit von Mittelwertvergleichen führen muss. Vielmehr kann ein Abweichen der Intercepts gerade Expertiseunterschiede abbilden (Krafft und Litfin 2002), (partielle) metrische Invarianz also eine ausreichende Voraussetzung für die Durchführung von Mittelwertvergleichen sein.

Annahme 9: Der Frage, ob die mit den Skalen repräsentierten Konstrukte hinreichend gegeneinander abgrenzbar sind, soll mittels Modellvergleichen im Rahmen der KFA und durch Korrelationsanalysen nachgegangen werden. Im Rahmen der KFA werden eindimensionale Modelle (Planungs- und Durchführungsskala im Handlungsfeld Experimentieren werden zusammengefasst) mit zweidimensionalen Modellen verglichen, in denen die Dimensionen Planung und Durchführung als getrennte, aber miteinander korrelierende Dimensionen modelliert werden. In der Regel bevorzugen Satorra-Bentler-korrigierte χ^2-Differenztests die zweidimensionalen Modelle bei allerdings hoher latenter Korrelation der Dimensionen (für das Handlungsfeld Experimentieren $0{,}89 \leq r \leq 0{,}96$ in den drei Befragungsgruppen).

Korrelationsvergleiche über Spearman's ρ weisen ebenfalls auf hohe Korrelationen hin (für das Handlungsfeld Experimentieren: $0{,}67^{**} \leq \rho \leq 0{,}81^{**}$ in den drei Befragungsgruppen). Vorläufig werden diese Befunde so gedeutet, dass sie die Annahme zur Trennbarkeit der Dimensionen eher stützen, aber sie verdient in jedem Fall auch in weiteren Validierungsanstrengungen Aufmerksamkeit. Eine Konsequenz, die von Seiten der Testentwickler gezogen wurde, ist der Vorschlag von Kurzskalen, die die Abbildung eines Handlungsfeldes ohne Berücksichtigung der Dimensionen Planung und Durchführung erlaubt (vgl. Meinhardt et al. 2016). Es wird vom jeweiligen Anwendungsfall abhängen, ob die Ausprägungen von physikdidaktischen SWE in einer Dimension oder in beiden Dimensionen interessieren oder ob ggf. eine Ausprägung der physikdidaktischen SWE für das Handlungsfeld an sich ausreicht. Hier sind also potentielle Testanwender in der Verantwortung ggf. weitere Validierungsmaßnahmen zu ergreifen.

Annahme 10: Ob die physikdidaktischen SWE gegenüber anderen Konstrukten abgrenzbar sind, wird mit Hilfe von Korrelationsanalysen geprüft. Betrachtet werden hier exemplarisch die Korrelationen zwischen den revidierten Skalen zum Handlungsfeld Experimentieren und den zusätzlich erhobenen Skalen zu allgemeinen SWE (Schwarzer und Jerusalem 1999, S. 60), Lehrer-SWE (Schmitz und Schwarzer 2000) und dem physikalischen Selbstkonzept (nur für die Studierendenkohorte, adaptiert nach Hoffmann et al. 1998). Es ergeben sich theoriekonform schwache bis mittlere Korrelationen zu den allgemeinen SWE, mittlere (also etwas größere) Korrelationen zu den L-SWE und sehr schwache Korrelationen zum physikalischen Selbstkonzept (vgl. Tab. 8 exemplarisch für die exp-Skalen).

Insofern wird davon ausgegangen, dass sich die neuen Skalen zu physikdidaktischen SWE hinreichend gut gegen andere, benachbarte Konstrukte abgrenzen lassen. Damit sind auch die konkurrierenden Testwertinterpretationen betroffen: Sollten die neuen Skalen eines der benachbarten Konstrukte erfassen, wäre zu erwarten, dass die Korrelationen zu den entsprechenden Skalen deutlich höher ausfallen.

Weitere Korrelationsanalysen, die begründete Hypothesen bzgl. der Zusammenhänge der neu entwickelten Skalen untereinander aufgreifen, können bei Meinhardt (2018) nachgelesen werden.

Annahme 12: Bei den vorliegenden Daten handelt es sich nicht um längsschnittliche Daten, so dass Entwicklungen von physikdidaktischen SWE im eigentlichen Sinne nicht betrachtet werden können. Trotzdem können Mittelwertvergleiche für die Befragungsgruppen erste Evidenzen dazu liefern, ob die Mittelwerte der Befragungsgruppen solchen Erwartungen entsprechen, die aufgrund theoretischer Betrachtungen und der Durchsicht bisheriger Studien zu SWE

Tab. 8 Korrelationen (Spearman's ρ) zwischen den neu entwickelten Skalen und benachbarten Konstrukten

		A-SWE	L-SWE	PS
Studierende	SWE-Ex-P	0,316**	0,465**	0,158**
	SWE-Ex-D	0,370**	0,486**	0,180**
Referendare	SWE-Ex-P	0,409**	0,500**	k. A.
	SWE-Ex-D	0,492**	0,463**	k. A.
Lehrkräfte	SWE-Ex-P	0,297**	0,446**	k. A.
	SWE-Ex-D	0,348**	0,495**	k. A.

Tab. 9 Geschätzte Mittelwertdifferenzen im Rahmen der KFA nach Personengruppe, Referenzgruppe jeweils an erster Stelle genannt

Skala	Gruppenvergleich	ΔMW	S. E.	ΔMW/S. E.	*p*
SWE-Ex-P	Studierende/Referendare	–0,242	0,092	–2,632	0,01
	Studierende/Lehrkräfte	0,449	0,107	4,189	0,00
	Referendare/Lehrkräfte	0,690	0,123	5,168	0,00
SWE-Ex-D	Studierende/Referendare	–0,021	0,090	–0,229	0,82
	Studierende/Lehrkräfte	0,841	0,105	8,020	0,00
	Referendare/Lehrkräfte	0,862	0,121	7,138	0,00

Δ*MW* Schätzwert der Mittelwertdifferenz, *S. E.* Standardfehler des Schätzwertes

(wenn auch auf anderen Spezifitätsniveaus) aufgestellt wurden. Weitere Subgruppenvergleiche (nach Geschlecht, Schulform, Studiengang und Praxiserfahrung) wurden zwar durchgeführt, sollen hier aber nicht dargestellt werden. Die Ergebnisse dieser Analysen unterstützen allerdings nur bedingt ein Validitätsargument, da belastbare theoretische Annahmen fehlen, können aber für die Generierung von Anschlussfragestellungen genutzt werden (vgl. dazu Meinhardt 2018).

Berechnet wurden die latenten Mittelwertdifferenzen im Zuge der Mehrgruppenvergleiche im Rahmen der KFA. Es zeigt sich, dass erwartungskonform die Mittelwerte der Lehrkräfte für alle Skalen höher geschätzt werden als die der Studierenden und der Referendare. Allerdings werden auch für vier von acht Skalen die Mittelwerte der Referendare signifikant niedriger geschätzt als für die Studierenden. Dieser Befund kann, wie bereits erwähnt, nicht vor einem belastbaren theoretischen Hintergrund interpretiert und deshalb nicht für die Ableitung eines Validitätsarguments herangezogen werden. Gegebenenfalls können die vorliegenden Daten jedoch als Indiz für einen Praxisschock (Lamote und Engels 2010) gewertet werden und sind daher mit Blick auf sich anschließende Forschungsbemühungen von Interesse. In diesem Zusammenhang sollte auch der in der Literatur beschriebene Mentoreneffekt berücksichtigt werden (Woolfolk Hoy und Spero 2005; Fives et al. 2007; Richter et al. 2011), der im Studium u. U. dafür sorgt, dass aufgrund der gut betreuten Praxiserfahrungen der Praxisschock ausbleibt oder abgemildert wird. In Tab. 9 sind die Werte exemplarisch für die Skalen zum Handlungsfeld Experimentieren dargestellt.

Abschließend soll zunächst die Frage der Repräsentativität der oben beschriebenen Stichprobe thematisiert und kurz auf die Stichprobengröße eingegangen werden. Für die Gruppe der Studierenden und Referendare, die bundesweit für die Befragung gewonnen wurden, liegen derzeit keine plausiblen Gründe oder Evidenzen dafür vor, dass die Testwertinterpretation nicht auf die Gesamtheit der deutschen Physiklehramtsstudierenden und -referendare verallgemeinert werden sollte, auch wenn es sich streng genommen nicht um eine repräsentative Stichprobe handelt. Bei der Befragungsgruppe der Physiklehrkräfte ist insofern bei der Nutzung der Skalen weitere Aufmerksamkeit angeraten, als die Stichprobe möglicherweise eine Positivauswahl darstellt, da die Lehrkräfte über den Online-Fragebogen für die Untersuchung gewonnen wurden. Es ist nicht unwahrscheinlich, dass dadurch die Repräsentativität der Stichprobe Einschränkungen unterliegt.

Die Stichprobengröße, insbesondere der Gruppe der Studierenden ($N = 525$), mag auf den ersten Blick überraschen – auch bei der vorgesehenen Verwendung von Strukturgleichungsmodellen bzw. konfirmatorischen Faktorenanalysen. Neuere Analysen legen nahe, dass die Frage der Stichprobengröße nur in Abhängigkeit von der Anzahl der verwendeten Faktoren, der Anzahl der Indikatoren Indikatoren je Faktor und deren Ladungen, der gewünschten Testpower, Art und Anzahl fehlender Daten, des verwendeten Schätzers etc. zu beantworten ist (vgl. z. B. Wolf et al. 2013). Die Ergebnisse weichen dann mehr oder weniger stark von den typischen Faustregeln (z. B. Bentler und Chou 1987, S. 91) ab. Für zweifaktorielle Modelle (hier: Planung, Durchführung) würden demnach bei Faktorladungen von 0,65 und 8 Items je Faktor (ohne fehlende Werte), einer minimalen Testpower von 0,8 ($\alpha = 0{,}05$) und bei Verwendung eines ML-Schätzers ca. 120 Probanden für eine Einzelgruppenanalyse genügen. Diese Größenordnung erreichen wir für alle Subgruppen der Studierendenkohorte, die im Rahmen der durchgeführten Mehrgruppenvergleiche notwendig sind

nur knapp. In einzelnen Subgruppen wird diese Zahl (insbesondere wegen des hohen Anteils männlicher Lehramtstudierender im Fach Physik für Gymnasium) deutlich übertroffen, was in Zeiten studentischer Übertestung auch auf einen unerwartet hohen Rücklauf an Fragebögen zurückzuführen ist. Besteht also anders als in der hier vorgestellten Studie in künftigen Untersuchungen die Möglichkeit, diesen Rücklauf genauer zu prognostizieren oder gar zu steuern, ließe sich so der Stichprobenumfang und damit der Aufwand gezielt reduzieren.

Die zum jetzigen Zeitpunkt zur Verwendung empfohlenen Skalen zum Handlungsfeld Experimentieren können Tab. 3, ansonsten aber auch dem Skalenreport (Meinhardt et al. 2016) entnommen werden, wo Details zur Veränderung von Items aufgrund der Hauptstudie offengelegt werden.

Zusammenfassung und Diskussion

An die aktuelle Diskussion um die Interpretation von erhobenen Daten im Kontext der fachdidaktischen Forschung zur Lehrerprofessionalität anknüpfend war es Ziel der Darstellungen, eine konkrete Vorgehensweise für die Durchführung einer Validierungsstudie zur Diskussion zu stellen, die ggf. als Beispiel dienen kann. Beispielhaft kann dieses Vorgehen jedoch nicht im Sinne eines abzuarbeitenden Routineverfahrens werden, welches eins-zu-eins auf andere Fragestellungen übertragen werden kann. Dies widerspricht auch per se dem vorgestellten Validierungskonzept. Vielmehr meint „beispielhaft“ in diesem Zusammenhang ein systematisches und an einer theoretisch fundierten Testwertinterpretation orientiertes Vorgehen, welches jedoch für jeden Forschungskontext individuell zugeschnitten werden muss. In den dargelegten Validierungsschritten konnten eine Vielzahl von Indizien für die vorläufige Gültigkeit der festgelegten Testwertinterpretation gesammelt werden (Validitätsargument). Das präsentierte Verfahren erscheint jedoch nicht aufgrund dieser wahrscheinlichen Gültigkeit als tragfähig. Vielmehr ist davon auszugehen, dass die durchgeführten Schritte bei mangelhafter Operationalisierung die Ablehnung des intendierten Interpretationsansatzes nahegelegt hätten. Nicht anhand des Ergebnisses, sondern anhand des Prozesses muss also über die Qualität der Validierungsstudie entschieden werden. Da Instrumente zur Erfassung benachbarter Konstrukte auf einem ähnlichen Spezifitätsniveau teils gänzlich fehlen oder aber keinen hinreichenden Validierungsprozess durchlaufen haben, stoßen strukturelle wie externale Validierungsfacetten (z. B. Korrelationsanalysen) teilweise an ihre Grenzen. Beispielsweise gibt es keine Skala, die physikdidaktische Handlungsergebniserwartungen erfasst. Insofern konnten alternative, konkurrierende Testwertinterpretationen hauptsächlich durch inhaltliche Überlegungen und Argumente entkräftet werden. Zu Validierungszwecken wäre es ebenso wünschenswert, mit Hilfe des neu entwickelten Instruments theoriekonforme Mittelwertunterschiede hinsichtlich der Ausprägung der physikdidaktischen SWE bei verschiedenen Probandengruppen abbilden zu können. Eine unzureichende theoretische Basis sowie wenige oder widersprüchliche Studienergebnisse insbesondere auf dem relevanten Spezifitätsniveau erlauben es jedoch nicht, Mittelwertvergleiche für die Validierung zu nutzen. Das Validitätsargument könnte durch empirische Belege in den genannten Bereichen jedoch weiter gestärkt werden.

Insgesamt kann nach dem Durchführen der Einzelstudien und Abwägen der Ergebnisse argumentiert werden, dass mit Hilfe des Instrumentariums beschreibende und verallgemeinernde Schlüsse bezüglich vorliegender physikdidaktischer SWE vertreten werden können, andere Schlussfolgerungen (z. B. Prognosen) jedoch nicht durch die Validierung abgesichert sind. Dies ist ein Resultat der auf deskriptive Aspekte eingeschränkten Testwertinterpretation. In dieser Begrenzung liegen Stärke und Schwäche zugleich, denn mögliche Einsatzszenarien sind folgerichtig limitiert; wenige Anwendungskontexte können dafür guten Gewissens empfohlen werden. Insbesondere scheint das Instrument geeignet, um Kohorten von Physiklehramtsstudierenden und Physiklehrpersonen (im Vorbereitungsdienst) miteinander zu vergleichen, Entwicklungsverläufe zu untersuchen sowie mögliche Quellen oder Einflussmöglichkeiten auf das domänenspezifische Konstrukt zu erforschen. Abhängig vom jeweiligen Erkenntnisinteresse sind die Skalen darüber hinaus geeignet, um für spezifischere Fragestellungen angepasst zu werden (z. B. auf den Bereich der Mechanik). Offen muss zunächst bleiben, welches Spezifitätsniveau für bestimmte Fragestellungen am geeignetsten ist.

Die erreichte Qualität der Skalen resultiert aus unserer Sicht zu einem guten Teil aus der ausführlichen Vorbereitung der Haupterhebung durch die Pilotstudien, da die Qualität der einzelnen Items und damit auch der Skalen insgesamt dadurch erheblich gesteigert bzw. abgesichert werden konnte. Ein weiteres Qualitätsargument stellt die ausführliche Dokumentation der Validierungsergebnisse in Form des Skalenreports dar, der es zukünftigen Testnutzern erlaubt, an die bisherigen Validierungsschritte anzuknüpfen und ihre Qualität zu beurteilen.

Nachfolgend sollen einige Merkmale des beschriebenen Prozesses detaillierter betrachtet, sowie aus dem argumentbasierten Ansatz resultierende Implikationen für Forschungsprozesse diskutiert werden. Auch wenn dieser Aspekt hier nur angedeutet werden konnte, so ist für eine sinnstiftende Validierungsarbeit eine tiefgreifende Aufarbeitung des Forschungsgegenstandes vonnöten. Aus dieser resultiert eine inhaltliche Klarheit über die zu erhebenden Konstrukte, die als Reflexionsfolie in vielerlei Hinsicht un-

abdingbar ist, z. B. bei der Evaluation eigener/fremder Operationalisierungen oder der theoriegeleiteten Überarbeitung von Items. Für den Forschungsprozess ist diese umfassende Theoriearbeit mühsam, weil in der Regel schnell ein intuitiver Zugang zu interessierenden Konstrukten vorliegt und aufgrund immer komplexerer Fragestellungen häufig eine Vielzahl von Konstrukten betrachtet werden müssen. Hinzu kommt, dass der vorgestellte Validierungszugang relativ kleinschrittig und in jedem Fall mehrstufig ist. Damit geht aber auch ein erhöhtes Maß an Transparenz einher, da Annahmen und Maßnahmen aufeinander aufbauen und offengelegt sind. Einzelne Überarbeitungen und Verbesserungen können nachvollzogen werden, wobei der Dokumentationsaufwand teils erheblich ist. Die Mehrstufigkeit erlaubt es zusätzlich, die Vorteile der Wechselbeziehung von Theoriearbeit und Instrumentenentwicklung zu nutzen. Oft wird erst im Überarbeitungsprozess die Ausschärfung festgelegter Merkmale und Kriterien von Bedeutung, was u. a. zu der beschriebenen Qualitätssteigerung in den Itemformulierungen beiträgt.

Zu dieser trägt unserer Meinung auch der Fokus der Pilotierungen auf qualitative Forschungsansätze bei. Deren Beitrag zur Qualitätssicherung der Items und Skalen liegt vor allem in der Zugänglichkeit der Interpretations- und Abwägungsprozesse der Probanden und dem damit erreichten tieferen Einblick in „das Funktionieren“ der Items begründet. Die Absicherung der Itemqualität auf dieser inhaltlichen Ebene scheint uns notwendige Voraussetzung für eine sinnvolle Qualitätssicherung auf der quantitativen Ebene zu sein. Ohne sie verkommt das Berichten von Kennwerten der Modellpassung oder Konsistenzmaßen zu einer formalen Angelegenheit, die wenig dazu beiträgt, eine bestimmte Testwertinterpretation zu stützen. Qualitative Studien können gleichzeitig die Grenzen der Testwertinterpretation sichtbar machen, wie die letzte vorgestellte Teilstudie gezeigt hat. Die verbreitete Praxis der Instrumentenentwicklung, bei der entweder bekannte Items angepasst beziehungsweise ergänzt oder gar ganze Skalen ad hoc generiert werden („quick-and-dirty-Verfahren“), fokussiert dagegen häufig schwerpunktmäßig auf eine „quantitative“ Analyse derart zusammengestellter Instrumente, was sich vor dem Hintergrund der in diesem Projekt gemachten Erfahrungen als problematisch erweist.

Insgesamt ist zu betonen, dass die genannten Merkmale des beschriebenen Validierungsprozesses wie Theoriearbeit, Mehrstufigkeit, Transparenz, Systematik, Verzahnung von qualitativen und quantitativen Ansätzen etc. formal vorhanden sein können, ihr Gewinn jedoch erst dadurch zustande kommt, dass auch und gerade unerwartete, widersprüchliche und unbequeme Befunde ernst genommen und in die weitere Arbeit integriert werden. Letztlich hängt die Qualität der Validierung maßgeblich von der Fähigkeit und dem Willen der Forschenden ab, sich immer wieder gegenseitig zu widerlegen und in Frage zu stellen.

Abschließend werden Grenzen des dargestellten Validierungsunterfangens aufgezeigt und diskutiert. Das gewählte Vorgehen führt aufgrund des damit verbundenen zeitlichen wie personellen Aufwands zu einer „Entschleunigung des Forschungsprozesses“. In einer Kosten-Nutzen-Abwägung steigern sich somit auf den ersten Blick die Kosten. Für die Autoren erscheint der gewonnene Nutzen jedoch ungleich höher, weil Forschungstraditionen auf tragfähigeren Fundamenten aufbauen können. Die Investitionskosten zahlen sich perspektivisch damit um ein Vielfaches aus. Im mittlerweile schnelllebigen Forschungsbetrieb stellt die Kostenerhöhung jedoch ein nicht zu unterschätzendes Problem dar, weil Forschungsszenarien in der Regel eher kurze Zeitskalen umfassen, sodass schneller hohe Erträge abgerechnet werden können.

Die Verlangsamung des Forschungsprozesses wird zusätzlich dadurch hervorgerufen, dass – wie im vorliegenden Fall – zunächst nur deskriptive Testwertinterpretationen geprüft werden können. Erwartungshaltungen an die Instrumentenentwicklung müssen damit grundsätzlich in Frage gestellt und überdacht werden. Was kann der Testentwickler leisten? Welche Verantwortung muss ggf. zukünftig von Testanwendern übernommen werden? Inwiefern wäre dieses Prinzip der „joint responsibility“ praktikabel? Hierbei handelt es sich um eine Herausforderung für etablierte Forschungspraxen, die unserer Ansicht nach eine breite Diskussion in der Forschungsgemeinschaft erfordert. Wie soll in Zukunft an Forschungsfragen herangegangen werden? Wann wäre es beispielsweis gerechtfertigt, neue Instrumente zu entwickeln? Welche (Qualitäts)Kriterien legen wir uns als community auf? Welche Konsequenzen sind aufgrund einer anderen Forschungspraxis einzukalkulieren? Solche oder ähnliche Fragen werden u. a. bedeutsam, wenn man das derzeit diskutierte „Validitäts-Performanz-Dilemma“ (Fischer 2017, S. 251) genauer betrachtet, da sich in diesem Kontext augenscheinlich die Frage stellt, ob und inwiefern Validierungen deskriptiver Testwertinterpretationen präskriptiven vorausgehen müssen oder ob und inwiefern präskriptive Interpretationen theoretisch hinreichend untermauert sind.

Zusammenfassend scheint uns die damit aufgeworfene Frage nach dem Umgang mit dem Validierungsanspruch an Testinstrumente sowohl eine innerwissenschaftliche als auch eine wissenschaftspolitische Facette zu umfassen. Formal ist der Anspruch an valide Testwertinterpretationen auf wissenschaftlicher Ebene formuliert (AERA 2014) und in der Folge ernst zu nehmen. Die Forschungspraxis vermittelt dagegen häufig ein anderes Bild. Dass Studien im Bereich der Lehrerprofessionalität im deutschsprachigen Raum immer wieder auf die Problematik einer validen Testwertinterpretation zurückgeworfen werden, zeigt u. a. der aktuelle

Sammelband zur professionellen Kompetenz von Lehrkräften der Chemie und Physik (Fischler und Sumfleth 2017). Auf wissenschaftspolitischer Ebene besteht gerade in einem Feld wie der Lehrerprofessionalisierung ein durchaus berechtigtes Interesse daran, möglichst zügig zu Erkenntnissen über das Untersuchungsfeld zu gelangen, da für dieses Interesse teils erhebliche finanzielle Mittel zur Verfügung gestellt werden. Andererseits sollen sich die Ressourcen möglichst nachhaltig in validen Testwertinterpretationen und damit verlässlichen Studienergebnissen niederschlagen, die der Forschungsgemeinschaft und/oder politischen Entscheidungsträgern zur Verfügung gestellt werden.

Wir stellen uns keineswegs auf den Standpunkt, dass das angedeutete Dilemma zwischen Effizienz- und Validitätsansprüchen einfach aufzulösen wäre. Eine sorgfältige Abwägung der Ansprüche scheint uns allerdings unumgänglich und eine wissenschaftliche Debatte, zu diesem Dilemma würden wir uns wünschen. Eine Debatte, an deren Ende Cronbachs Beschreibung der Validierungsbemühungen seiner Generation in unserer Forschungsgemeinschaft endgültig der Vergangenheit angehört und einer wissenschaftsethisch-rational reflektierten Normalität gewichen ist.

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

Danksagung Die Autoren danken den Reviewern und Herausgebern für zum Teil harte Diskussionen, die uns dazu bewegten, Argumente auszuschärfen und die so die Qualität des Beitrages positiv beeinflusst haben. Wir hoffen auf eine ähnlich intensive öffentliche Fortsetzung der Debatte.

Literatur

AERA (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Bandura, A. (1997). *Self-efficacy. The exercise of control*. New York: W. H. Freeman and Company.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Hrsg.), *Self-efficacy beliefs of adolescents* (S. 307–337). Greenwich: Information Age.

Bentler, P.M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods and Research*, *16*, 78–117.

Brouwers, A., & Tomic, W. (2003). A test of the factorial validity of the teacher efficacy scale. *Research in Education*, *69*, 67–79.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York, London: The Guilford Press.

Cakiroglu, J., Capa-Aydin, Y., & Woolfolk Hoy, A. (2012). Science teaching self efficacy beliefs. In B.J. Fraser, K.G. Tobin & C.J. McRobbie (Hrsg.), *Second international handbook of science education* (S. 449–462). Dordrecht: Springer.

Cauet, E., Liepertz, S., Borowski, A., & Fischer, H.E. (2015). Does it matter what we measure? Domain-specific professional knowledge of physics teachers. *Revue Suisse des Sciences De L'éducation*, *37*(3), 462–479.

Coladarci, T., & Fink, D.R. (1995). *Correlations among measures of teacher efficacy: Are they measuring the same thing?* Annual Meeting of the American Educational Research Association, San Francisco.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Hrsg.), *Test validity* (S. 3–17). Hillsday: Lawrence Erlbaum.

Dellinger, A.B., Bobbett, J.J., Olivier, D.F., & Ellett, C.D. (2008). Measuring teachers' self-efficacy beliefs: Development and use of the TEBS-Self. *Teaching and Teacher Education*, *24*, 751–766. https://doi.org/10.1016/j.tate.2007.02.010.

Denzine, G.M., Cooney, J.B., & McKenzie, R. (2005). Confirmatory factor analysis of the teacher efficacy scale for prospective teachers. *The British Journal of Educational Psychology*, *75*(4), 689–708. https://doi.org/10.1348/000709905X37253.

Dickmann, M. (2016). *Messung von Experimentierfähigkeiten. Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*. Berlin: Logos.

Enochs, L.G., & Riggs, I.M. (1990). Further development of an elementary science teaching efficacy belief instrument: a preservice elementary scale. *School Science and Mathematics*, *90*(8), 694–706.

Fischer, H.E. (2017). Professionskompetenz und Handeln von Lehrpersonen – Ein Referenzrahmen und kritische Bemerkungen. In H. Fischler & E. Sumfleth (Hrsg.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik* (S. 237–255). Berlin: Logos.

Fischler, H., & Sumfleth, E. (Hrsg.). (2017). *Professionelle Kompetenz von Lehrkräften der Chemie und Physik*. Berlin: Logos.

Fives, H., Hamman, D., & Olivarez, A. (2007). Does burnout begin with student-teaching? Analyzing efficacy, burnout, and support during the student-teaching semester. *Teaching and Teacher Education*, *23*(6), 916–934. https://doi.org/10.1016/j.tate.2006.03.013.

Franzen, A. (2014). Antwortskalen in standardisierten Befragungen. In N. Baur & J. Blasius (Hrsg.), *Handbuch Methoden der empirischen Sozialforschung* (S. 701–711). Wiesbaden: Springer.

Gibson, S., & Dembo, M.H. (1984). Teacher efficacy: a construct validation. *Journal of Educational Psychology*, *76*(4), 569–582. https://doi.org/10.1037//0022-0663.76.4.569.

Gramzow, Y. (2015). *Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik. Modellierung und Testkonstruktion*. Berlin: Logos.

Guskey, T.R. (1988). Teacher efficacy, self-concept, and attitudes toward the implementation of instructional innovation. *Teaching and Teacher Education*, *4*(1), 63–69. https://doi.org/10.1016/0742-051X(88)90025-X.

Guskey, T.R., & Passaro, P.D. (1994). Teacher efficacy: a study of construct dimensions. *American Educational Research Journal*, *31*(3), 627–643.

Hadenfeldt, J.C., & Neumann, K. (2012). Die Erfassung des Verständnisses von Materie durch Ordered Multiple Choice Aufgaben. *Zeitschrift für Didaktik der Naturwissenschaften*, *18*, 317–338.

Henson, R.K. (2001). *Teacher self-efficacy: substantive implications and measurement dilemmas*. Annual meeting of the Educational Research Exchange, Texas, 26.1.2001.

Hoffmann, L., Häussler, P., & Lehrke, M. (1998). *Die IPN-Interessenstudie*. Kiel: IPN.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12001.

Klassen, R. M., Tze, V. M. C., Betts, S. M., & Gordon, K. A. (2011). Teacher efficacy research 1998–2009: signs of progress or unfulfilled promise? *Educational Psychology Review*, *23*(1), 21–43. https://doi.org/10.1007/s10648-010-9141-8.

Krafft, M., & Litfin, T. (2002). Adoption innovativer Telekommunikationsdienste: Validierung der Rogers-Kriterien bei Vorliegen potenziell heterogener Gruppen. *Zeitschrift für betriebswirtschaftliche Forschung*, *54*(2), 64–83.

Kushner, S. N. (1993). *Teacher efficacy and Preservice teachers: a construct validation*. Annual Meeting of the Eastern Educational Research Association, Clearwater Beach, Februar: 17–22, 1993.

Lamote, C., & Engels, N. (2010). The development of student teachers' professional identity. *European Journal of Teacher Education*, *33*(1), 3–18. https://doi.org/10.1080/02619760903457735.

Leuders, T. (2014). Modellierungen mathematischer Kompetenzen – Kriterien für eine Validitätsprüfung aus fachdidaktischer Sicht. *Journal für Mathematik-Didaktik*, *35*(1), 7–48. https://doi.org/10.1007/s13138-013-0060-3.

Meinhardt, C. (2018). *Entwicklung und Validierung eines Testinstruments zu Selbstwirksamkeitserwartungen von (angehenden) Physiklehrkräften in physikdidaktischen Handlungsfeldern*. Berlin: Logos.

Meinhardt, C., Rabe, T., & Krey, O. (2016). Selbstwirksamkeitserwartungen in physikdidaktischen Handlungsfeldern. Skalendokumentation. http://www.pedocs.de/frontdoor.php?source_opus=11818

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749.

Miller, G. A. (1956). The magical number of seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*, 81–97.

Moosbrugger, H., & Kelava, A. (Hrsg.). (2008). *Testtheorie und Fragebogenkonstruktion*. Berlin, Heidelberg, New York: Springer.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251) https://doi.org/10.1126/science.aac4716.

Pruski, L. A., Blanco, S. L., Riggs, R. A., Grimes, K. K., Fordtran, C. W., Barbola, G. M., et al. (2013). Construct validation of the self-efficacy teaching and knowledge instrument for science teachers-revised (SETAKIST-R): lessons learned. *Journal of Science Teacher Education*, *24*(7), 1133–1156. https://doi.org/10.1007/s10972-013-9351-2.

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

Reinhold, P., Riese, J., & Gramzow, Y. (2017). Fachdidaktisches Wissen im Lehramtsstudium. In H. Fischler & E. Sumfleth (Hrsg.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik* (S. 39–56). Berlin: Logos.

Richter, D., Kunter, M., Lüdtke, O., Klusmann, U., & Baumert, J. (2011). Soziale Unterstützung beim Berufseinstieg ins Lehramt. *Zeitschrift für Erziehungswissenschaft*, *14*(1), 35–59. https://doi.org/10.1007/s11618-011-0173-8.

Riggs, I. M., & Enochs, L. G. (1990). Toward the development of an elementary teacher's science teaching efficacy belief instrument. *Science Education*, *74*(6), 625–637.

Roberts, J. K., & Henson, R. K. (2000). *Self-efficacy teaching and knowledge instrument for science teachers (SETAKIST): a proposal for a new efficacy instrument*. Annual Meeting of the Mid-South Educational Research Association, Bowling Green, November 17–19, 2000.

Sandmann, A. (2014). Lautes Denken – die Analyse von Denk-, Lern- und Problemlöseprozessen. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 179–188). Berlin, Heidelberg: Springer Spektrum.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology*, *29*, 65–88.

Schmiemann, P., & Lücken, M. (2014). Validität – Misst mein Test, was er soll? In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 107–118). Berlin, Heidelberg: Springer Spektrum.

Schmitz, G. S., & Schwarzer, R. (2000). Selbstwirksamkeitserwartung von Lehrern: Längsschnittbefunde mit einem neuen Instrument. *Zeitschrift Für Pädagogische Psychologie*, *14*(1), 12–25. https://doi.org/10.1024//1010-0652.14.1.12.

Schwarzer, R., & Jerusalem, M. (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin: Humboldt-Universität.

Shroyer, G., Riggs, I., & Enochs, L. (2014). Measurement of science teachers' efficacy beliefs. In R. Evans, J. Luft, C. Czerniak & C. Pea (Hrsg.), *The role of science teachers' beliefs in international classrooms. From teacher actions to student learning* (S. 103–118). Rotterdam: Sense Publishers.

Skinner, E. A., Chapman, M., & Baltes, P. B. (1988). Control, means-ends, and agency beliefs: a new conceptualization and its measurement during childhood. *Journal of Personality and Social Psychology*, *54*(1), 117–133.

Tschannen-Moran, M., Woolfolk Hoy, A., & Hoy, W. K. (1998). Teacher efficacy: its meaning and measure. *Review of Educational Research*, *68*(2), 202–248. https://doi.org/10.3102/00346543068002202.

Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften. Zusammenhangsanalysen zwischen Lehrerkompetenz und Lehrerperformanz*. Berlin: Logos.

Vogelsang, C., & Cauet, E. (2017). Wie valide sind Professionswissenstests? – Zum Zusammenhang von erfasstem Wissen, Unterrichtshandeln und Unterrichtserfolg. In H. Fischler & E. Sumfleth (Hrsg.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik* (S. 77–96). Berlin: Logos.

Wallace, C. S. (2014). Overview of the role of teacher beliefs in science education. In R. Evans, J. Luft, C. Czerniak & C. Pea (Hrsg.), *The role of science teachers' beliefs in international classrooms. From teacher actions to student learning* (S. 17–31). Rotterdam: Sense Publishers.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: an evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237.

Woolfolk Hoy, A., & Davis, H. A. (2006). Teacher self-efficacy and its influence on the achievement of adolescents. In F. Pajares & T. Urdan (Hrsg.), *Self-efficacy beliefs of adolescents* (S. 117–137). Greenwich: Information Age Publishing.

Woolfolk Hoy, A., & Spero, R. B. (2005). Changes in teacher efficacy during the early years of teaching: a comparison of four measures. *Teaching and Teacher Education*, *21*(4), 343–356. https://doi.org/10.1016/j.tate.2005.01.007.

Woolfolk Hoy, A., Hoy, W. K., & Davis, H. A. (2009). Teachers' self-efficacy beliefs. In K. R. Wentzel & A. Wigfield (Hrsg.), *Handbook of motivation at school* (S. 627–653). New York: Routledge.