Unterrichtswissenschaft (2024) 52:135–161
https://doi.org/10.1007/s42010-024-00195-4

ALLGEMEINER TEIL

Reflexionsfähigkeit von Physiklehramtsstudierenden: Ein Online-Self-Assessment mit Feedback

Anna Weißbach · Christoph Kulgemeyer

Eingegangen: 1. Juni 2023 / Überarbeitet: 16. November 2023 / Angenommen: 6. Januar 2024 / Online publiziert: 8. Februar 2024

Zusammenfassung Unterrichtsreflexion ist von zentraler Bedeutung für die Verbesserung von Unterricht und die Professionalisierung von Lehrkräften. Die Reflexionsfähigkeit Studierender ist allerdings häufig nur schwach ausgeprägt. In diesem Beitrag wird ein Testinstrument mit Assessment-Feedback vorgestellt, das Studierenden eine Selbsteinschätzung ihrer Reflexionsfähigkeit ermöglicht und das Ziel hat, Assessment-Feedback als Ausgangspunkt für Förderung zu nutzen. Dazu wird die Interpretation der Testwerte als Maß für die Reflexionsfähigkeit argumentbasiert validiert. In Teilstudie 1 (inhaltliche Validität) wird die Passung zwischen Testinstrument und theoretischer Modellierung des Konstrukts diskutiert. In Teilstudie 2 wird die kognitive Validität mit Think-Aloud-Protokollen der Testbearbeitung von $N = 7$ Studierenden untersucht. In Teilstudie 3 (externe Validität) führen $N = 39$ Studierende zusätzlich einen etablierten Reflexionsperformanztest durch, der es ermöglicht, Reflexionen in einem authentischeren, aber standardisierten Format zu erfassen. Teilstudie 4 beschäftigt sich mit der Generalisierbarkeit der Ergebnisse über das Bachelor- und Masterstudium. Dazu werden $N = 136$ Bearbeitungen von Studierenden dieser Studienphasen berücksichtigt und in Bezug auf statistische Kenngrößen (z. B. Lösungshäufigkeit) verglichen. In Teilstudie 5 werden zur Untersuchung der konsequentiellen Validität $N = 12$ Studierende in leitfadengestützten Interviews zu ihrer Interpretation des Assessment-Feedbacks befragt. Insgesamt können starke Argumente für die Validität der Testwertinterpretation und die Eignung des Instruments gefunden werden. Geringe Einschränkungen liegen bezüglich der inhaltlichen Validität in der Beschränkung auf den Themenbereich Mechanik und bezüglich der konsequentiellen Validität in der Vorstellung weniger Studierender, das Testinstrument biete Beispielaufgaben für universitäre Prüfungen. Das Instrument wird online

✉ Anna Weißbach · Christoph Kulgemeyer
Institut für Didaktik der Naturwissenschaften, Abteilung Physikdidaktik, Universität Bremen, Otto-Hahn-Allee 1, 28359 Bremen, Deutschland
E-Mail: anna.weissbach@uni-bremen.de

frei zur Verfügung gestellt, um insbesondere in Praxisphasen der Lehramtsausbildung genutzt werden zu können.

Schlüsselwörter Reflexion · Physikunterricht · Testinstrument · Assessment-Feedback · Validierung

Reflection on teaching by physics teacher students: an online-Self-Assessment offering feedback

Abstract Reflection on teaching is of particular importance for the development of teaching and teachers' professional skills. However, students' reflection skills are often weakly pronounced, so there is a need for further support. We present a multiple-choice instrument accompanied by assessment feedback for measuring and self-assessing reflection skills. We then validate the interpretation of test scores as a measure for reflection. First, content validity is evaluated by discussing the fit between the test instrument and the modeling of reflection on teaching (Study 1). In Study 2, cognitive validity is examined using think-aloud protocols from a sample of seven students during test completion. In Study 3, external validity is examined. For this purpose, a sample of $N = 39$ students additionally completed a performance test that captures reflections in a more authentic but still standardized setting. Study 4 deals with the generalizability of the test scores across bachelor's and master's degree students. For that purpose, a total of $N = 136$ completed instruments from both degree programs are compared with regard to statistical parameters (e.g., task difficulty). To examine concurrent validity, in Study 5, $N = 12$ semi-structured interviews were conducted to record students' interpretations of the assessment feedback. The results of the studies support the validity of the test score interpretation. Limitations need to be considered concerning the representation of physics lessons, which is restricted to the subject area of mechanics (content validity), and the idea of students that the test instrument offers sample tasks for university examinations (consequential validity).

Keywords Reflection on teaching · Physics instruction · Test instrument · Assessment Feedback · Validity

1 Einleitung

Die Reflexion von Unterricht ist eine Kernaufgabe von Lehrkräften (KMK 2004) und spielt somit eine zentrale Rolle für Lehrkräfte sowie in der Lehramtsausbildung (Klewin et al. 2020; Abels 2011, S. 59). Ihr wird zugesprochen, wesentlich zur Professionalisierung von Lehrkräften beizutragen, indem durch Reflexion professionelles Wissen aus Erfahrungen generiert wird (Carlson et al. 2019; McAlpine et al. 1999). Gleichzeitig wurde im Verlauf der letzten Jahrzehnte wiederholt festgestellt, dass die Reflexionsfähigkeit angehender Lehrkräfte nur schwach ausgeprägt ist. Lehramtsstudierende reflektieren überwiegend auf einem deskriptiven Niveau

(Hatton und Smith 1995) – zumindest sofern nicht durch die Aufgabenstellung explizit Aspekte höherer Niveaus eingefordert werden (Stender et al. 2019) – und ihre Reflexionen sind häufig nicht systematisch (Rothland und Boecker 2015). Der Bedarf an Förderung von Reflexionsfähigkeit wird oft festgestellt und insbesondere mit Praxisphasen verbunden (Kulgemeyer et al. 2021). In der Physikdidaktik wurde dazu ein Performanztest zur Messung der Reflexionsfähigkeit entwickelt, der es ermöglicht, die Reflexionsfähigkeit Physiklehramtsstudierender im Praxissemester in einer möglichst authentischen Peer-Feedback-Situation bei gleichzeitig hoher Standardisierung zu erfassen (Kulgemeyer et al. 2021). Untersuchungen mit diesem Performanztest zeigen, dass das Praxissemester nicht per se zu einer Verbesserung der Reflexionsfähigkeit führt, sondern dazu bestimmte Lerngelegenheiten – insbesondere gemeinsame Reflexionsgespräche mit universitären Mentorierenden – effektiv genutzt werden müssen (Kulgemeyer et al. 2021). Um die der Reflexion zugeschriebenen positiven Beiträge zur Professionalisierung auszunutzen, kann argumentiert werden, dass vor dem Hintergrund der üblicherweise nur gering ausgeprägten Reflexionsfähigkeit der Studierenden Bedarf für deren Förderung in der Lehramtsausbildung besteht – und zwar auch über Praxisphasen hinaus. Um Förderung anzubahnen, ist an die Diagnostik adaptierte Lehre vielversprechend, doch auch der Diagnoseprozess selbst bietet Chancen: Insbesondere in Verbindung mit einem gezielten Assessment-Feedback könnte Diagnostik Anlass für Studierende bieten, Reflexionsfähigkeit zu entwickeln, da solch ein Feedback schließlich potenziell eine realistische Einschätzung des Leistungsvermögens erreichen und positiv auf Selbstwirksamkeit sowie Leistungen wirken kann (z. B. Hattie 2009, S. 173). Regelmäßiges Testen ohne adäquate Umsetzung des Feedbacks gilt im Vergleich als bedeutsam weniger wirksam hinsichtlich einer Leistungsentwicklung (Hattie 2009, S. 178). Voraussetzung zur Förderung ist zunächst die valide Diagnostik von Reflexionsfähigkeit – die in diesem Feld verbreiteten Selbsteinschätzungsbögen werden bezüglich ihrer Validität häufig kritisiert (Holtz und Gnambs 2017).

Ziel der hier vorgestellten Studie ist daher, ein Instrument zur Erfassung der Reflexionsfähigkeit zu entwickeln, das valide Rückmeldungen über die Ausprägung der Fähigkeit erlaubt und Studierenden so eine valide Selbsteinschätzung ihrer Reflexionsfähigkeit ermöglicht. Da der Fokus auf der Fähigkeit, Physikunterricht zu reflektieren, liegt, werden fachlich und fachdidaktisch motivierte Reflexionen im Instrument besonders adressiert. Im Folgenden wird das zu diesem Ziel entwickelte Instrument vorgestellt und es werden zentrale Validitätsargumente bezüglich der Interpretation der Testwerte als Ausdruck der Reflexionsfähigkeit diskutiert. Dazu werden Aspekte der inhaltlichen, kognitiven, externen und konsequentiellen Validität sowie der Generalisierbarkeit berücksichtigt. Das Instrument steht unter (https://www.unterrichtsreflexion.de) allen Interessierten zur freien Verfügung und wird bspw. zum Einsatz in Lerngruppen in Praxisphasen empfohlen.

2 Theoretischer Hintergrund

2.1 Unterrichtsreflexion

(Unterrichts-)Reflexion ist seit geraumer Zeit Gegenstand der Forschung (Abels 2011, S. 59) und ihre Relevanz für die professionelle Weiterentwicklung von Lehrkräften ist unumstritten (von Aufschnaiter et al. 2019a). Dennoch lässt sich in der Literatur kein einheitliches Begriffsverständnis des Reflexionsbegriffs ausmachen (z. B. Clarà 2015; Beauchamp 2006, S. 5). Vielmehr werden unzureichende Definitionen (Hatton und Smith 1995) und die „fast schon inflationär[e]“ (von Aufschnaiter et al. 2019a, S. 145) Verwendung des Begriffs kritisiert.

Die Verwendung des Reflexionsbegriffs geht zurück auf die Arbeiten von Dewey (1910/2002) und Schön (1983). Dewey (1910/2002, S. 13) führte den Begriff der Reflexion als eine bestimmte Art des Denkens ein, die mit dem Ziel verbunden ist, eine vorliegende problembehaftete Situation aufzulösen. Schön (1983, S. 278) unterscheidet verschiedene Formen von Reflexion unter Berücksichtigung des zeitlichen Bezugs zur reflektierten Handlung. Reflection-in-Action beschreibt die Reflexion während des Handelns, Reflection-on-Action die Reflexion im Anschluss an das Handeln. Szogs et al. (2019) benennen zusätzliche Unterscheidungsmerkmale verschiedener Definitionen des Reflexionsbegriffs wie den „Reflexionsinhalt, die Reflexionstiefe, die teilnehmenden Personen, [...] sowie das reflektierte Medium“ (Szogs et al. 2019, S. 318). Umfassende Einigkeit besteht letztlich lediglich darin, dass die verschiedenen Autoren Reflexion als bestimmte Form des Denkens beschreiben (von Aufschnaiter et al. 2019b, Aeppli und Lötscher 2016).

Der hier vorgestellten Untersuchung liegt ein auf den Unterricht fokussiertes und pragmatisches Verständnis des Reflexionsbegriffs zugrunde. Unterrichtsreflexion wird verstanden als die „theoriegeleitete Analyse von Unterricht mit dem Ziel der Verbesserung der Unterrichtsqualität und der Entwicklung der Professionalität von Lehrkräften“ (Kempin et al. 2020). Diese Definition inkludiert also sowohl die Reflexion des eigenen Unterrichts (Selbstreflexion) als auch von beobachtetem Unterricht (Fremdreflexion) und ist abgesehen von der Einschränkung der zu reflektierenden Gegenstände auf solche mit Bezug zum beobachteten bzw. durchgeführten Unterricht offen gefasst. So ist bspw. nicht festgelegt, welche Personen an der Reflexion teilnehmen, in welcher Form der reflektierte Unterricht dargestellt ist (reale Beobachtung, Unterrichtsvideos, schriftliche Vignetten, etc.) und in welcher Form die Reflexion durchgeführt wird (verbal, schriftlich, etc.). Insbesondere gehen Reflexionen in diesem Verständnis nicht zwangsläufig mit dem Ziel einher, „an der eigenen Professionalität zu arbeiten“ (von Aufschnaiter et al. 2019b, S. 148), sondern umfassen z. B. auch die Analyse von beobachtetem Unterricht, mit dem Ziel, diesen zu verbessern bzw. die entsprechende Lehrkraft bei ihrer Professionalisierung zu unterstützen.

2.2 Modellierung des Reflexionsbegriffs

In der Literatur lassen sich verschiedene Modellierungen zum Reflexionsbegriff finden, die jeweils Bestandteile bzw. Stufen von Reflexionsprozessen beinhalten und als Grundlage für die Bewertung von Reflexionen dienen können (einen Überblick über 40 Modellierungen des Reflexionsbegriffs liefern Poldner et al. 2014). Ausgangspunkt für eine Unterrichtsreflexion ist stets eine erlebte oder beobachtete Situation (z. B. von Aufschnaiter et al. 2019b; Korthagen 1985). Diese Situation wird zunächst rekapituliert (z. B. „looking back", Korthagen 1985, S. 12) bzw. beschrieben (z. B. durch „die Beschreibung der Schlüsselsequenzen des Unterrichts", Windt und Lenske 2016, S. 285). Anschließend erfolgt die Analyse bzw. Interpretation der Situation, die – abhängig vom konkreten Modell – z. B. die Deutung der Beobachtung und Identifikation von Ursachen (von Aufschnaiter et al. 2019b; Aeppli und Lötscher 2016) bzw. die begründete Bewertung der Situation (Windt et al. 2017) umfasst. Darauf aufbauend werden alternative Handlungsoptionen entwickelt sowie Konsequenzen für die Professionalisierung formuliert (z. B. von Aufschnaiter et al. 2019b; Windt et al. 2017; Korthagen 1985).

Grundlegend unterschieden werden kann allerdings zwischen Modellen, die das Erleben der reflektierten Situation als Teil des Reflexionsprozesses verstehen und/oder im Sinne eines Reflexionszyklus mit der Erprobung von entwickelten Handlungsoptionen abschließen bzw. einen neuen Zyklus beginnen (z. B. von Aufschnaiter et al. 2019b; Aeppli und Lötscher 2016; Korthagen 1985) und solchen, die ausschließlich die analytischen Tätigkeiten umfassen (z. B. Windt et al. 2017). Auch bestehen Unterschiede in Bezug darauf, ob die Elemente als hierarchische oder nicht geordnete Stufen zu verstehen sind (von Aufschnaiter et al. 2019b; Plöger et al. 2015; Poldner et al. 2014).

Basierend auf den Modellen von Plöger et al. (2015) und Windt und Lenske (2016) entwickelten Nowak et al. (2019) ein hierarchisches Modell zur Reflexion von Physikunterricht (s. Abb. 1). Dieses Modell ermöglicht es, (Teile von) Reflexionen nicht nur in Bezug auf die Art der Aussage zu klassifizieren (das „Element der

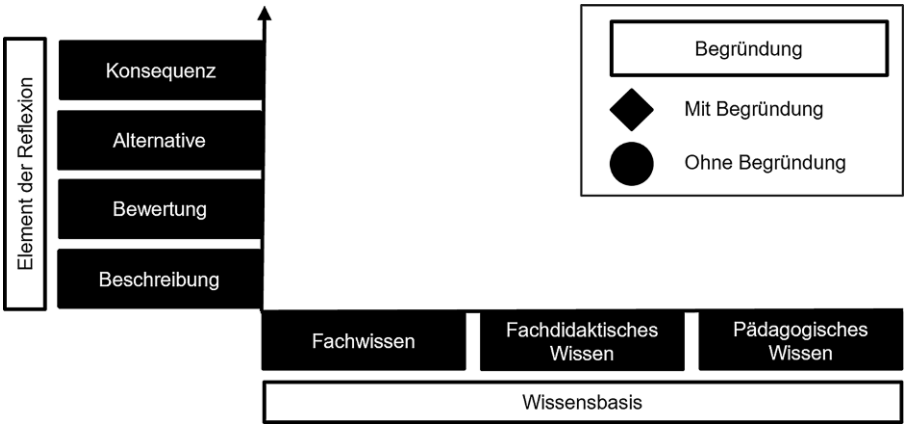

Abb. 1 Modell zur Reflexion von Physikunterricht. (Nach Nowak et al. (2019))

Reflexion“: Wird eine Unterrichtssituation beschrieben oder bewertet, werden alternative Handlungsoptionen vorgeschlagen oder Konsequenzen für den Folgeunterricht und/oder die Professionalisierung der Lehrkraft formuliert?), sondern auch die einer Aussage zugrundeliegende Wissensbasis zuzuordnen (Wird für diese Aussage fachliches, fachdidaktisches oder pädagogisches Wissen benötigt?). Hervorzuheben ist außerdem, dass dieses Modell Begründungen nicht als Elemente der Reflexion auffasst, sondern das Vorliegen einer Begründung quer zu diesen eingestuft wird. Somit werden Begründungen nicht nur für Bewertungen (Windt et al. 2017) oder Alternativen (Plöger et al. 2015) berücksichtigt, sondern können grundsätzlich zu allen Reflexionselementen geliefert werden (Nowak et al. 2019).

Das Professionswissen stellt – wie z. B. im Modell der Kompetenz als Kontinuum zusammengefasst (Blömeke et al. 2015) – eine zentrale Disposition für Reflexionsperformanz, d. h. die beobachtbare Ausführung einer Reflexion als Ausdruck der Reflexionskompetenz, dar: Um Unterricht theoretisch fundiert einzuschätzen, braucht es Wissensbestände, auf die dafür zurückgegriffen werden kann. Ausgehend von Shulmans (1987) Taxonomie des professionellen Wissens hat sich die Berücksichtigung und Unterscheidung der drei Facetten fachliches Wissen, fachdidaktisches Wissen und pädagogisches Wissen für das Professionswissen etabliert (Baumert und Kunter 2006), die so auch im Reflexionsmodell von Nowak et al. (2019) in der Facette *Wissensbasis* abgebildet ist. von Aufschnaiter et al. (2019b) benennen in ihrer Adaption des Kompetenzmodells von Blömeke et al. (2015) zusätzliche Dispositionen, wie Einstellungen und Überzeugungen zur Relevanz von Reflexion sowie die Fähigkeit und Bereitschaft, eine Reflexion durchzuführen; das Modell in Abb. 1 fokussiert auf die kognitiven Facetten.

Im Gegensatz zu Modellierungen des Reflexionsbegriffs, die das Erleben der Situation bzw. der „Auslöser für den Reflexionsanlass“ (Aeppli und Lötscher 2016, S. 88) als Teil des Reflexionsprozesses verstehen, ist die Anforderung, zentrale Situationen im Unterricht zu erkennen, zwar notwendig für eine Reflexion, dem Reflexionsprozess nach dem von Nowak et al. (2019) entwickelten Modell allerdings vorgeschaltet und damit kein inhärenter Teil der Reflexionsfähigkeit (Wöhlke 2020, S. 31). Diese Anforderung entspricht dem als „Noticing“ (van Es und Sherin 2008, Seidel et al. 2010) bzw. „Wahrnehmung oder Informationsbeschaffung“ (Wöhlke 2020, S. 31) bezeichneten Bestandteil der professionellen Unterrichtswahrnehmung, welcher die „Identifikation relevanter Situationen und Ereignisse im Unterrichtsgeschehen“ (Seidel et al. 2010) umfasst.

2.3 Assessment-Feedback zur Unterstützung von Lernprozessen

Self-Assessment wird generell als lernförderlich betrachtet (z. B. Max et al. 2022), aber Lernende neigen dazu, ihren Fähigkeitsstand zu überschätzen, was das Self-Assessment ineffektiv machen könnte (z. B. Nowell und Alston 2007). Das nachfolgend vorgestellte Instrument greift deshalb auf Assessment-Feedback als Werkzeug zur Regulation der Selbsteinschätzung und Förderung von Reflexionsfähigkeit zurück. Assessment-Feedback ist eine spezifische Form generellen Feedbacks. Im Kontext des (universitären) Lernens kann Feedback verstanden werden als von außen (z. B. durch Dozierende) bereitgestellte Informationen über die individuelle Performanz,

das Erreichen von Lernzielen o. ä. (vgl. Hattie und Timperley 2007). Hattie (2009, S. 175ff.) beschreibt, dass effektives Feedback Informationen über die Ziele des Lernprozesses gibt („Feed Up“) und den aktuellen Entwicklungsstand, der eine Selbsteinschätzung ermöglicht („Feed Back“) sowie Hinweise für nächste Schritte oder Ziele („Feed Forward“) umfasst. Die Unterstützung von Lernprozessen durch Feedback ist besonders dann effektiv, wenn es sachliche Informationen über die Aufgabe bzw. deren Bearbeitung beinhaltet und nicht auf persönliche Eigenschaften eingegangen oder z. B. gelobt wird (Hattie 2009, S. 177). Sippel (2009) spricht sich für eine Stärkung „lernförderlicher Assessments“ (Sippel 2009, S. 2) in der universitären Lehre aus. Sie betont die Bedeutung von Assessment-Feedback (d. h. Feedback zu Ergebnissen eines durchgeführten Tests) mit Verbesserungshinweisen dafür und für eine transparente Bewertung.

2.4 Validitätskonzept

Zur Untersuchung der Validität wird auf den argumentbasierten Ansatz nach Kane (2013, 1992) zurückgegriffen. Demnach wird die Validität nicht als Eigenschaft des Testinstruments verstanden (wie z. B. von Bühner 2011, S. 61), sondern als Kriterium für die Gültigkeit einer bestimmten Interpretation der Testergebnisse vor dem Hintergrund einer bestimmten Nutzungsabsicht (Kane 2013, S. 16). Da das Testinstrument in Verbindung mit dem Assessment-Feedback dafür konzipiert wurde, Studierenden eine fundierte Rückmeldung über ihre Reflexionsfähigkeit zu geben, dient die Validierung dazu, sicherzustellen, dass auch die Schlussfolgerungen bzgl. ihrer Reflexionsfähigkeit zu Testinstrument und Assessment-Feedback passen. Die Argumentation für Validität der Testwertinterpretation kann dabei auf ein breites Spektrum verschiedener Evidenzen gestützt werden. Um diese Breite angemessen zu berücksichtigen, wird für die Sammlung (empirischer) Evidenzen auf Messicks (1995) Unterscheidung von sechs Validitätskriterien zurückgegriffen, die aus seiner Perspektive im Rahmen einer umfassenden Untersuchung der Validität berücksichtigt werden sollen:

1. Die *inhaltliche Validität* beschreibt die Güte der Abbildung des zu untersuchenden Konstrukts durch die Items im Testinstrument (Messick 1995, S. 745). Als Kriterien für die inhaltliche Validität benennen Hartig et al. (2012), dass die Items insgesamt „eine repräsentative Auswahl der interessierenden Gesamtheit möglicher Items“ (ebd., S. 152) darstellen, d. h. die Items alle relevanten Inhalte des Konstrukts abbilden, ohne irrelevante Inhalte abzubilden und verschiedene Testinhalte durch ein dem Konstrukt angemessenes Verhältnis an Items abgebildet werden.
2. Die Untersuchung der *kognitiven Validität* („substantive aspect of construct validity“, Messick 1995, S. 745) zielt auf eine Untersuchung der kognitiven Prozesse während der Testbearbeitung ab, um konstuktrelevante Überlegungen bei der Bearbeitung sicherzustellen.
3. Die *strukturelle Validität* beschäftigt sich mit der Passung zwischen der Struktur des untersuchten Konstrukts und dem gewählten Messmodell, also der gewählten Anzahl an Variablen bzw. Dimensionen und dem Bewertungssystem (Messick 1995, S. 745f.)

4. Die *externe Validität* bezieht sich auf Zusammenhänge zwischen den Testergebnissen und externen Kriterien wie Ergebnissen der Bearbeitung anderer Testinstrumente. Messick (1995, S. 746) unterscheidet zwischen konvergenter Validität, wenn hohe Zusammenhänge zwischen verschiedenen Testergebnissen erwartet werden und diskriminanter Validität, wenn keine oder nur eine geringe Korrelation zu erwarten ist (s. auch Schmiemann und Lücken 2014).
5. Die Untersuchung der *Generalisierbarkeit* soll sicherstellen, dass die Interpretation der Testwerte unabhängig von der konkreten Itemauswahl, der Stichprobe oder der auswertenden Person gültig ist (Messick 1995, S. 746).
6. Die *konsequentielle Validität* bezieht sich auf Konsequenzen, die sich aus der Testwertinterpretation ergeben; insbesondere sollen negative Konsequenzen, die aus mangelnder Validität resultieren, vermieden werden (Messick 1995, S. 746).

Aufgrund des Aufwands einer extensiven Prüfung aller Validitätsaspekte, wird empfohlen, abhängig von der konkreten Interpretation der Testwerte relevante Aspekte auszuwählen (Kane 2013, S. 34). Die verschiedenen Argumente für die Validität einer Testwertinterpretation sollen dann diskutiert und zu einer kohärenten und durch Evidenz gestützten Validitätsargumentation zusammengefügt werden (Kane 2013, S. 3). Zur praktischen Anwendung dieses argumentbasierten Validitätskonzepts schlägt Dickmann (2016, S. 67) vor, sich an den Übersetzungsschritten zu orientieren, die notwendig sind, um der Ausprägung eines Konstrukts Testwerte zuweisen zu können:

1. Das Konstrukt von Interesse (hier die Fähigkeit, Physikunterricht zu reflektieren) wird dazu zunächst in Aufgaben übersetzt.
2. Diese Aufgaben führen beim Einsatz des Testinstruments zu beobachtbarem Verhalten der Teilnehmenden.
3. Das Verhalten während der Testbearbeitung wird dann in Testwerte überführt.

Das Ziel des hier vorgestellten Instruments, Rückmeldungen über Reflexionsfähigkeit bereitzustellen, erfordert außerdem eine sorgfältige Evaluation der Konsequenzen, die sich für Studierende aus der Testbearbeitung ergeben. Dazu kann ein weiterer Übersetzungsschritt betrachtet werden:

4. Aus den Testwerten ziehen Studierende Konsequenzen mit Bezug auf ihre individuelle Professionalisierung.

Können alle Übersetzungsschritte als gelungen evaluiert werden, kann davon ausgegangen werden, dass die Testwerte valider Ausdruck des untersuchten Konstrukts sind (Bartels 2018, S. 20; Dickmann 2016, S. 67).

3 Methode

3.1 Das entwickelte Multiple-Choice-Testinstrument mit Assessment-Feedback zur Reflexionsfähigkeit

3.1.1 Testinstrument

Kempin et al. (2018) haben einen Performanztest zur Messung der Reflexionsfähigkeit von Physiklehramtsstudierenden entwickelt. In diesem Instrument werden Studierende in eine fiktive kollegiale Beratungs- und Reflexionssituation versetzt und geben Robert, einem fiktiven Mitpraktikanten, Rückmeldungen zu seinem Unterricht. Die Rückmeldungen werden direkt in ein Mikrofon gesprochen, mitgeschnitten und nach einem Kodiermanual ausgewertet. Basierend auf diesem Performanztest von Kempin et al. (2018) wurde ein aus Multiple-Choice-(MC-)Aufgaben bestehendes Instrument entwickelt, sodass eine automatisierte Auswertung ermöglicht wird. Auch dieses Instrument simuliert die fiktive kollegiale Beratungs- und Reflexionssituation: Robert, ein fiktiver Mitpraktikant im Praxissemester, bittet die Studierenden um Feedback zu einer von ihm durchgeführten Doppelstunde Physik-

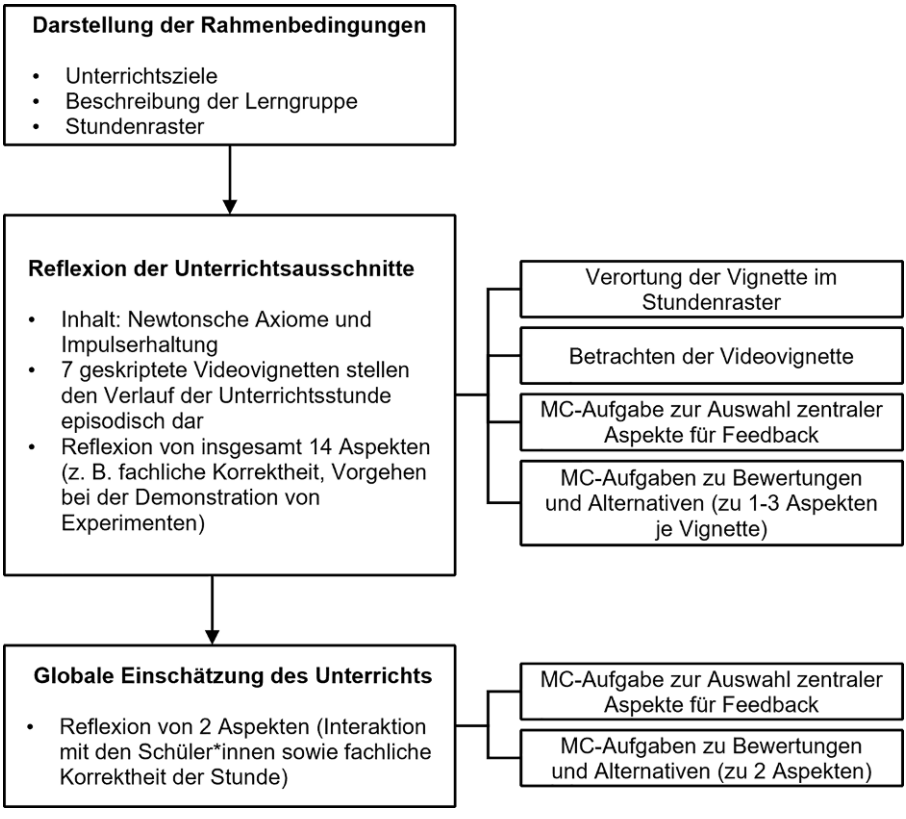

Abb. 2 Schematische Darstellung des Aufbaus des Testinstruments

unterricht. Der Unterricht wird mit Hilfe von sieben geskripteten Videovignetten zugänglich gemacht, die jeweils in den Stundenverlauf eingeordnet werden, sodass gezielt (problematische) Situationen dargestellt werden können, die von den Studierenden reflektiert werden sollen. Fachlicher Inhalt der Doppelstunde sind die Newtonschen Axiome sowie die Impulserhaltung, zentrale Inhalte des Themenbereichs Mechanik. Nach jeder gezeigten Unterrichtsvignette werden die Studierenden aufgefordert, eine Rückmeldung zu geben. Die Erfassung der Reflexionen erfolgt mit MC-Aufgaben. Nach jeder Videovignette sowie zur globalen Betrachtung der Unterrichtsstunde werden die Studierenden aufgefordert, aus einer Auswahl an Reflexionsanlässen diejenigen auszuwählen, zu denen Robert eine (positive oder kritische) Rückmeldung erhalten sollte. Anschließend wird bzw. werden zu jeweils einem bis drei vorgegebenen Aspekt(en) je eine Aufgabe zur Bewertung des Aspekts und zum Vorschlagen von alternativen Handlungsoptionen bearbeitet – die zuvor getroffene Auswahl relevanter Aspekte hat dabei keinen Einfluss auf die nachfolgend zu bearbeitenden Aufgaben; das Testinstrument weist eine lineare Struktur auf, sodass alle Teilnehmenden dieselben Aufgaben bearbeiten. Ein Überblick über den Aufbau des Instruments ist in Abb. 2 dargestellt. Insgesamt werden im Verlauf des Tests so 16 Aspekte näher betrachtet (insgesamt 14 Aspekte direkt zu den sieben Videovignetten und zwei Aspekte als globale Rückmeldung zur Unterrichtsstunde, darunter u. a. die fachliche Korrektheit formulierter Merksätze, die Verwendung von Fachbegriffen und der stereotypische Umgang mit den Schüler*innen). Die jeweils erste MC-Aufgabe zur Auswahl der Aspekte, zu denen Robert eine Rückmeldung erhalten sollte, dient dazu, die Studierenden zur selbstständigen Reflexion des beobachteten Unterrichts anzuregen, bevor sie die MC-Items zu Bewertungen und Alternativen anwählen können. Sie resultiert aus der Anforderung des ursprünglich offenen Antwortformats im Performanztest, relevante Aspekte zu erkennen, um diese anschließend reflektieren zu können. Diese erste MC-Aufgabe zu jedem Video entspricht einem der Reflexion vorgeschalteten Noticing und wird für die Berechnung der Testwerte nicht berücksichtigt.

Um die MC-Items möglichst authentisch zu halten, wurden diese aus den realen Reflexionen entwickelt, die Studierende während der Bearbeitung des Performanztests von Kempin et al. (2018) zu denselben Videovignetten vorgenommen haben. Je Aufgabe liegen vier bis maximal acht (meist fünf bis sieben) MC-Items vor, von denen jeweils die Hälfte der Optionen (bzw. bei ungraden Itemzahlen etwa die Hälfte der Optionen) korrekte und relevante Rückmeldungen darstellen. Die Beurteilung der Reflexionsfähigkeit erfolgt, indem die richtig angewählten MC-Items aufsummiert und die maximal erreichbare Punktezahl je Aufgabe auf eins normiert werden. Dadurch wird eine unterschiedliche Gewichtung der verschiedenen reflektierten Aspekte bzw. Reflexionsstufen vermieden, da diese theoretisch nicht gerechtfertigt werden könnte. Zudem wird so eine damit einhergehende Beeinflussung der Testwerte aufgrund der unterschiedlichen Itemzahlen je Aufgabe verhindert. So können insgesamt 32 Punkte für richtig angewählte Bewertungen und Alternativen erreicht werden. Die Beurteilung der Korrektheit einer Antwortoption basiert auf dem Kodiermanual von Kempin et al. (2018), welches das relevante Feedback zu den jeweiligen Unterrichtsvignetten und den darin dargestellten Problemen umfasst und umfangreichen Validierungsstudien unterzogen wurde (Expertenbefragung, s. Teilstudie 1): Würde

die angewählte Aussage nach diesem Manual kodiert werden, stellt sie einen richtigen und wichtigen Hinweis dar und ist als korrekt einzuordnen. So wird bspw. zur Präsentation eines Demonstrationsexperiments in einer Unterrichtsvignette im Kodiermanual als relevanter Aspekt festgehalten: „Robert (Lehrer) vermittelt ein falsches Bild von Messunsicherheiten." Auf die Frage im MC-Instrument „Welches Feedback möchten Sie Robert zu seinem Umgang mit Messunsicherheiten am ehesten geben?" wird auf dieser Grundlage das Item „Es ist nicht gut, dass du sagst, du könntest ‚exakt' oder ‚genau' messen, da du so die Schülervorstellung beförderst, man könnte Größen ohne Messunsicherheiten bestimmen." als korrekt bewertet. Die Aussage „Es ist gut, dass du darauf verzichtest, auch auf Messunsicherheiten einzugehen, um die Komplexität des Experiments überschaubar zu halten." wird hingegen als falsch bewertet, da der problematische Umgang mit Messunsicherheiten nicht adressiert wird. Folgerichtig wird bspw. die Handlungsalternative „Es wäre besser, beide Wagen zu wiegen, da du im Vergleich der Massen auf Messunsicherheiten bzw. Ungenauigkeiten eingehen kannst." als korrekt bewertet.

Das Testinstrument ist vollständig in der freien Software *LimeSurvey* umgesetzt und kann jederzeit selbstständig online bearbeitet werden. Die Testwerte weisen eine hohe interne Konsistenz von $\alpha_{\text{Cronbach}} = 0{,}95$ sowie eine angemessene durchschnittliche Lösungshäufigkeit von 0,66 und eine durchschnittlich hohe Trennschärfe von 0,56 (Döring und Bortz 2016, S. 443, 478) auf. Alle Testaufgaben können online unter www.unterrichtsreflexion.de eingesehen werden.

3.1.2 Assessment-Feedback

Nach Bearbeitung des Testinstruments erhalten Studierende eine Rückmeldung über ihr individuelles Abschneiden. Die Auswertung der Bearbeitung und Erstellung der Rückmeldungen erfolgt automatisiert mit einem in R umgesetzten Skript, das die statistische Auswertung vornimmt und den Feedbackbogen erstellt, der dann auf einem Server zum Download bereitstellt wird. Zur Gewährleistung der Anonymität kann die Rückmeldung von Studierenden mit Hilfe eines persönlichen Codes online abgerufen werden. Wenn ein Kurs oder eine Person die Bearbeitung vollständig abgeschlossen hat, genügt ein Hinweis an die online genannten Kontakte, um das Assessment-Feedback online abrufen zu können. Diese Rückmeldung bei der Testleitung ist derzeit weiterhin notwendig, damit Kurse vollständig im Verbund teilnehmen können und um auch Dozierenden Rückmeldungen zu ihrem Gesamtkurs zu geben. Das Assessment-Feedback beinhaltet neben einer der Transparenz dienenden Klärung des Begriffs der Unterrichtsreflexion und dessen Umsetzung im Testinstrument die erzielten Ergebnisse (Gesamtergebnis und Teilergebnisse für Bewertungen, Alternativen und Fehlvorstellungen) zur Selbsteinschätzung und -evaluation im Sinne des „Feed Back" nach Hattie (2009, S. 176f.) und den Hinweis auf ein Fördermaterial zur vertieften Beschäftigung mit Unterrichtsreflexion (als „Feed Forward", Hattie 2009, S. 176f.), in welchem den Studierenden Leitfragen zur Unterrichtsreflexion zur Verfügung gestellt werden und drei weitere Szenen selbstständig, aber gemäß dem Scaffolding-Prinzip angeleitet mit Hinweisen und Musterlösungen, reflektiert werden können. Die erzielten Ergebnisse werden als Prozentwerte dargestellt und mit Hilfe von Boxplots in eine Vergleichsgruppe (bestehend aus anderen

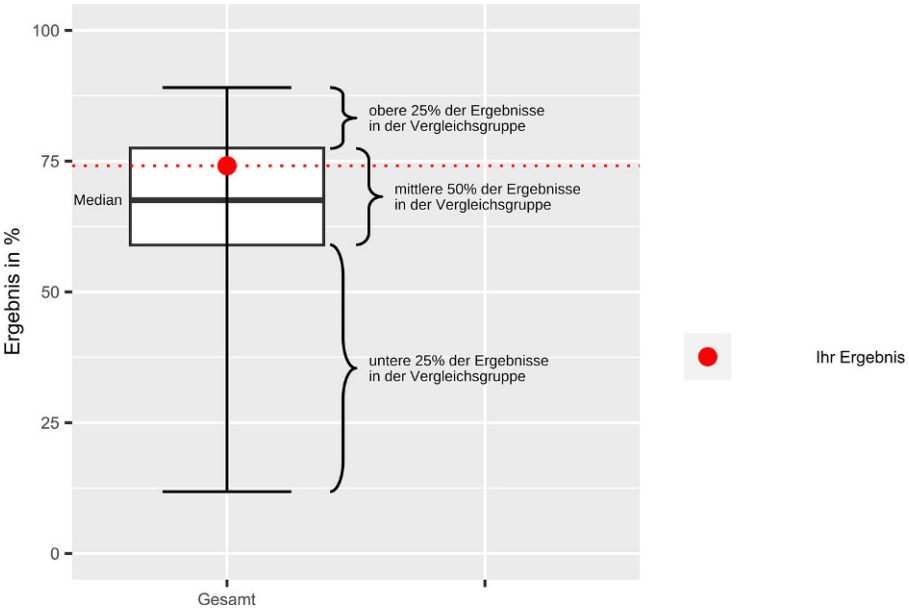

Abb. 3 Auszug aus dem Assessment-Feedback: Boxplot zur Rückmeldung des Gesamtergebnisses

Studierenden, die das Testinstrument bearbeitet haben) eingeordnet. Ein Boxplot zur Rückmeldung eines Gesamtergebnisses ist exemplarisch in Abb. 3 dargestellt.

3.2 Ziele der Untersuchung verschiedener Validitätsaspekte

Die Untersuchung der Validität erfolgt angelehnt an das von Dickmann (2016, S. 67) umgesetzte Vorgehen entlang der verschiedenen Übersetzungsschritte ausgehend vom interessierenden Konstrukt hin zu den Testwerten, wobei als zusätzlicher Übersetzungsschritt die auf Grundlage des Assessment-Feedbacks gezogenen Schlussfolgerungen der Studierenden berücksichtigt werden (s. Abb. 4).

Die Evaluation dieser Übersetzungsschritte erfolgt mit unterschiedlichen Teilstudien, die im Folgenden vorgestellt und den einzelnen Validitätsaspekten zugeordnet werden. Von einer gelungenen Übersetzung kann ausgegangen werden, wenn theoretisch nachvollziehbare Anforderungen erfüllt sind, die in den jeweiligen Teilstudien überprüft werden. In allen Teilstudien werden Gelegenheitsstichproben genutzt. Mindestens werden Personen aus verschiedenen universitären Veranstaltungen und Studienjahren rekrutiert, um Kohorteneffekte zu minimieren. In Teilstudie 4 werden

Abb. 4 Übersetzungsschritte bei der Messung und Rückmeldung der Fähigkeit, Physikunterricht zu Reflexionen in Anlehnung an Dickmann (2016, S. 67) mit Zuordnung der Teilstudien (*TS*) zur Validierung

Personen aus verschiedenen Universitäten rekrutiert, um curricular bedingte Standorteffekte zu minimieren.

3.2.1 Teilstudie 1 (inhaltliche Validität): Übersetzung des Konstrukts Unterrichtsreflexion in das Testinstrument

Die Prüfung, inwiefern die Aufgaben im Testinstrument das Konstrukt der Reflexion von Physikunterricht angemessen abbilden, ist der inhaltlichen Validität zuzuordnen. Dazu werden theoretische Betrachtungen von Aufbau und Inhalt des Testinstruments zu folgenden Anforderungen angestellt:

- A1.1: Die Aufgaben im Testinstrument bilden das Konstrukt der Reflexion von Physikunterricht umfassend ab.
- A1.2: Die als richtig eingestuften Antworten bilden inhaltlich relevante Rückmeldungen ab.

3.2.2 Teilstudie 2 (kognitive Validität): Übersetzung der Aufgaben im Testinstrument in Verhalten

Für das Verhalten der Studierenden während der Testbearbeitung spielt die kognitive Validität eine zentrale Rolle. Für eine valide Testwertinterpretation muss sichergestellt sein, dass während der Bearbeitung Reflexionsprozesse bei den Studierenden ausgelöst werden und inhaltliche Überlegungen die Bearbeitung der Aufgaben bestimmen, was durch folgende Anforderungen repräsentiert wird:

- A2.1: Die Studierenden verstehen die MC-Aufgaben wie intendiert.
- A2.2: Die Überlegungen der Studierenden während der Testbearbeitung passen zur Aufgabenstellung und beziehen sich auf die Qualität des Unterrichts bzw. auf die Professionalisierung als Lehrperson.

Dazu wurde eine Think-Aloud-Studie mit $N = 7$ Physiklehramtsstudierenden im Bachelor- ($N_{BA} = 3$) bzw. Masterstudium ($N_{MA} = 4$) durchgeführt. Die Studierenden waren aufgefordert, alle ihre Überlegungen während der Testbearbeitung laut auszusprechen. Diese Überlegungen wurden anschließend transkribiert und vor dem Hintergrund der Anforderungen inhaltsanalytisch ausgewertet.

3.2.3 Teilstudien 3 (externe Validität) und 4 (Generalisierbarkeit): Übersetzung des Verhaltens in Testwerte

Die Generierung von Testwerten aus dem Ankreuzverhalten der Studierenden ist Grundlage für die Rückmeldungen, die Studierende nach Bearbeitung des Testinstruments erhalten. Für die valide Interpretation der Testwerte ist bedeutsam, dass sie Zusammenhänge zu anderen Messungen der Reflexionsfähigkeit aufweisen (externe Validität). Dafür wurden von $N = 39$ Studierenden (davon studierten $N_{BA} = 27$ im Bachelor- und $N_{MA} = 12$ im Masterstudium) Bearbeitungen des vorgestellten Testinstruments sowie vom Reflexionsperformanztest erhoben, um folgende Anforderung zu untersuchen:

- A3: Die Testwerte weisen Zusammenhänge zu authentischerem Reflexionshandeln (repräsentiert durch die Bearbeitung des Performanztests) auf.

Auch die Generalisierbarkeit der Testwerte ist bedeutsam, da das Instrument nicht für eine bestimmte Studienphase konzipiert ist, sondern übergreifend einsetzbar sein soll. Von den insgesamt $N = 136$ Bearbeitungen stammen $N_{BA} = 77$ bzw. $N_{MA} = 59$ Bearbeitungen von Physiklehramtsstudierenden im Bachelor- bzw. Masterstudium, was einen Vergleich dieser beiden Gruppen zur Prüfung der nachfolgenden Anforderung an die Testwerte ermöglicht.

- A4: Die statistischen Kennwerte zu den Testwerten (interne Konsistenz, Lösungshäufigkeit und Trennschärfe) sind für die Studienphasen Bachelor und Master annehmbar.

Die Prüfung der statistischen Kennwerte erfolgt für beide Studienphasen getrennt entlang der Einordnungen von Döring und Bortz (2016), da das Instrument für beide Stichproben tauglich sein soll. Demnach ist die interne Konsistenz ab einem Wert von 0,80 als ausreichend (ab 0,90 als hoch) einzustufen (Döring und Bortz 2016, S. 443), Lösungshäufigkeiten sollten in einem Bereich zwischen 0,20 und 0,80 liegen (Döring und Bortz 2016, S. 477) und die Trennschärfen mindestens bei 0,30 (Döring und Bortz 2016, S. 478). Aufgrund des Studienfortschritts des Masterstudierenden, der mit einem größeren Umfang an Praxiserfahrung und absolvierten Lehrveranstaltungen einhergeht, können insgesamt höhere Testwerte für die Masterstudierenden erwartet werden.

3.2.4 Teilstudie 5 (konsequentielle Validität): Übersetzung der Testwerte in Konsequenzen

Die Berücksichtigung der konsequentiellen Validität ist hier von besonderer Bedeutung, da die Testwerte im Rahmen des Assessment-Feedback an Studierende transparent gemacht und als möglicher Ausgangspunkt für Professionalisierungsprozesse eingestuft werden. Im Rahmen einer Interviewstudie ($N = 12$ Bachelor-Physiklehramtsstudierende) werden die Interpretationen des Assessment-Feedback durch die Studierenden erhoben, um Anforderung A5 zu prüfen.

- A5: Die Konsequenzen, die sich für die Studierenden aus dem Assessment-Feedback ergeben, entsprechen dem Konstrukt der Unterrichtsreflexion und Entwicklungsbedarf der Studierenden.

3.2.5 Ergebnisse

Mit den hier dargestellten Teilstudien werden die wichtigsten Validitätsargumente untersucht, denn es werden sowohl alle Übersetzungsschritte abgedeckt, als auch fünf der sechs Validitätsaspekte einbezogen. Der sechste Aspekt, die strukturelle Validität, wird nicht explizit faktorenanalytisch untersucht, da die Fähigkeit, Unterricht zu reflektieren als eine (eindimensionale) Facette der professionellen Handlungskompetenz aufgefasst wird, die im Wechselspiel mit verschiedenen anderen Kompetenzen die professionelle Handlungskompetenz einer Lehrkraft bildet. Als solche umfasst sie wiederum „ein Bündel von [...] geistigen Fähigkeiten" (Frey 2006, S. 31), wie das Bewerten von vielfältigen Unterrichtssituationen oder das Finden adäquater Handlungsalternativen. Die inhaltliche Breite der Reflexionsfähigkeit wird durch das oben dargestellte Reflexionsmodell abgebildet. Dieses zugrundeliegende Verständnis von Unterrichtsreflexion als eindimensionales Konstrukt wird durch die sehr hohe interne Konsistenz des Instruments ($\alpha_{Cronbach} = 0{,}95$ bei einer angemessenen durchschnittlichen Inter-Item-Korrelation von 0,36; Döring und Bortz 2016, S. 481) gestützt.

4 Ergebnisse/Validitätsargumente

4.1 Teilstudie 1: Inhaltliche Validität (Übersetzung des Konstrukts Unterrichtsreflexion in das Testinstrument)

Das entwickelte Testinstrument soll die Fähigkeit Studierender erfassen, Physikunterricht zu reflektieren. Nach Hartig et al. (2012) ist sicherzustellen, dass das Testinstrument (1) das adressierte Konstrukt in allen relevanten Aspekten abbildet, ohne irrelevante Inhalte darzustellen und (2) diese Inhalte in einem angemessenen Verhältnis zueinander abgebildet werden. Die Abbildung des Konstrukts der Reflexion von Physikunterricht ist strukturell am Reflexionsmodell von Nowak et al. (2019) orientiert, nach welchem eine ideale Reflexion vollständig und geordnet in Bezug auf die Reflexionselemente und vorliegenden Begründungen ist. Die erste Facette des Reflexionsmodells beinhaltet die Unterscheidung der den Items zugrundeliegenden Wissensbasen. Aufgrund seines begrenzten Umfangs kann das Testinstrument Unterricht nicht vollumfänglich abbilden. Physikalischer Inhalt der Doppelstunde ist die Mechanik, da diese im Rahmen der universitären Ausbildung früh unterrichtet und das Instrument somit schon in unteren Semestern sinnvoll eingesetzt werden kann. Außerdem hat sich das Fachwissen im Bereich der klassischen Mechanik als guter Prädiktor für das physikalische Fachwissen allgemein herausgestellt (Riese und Reinhold 2008; Friege und Lind 2004). Die 16 Aspekte, die im Verlauf des Testinstruments näher betrachtet werden, lassen sich immer (mindestens) einer der Wissensbasen zuordnen (der Umgang mit Fehlvorstellungen bspw. dem fachdidaktischen Wissen). Alle drei Wissensbasen werden im Testinstrument angesprochen, der Schwerpunkt liegt jedoch auf Fachwissen und fachdidaktischem Wissen (in neun bzw. zwölf Aspekten angesprochen). Dies spiegelt den Fokus des Testinstruments auf die Reflexion von *Physik*unterricht wider. Durch die aufeinanderfolgenden MC-

Aufgaben zu Bewertungen und alternativen Handlungsoptionen (in denen teilweise Konsequenzen als längerfristige Handlungsoptionen zur Professionalisierung bzw. Sicherstellung der Unterrichtsqualität inkludiert sind; z. B. die Empfehlung „Du solltest das Thema Messunsicherheiten für dich selbst noch einmal aufbereiten […].“) werden die Elemente des Reflexionsmodells (zweite Facette) in komprimierter Form durchlaufen. Beschreibungen werden nicht explizit aufgeführt bzw. eingefordert, da in den bewertenden MC-Items jeweils beschreibende Elemente vorhanden sind (z. B. wird im Item „Es ist gut, dass die Schüler*innen zunächst ihre Alltagssprache verwenden können um zu erklären, warum es zur Kollision gekommen ist. […]“ beschrieben, dass Alltagssprache verwendet wird). Somit sind alle Reflexionselemente abgebildet. Zudem beruhen die Distraktoren und Attraktoren auf realen Reflexionen von Studierenden. Dass nicht alle Aufgaben zu Alternativen auch Konsequenzen beinhalten, entspricht also tatsächlichen Reflexionen. Da alle Items Begründungen beinhalten, wird diese Facette des Reflexionsmodells zwar nicht differenziert, aber im Sinne einer vollständigen Reflexion abgebildet (Nowak et al. 2019). Abhängig von dem in der jeweiligen Antwortoption diskutierten Aspekt, beziehen sich auch die Begründungen auf unterschiedliche Wissensbasen und sind bspw. fachlicher (z. B. „Es ist nicht gut, dass du sagst, das Thema ist ‚komplett neu‘, da es einen fachlichen Bezug zwischen den Newtonschen Axiomen und dem Impulsbegriff gibt.“) oder pädagogischer Natur (z. B. „Es ist nicht gut, dass du deinen vorgefertigten Merksatz nutzt und die Merksätze der Schüler*innen nicht einbeziehst, da das den Schüler*innen gegenüber nicht wertschätzend ist.“). Die Aufgaben des Testinstruments bilden das Konstrukt der Reflexion von Physikunterricht vor dem Hintergrund der Modellierung also umfangreich ab. Anforderung A1.1 kann also in Bezug auf das Ziel, einen Test zur Messung der Reflexionsfähigkeit Physiklehramtsstudierender zu entwickeln, als erfüllt bewertet werden.

Die Klassifizierung der Antwortoptionen in richtige und falsche basiert auf dem Kodiermanual zum Performanztest, der dem hier vorgestellten Testinstrument zugrunde liegt. Antwortoptionen werden dann als richtig eingestuft, wenn diese laut Kodiermanual bedeutsame und daher zu reflektierende Aspekte ansprechen. Die Inhalte der (geskripteten) Unterrichtsvignetten und das Kodiermanual wurden in einer Expertenrunde (bestehend aus neun Wissenschaftler*innen mit einem Forschungsschwerpunkt in der physikdidaktischen Professionsforschung) in mehreren Feedbackschleifen kommunikativ validiert (Kempin et al. 2019). Das Kodiermanual gibt Aufschluss darüber, welches die relevanten Hinweise zur entsprechenden Vignette sind, sodass davon ausgegangen werden kann, dass die als richtig eingestuften Antworten inhaltlich relevante Rückmeldungen abbilden, Anforderung A1.2 also erfüllt ist.

4.2 Teilstudie 2: Kognitive Validität (Übersetzung der Aufgaben im Testinstrument in Verhalten)

4.2.1 Anlage der Teilstudie

Die Untersuchung der kognitiven Validität des Testinstruments soll sicherstellen, dass die Bearbeitung der Aufgaben auf Grundlage inhaltlicher Überlegungen erfolgt und Antwortoptionen auf dieser Grundlage (und nicht bspw. aufgrund der Satzlänge) angewählt werden. Dazu wurden die kognitiven Prozesse während der Bearbeitung des Tests mit Think-Aloud-Interviews erhoben und analysiert (Konrad 2010). Durchgeführt wurden sie in Videokonferenzen mit $N = 7$ Physiklehramtsstudierenden im Bachelor- ($N_{BA} = 3$) und Masterstudium ($N_{MA} = 4$). So konnten sowohl die verbalen Äußerungen als auch die Interaktion mit dem Testinstrument zur besseren Nachvollziehbarkeit während der Auswertung aufgezeichnet werden. Im Rahmen der Think-Aloud-Interviews wurden die Studierenden dazu aufgefordert, ihre Gedanken während der gesamten Bearbeitung des Testinstruments zu verbalisieren. Die aufgezeichneten Interviews spannen im Umfang zwischen 76 und 145 min. Sie wurden vollständig transkribiert und mit Hilfe eines Kategoriensystems ausgewertet. Unterschieden wurden (1) das Vorlesen von Inhalten des Testinstruments, geäußerte Reflexionselemente wie (2) Beschreibungen, (3) Bewertungen und (4) Alternativen, (5) explizit geäußerte Zustimmung oder Ablehnung von Antwortoptionen des Instruments, (6) sonstige Überlegungen mit direktem Bezug zum beobachteten Unterricht und (7) sonstige Aussagen. Zusätzlich wurden jegliche Hinweise auf un- oder missverständliche Formulierungen im Testinstrument kodiert (z. B. das wiederholte Vorlesen einer Antwortoption).

4.2.2 Ergebnisse und Diskussion der Teilstudie[1]

Die Kodierung der Think-Aloud-Protokolle führt zu insgesamt 26 unterschiedlichen potenziellen Missverständnissen bzw. Unklarheiten bezüglich der Antwortoptionen im Testinstrument. Eine nähere Analyse der Protokolle zeigt auf, dass diese überwiegend individuell sind und teilweise durch erneutes Lesen einer Antwortoption von den Studierenden selbst aufgelöst werden. Übrig bleiben lediglich drei Unklarheiten, wobei die Ursache der ersten beiden vermutlich in der Zusammenstellung der Vignetten und Kürzungen gegenüber dem ursprünglichen Performanztest zu finden ist: Eine Person übersah in den Unterrichtsvideos (Vignette 3) eine beschriebene Tafel, sodass ihr unklar war, worauf sich das Item „Es wäre besser, vor dem Einstieg in das neue Thema auch die Tafel zu wischen [...]“ bezieht. Die beschriebene Tafel ist zwar etwa 55 s lang im Hintergrund und damit über die Hälfte der Vignette sichtbar, aufgrund eines Aufnahme- bzw. Schnittfehlers allerdings auch etwa in den letzten 45 s der Vignette nicht mehr. Da diese Antwortoption allerdings kein zentrales Feedback zur beobachteten Szene und somit keine richtige Antwortoption darstellt, entsteht für Studierende, die die Skizze übersehen, kein Nachteil in der Testbearbeitung. Auch zur Antwortoption „Es ist nicht gut, dass du die Masse des Wagens und ins-

[1] Eine ausführlichere Darstellung der Ergebnisse findet sich in Weißbach & Kulgemeyer (2022).

besondere die Geschwindigkeit selbst bestimmst, da du den Schüler*innen vorher schon erklärt hast, wie sie das selbst machen könnten.“ im Kontext der Auswertung eines Demonstrationsexperiments (zu Vignette 5) merkt eine Person an, dass diese Erläuterung in der Vignette nicht zu sehen ist. Diese Information wird daher explizit in der Einleitung der Vignette im Testinstrument ergänzt. Das dritte unklare Item (ein Distraktor zur Erkennung aufgetretener Fehlvorstellungen zu Vignette 2) wird umformuliert und im Nachgang vier weiteren Studierenden vorgelegt, wobei keine erneuten Unklarheiten deutlich wurden. Insgesamt kann also davon ausgegangen werden, dass keine die Validität gefährdenden Unklarheiten vorliegen und damit Anforderung A2.1 erfüllt ist.

In den sieben Think-Aloud-Protokollen wurden durchschnittlich 677 Kodierungen vorgenommen (zwischen 416 und 846 Kodierungen je Interview). Die in allen Interviews am häufigsten kodierte Kategorie ist das Vorlesen z. B. einzelner Items mit durchschnittlich 39 %. Die konstruktrelevanten Kategorien (2–6), also Reflexionen, die explizite Zustimmung oder Ablehnung von Antwortoptionen sowie weitere Überlegungen mit direktem Bezug zum Unterricht, machen insgesamt durchschnittlich 53 % der Kodierungen aus. Der Anteil sonstiger Äußerungen liegt in sechs der sieben Interviews bei unter 10 % (im siebten bei 26 %). Etwa 15 % der vorliegenden Segmente wurden durch eine zweite Person kodiert. Für die unterschiedlichen Kategorien ergeben sich Übereinstimmungswerte von $0{,}59 \leq \kappa_{Cohen} \leq 0{,}91$, die mit Ausnahme der Kategorie „Beschreiben“ ($\kappa_{Cohen} = 0{,}59$) mittelmäßigen bis sehr guten Übereinstimmungen entsprechen (Döring und Bortz 2016, S. 569).

Der niedrige Anteil an konstruktirrelevanten Aussagen (Kategorie (7) „Sonstiges“, in 6 von 7 Interviews unter 10 %) lässt den Schluss zu, dass Anforderung A2.2 insgesamt erfüllt ist. Alle übrigen Aussagen weisen Bezug zu den Aufgabenstellungen des Testinstruments auf – entweder unmittelbar durch das Vorlesen von Items bzw. die Zustimmung zu oder Ablehnung von diesen oder aufgrund des inhaltlichen Bezugs der Überlegungen zum Unterricht bzw. der Professionalisierung als Lehrperson, wenn z. B. unabhängig von Antwortoptionen alternative Handlungsoptionen für die Lehrkraft vorgeschlagen werden.

4.3 Teilstudie 3: Externe Validität (Übersetzung des Verhaltens in Testwerte)

4.3.1 Anlage der Teilstudie

Studie 3 dient der Untersuchung der externen Validität. Es sollen Zusammenhänge zwischen den Testwerten des entwickelten MC-Tests und denen des Performanztests zur Messung der Reflexionsfähigkeit Physiklehramtsstudierender untersucht werden. Beide Instrumente fordern die Fremdreflexion einer Physik-Doppelstunde ein; die unterschiedlichen Formate zur Erfassung der Reflexionen (frei verbalisierte Reflexionen im Performanztest vs. MC-Aufgaben im hier vorgestellten Test) gehen mit einem deutlichen Unterschied in der Authentizität der Reflexionssituation einher. Dazu werden in drei universitären Veranstaltungen (insgesamt $N = 39$ Physik- und Sachunterrichtslehramtsstudierende mit Vertiefung im Fach Physik, davon $N_{BA} = 27$ im Bachelor- und $N_{MA} = 12$ im Masterstudium) beide Instrumente nacheinander im Abstand von drei bis acht Wochen (abhängig von den Gegebenheiten der je-

weiligen Lehrveranstaltung) eingesetzt. Um die freien Reflexionen der Studierenden im Performanztest nicht durch die vorgeschlagenen Antwortoptionen im MC-Test zu beeinflussen, wird der Performanztest zuerst durchgeführt. Eine Rückmeldung über das Abschneiden wird den Studierenden nach Bearbeitung des MC-Tests zur Verfügung gestellt. Zur Auswertung des Performanztests liegt ein Kodiermanual vor (Kempin et al. 2018).

4.3.2 Ergebnisse und Diskussion der Teilstudie

Für beide Instrumente ist in dieser Stichprobe nach dem Kolmogorow-Smirnow-Test von einer Normalverteilung der Testwerte auszugehen ($p = 0{,}96$ für den Performanztest und $p = 0{,}98$ für den MC-Test) und beide Instrumente weisen hohe interne Konsistenzen von $\alpha_{\text{Cronbach},P} = 0{,}82$ (Performanztest) und $\alpha_{\text{Cronbach,MC}} = 0{,}94$ (MC-Test) auf (Döring und Bortz 2016, S. 443). Die Studierenden erreichen durchschnittliche Punktzahlen von 57,5 (SD: 27,0) im Performanztest und 21,3 (SD: 4,6) im MC-Test. Um Zusammenhänge zwischen beiden Testwerten zu untersuchen wird der Korrelationskoeffizient nach Pearson berechnet. Da beide Instrumente das gleiche Konstrukt messen, ist grundsätzlich ein hoher Zusammenhang zu erwarten. Es ergibt sich eine signifikante Korrelation mit mittlerer Effektstärke an der Grenze zu einer großen Effektstärke ($r = 0{,}49$; $p < 0{,}01$) (Döring und Bortz 2016, S. 669). Aufgrund der inhaltlichen Vergleichbarkeit beider Instrumente, deutet die Korrelation von nur mittlerer Effektstärke darauf hin, dass das Antwortformat Einfluss auf die Erfassung der Reflexionsfähigkeit nimmt. Auch die Streuung der erreichten Punktzahlen ist im Performanztest mit offenem Antwortformat deutlich größer, als im MC-Test. Ursächlich hierfür könnte sein, dass der offene Performanztest zu einem bedeutsamen Teil auch Noticing zur Reflexion bedeutsamer Situationen integriert: Es werden nur die Aspekte reflektiert, die den Studierenden selbst auffallen, während die zu reflektierenden Aspekte im MC-Test vorgegeben sind und deren Anzahl gegenüber dem Performanztest reduziert ist, um einen handhabbaren Umfang des Testinstruments sicherzustellen. Bartels (2018) entwickelte einen performanznahen MC-Test zur Erklärfähigkeit Physiklehramtsstudierender und setzte diesen im Vergleich zu einer authentischeren, aber weiterhin standardisierten Erklärsituation ein, führte also eine methodisch vergleichbare Untersuchung für eine andere Fähigkeit durch. Hier ergab sich – wenn auch für eine kleine Stichprobe von $N = 12$ – mit $r = 0{,}54$ ($p < 0{,}05$) ein Zusammenhang in vergleichbarer Höhe (ebd., S. 139f.).

Insgesamt fällt der beobachtete Zusammenhang zwischen beiden Instrumenten etwas geringer aus, als erwartet. Er kann aber vor dem Hintergrund der Ergebnisse von Bartels (2018) in ähnlicher Höhe durchaus als Hinweis auf die Erfüllung von Anforderung A3, dass ein Zusammenhang zwischen den Ergebnissen im Reflexionstest und authentischerem Reflexionshandeln besteht, gewertet werden.

4.4 Teilstudie 4: Generalisierbarkeit (Übersetzung des Verhaltens in Testwerte)

4.4.1 Anlage der Teilstudie

Ziel von Studie 4 ist, die Generalisierbarkeit der Testwerte über die beiden Studienphasen Bachelor- und Masterstudium zu prüfen. Dazu wurde das Testinstrument über drei Semester hinweg in verschiedenen universitären Veranstaltungen an drei verschiedenen Universitäten im eingesetzt. Insgesamt liegen $N_{BA} = 77$ Bearbeitungen von Physiklehramtsstudierenden im Bachelor- und $N_{MA} = 59$ im Masterstudium vor. Diese beiden Stichproben wurden in Bezug auf statistische Kennwerte, wie die interne Konsistenz (Cronbachs-α), sowie Lösungshäufigkeiten und Trennschärfe als typische Aspekte der Itemanalyse (Döring und Bortz 2016, S. 476) verglichen.

4.4.2 Ergebnisse und Diskussion der Teilstudie

Für beide Teilgruppen ergeben sich sehr hohe interne Konsistenzen von $\alpha_{\text{Cronbach,BA}} = 0{,}95$ für die Kohorte der Bachelorstudierenden und $\alpha_{\text{Cronbach,MA}} = 0{,}91$ für die Kohorte der Masterstudierenden (Döring und Bortz 2016, S. 443). Die Trennschärfe ist für beide Kohorten akzeptabel und liegt durchschnittlich in einem hohen Bereich für Bachelorstudierende (0,61) bzw. mittleren Bereich für Masterstudierende (0,48) (Döring und Bortz 2016, S. 478). Der Mann-Whithney-U-Test zeigt, dass die Gruppe der Masterstudierenden im Testinstrument insgesamt besser abschneidet ($p < 0{,}01$), was sich in den unterschiedlichen durchschnittlichen Lösungshäufigkeiten für beide Kohorten widerspiegelt; diese liegen aber beide in einem akzeptablen Bereich (0,60 für Bachelor- und 0,71 für Masterstudierende, Döring und Bortz 2016, S. 477). Auch zeigt sich eine fast identische Rangfolge der Aufgaben in Bezug auf die Lösungshäufigkeit für beide Teilgruppen ($r_{\text{Spearman}} = 0{,}96$).

Die betrachteten statistischen Kennwerte liegen also sowohl für die Kohorte der Bachelor- als auch der Masterstudierenden in den oben benannten annehmbaren Bereichen nach Döring und Bortz (2016), sodass Anforderung A4 als Indikator für die Generalisierbarkeit der Testwertinterpretation über die verschiedenen Studienphasen hinweg als erfüllt angesehen werden kann. Ursachen für die auftretenden Unterschiede in Bezug auf die besseren Testwerte bzw. niedrigeren Lösungshäufigkeiten und Trennschärfen der Masterstudierenden werden im Studienfortschritt vermutet, da tendenziell mehr universitäre Veranstaltungen bestanden und Praxiserfahrungen gesammelt wurden. Die zusätzlich durchschnittlich niedrigere Trennschärfe der Testaufgaben für die Masterkohorte könnte als Hinweis darauf gewertet werden, dass das Testinstrument für viele Masterstudierende vergleichsweise einfach ist. Da die Trennschärfe aber weiterhin in einem annehmbaren Bereich liegt, ist das unproblematisch. Dass für beide Kohorten überwiegend die gleichen Aufgaben hohe bzw. niedrige Lösungshäufigkeiten aufweisen, stützt die Generalisierbarkeit zusätzlich auf Aufgabenebene.

4.5 Teilstudie 5: Konsequentielle Validität (Übersetzung der Testwerte in Konsequenzen)

4.5.1 Anlage der Teilstudie

Studie 5 dient der Erfassung der Studierendenperspektive auf das erhaltene Assessment-Feedback. Es soll die Angemessenheit der Konsequenzen evaluiert werden, die sich für sie aus der Darstellung der Testwertinterpretation im Assessment-Feedback ergeben. Dazu wurden insgesamt $N = 12$ leitfadengestützte Interviews mit Physiklehramtsstudierenden im Bachelorstudium zu zwei Versionen des Assessment-Feedbacks geführt. Sechs der Interviews bezogen sich allerdings auf eine frühere Version des Assessment-Feedback und dienten damit auch als Grundlage für die Überarbeitung dessen. Aufgrund der inhaltlichen Ähnlichkeit beider Versionen des Assessment-Feedback können zentrale Informationen v. a. über die Interpretation der Boxplots aus Studierendenperspektive zusammengefasst werden.

Die Interviews weisen Längen zwischen elf und 34 min auf. Sie wurden transkribiert und mit Hilfe eines Kategoriensystems inhaltsanalytisch ausgewertet (vgl. Krüger und Riemeier 2014). Das Kategoriensystem wurde deduktiv auf Grundlage des Interviewleitfadens entwickelt (6 Kategorien) und induktiv auf 18 Subkategorien ausdifferenziert. Die zentralen Oberkategorien beziehen sich auf die Beschreibung bzw. Interpretation der Boxplots (z. B. die Bezugnahme auf die Vergleichsgruppe bei der Einordnung des Ergebnisses), die aus dem Assessment-Feedback gezogenen Schlussfolgerungen (z. B. die Identifikation von Verbesserungspotentialen) sowie aus der Perspektive der Studierenden vorliegende Änderungsbedarfe.

4.5.2 Ergebnisse und Diskussion der Teilstudie

Aus den Interviews wird deutlich, dass sich alle Studierenden bei der Einordnung der Ergebnisse vorwiegend an denen der Vergleichsgruppe orientieren. Sie nutzen den Median als Richtwert, um die individuellen Ergebnisse als positiv zu bewerten (wenn sie über dem Median liegen, z. B. „[D]iese Rückmeldung ist ja sehr positiv[, w]eil das Ergebnis hier im oberen Mittel liegt“, I12) bzw. Verbesserungspotential zu diagnostizieren (wenn sie darunter liegen, z. B. „[Fehl]vorstellungen würde ich mir dann wahrscheinlich nochmal angucken, hätte ich die Zeit“, I11). Gleichzeitig fällt auf, dass knapp die Hälfte der Studierenden den dargestellten Median im Boxplot für das arithmetische Mittel zu halten scheint (z. B. „genau hier sieht man nochmal den Mittelwert […]“, I8). Jeweils eine Person nimmt das Testinstrument als Beispiel für Prüfungsleistungen wahr (I9) und die erreichten Testwerte als Bestätigung dafür, „gar nicht so falsch in dem Job“ (I1) zu sein. Fünf der Studierenden benennen zusätzlich mögliche Ursachen für die erzielten Ergebnisse, in der Regel eine mangelnde Auseinandersetzung mit dem Testinstrument während der Bearbeitung (z. B. „ich glaube, das lag einfach nur noch daran, dass ich mich halt nicht so doll angestrengt […] habe, wie ich es hätte tun sollen“, I8). Konkrete Konsequenzen zur Umsetzung des identifizierten Verbesserungspotentials werden von keiner Person benannt.

Im Rahmen der Interviews werden keine mehrheitlich geteilten Änderungsbedarfe in Bezug auf die Rückmeldung deutlich. Auffällig ist allerdings, dass die Rückmeldung als „umfangreich“ (I12) bezeichnet und von mindestens drei Studierenden nicht vollständig gelesen wird, während auf Nachfrage keine Inhalte der Rückmeldung als überflüssig bewertet werden.

Insgesamt kann die Interpretation der Ergebnisse durch die Studierenden als überwiegend inhaltlich angemessen bewertet werden. Die Orientierung an den Ergebnissen der Vergleichsgruppe ist durch die Darstellung mit Boxplots angelegt, durch die Testentwicklung intendiert und wird entsprechend angenommen. Das stützt die Erfüllung der Anforderung A5, dass die Konsequenzen, die sich für die Studierenden aus den im erhaltenen Assessment-Feedback dargestellten Testwerten ergeben, zum Konstrukt und den individuellen Entwicklungsbedarfen passen. Einschränkend ist festzustellen, dass die Interpretationen der Studierenden nicht umfassend korrekt sind (z. B. aufgrund der falschen Beschreibung des Medians als Mittelwert) und das Assessment-Feedback nicht von allen Studierenden gelesen wird. Dies ist nicht als Einschränkung der Validität zu bewerten, da die Studierenden sich nur an Gruppentendenzen und nicht an konkreten Prozentwerten orientieren und die Schlussfolgerungen dadurch nicht grundsätzlich verfälscht werden. Problematisch vor dem Hintergrund der Konzeption des Tests als Selbsteinschätzungsinstrument ist allerdings die jeweils von einer Person geäußerte Interpretation des Instruments als Beispiel für Prüfungsleistungen (I9) oder Bestätigung der Eignung für den Lehrerberuf (I1). Um dem entgegenzuwirken, muss ein Einsatz des Testinstruments im Rahmen von Veranstaltungen mit einer kurzen Erläuterung der Funktion des Testinstruments einhergehen.

Dass die Studierenden das Assessment-Feedback nicht (vollständig) lesen, steht der Zielstellung, den Studierenden eine Rückmeldung über die Ausprägung ihrer Fähigkeiten zu geben, entgegen. Dies betrifft allerdings nicht die konsequentiellen Validität, da hieraus keine Erkenntnisse zur Evaluation der inhaltlichen Angemessenheit der von den Studierenden aus dem Assessment-Feedback gezogenen Schlussfolgerungen gewonnen werden können. Mögliche Ursachen hierfür sind ein mangelndes Interesse am eigenen Ergebnis oder eine unzureichende Akzeptanz des Test- und Rückmeldeverfahrens.

5 Diskussion der Validitätsargumentation und Empfehlungen für den Testeinsatz in der Praxis

Die hier vorgestellte Kombination von Testinstrument und Assessment-Feedback soll Studierenden eine fundierte Selbsteinschätzung ihrer Fähigkeit, Physikunterricht zu reflektieren, ermöglichen. Insofern ist sicherzustellen, dass die Testwerte valide als Maß für die Reflexionsfähigkeit interpretiert werden können und die Studierenden valide Schlüsse aus dem erhaltenen Assessment-Feedback ziehen. Im Sinne des argumentbasierten Ansatzes von Kane (2013, 1992) werden hierfür vorwiegend empirische Argumente gesammelt, die die angemessene Umsetzung der nötigen Schritte vom Konstrukt bis hin zur Interpretation des Assessment-Feedback stützen.

Zusammenfassend stützen die Ergebnisse die Validität der Interpretation der Testwerte als Maß für die Fähigkeit, Physikunterricht zur reflektieren mit obigen Einschränkungen vorrangig bzgl. der Reduktion der Darstellung von Unterricht auf den Themenbereich der Mechanik (inhaltliche Validität) und falscher Vorstellungen über den Beispielcharakter der Testaufgaben für Prüfungsleistungen oder das Referendariat (konsequentielle Validität). Für den Einsatz des Testinstruments im Rahmen von universitären Veranstaltungen sollten diese falschen Vorstellungen präventiv adressiert und ausgeräumt werden. Eine Einordnung und Erläuterung des Instruments erscheint auch vor dem Hintergrund der eher kleinen Interviewstichprobe (insgesamt $N = 12$) sinnvoll, um weitere möglicherweise abhängig vom jeweiligen Einsatzkontext auftretende Fehlinterpretationen erkennen und auflösen zu können. Auf die ebenfalls geringe Stichprobengröße von $N = 7$ im Rahmen der Think-Aloud-Interviews wird reagiert, indem weiterhin ein Kommentarfeld im Testinstrument inkludiert ist, dessen Eingaben regelmäßig auf aufgetretene Unklarheiten oder übrige Hürden in der Testbearbeitung überprüft werden. Auch die Korrelation zu authentischerem Reflexionshandeln (repräsentiert durch den Reflexionsperformanztest, der natürlich seinerseits auch nur eine möglichst authentische Abbildung von realem Reflexionshandeln darstellt) weist nur eine mittlere Effektstärke auf, die vermutlich teilweise auf die unterschiedlichen Erhebungsmethoden in beiden Instrumenten zurückgeführt werden kann und zudem daraus resultieren könnte, dass im freien Performanztest der Aspekt des Noticing eine stärkere Gewichtung erhält. Gleichzeitig ist einschränkend festzuhalten, dass das zur externen Validierung gewählte Instrument hohe inhaltliche Überschneidungen zum entwickelten Instrument ausweist. Für eine Ausweitung der Validitätsargumentation in diesem Aspekt könnten weitere Instrumente eingesetzt werden (z. B. das Testinstrument zur Erfassung der Reflexionskompetenz von Kobl (2021) – das allerdings explizites Lernen über Reflexionsmodelle voraussetzt – oder die Bewertung von Reflexionstexten anhand von Kodiermanualen, wie z. B. von Kobl (2021) oder Abels (2011) entwickelt). In Summe wurden Validitätsargumente für alle vier Übersetzungsschritte gefunden, sodass trotz dieser Einschränkungen davon ausgegangen werden kann, dass das Instrument geeignet ist, Studierenden eine fundierte Selbsteinschätzung ihrer Fähigkeiten zu ermöglichen.

Als wichtige Einschränkung muss erwähnt werden, dass derzeit zwar argumentiert werden kann, dass die Testwerte des Instruments Auskunft über die Reflexionsfähigkeit der Studierenden geben – aber unklar ist, ob durch das Assessment-Feedback auch Lernprozesse angestoßen werden. Dies erscheint angesichts der Vorarbeiten zu Feedback zwar plausibel, war aber nicht Gegenstand der hier vorgestellten Studien. Wir empfehlen deshalb eine gute Einbettung des Instruments in Lehrveranstaltungen, in denen das Thema „Reflexion“ aufgegriffen und inhaltlich behandelt wird.

Erstrebenswert ist, auch Dozierenden Informationen zum Fähigkeitsstand ihrer Lerngruppe bereitzustellen, um z. B. eine adaptive Begleitung von Praxisphasen zu ermöglichen. Hier besteht Forschungsbedarf in Bezug auf die valide Interpretation der entsprechenden Rückmeldung durch Dozierende. Der letzte Übersetzungsschritt, welcher die Interpretation der Ergebnisse durch die Studierenden abbildet, sollte um diese Perspektive ergänzt werden: „Aus den Testwerten werden Konsequenzen für die individuelle Professionalisierung (Studierende) bzw. die Gestaltung von Lehrver-

anstaltungen (Dozierende) gezogen.“ Auch sonstige neue Einsatzszenarien sollten mit neuen Argumenten für die valide Interpretation der Testwerte entwickelt werden (Kane 2013, S. 45); die Validitätsargumentation kann also grundsätzlich nicht als abgeschlossen betrachtet werden (Kane 1992). Ergänzend erscheint die Untersuchung der Akzeptanz Studierender und Dozierender gegenüber dem Test- und Feedbackformat vielversprechend, da davon auszugehen ist, dass sie Einfluss auf die Bearbeitung des Testinstruments sowie die Interpretation des Assessment-Feedbacks nimmt und insbesondere mangelnde Akzeptanz einem Transfer des Instruments in die universitäre Lehr- und Lernpraxis entgegensteht.

Vor dem Hintergrund der gewonnenen Erkenntnisse über die valide Interpretation der Testwerte des Instruments als Maß für die Fähigkeit Studierender, Unterricht zu reflektieren, kann ein Einsatz des Testinstruments im Rahmen von universitären, physikdidaktischen Lehrveranstaltungen empfohlen werden. Inhaltlich bietet sich dabei insbesondere die Vorbereitung oder Begleitung von Praxisphasen (z. B. das Praxissemester) an. Bedeutsam ist die transparente Einbettung des Instruments gegenüber den Studierenden als Möglichkeit, eine fundierte Rückmeldung unabhängig von (künftigen) Prüfungssituationen zu erhalten. Das Instrument ist frei verfügbar und steht unter www.unterrichtsreflexion.de zum Einsatz zur Verfügung.

Funding Open Access funding enabled and organized by Projekt DEAL.

Literatur

Abels, S. (2011). *LehrerInnen als „Reflective Practitioner“. Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht*. Wiesbaden: VS.

Aeppli, J., & Lötscher, H. (2016). EDAMA – Ein Rahmenmodell für Reflexion. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *34*(1), 78–97.

von Aufschnaiter, C., Hofmann, C., Geisler, M., & Kirschner, S. (2019a). Möglichkeiten und Herausforderungen der Förderung von Reflexivität in der Lehrerbildung. *SEMINAR*, *25*(1), 49–60.

von Aufschnaiter, C., Fraij, A., & Kost, D. (2019b). Reflexion und Reflexivität in der Lehrerbildung. *Herausforderung Lehrer_innenbildung – Zeitschrift zur Konzeption, Gestaltung und Diskussion*, *2*(1), 144–159. https://doi.org/10.4119/UNIBI/hlz-144.

Bartels, H. (2018). *Entwicklung und Bewertung eines performanznahen Videovignettentests zur Messung der Erklärfähigkeit von Physiklehrkräften*. Studien zum Physik- und Chemielernen, Bd. 264. Berlin: Logos.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*(4), 469–520. https://doi.org/10.1007/s11618-006-0165-2.

Beauchamp, C. (2006). *Understanding reflection in teaching. A framework for analysing the literature*. Ottawa: Library and Archives Canada.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond Dichotomies. *Zeitschrift für Psychologie*, *223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion*. München: Pearson.

Carlson, J., Daehler, K. R., Alonzo, A. C., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K. K. H., Cooper, R., Friedchrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepert, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., & Wilson, C. D. (2019). The refined consensus model of pedagogical content knowledge in science education. In: A. Hume, R. Cooper & A. Borowski (Hrsg.), *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science* (S. 77–94). Singapore: Springer

Clarà, M. (2015). What is reflection? Looking for clarity in an ambiguous notion. *Journal of Teacher Education*, *66*(3), 261–271. https://doi.org/10.1177/0022487114552028.

Dewey, J. (2002). *Wie wir denken*. (A. Burgeni, Übers., überarbeitete Ausgabe von R. Horlacher, & J. Oelkers, Hrsg.). Zürich: Pestalozzianum. (Zuerst veröffentlicht 1910)

Dickmann, M. (2016). *Messung von Experimentierfähigkeiten. Validierungsstudien zur Qualität eines computerbasierten Testverfahrens*. Studien zum Physik- und Chemielernen, Bd. 210. Berlin: Logos.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin, Heidelberg: Springer.

van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers' „learning to notice" in the context of a video club. *Teaching and Teacher Education*, *24*(2), 244–276. https://doi.org/10.1016/j.tate.2006.11.005.

Frey, A. (2006). Methoden und Instrumente zur Diagnose beruflicher Kompetenzen von Lehrkräften. Eine erste Standortbestimmung zu bereits publizierten Instrumenten. In: C. Allemann-Ghionda & E. Terhart (Hrsg.), *Kompetenzen und Kompetenzentwicklung von Lehrerinnen und Lehrern* (S. 30–46). Weinheim: Beltz.

Friege, G., & Lind, G. (2004). Leistungsmessung im Leistungskurs. *Der Mathematische und Naturwissenschaftliche Unterricht. MNU*, *57*(5), 259–265.

Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 143–171). Berlin Heidelberg New York: Springer.

Hattie, J. A. C. (2009). *Visible Learning. A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hattie, J. A. C., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487.

Hatton, N., & Smith, D. (1995). Reflection in teacher education: towards definition and implementation. *Teaching and Teacher Education*, *11*(1), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U.

Holtz, P., & Gnambs, T. (2017). The improvement of student teachers' instructional quality during a 15-week field experience: a latent multimethod change analysis. *Higher Education*, *74*(4), 669–685. https://doi.org/10.1007/s10734-016-0071-3.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kempin, M., Kulgemeyer, C. & Schecker, H. (2018). Reflexion von Physikunterricht: Ein Performanztest. In: C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht- normative und empirische Dimensionen. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Regensburg* 2017 (S. 867-870). Regensburg: Universität Regensburg.

Kempin, M., Kulgemeyer, C. & Schecker, H. (2019). Erste Einblicke in die Entwicklung der Reflexionsfähigkeit von Physiklehramtsstudierenden im Praxissemester. In: C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Kiel 2018* (S. 357-360). Regensburg: Universität Regensburg.

Kempin, M., Kulgemeyer, C. & Schecker, H. (2020). Wirkung von Professionswissen und Praxisphasen auf die Reflexionsfähigkeit von Physiklehramtsstudierenden. In: S. Habig (Hrsg.): *Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung 2019 in Wien* (S. 439-442). Essen: Universität Duisburg-Essen.

Klewin, G., Köker, A., & Störtländer, J. C. (2020). Veranlasste und benotete Reflexion: ein unmögliches Prüfungsformat? *Herausforderung Lehrer*innenbildung – Zeitschrift Zur Konzeption, Gestaltung Und Diskussion*, *3*(2), 108–121. https://doi.org/10.4119/hlz-2530.

Kobl, C. (2021). *Förderung und Erfassung der Reflexionskompetenz im Fach Chemie*. Studien zum Physik- und Chemielernen, Bd. 312. Berlin: Logos.

Konrad, K. (2010). Lautes Denken. In: G. Mey & K. Mruck (Hrsg.), *Handbuch qualitative Forschung in der Psychologie* (1. Aufl., S. 476–490). Wiesbaden: VS.

Korthagen, F. A. J. (1985). Reflective teaching and Preservice teacher education in the Netherlands. *Journal of Teacher Education*, *36*(5), 11–15.

Krüger, D., & Riemeier, T. (2014). Die qualitative Inhaltsanalyse – eine Methode zur Auswertung von Interviews. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 135–145). Berlin, Heidelberg: Springer Spektrum. https://doi.org/10.1007/978-3-642-37827-0_11.

Kulgemeyer, C., Kempin, M., Weißbach, A., Borowski, A., Buschhüter, D., Enkrott, P., Riese, J., Schecker, H., Schröder, J. & Vogelsang, C. (2021). Exploring the impact of pre-service science teachers' reflection skills on the development of professional knowledge during a field experience. *International Journal of Science Education 43*(18), 3035-3057. https://doi.org/10.1080/09500693.2021.2006820

Kultusministerkonferenz (KMK) (2004i). Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2009. https://www.kmk.org/fileadmin/veroeffentli-chungen_beschluesse/2004/2004_12_16-Stan-dards-Lehrerbildung-Bildungswissenschaf-ten.pdf

Max, A., Lukas, S., & Weitzel, H. (2022). The relationship between self-assessment and performance in learning TPACK: Are self-assessments a good way to support preservice teachers' learning? *Journal of Computer Assisted Learning*, *38*(4), 1160–1172. https://doi.org/10.1111/jcal.12674.

McAlpine, L., Weston, C., Beauchamp, C., Wiseman, C., & Beauchamp, J. (1999). Monitoring student cues: tracking student behaviour in order to improve instruction in higher education. *The Canadian Journal of Higher Education*, *29*(3), 113–144. https://doi.org/10.47678/cjhe.v29i3.183335.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Nowak, A., Kempin, M., Kulgemeyer, C. & Borowski, A. (2019). Reflexion von Physikunterricht. In: C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung in Kiel 2018* (S. 838-841). Regensburg: Universität Regensburg.

Nowell, C., & Alston, R. M. (2007). I thought I got an A! Overconfidence across the economics curriculum. *Journal of Economic Education*, *38*, 131–142. https://doi.org/10.3200/JECE.38.2.131-142.

Plöger, W., Scholl, D., & Seifert, A. (2015). Analysekompetenz – ein zweidimensionales Konstrukt?! *Unterrichtswissenschaft*, *43*(2), 166–184.

Poldner, E., van der Schaaf, M., Simons, P. R.-J., van Tartwijk, J., & Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education*, *37*(3), 348–373. https://doi.org/10.1080/02619768.2014.892479.

Riese, J., & Reinhold, P. (2008). Entwicklung und Validierung eines Instruments zur Messung professioneller Handlungskompetenz bei (angehenden) Physiklehrkräften. *Lehrerbildung auf dem Prüfstand*, *1*(2), 625–640.

Rothland, M., & Boecker, S. K. (2015). Viel hilft viel? Forschungsbefunde und -perspektiven zum Praxissemester in der Lehrerbildung. *Lehrerbildung auf dem Prüfstand*, *8*(2), 112–134.

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. https://doi.org/10.1007/978-3-642-37827-0_9.

Schön, D. A. (1983). *The reflective practicioner*. New York: Basic Books.

Seidel, T., Blomberg, G., & Stürmer, K. (2010). „Oberserver" – Validierung eines videobasierten Instruments zur Erfassung der professionellen Wahrnehmung von Unterricht. Projekt OBSERVE. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (S. 196–306). Weinheim, Basel: Beltz.

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–21.

Sippel, S. (2009). Zur Relevanz von Assessment-Feedback in der Hochschullehre. *Zeitschrift für Hochschulentwicklung*, *4*(1), 1–22. https://doi.org/10.3217/zfhe-4-01/02.

Stender, J., Schaper, N., Vogelsang, C., & Watson, C. (2019). Professionalisierung durch Portfolioarbeit – Ein Beitrag zur Erfassung der Reflexionskompetenz von Lehramtsstudierenden. *Lehrerbildung auf dem Prüfstand*, *12*(2), 181–206.

Szogs, M., Kobl, C., Volmer, M., & Korneck, F. (2019). Bedeutsamkeit von Reflexion und Reflexivität in der Professionalisierung von Lehrkräften sowie ihre Beziehung zu anderen Prozessen und Konstrukten. In C. Maurer (Hrsg.), *Naturwissenschaftliche Bildung als Grundlage für berufliche und ge-*

sellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung, Kiel, 2018. (S. 317–320). Regensburg: Universität Regensburg.

Weißbach, A. & Kulgemeyer, C. (2022). Reflexion von Physikunterricht - ein Online-Assessment mit Feedback. *PhyDid B - Didaktik Der Physik - Beiträge Zur DPG-Frühjahrstagung*, 203–210.

Windt, A., & Lenske, G. (2016). Qualität der Sachunterrichtsreflexion im Vorbereitungsdienst. In: C. Maurer (Hrsg.), *Authentizität und Lernen – das Fach in der Fachdidaktik. Gesellschaft für Didaktik der Chemie und Physik Jahrestagung, Berlin, 2015* (S. 284–286). Regensburg: Universität Regensburg.

Windt, A., Rau, S., Hasenkamp, A., Lenske, G., & Rumann, S. (2017). Mit welchen Kompetenzen starten angehende Lehrkräfte in den Vorbereitungsdienst? In: H. Fischler & E. Sumfleth (Hrsg.), *Professionelle Kompetenz von Lehrkräften der Chemie und Physik*. Studien zum Physik- und Chemielernen, Bd. 200 (S. 185–199). Berlin: Logos.

Wöhlke, C. (2020). *Entwicklung und Validierung eines Instruments zur Erfassung der professionellen Unterrichtswahrnehmung angehender Physiklehrkräfte*. Studien zum Physik- und Chemielernen, Bd. 298. Berlin: Logos.