Reflexion von Physikunterricht – ein Online Assessment mit Feedback

Anna Weißbach*, Christoph Kulgemeyer*

* Universität Bremen, Institut für Didaktik der Naturwissenschaften – Abt. Physikdidaktik, Bibliothekstraße 1, 28234 Bremen
anna.weissbach@uni-bremen.de

Kurzfassung

Die Reflexion von Physikunterricht spielt im Unterricht eine zentrale Rolle und dient der Professionalisierung von Lehrpersonen sowie der Weiterentwicklung des Unterrichts (von Aufschnaiter, Hofmann, Geisler & Kirschner, 2019, S. 49). Gleichzeitig reflektieren Studierende häufig auf einem niedrigen, beschreibenden Niveau (Hatton & Smith, 1995, S. 41). Die Förderung der Reflexionsfähigkeit ist daher auch schon in der ersten Phase der Lehrkräftebildung von Bedeutung. Ausgehend von einem bestehenden Performanztest zur Messung der Reflexionsfähigkeit von Studierenden wird ein geschlossenes Diagnose-Instrument entwickelt, in welchem Studierende einem fiktiven Mitpraktikanten Feedback zu Ausschnitten seines Unterrichts (in Form von Videovignetten) geben. Das Diagnose-Instrument wird mit teilautomatisiertem Assessment-Feedback sowie Hinweisen zur Förderung der Reflexionsfähigkeit versehen. Das Instrument weist eine hohe interne Konsistenz ($\alpha_{Cronbach} = 0{,}955$) auf. Sieben im Rahmen der Pilotierung des Diagnose-Instruments durchgeführte Think-Aloud-Interviews liefern Hinweise auf die Validität der Interpretation der Testwerte als Maß für die Reflexionsfähigkeit. Die Auswertung der Interviews macht deutlich, dass das Instrument keine zentralen Verständnishürden aufweist und die Studierenden während dessen Bearbeitung hauptsächlich Überlegungen mit Bezug zum Unterricht bzw. die Professionalisierung als Lehrperson anstellen.

1. Motivation

Die Reflexion von Unterricht gilt als eine „Kernaufgabe“ (KMK, 2004, S. 3) und „entscheidende Kompetenz“ (Abels, 2011, S. 59) von Lehrkräften. Sie dient dabei sowohl der Verbesserung der Unterrichtsqualität als auch der weiteren Professionalisierung als Lehrkraft (von Aufschnaiter et al., 2019, S. 49). Gleichzeitig reflektieren Studierende Unterricht allerdings oft nicht systematisch (Rothland & Boecker, 2015, S. 115) und hauptsächlich auf einem deskriptiven und kaum kritischen Niveau (Hatton & Smith, 1995, S. 41). Die Reflexionskompetenz Lehramtsstudierender ist insgesamt also „eher schwach ausgebildet“ (Wyss & Mahler, 2021, S. 17).

Im Rahmen des vorausgegangenen Projekts ProfiLeP+ wurde ein Performanztest zur Messung der Reflexionsfähigkeit von Physiklehramtsstudierenden entwickelt (Kempin, Kulgemeyer & Schecker, 2018, S. 868f.). Untersuchungen mit diesem Performanztest zeigen, dass das Praxissemester als Lerngelegenheit für Unterrichtsreflexion nicht grundsätzlich zu einer Verbesserung der Reflexionsfähigkeit führt, sondern dafür die Nutzung bestimmter Lerngelegenheiten, dazu zählen insbesondere Reflexionsgespräche mit universitären Mentorierenden, nötig sind (Kulgemeyer et al., 2021, S. 3051).

Hier schließt das Projekt ProfiLeP-Transfer an, in dem der Ansatz einer systematischen Förderung der Reflexionsfähigkeit verfolgt wird. Dazu werden im Rahmen einer Lernumgebung ein Diagnose-Instrument zur Reflexionsfähigkeit mit Assessment-Feedback und einem Fördermaterial gekoppelt, um Studierenden die selbstständige und aktive Auseinandersetzung mit ihrem Assessment-Feedback zu ermöglichen (vgl. Casanova, Alsop & Huet, 2021, S. 2). Das Diagnose-Instrument wird auf Grundlage des vorliegenden Performanztests als geschlossenes Online-Diagnose-Instrument entwickelt und mit halb-automatisiertem Feedback versehen. Im Rahmen des Projekts werden die Ziele verfolgt, (1) Argumente für die Validität des geschlossenen Online-Diagnose-Instrument zu sammeln und (2) ein valides Rückmeldeformat zu entwickeln, welches Studierenden die fundierte Selbsteinschätzung und Förderung der Reflexionsfähigkeit bzw. Dozierenden einen fundierten Einblick in die Reflexionsfähigkeit der Studierenden ihres Kurses ermöglicht. Darüber hinaus werden die entwickelten Materialien Dozierenden zur Verfügung gestellt, um (3) den nachhaltigen Transfer in die Lehrpraxis zu ermöglichen.

In diesem Beitrag werden die entwickelte Lernumgebung bestehend aus Diagnose-Instrument, Assessment-Feedback und Fördermaterial sowie Ergebnisse einer Think-Aloud-Studie zur Evaluation der kognitiven Validität des Diagnose-Instruments vorgestellt.

2. Reflexion von Physikunterricht

Die Relevanz der Reflexion im Lehrberuf scheint unumstritten (von Aufschnaiter et al., 2019, S. 49), ins-

besondere da sie die Entwicklung des Professionswissens auf Grundlage von individuellen Erfahrungen ermöglicht (Carlson et al., 2019, S. 84). Ebenso unstrittig ist das Verständnis von Reflexion als eine eigene Art des Denkens (Wyss, 2013, S. 38). Darüber hinaus wird der Reflexionsbegriff allerdings nicht einheitlich verwendet. So kann beispielsweise unterschieden werden, ob Reflexionen ausschließlich auf die eigene Person bezogen sind (Selbstreflexion, von Aufschnaiter, Fraij & Kost, 2019, S. 148f.) oder auch auf andere Personen (Fremdreflexion, Wyss, 2008, S. 3). Schön (1983) betrachtet den zeitlichen Bezug der Reflexion zur Handlung und unterscheidet zwischen der Reflection-*in*-action (Reflexion während der Handlung) und Reflection-*on*-action (Reflexion im Anschluss an die Handlung). Weitere Unterscheidungsmerkmale sind unter anderem die Zielstellung und Inhalte einer Reflexion (von Aufschnaiter et al., 2019, S. 146f.) sowie die Frage danach, in welchem Medium die zu reflektierende Situation vorliegt (z. B. Videoaufzeichnung), in welcher Form eine Reflexion vorliegt (z. B. in Textform oder im Dialog) oder ob weitere Personen (z. B. Mitstudierende) anwesend sind (Szogs, Kobl, Volmer & Korneck, 2019, S. 318). Modellierungen des Reflexionsbegriffs beinhalten außerdem häufig Stufenmodelle, die es ermöglichen, die Qualität von Reflexionen zu bewerten (z. B. Plöger, Scholl & Seifert, 2015; Windt & Lenske, 2016).

Im Rahmen des hier vorgestellten Projekts wird eine eher breite Definition des Reflexionsbegriffs gewählt. Reflexion wird verstanden als die „theoriegeleitete Analyse von Unterricht mit dem Ziel der Verbesserung der Unterrichtsqualität und der Entwicklung der Professionalität von Lehrpersonen“ (Kempin, Kulgemeyer & Schecker, 2020, S.439). Ergänzt wird diese Definition durch das Modell zur Bewertung der Güte von Reflexionen von Nowak, Kempin, Kulgemeyer und Borowski (2019), welches auf Grundlage der oben genannten Stufenmodelle entwickelt wurde (ebd.). Aussagen, die im Rahmen einer Reflexion getätigt werden, können in diesem Modell in drei Dimensionen eingeordnet werden (s. Abb. 1): Kategorisiert wird, (1) welchem Reflexionselement eine Aussage zugeordnet werden kann (wird eine Situation oder ein Verhalten beschrieben, bewertet, werden alternative Handlungsoptionen vorgeschlagen oder Konsequenzen für nachfolgenden Unterricht bzw. die Professionalisierung der Lehrperson formuliert?), (2) welche Wissensbasis einer Aussage zugrunde liegt (handelt es sich um fachwissenschaftliche, fachdidaktische oder pädagogische Überlegungen?) und (3) ob Aussagen begründet oder unbegründet vorliegen (nicht möglich bei Beschreibungen).

Aufgrund der zentralen Bedeutung, die der Reflexionsfähigkeit im Lehrberuf zukommt, liegt es nahe, diese auch im Rahmen der Lehramtsausbildung zu fördern. Klempin (2019, S. 109-132) gibt einen Überblick über verschiedene Ansätze zur Förderung der Reflexionsfähigkeit Physiklehramtsstudierender und unterscheidet zwischen „individuell-monologischen Formaten“ (ebd., S. 111, z. B. Reflexionstagebücher oder Portfolios), „visualisierenden“ Formaten (ebd., S. 113, z. B. Concept Maps oder Reflexionen basierend auf Videomaterial), „[k]ollegial-dialogische[n] Ansätze[n]“ (ebd., S. 119), die sich „durch die intensive Begleitung, kognitive Herausforderung und Unterstützung des Reflexionsprozesses“ (ebd., S. 199) auszeichnen und „experimentelle[n] Ansätzen“ (ebd., S. 121, z. B. in Lehr-Lern-Laboren).

3. **Die entwickelte Lernumgebung**

3.1. **Diagnose-Instrument**

Auf Grundlage des bestehenden Performanztests zur Messung der Reflexionsfähigkeit (Kempin, Kulgemeyer & Schecker, 2018) sowie des beschriebenen Reflexionsmodells (Nowak et al., 2019) wurde ein geschlossenes Online-Diagnose-Instrument entwickelt. In dessen Rahmen werden Studierende dazu aufgefordert, einem fiktiven Mitpraktikanten „Robert“ Feedback zu sieben inhaltliche zusammenhängenden Ausschnitten (Videovignetten) einer Physik-Doppelstunde zu geben. Abschließend wird außerdem ein Feedback zur gesamten Doppelstunde eingefordert. Inhaltlich ist die Doppelstunde im Bereich der Mechanik angesiedelt, behandelt werden die Newtonschen Axiome sowie die Einführung des Impulsbegriffs. Die Unterrichtsausschnitte sind dabei so konzipiert, dass verschiedene, in der Regel problematische Situationen zu beobachten sind, die im Feedback aufgegriffen werden (z. B. fachliche Fehler, der Umgang mit Schülervorstellungen oder ein stark stereotypischer Umgang mit Mädchen und Jungen im

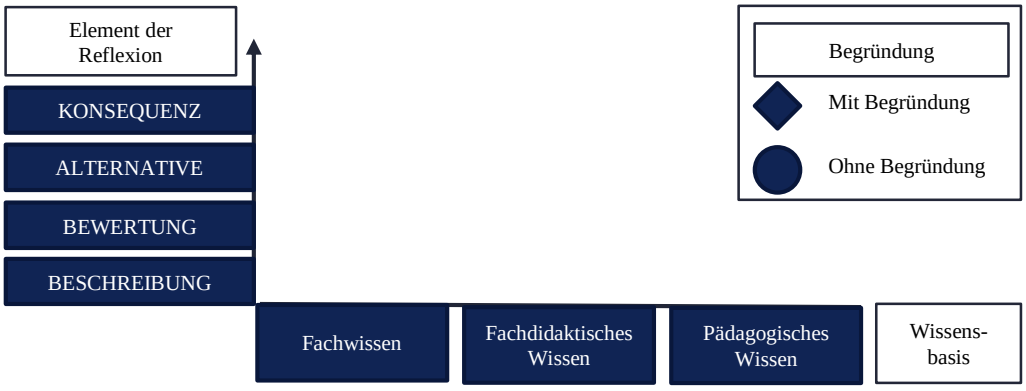

Abb. 1: Modell zur Reflexionen von Physikunterricht

Unterricht). So werden insgesamt 16 unterschiedliche Aspekte berücksichtigt. Das Feedback, welches Studierende im Rahmen der Durchführung zu diesen Aspekten geben, setzt sich basierend auf dem oben vorgestellten Reflexionsmodell (s. Abb. 1) jeweils zusammen aus einer Multiple-Choice-Aufgabe zur Bewertung der Situation und einer Multiple-Choice-Aufgabe zu alternativen Handlungsoptionen (bzw. der Empfehlung, das Vorgehen so beizubehalten), insgesamt werden im Rahmen des Feedbacks also 32 Aufgaben bearbeitet. Indem Studierende die beobachteten Situationen im Rahmen dieser Multiple-Choice-Items bewerten und alternative Handlungsoptionen vorschlagen, führen sie eine (bzw. Teile einer) Fremdreflexion durch.

Im Vergleich zum hier vorgestellten Diagnose-Instrument ist die Bearbeitung des zugrundeliegenden Performanztests näher am tatsächlichen beruflichen Handeln und stellt somit eine authentischere Repräsentation der Reflexionssituation dar. In Anlehnung an Miller (1980) kann hier zwischen dem Performanztest („shows how“, Miller, 1980, S. 63) und dem Konpetenztest („knows how“, ebd., S. 63) unterschieden werden. Um trotzdem eine möglichst hohe Authentizität der in den Multiple-Choice-Aufgaben zur Auswahl stehenden Antwortoptionen zu gewährleisten, basieren diese (mit Ausnahme weniger Ergänzungen) auf Reflexionen, die im Rahmen der der Bearbeitung des Performanztests zu den entsprechenden Unterrichtsausschnitten von Studierenden geäußert wurden.

Für die 32 Aufgaben, die während der Durchführung bearbeitet werden, weist das Diagnose-Instrument auf Basis von 89 aktuell vorliegenden Bearbeitungen eine sehr hohe interne Konsistenz von $\alpha_{Cronbach} = 0{,}955$ bei einer durchschnittlichen korrigierten Item-Skala-Korrelation von 0,61 auf.

Das Diagnose-Instrument ist vollständig in der Software LimeSurvey umgesetzt und kann so unabhängig vom Betriebssystem online aufgerufen und bearbeitet werden. Es kann online unter www.unterrichtsreflexion.de eingesehen und bearbeitet werden.

3.2. **Assessment-Feedback**

Die Bearbeitungen des Diagnose-Instruments werden halb-automatisiert ausgewertet. Dabei werden wiederum basierend auf dem Performanztest jene Antwortoptionen als „richtig“ bewertet, die sinnvolle Rückmeldungen zu relevanten Aspekten des Unterrichts darstellen. Die bei den insgesamt 32 bewerteten Multiple-Choice-Aufgaben maximal erreichbaren Punktzahlen werden dabei so normiert, dass sie jeweils mit gleichem Gewicht in die Berechnung des Gesamtergebnisses eingehen.

Studierende erhalten anschließend eine individuelle Rückmeldung zur Einordnung ihrer eigenen Bearbeitung, Dozierende eine Rückmeldung zur Einordnung der Bearbeitungen aller Kursteilnehmenden im Überblick. Um die Anonymität zu gewährleisten, können Studierende ihre Rückmeldungen mit einem persönlichen Code online abrufen. Im Rahmen dieses Assessment-Feedbacks werden die individuellen Ergebnisse bzw. die Ergebnisse des Kurses in einer Referenzgruppe bestehend aus anderen Physik-Lehramtsstudierenden mittels Boxplots verortet. Zusätzlich zum Gesamtergebnis wird außerdem eine in die Teilfähigkeiten „Bewerten“, „Alternativen vorschlagen“ sowie „Erkennen von und Umgang mit aufgetretenen Schülervorstellungen“ aufgeschlüsselte Rückmeldung gegeben. Zur Unterstützung der weiteren Professionalisierung werden die Studierenden im Rahmen der Rückmeldung außerdem auf das entwickelte Fördermaterial als eine Möglichkeit, das Reflektieren angeleitet zu üben, hingewiesen.

3.3. **Fördermaterial**

Das Fördermaterial ergänzt das Assessment Feedback um die Komponente des „feed forward“ (Hattie, Timperley, 2007, S. 86) und adressiert die Frage, wie vorgegangen werden kann, um die eigene Reflexionsfähigkeit zu verbessern. Das Fördermaterial beinhaltet einerseits einen kurzen (theoretischen) Input zum Begriff der Reflexion, dem Ziel des Fördermaterials sowie konkrete Hinweise zum weiteren Üben von Reflexion (insb. Leitfragen zur Unterrichtsreflexion, vgl. Nowak, Ackermann & Borowski, 2018, S. 223) und fachdidaktische Literaturhinweise. Andererseits können drei weitere Szenen der Unterrichtsstunde, die bereits im Testinstrument betrachtet wurde, angeleitet reflektiert werden. Dazu wird (1) eine am Reflexionsmodell (s. Abb. 1) ausgerichtete Struktur vorgegeben, es werden (2) konkret auf den Unterrichtsausschnitt bezogene Hinweise gegeben sowie (3) anhand der Leitfragen ausformulierte Beispielreflexionen über die zentralen Elemente des Unterrichtsausschnitts zur Verfügung gestellt, um Studierenden einen Abgleich mit den eigenen Überlegungen zu ermöglichen. Ziel ist die Simulation eines kollegial-dialogischen Ansatzes zur Förderung der Reflexionsfähigkeit, in welchem die Studierenden – wenn auch nicht interaktiv – bei der Reflexion angeleitet und unterstützt werden (Klempin, 2019, S. 119).

Das Fördermaterial ist wie das Diagnose-Instrument in der Software LimeSurvey umgesetzt und kann so selbständig von Studierenden bearbeitet werden, eine Anleitung oder Begleitung ist nicht notwendig.

4. **Untersuchungsdesign**

In Rahmen des Projekts soll die Validität der Interpretation der Testwerte als Maß für die Reflexionsfähigkeit der Studierenden einerseits und als Ausgangspunkt für einen weiteren Professionalisierungsprozess andererseits evaluiert werden. Dazu werden im Sinne des Argument-based-Approach nach Kane (2013) empirisch fundierte Argumente für (bzw. gegen) die Validität gesammelt. In Anlehnung an Dickmann (2016) werden dafür die Übersetzungsschritte betrachtet, die ausgehend vom Konstrukt der Reflexionsfähigkeit über die entwickelten Aufgaben und

Testwerte bis hin zu den aus der Bearbeitung des Instruments folgenden Konsequenzen gegangen werden (s. Abb. 2).

Die Produkte dieser Übersetzungsschritte (z. B. die Multiple-Choice-Aufgaben als Produkt der Übersetzung der Reflexionsfähigkeit (I) in ein Diagnose-Instrument (II) oder die Testwerte als Produkt der Bewertung bzw. Übersetzung der Bearbeitung des Diagnose-Instruments (III) in quantitative Werte (IV)) können jeweils in Bezug auf konkrete Anforderungen evaluiert werden.

Abhängig vom betrachteten Produkt kann dessen Qualität theoretisch begründet werden (Übersetzungen von (I) in (II) und (IV) in (V) in Abb. 2), oder durch verschiedene Interviewformen (Übersetzungen von (II) in (III) sowie (V) in (VI) in Abb. 2) und statistische Analysen wie u.a. die Untersuchung der Reliabilität des Diagnose-Instruments (Übersetzung von (III) in (IV) in Abb. 2) empirisch untersucht werden.

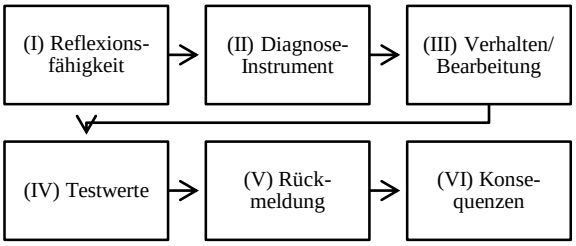

Abb.2: Zu evaluierende Übersetzungsschritte vom betrachteten Konstrukt Reflexionsfähigkeit bis zu Konsequenzen aus der Arbeit mit der Lernumgebung.

5. Think-Aloud-Studie zur Evaluation der kognitiven Validität des Diagnose-Instruments

Die Untersuchung des durch die Bearbeitung des Testinstruments evozierten Verhaltens auf Seiten der Studierenden (Übersetzung vom Testinstrument (II) in Verhalten (III) in Abb. 2) soll sicherstellen, dass die Anforderung „Das Verhalten der Teilnehmenden passt zum bei der Testkonzeption beabsichtigten Verhalten“ erfüllt ist. Dazu werden im Rahmen der Pilotierung des Testinstruments zwei Indikatoren geprüft: (1) Die Studierenden, die das Diagnose-Instrument bearbeiten, verstehen die Inhalte des Instruments (insbesondere die Multiple-Choice-Items) wie intendiert und (2) ihre Überlegungen passen zur Aufgabenstellung und beziehen sich auf die Qualität des Unterrichts bzw. auf die Professionalisierung als Lehrperson.

5.1. Methodisches Vorgehen

Zur Untersuchung des durch die Bearbeitung des Diagnose-Instrument auf Seiten der teilnehmenden Studierenden evozierten Verhaltens wurden $N = 7$ Think-Aloud-Interviews mit Bachelorstudierenden geführt. Um einen Einblick in die gedanklichen Prozesse während der Testbearbeitung zu erhalten (Konrad, 2010, S. 373) und möglicherweise auftretende Hürden für das Verständnis aufzudecken (Sandmann, 2014, S. 182), wurden die Teilnehmenden dazu aufgefordert, das Diagnose-Instrument zu bearbeiten und parallel dazu ihre Gedanken zu verbalisieren.

Die Interviews (von durchschnittlich 114 Minuten Länge) wurden transkribiert und zunächst event-basiert segmentiert. Ein neues Segment beginnt immer dann, wenn ein Sprecherwechsel vorliegt oder Äußerungen sich auf einen neuen Inhalt (z.B. eine neue Antwortoption in den Multiple-Choice-Antworten) beziehen. Diese Form der Segmentierung wurde gewählt, um Sinneinheiten für die Kodierung beizubehalten (vgl. Brückmann & Duit, 2014, S. 192). Anschließend wurden die einzelnen Aussagen in den Segmenten inhaltlich kodiert. Hierbei wurde unterschieden zwischen dem reinen Vorlesen von Inhalten des Testinstruments, Aussagen, die Reflexionen darstellen bzw. sich auf den Unterricht beziehen und sonstigen Aussagen (z. B. zur Methode der Think-Aloud-Interviews oder technischen Problemen). Insgesamt wurden sieben Kategorien differenziert (s. Tab. 1).

5.2. Ergebnisse

5.2.1. Indikator 1: Studierende verstehen die Multiple-Choice-Aufgaben wie intendiert

Im Rahmen der Kodierung wurden Aussagen, die Hinweise auf un- oder missverständliche Formulierungen im Textinstrument liefern (z. B. das wiederholte Vorlesen von Antwortoptionen) oder ein vom Intendierten abweichendes Verständnis in der Kategorie „Sonstiges“ mit zusätzlichem Vermerk kodiert. So konnten die Stellen im Diagnose-Instrument identifiziert werden, die potentiell ein anderes als das intendierte Verständnis hervorrufen. Dies betrifft insgesamt 44 Kodierungen zu 26 verschiedenen Aspekten, wobei auch solche Hinweise kodiert wurden, die verdeutlichen, dass Studierende über das korrekte Verständnis von Inhalten nachgedacht und diese aber wie intendiert verstanden haben (z. B. „Du solltest hinterfragen, woher die geschlechterspezifischen Zuschreibungen stammen, die du vertrittst, um sie zu überwinden. (..) Ah da geht es um Mädels und sowas, ne? […]“, S2, Z. 260). Eine nähere Betrachtung dieser Aspekte verdeutlicht, dass sich die in den verschiedenen Interviews identifizierten Hürden mit Ausnahme zweier Antwortoptionen nicht überschneiden, sodass angenommen werden kann, dass hier keine grundsätzlichen Verständnishürden vorliegen. Exemplarisch soll auf eine dieser Antwortoptionen eingegangen werden: In zwei der sieben Interviews wurde in Bezug auf die Antwortoption „Es wäre besser, vor dem Einstieg in das neue Thema auch die Tafel zu wischen […]“ aufgeführt, dass in den Videos nicht beobachtet werden konnte, dass die Tafel beschrieben wurde. Das ist so zwar korrekt, allerdings ist die beschriebene Tafel im Rahmen der Unterrichtsausschnitte sichtbar und wird von anderen Studierenden auch bemerkt („Und das mit der Tafel ist mir vorhin auch schon aufgefallen, dass die hätte gewischt wer-

Kategorien der Kodierung	Interview							Cohens κ
	S1	S2	S3	S4	S5	S6	S7	
Vorlesen	0,37	0,38	0,48	0,45	0,39	0,43	0,23	0,91
Beschreibung	0,02	0,06	0,04	0,02	0,06	0,04	0,1	0,59
Bewertung	0,03	0,06	0,02	0,04	0,07	0,07	0,07	0,84
Alternativen	0,04	0,09	0,11	0,06	0,13	0,09	0,08	0,62
Zustimmung/Ablehnung von Antwortoptionen	0,21	0,25	0,31	0,31	0,27	0,25	0,42	0,86
Andere Überlegungen zum Unterricht	0,07	0,07	0,02	0,04	0,04	0,06	0,07	0,72
Konstruktrelevante Aussagen (Kodierungen ohne Vorlesen und Sonstiges) insgesamt	0,37	0,53	0,49	0,46	0,58	0,51	0,74	0,80
Sonstiges	0,26	0,09	0,04	0,09	0,03	0,06	0,03	0,74

Tab. 1: Übersicht über Anteile der einzelnen Kategorien an allen Kodierungen in den Interviews sowie die Beurteilerübereinstimmung für die verschiedenen Kategorien (Cohens κ)

den sollen, genau. (.)“, S7, Z. 90). Die Antwortoptionen stellen also nicht für alle Studierenden Schwierigkeiten dar.

Die potentiellen Hürden in beiden Antwortoptionen sind darüber hinaus insofern unproblematisch, als dass sie beide „falsch“ sind (d. h. keine relevante Rückmeldung darstellen und daher keine Punkte für eine Auswahl vergeben werden). So entstehen für die Studierenden keinerlei Nachteile in der Testbewertung, wenn diese Antwortoptionen missverstanden anwählen oder auslassen, sodass keine Gefährdung für die Validität vorhanden ist.

5.2.2. Indikator 2 – Die Überlegungen der Studierenden passen zur Aufgabenstellung und beziehen sich auf die Qualität das Unterrichts bzw. auf die Professionalisierung als Lehrperson

In den geführten Think-Aloud-Interviews wurden zwischen 416 und 846 (Mittelwert (MW): 677, Standardabweichung (SD): 147) Kodierungen vorgenommen, wobei zwischen 23 % und 48 % (MW: 39 %, SD: 7 %) der Aussagen auf das Vorlesen von Inhalten des Testinstruments entfallen. Durchschnittlich sind etwa die Hälfte (MW: 53 %, SD: 11 %) der Kodierungen Reflexionen bzw. anderweitig auf den Unterricht bezogene Aussagen. In sechs der sieben Interviews sind jeweils unter zehn Prozent der Kodierungen der Kategorie „Sonstiges“ zuzuordnen; im siebten Interview (S1) sind es 26 %. Ein Überblick über die Anteile der verschiedenen Kategorien an allen Kodierungen je Interview findet sich in Tabelle 1.

Zusätzlich wurden 346 Segmente (14,5 % der vorliegenden Segmente) durch eine zweite Person kodiert. Für die unterschiedlichen Kategorien ergeben sich Übereinstimmungswerte zwischen beiden Kodierenden (bestimmt wurde jeweils Cohens κ) zwischen $\kappa = 0{,}59$ für die Kategorie „Beschreibung“ und $\kappa = 0{,}91$ für die Kategorie „Vorlesen“ (s. Tabelle 1). Die Werte der Beurteilerübereinstimmung liegen damit überwiegend in einem ausreichenden bis guten, teilweise auch sehr guten, Bereich; lediglich die Übereinstimmung in der Kategorie „Beschreibung“ ist mit einem $\kappa < 0{,}60$ als „mittelmäßig“ zu bewerten (Döring & Bortz, 2016, S. 346). Zur Bestimmung der Übereinstimmung für die Kategorie der „konstruktrelevanten Aussagen insgesamt“ wurden die Kodierungen aller Kategorien ohne „Vorlesen“ und „Sonstiges“ zusammengefasst.

Grundsätzlich beziehen sich alle Kodierungen, die nicht der Kategorie „Sonstiges“ zugeordnet werden, auf den durchgeführten Unterricht, die Inhalte des Unterrichts beziehungsweise auf die Professionalität der eigenen Person oder von Robert (dem fiktiven Mitpraktikanten, dessen Unterricht in den Videovignetten dargestellt ist). Diese Unterscheidung zeigt, dass der Anteil der Kodierungen in der Kategorie „Sonstiges“ mit zumeist unter 10 % sehr gering ist und sich der Großteil der Äußerungen auf den Unterricht beziehungsweise die Professionalisierung als Lehrperson bezieht.

In einem zweiten Schritt soll betrachtet werden, inwiefern die Überlegungen der Studierenden zur Aufgabenstellung passen. Im Rahmen der Think-Aloud-Interviews haben die Studierenden die im Diagnose-Instrument enthaltenen Texte laut vorgelesen. Hier ist der Bezug zum Diagnose-Instrument bzw. zu den einzelnen Aufgabenstellungen natürlich unmittelbar gegeben. Darüber hinaus stellen durchschnittlich 29 % (SD: 6 %) der Kodierungen Aussagen direkt auf die Antwortoptionen bezogene Zustimmungen oder Ablehnungen dar, teilweise mit Begründungen, auch hier ist jeweils ein Bezug zur Aufgabenstellung gegeben. Die übrigen Kodierungen enthalten zum Beispiel Bewertungen des Unterrichts, Vorschläge für alternative Handlungsoptionen oder fachliche Überlegungen, die durch den gesehenen Unterrichtsausschnitt bzw. die gelesenen Aufgabenstellungen und Antwortoptionen angestoßen wurden und daher einen unterschiedlich starken Bezug zur jeweiligen Aufgabe aufweisen.

Die Zweiteilung der Aufgaben jeweils in die Bewertung des Gesehenen und das Vorschlagen von alternativen Handlungsoptionen legen nahe, dass auch die Zahl der von den Studierenden geäußerten Bewertungen bei der Bearbeitung von Aufgaben zur Bewertung größer sein sollte, als bei der Bearbeitung der Aufgaben zu Alternativen und sich die Zahl der geäußerten Alternativen sich genau entgegengesetzt verhält. Insgesamt wurden – sowohl während der Bearbeitung von Aufgaben zu Bewertungen als auch zu Alternativen – deutlich mehr Alternativen geäußert, als Bewertungen. Das Verhältnis variiert bei den einzelnen Studierenden allerdings stark (zwischen 1,7 und 18,5 Mal so vielen Alternativen, wie Bewertungen während der Bearbeitung zu Aufgaben zur Bewertung und in sechs der Interviews zwischen 2,3 und 6,6 Mal so vielen Alternativen, wie Bewertungen während der Bearbeitung von Aufgaben zu Alternativen, eine Person äußert hier ausschließlich Alternativen und keine Bewertung). Dies könnte ein Hinweis darauf sein, dass die Überlegungen der Studierenden zwar wie oben beschrieben in Bezug auf ihren Inhalt gut zu den Aufgabenstellungen und betrachteten Unterrichtsausschnitten passen, allerdings während der Bearbeitung von Aufgaben zu Bewertungen nicht in Bezug auf die Art der Überlegung bzw. das angesprochen Reflexionselement. Das Vorschlagen von alternativen Handlungsoptionen liegt im Stufenmodell nach Nowak et al. (2019) auf einer höheren Stufe als das reine Bewerten und kann auf dieses aufbauen, indem alternative Handlungsoptionen für als problematisch bewertete Situationen vorgeschlagen werden. Überlegungen zu alternativen Vorgehensweisen können also unmittelbar an Bewertungen anschließen und insofern also auch bei der Bearbeitung von Aufgaben zu Bewertungen als angemessen bewertet werden.

Dass der Großteil der Kodierungen einen Bezug zum in den Unterrichtsausschnitten dargestellten Unterricht oder zur Professionalisierung der Studierenden bzw. des beobachteten Mitpraktikanten aufweisen, kann als Hinweis für die Validität des Diagnose-Instruments gewertet werden. Dass die Überlegungen der Studierenden nicht immer den durch die Aufgaben angesprochenen Stufen gemäß dem Stufenmodell zu Reflexion von Physikunterricht (Nowak et al., 2019) entsprechen, erscheint vor dem Hintergrund der inhaltlichen Passung der Überlegungen und der höheren Einordnung von alternativen Handlungsoptionen gegenüber Bewertungen unproblematisch.

6. **Zusammenfassung**

Die Diagnose und Förderung der Reflexionsfähigkeit von Physiklehramtsstudierenden ist vor dem Hintergrund ihrer Bedeutung für die Unterrichtsqualität und Professionalisierung von Lehrpersonen besonders relevant. Im Rahmen des Projekts ProfiLeP-Transfer wird daher ein geschlossenes Online-Diagnose-Instrument mit Assessment-Feedback entwickelt, welches neben einer Messung und Rückmeldung dieser Fähigkeit auch einen Anstoß zur Weiterentwicklung liefern soll, u. a. indem ein Fördermaterial bereitgestellt wird, in welchem Studierende drei Unterrichtsausschnitte angeleitet reflektieren können. Die entwickelten Materialien sollen evaluiert werden, um im Sinne des Argument-based-Approach empirisch abgesicherte Argumente für (bzw. gegen) die Validität der Interpretation der Testwerte als Maß für die Reflexionsfähigkeit und des Materials als Anstoß für eine weitere Professionalisierung zu sammeln.

Im Rahmen der Pilotierung des Diagnose-Instruments wurden sieben Think-Aloud-Interviews geführt, die einen Einblick in die Überlegungen der Studierenden während der Bearbeitung des Instruments liefern sollen. So soll geprüft werden, (1) ob die Studierenden die Inhalte des Testinstruments ohne Schwierigkeiten und wie intendiert verstehen und (2) ob sich die Überlegungen der Studierenden auf den Unterricht bzw. die Professionalisierung des beobachteten Mitpraktikanten oder der eigenen Person beziehen und zu den Aufgaben im Diagnose-Instrument passen. Die Interviews zeigen, dass es kaum geteilte Verständnishürden im Diagnose-Instrument gibt und die vorliegenden Hürden weitgehend unproblematisch erscheinen, da sie mit keinerlei Nachteilen in der Bearbeitung des Instruments verbunden sind. Außerdem zeigt sich, dass der Großteil der von den Studierenden geäußerten Überlegungen einen Bezug zum Unterricht oder zur Professionalisierung aufweisen (zählt man die vorgelesenen, auf den Unterricht bezogenen Antwortoptionen hinzu, weisen in sechs der sieben Interviews über 90 % der Kodierungen einen solchen Bezug auf, ohne die vorgelesenen Inhalte sind es in allen Interviews durchschnittlich etwa 52 %). Darüber hinaus beinhaltet ein bedeutender Anteil (durchschnittlich etwa 29 %) aller kodierten Äußerungen eine direkte Zustimmung oder Ablehnung zu gegebenen Antwortoptionen, diese Äußerungen weisen also unmittelbaren Bezug zur Aufgabenstellung auf. Da die übrigen konstruktrelevanten Äußerungen auf Unterricht bzw. die Professionalisierung bezogene Überlegungen sowie (Teile von) Reflexionen darstellen, passen auch sie inhaltlich zu den Aufgabenstellungen, auch wenn die geäußerten Überlegungen teilweise nicht derselben Stufe im Reflexionsmodell entsprechen, wie die Antwortoptionen im Instrument.

Die vorgestellte Think-Aloud-Erhebung liefert also zusätzlich zur hohen internen Konsistenz des Diagnose-Instruments (von $\alpha_{Cronbach} = 0{,}955$) Argumente für die Validität der Interpretation der Testwerte des Diagnose-Instruments als Maß für die Reflexionsfähigkeit.

7. **Ausblick**

In den kommenden Semestern sollen weitere Untersuchungen der Validität durchgeführt werden. Dazu werden die restlichen der oben genannten Übersetzungsschritte ausgehend vom Konstrukt bis hin zu Konsequenzen aus der Arbeit mit der entwickelten

Lernumgebung betrachtet. Insbesondere stehen Untersuchungen zum Zusammenhang der gemessenen Reflexionsfähigkeit der Lehramtsstudierenden mit anderen Fähigkeiten (z. B. dem fachdidaktischen Wissen oder der Reflexionsfähigkeit gemessen mit dem Performanztest in einem realitätsnäheren Setting) im Fokus.

Zusätzlich soll das entwickelte Assessment-Feedback für Studierende und Dozierende weiter erprobt und evaluiert werden.

Parallel dazu wird ein Transfer in die Lehrpraxis angestrebt, der durch die Bereitstellung der Materialien für interessierte Lehrende und Studierende unterstützt werden soll.

8. Literatur

Abels, S. (2011). LehrerInnen als "Reflective Practitioner": Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht. Wiesbaden: VS Verlag für Sozialwissenschaften.

Brückmann, M. & Duit, R. (2014). Videobasierte Analyse unterrichtlicher Sachstrukturen. In: Krüger, D., Parchmann, I. & Schecker, H. (Hrsg.), Methoden in der naturwissenschaftsdidaktischen Forschung. Berlin, Heidelberg: Springer Berlin Heidelberg, S. 189-201.

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: Hume, A., Cooper, R., Borowski, A. (Hrsg.) Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science. Singapore: Springer. DOI: https://doi.org/10.1007/978-981-13-5898-2_2

Casanova, D., Alsop, G. & Huet, I. (2021). Giving away some of their powers! Towards learner agency in digital assessment and feedback. Research and Practice in Technology Enhanced Learning 16(20). DOI: https://doi.org/10.1186/s41039-021-00168-6

Dickmann, M. (2016). Messung Von Experimentierfähigkeiten. Validierungsstudien zur Qualität eines Computerbasierten Testverfahrens. DOI: https://doi.org/10.5281/zenodo.168540

Döring, N. & Bortz, J. (2016). Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-41089-5

Hattie, J. & Timperley, H. (2007). The Power of Feedback. In: Review of Educational Research 77(1), S. 81-112. DOI: https://doi.org/10.3102/003465430298487

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

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement 50(1), S. 1-73. DOI: https://doi.org/10.1111/jedm.12000

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

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

Klempin, C. (2019). Reflexionskompetenz von Englischlehramtsstudierenden im Lehr-Lern-Labor-Seminar: Eine Interventionsstudie zur Förderung und Messung. Stuttgart: J. B. Metzler. DOI: https://doi.org/10.1007/978-3-476-05120-2

Kultusministerkonferenz (KMK) (2004). Standards für die Lehrerbildung: Bildungswissenschaften. Beschluss der Kultusministerkonferenz vom 16.12.2004 i. d. F. vom 16.05.2009. URL: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf

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

Kulgemeyer, C., Kempin, M., Weißbach, A., Borowski, A., Buschhüter, D., Enkrott, P., Reinhold, 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), S. 3035-3057. DOI: https://doi.org/10.1080/09500693.2021.2006820

Kulgemeyer, C., Kempin, M. & Weißbach, A. (2021). Entwicklung von Professionswissen und Reflexionsfähigkeit im Praxissemester. In: Habig, S. (Hrsg.), Naturwissenschaftlicher Unterricht und Lehrerbildung im Umbruch? Gesellschaft für

Didaktik der Chemie und Physik online Jahrestagung 2020. Essen: Duisburg-Essen, S. 262-265.

Nowak, A., Ackermann, P. & Borowski, A. (2018). Rahmenthema "Reflexion" im Praxissemester. In Borowski, A., Ehlert, A. & Prechtl, H. (Hrsg.), PSI-Potsdam - Erlebnisbericht zu den Aktivitäten im Rahmen der Qualitätsoffensive Lehrerbildung (2015-2018), S. 217-230. URL: https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/deliver/index/docId/41454/file/pblbf01.pdf

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

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

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

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

Schön, D. A. (1983). The Reflective Practitioner. How Professionals Think in Action. New York: Basic Books.

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: Maurer, C. (Hrsg.), Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. Gesellschaft für Didaktik der Chemie und Physik, Jahrestagung in Kiel 2018. Regensburg: Universität Regensburg, S. 317-320.

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

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

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 in Berlin 2015. Regensburg: Universität Regensburg, S. 284-286.

Wyss, C. (2008). Zur Reflexionsfähigkeit und -praxis der Lehrperson. Bildungsforschung 5(2), S. 1-15.

Wyss, C. (2013). Unterricht und Reflexion: Eine Mehrperspektivesche Untersuchung der Unterrichts- und Reflexionskompetenz von Lehrkräften. Münster u.a.: Waxmann.

Wyss, C. & Mahler, S. (2021). Mythos Reflexion. Theoretische und praxisbezogene Erkenntnisse in der Lehrer*innenbildung. journal für lehrerInnenbildung 21(1), S. 16-25. DOI: https://doi.org/10.35468/jlb-01-2021-01

Förderung

Das Projekt *ProfiLeP-Transfer* wird gefördert vom BMBF unter dem Förderkennzeichen 16PK19005B.