Die Messung der Fähigkeit zur Unterrichtsplanung im Fach Physik mit Hilfe eines standardisierten Performanztests

Jan Schröder¹ · Josef Riese¹ · Christoph Vogelsang² · Andreas Borowski³ · David Buschhüter³ · Patrick Enkrott³ · Maren Kempin⁴ · Christoph Kulgemeyer⁴ · Peter Reinhold⁵ · Horst Schecker⁴

Eingegangen: 26. Januar 2020 / Angenommen: 9. Juli 2020 / Online publiziert: 27. Juli 2020
© Der/die Autor(en) 2020

Zusammenfassung

Die Unterrichtsplanung stellt eine zentrale berufliche Anforderung an (Physik-)Lehrkräfte dar. Bereits im Lehramtsstudium sollen grundlegende Fähigkeiten zur Unterrichtsplanung erworben und im Rahmen von schulpraktischen Phasen erprobt bzw. erweitert werden. Unklar ist jedoch, inwieweit das professionelle Wissen, das Lehramtsstudierende in ihren fachlichen, fachdidaktischen und erziehungswissenschaftlichen Studienteilen erwerben, tatsächlich bei der Planung von Fachunterricht genutzt wird. Ebenso gibt es nur punktuelle Erkenntnisse über die Entwicklung der Fähigkeit zur Unterrichtsplanung im Lehramtsstudium. Um diese Fragen zu klären, sind Messverfahren erforderlich, die die Fähigkeiten zur Unterrichtsplanung valide erfassen können und die gleichzeitig ökonomisch auswertbar sind, um Fallzahlen realisieren zu können, die auch kleine Effekte auflösen. Bisherige Ansätze, die reale Handlungsproben der Unterrichtsplanung (z. B. schriftliche Unterrichtsentwürfe) untersuchen, stehen vor der methodischen Herausforderung, dass die Generierung eines Fähigkeitsmaßes höchst anspruchsvoll ist, da reale Planungen individuell sehr unterschiedlich angelegt sind und vielen Randbedingungen unterliegen, die statistisch kaum kontrollierbar sind. In diesem Beitrag wird daher ein sogenannter Performanztest zur Erfassung der Fähigkeit zur Unterrichtsplanung vorgestellt und diskutiert. Ein Performanztest ist ein Verfahren, in dem in einem realitätsnahen Setting Handlungen unter standardisierten Rahmenbedingungen simuliert werden. Neben der theoriegeleiteten Entwicklung des Instruments und Vorstudien zur Sammlung von Validitätsargumenten werden auch erste Ergebnisse dargestellt, die aus Erhebungen vor und nach einem Praxissemester an vier Universitäten in drei Bundesländern resultieren ($N=174$ Unterrichtsplanungen). Dabei stellt sich das Testverfahren als ausreichend sensitiv heraus, um für $N=68$ Testpersonen (ca. 20% Dropout) erwartungstreu eine signifikante Zunahme des Testscores während des Praxissemesters, welches als Lerngelegenheit angenommen werden kann, zu messen.

✉ Jan Schröder
jkschroeder@physik.rwth-aachen.de

Josef Riese
riese@physik.rwth-aachen.de

Christoph Vogelsang
vogelsang@plaz.upb.de

Andreas Borowski
Andreas.borowski@uni-potsdam.de

David Buschhüter
buschhue@uni-potsdam.de

Patrick Enkrott
enkrott@uni-potsdam.de

Maren Kempin
Maren.kempin@uni-bremen.de

Christoph Kulgemeyer
kulgemeyer@physik.uni-bremen.de

Peter Reinhold
Peter.reinhold@upb.de

Horst Schecker
schecker@physik.uni-bremen.de

- ¹ Lehrstuhl für Experimentalphysik I A und I. Physikalisches Institut, Lehr- und Forschungsgebiet Didaktik der Physik und Technik, RWTH Aachen University, Aachen, Deutschland
- ² Zentrum für Bildungsforschung und Lehrerbildung – PLAZ Professional School, Universität Paderborn, Paderborn, Deutschland
- ³ Institut für Physik und Astronomie, Lehrstuhl Didaktik der Physik, Universität Potsdam, Potsdam-Golm, Deutschland
- ⁴ Institut für Didaktik der Naturwissenschaften, Abt. Physikdidaktik, Universität Bremen, Bremen, Deutschland
- ⁵ Department Physik, Didaktik der Physik, Universität Paderborn, Paderborn, Deutschland

Schlüsselwörter Lehramtsstudierende · Physik · Unterrichtsplanung · Performanztest

Measuring the Skills of Lesson Planning in the Subject of Physics Using a Standardized Performance Assessment

Abstract

For every (physics-)teacher, lesson planning is a crucial professional demand. Therefore, basic skills regarding lesson planning should already be acquired during university courses and expanded during school internships. However, it is still unclear to what extent the professional knowledge, which students acquire in terms of content knowledge, pedagogical content knowledge and pedagogical knowledge, is used when planning a lesson. Furthermore, there is little evidence regarding the development of lesson planning skills during university teacher education. In order to evaluate such a development, assessments capable of a valid measurement and allowing economic scoring are required. Recent approaches, evaluating real lesson planning in action (e.g. from written lesson drafts), are lacking the capacity to produce comparable scores of planning skills. This is due to the fact that real lesson plans are very individual and underlie numerous external conditions (e.g., different topics and classes), which cannot be accounted for statistically. In this paper, we present the development and trial use of a performance assessment as a new method of measuring the skills to plan a physics lesson in a standardized and authentic way. First, we explain the theory-based development of the instrument and cover some preliminary studies regarding validity. Furthermore, results from the test use before and after a long-term-internship at four German universities from three federal states are presented ($N=174$ lesson drafts). As expected, a significant increase in the test score during the internship was observed for $N=68$ test persons, pointing out the instructional sensitivity of the assessment.

Keywords student teachers · physics · lesson planning · performance assessment

Einleitung

Problemstellung

Die Unterrichtsplanung bzw. -vorbereitung gehört zu den grundlegenden Tätigkeiten von Lehrkräften (Carlson und Daehler 2019; KMK 2019). Ein erheblicher Teil der beruflich genutzten Zeit wird hierfür aufgebracht (vgl. Hardwig und Mußmann 2018), wobei davon ausgegangen wird, dass die Fähigkeit, Unterricht zu planen, grundsätzlich lehr- und erlernbar ist (z.B. Plöger 2008). Dementsprechend werden entsprechende Lerngelegenheiten in den fachdidaktischen und erziehungswissenschaftlichen Teilen des Lehramtsstudiums angeboten, damit Studierende die geforderten grundlegenden Kompetenzen und Fähigkeiten im Bereich der Planung und Gestaltung von Unterricht (für das Fach Physik vgl. KMK 2019, S. 50) erwerben können. Inwieweit dies gelingt, ist in den naturwissenschaftlichen Fächern bislang jedoch wenig systematisch untersucht worden. Während das Professionswissen (Fachwissen (FW), fachdidaktisches Wissen (FDW) und erziehungswissenschaftliches Wissen (EW) (vgl. Shulman 1986)) von Lehrkräften in den Naturwissenschaften in den letzten Jahren in vielen Forschungsprojekten erfasst worden ist (vgl. z.B. Riese und Reinhold 2012; Tepner et al. 2012; Lederman und Abell 2014; Großschedl et al. 2014; Sorge et al. 2019) und demzufolge sowohl Messverfahren als auch Hinweise zur Wirksamkeit

des Lehramtsstudiums hinsichtlich des Professionswissens vorliegen, befassen sich nur wenige Arbeiten (für die Primarstufe z.B. Baer et al. 2011) explizit mit der Messung von Wissenselementen, Kompetenzen oder Fähigkeiten zur Unterrichtsplanung im Lehramtsstudium.

Aber nicht nur Wirksamkeitsstudien hinsichtlich des Erwerbs von Kompetenzen und Fähigkeiten im Bereich der Unterrichtsplanung erfordern entsprechende Messverfahren. Auch die Beantwortung von grundlegenden Fragen der Bedeutung des an der Universität erworbenen Professionswissens für das spätere Lehrerhandeln setzt Messverfahren für einzelne Elemente des Lehrerhandelns – wie der Unterrichtsplanung – voraus. So ist die Rolle des professionellen Wissens für die Handlungsqualität in realen Unterrichtssituationen nach wie vor unklar: Weder Vogelsang (2014), Cauet et al. (2015), Keller et al. (2016) noch Liepertz und Borowski (2019) fanden einheitliche bzw. systematischen Zusammenhänge zwischen dem professionellen Wissen von Physiklehrkräften bzw. Lehramtsstudierenden und unterschiedlichen Facetten videographierter Unterrichtsqualität. Einen anderen Ansatz zur Klärung der Rolle des Professionswissens für das Lehrerhandeln verfolgen Kulgemeyer und Riese (2018), die die Qualität des Handelns von angehenden Physiklehrkräften in standardisierten Erklärsituationen analysieren und für diese sehr spezifische Situation eine positive Wirkung des Professionswissens auf die Qualität von mündlichen Erklärungen nachweisen kön-

nen. Es erscheint somit vielversprechend, die Bedeutung des Professionswissens zunächst für ausgewählte zentrale Anforderungs- bzw. Fähigkeitsbereiche von Lehrkräften zu untersuchen, für die eine Wirkung des Professionswissens erwartet werden kann. Dies erfordert zunächst die Entwicklung von Messverfahren zur Erfassung entsprechender Fähigkeiten, z. B. im Bereich der Unterrichtsplanung.

Zielformulierung

An dieser Stelle setzt die vorliegende Arbeit an, indem ein Performanztest zur Messung der Fähigkeit zur Unterrichtsplanung für (angehende) Physiklehrkräfte der Sekundarstufe in einem standardisierten Handlungssetting entwickelt wird. Als Performanztest wird dabei ein Messverfahren bezeichnet, welches die Erfassung handlungsnaher Fähigkeiten in simulierten Handlungssituationen unter standardisierten Rahmenbedingungen erlaubt. Das vorgestellte Messverfahren fokussiert damit auf eine zentrale Handlungsanforderung für Lehrkräfte und soll insbesondere die Konfundierung möglicher Ergebnisse mit unkontrollierten Randbedingungen minimieren. Dieser Beitrag beschreibt die Entwicklung des Performanztests zur Unterrichtsplanung sowie die Sammlung von Validitätsargumenten. Hierzu werden auch Ergebnisse herangezogen, die aus Erhebungen an vier Universitäten in drei Bundesländern ($N=174$ Unterrichtsplanungen, $N=68$ Personen im Längsschnitt) jeweils vor und nach einem Praxissemester resultieren, welches als Lerngelegenheit angenommen wird, um die Sensitivität des Testverfahrens und Zusammenhänge zum Professionswissen zu überprüfen.

Diese Arbeit ist eingebettet in den Projektverbund *ProfiLe-P+* (Vogelsang et al. 2019), in dem die Wirkung des Lehramtsstudiums in Bachelor- und Masterstudiengängen im Lehramt Physik im Längsschnitt untersucht wird. Darüber hinaus werden in *ProfiLe-P+* Zusammenhänge zwischen zentralen Bereichen des Professionswissens (Fachwissen Physik, physikdidaktisches Wissen und erziehungswissenschaftliches Wissen) und Ausschnitten des Lehrerhandelns (neben der Unterrichtsplanung reales Erklären physikalischer Sachverhalte und die Reflexion von Unterricht) untersucht. Darüberhinausgehende Darstellungen zum *ProfiLe-P+*, seiner theoretischen Anlage sowie den allgemeinen Zielen finden sich bei Vogelsang et al. (2019).

Theoretischer Hintergrund

Unterrichtsplanung im Fach Physik

Obwohl die Planung von Unterricht als eine der Kernaufgaben von Lehrkräften angesehen wird (Carlson und

Daehler 2019), besteht keine einheitliche Definition des Begriffs Unterrichtsplanung. Sandfuchs (2009) begreift unter der *Unterrichtsplanung* oder der *Unterrichtsvorbereitung* zunächst allgemein „alle dem Unterricht vorausgehenden Maßnahmen [...], die das Lehren und Lernen im Unterricht selbst optimieren sollen“ (S. 512). Je nach Autorin bzw. Autor werden die Begriffe allerdings nicht immer synonym verwendet und reichen in ihrer Tragweite von der groben inhaltlichen Planung eines Schuljahres bis zur detaillierten Planung einer einzelnen Unterrichtsstunde.

Planung lässt sich nach Shavelson und Stern (1981) als ein zirkulärer Prozess beschreiben, bei dem zunächst Vorbedingungen (z. B. Lernvoraussetzungen) analysiert werden müssen, bevor darauf aufbauend über Unterrichtsaktivitäten entschieden und nach der Durchführung darüber reflektiert werden kann. Als Kernelement der Planung muss ein Unterrichtsablauf entwickelt und eine antizipierte Abfolge von geeigneten Handlungsskripten ausgewählt werden (Stender et al. 2017). Unter einem Handlungsskript kann eine Art Drehbuch für den Unterricht im Sinne einer kognitiv abgespeicherten Abfolge von Handlungsschritten verstanden werden kann, welche durch zirkulär durchgeführte Unterrichtsplanungen und Abspeicherung von Unterrichtsabläufen entstehen (Stender 2014, S. 24). Unterrichtsplanung kann in diesem Sinne als mentales Handeln zur Lösung eines Problems verstanden werden. Dieses Problem ist offen und es müssen teilweise widersprüchliche Zielanforderungen integriert werden (z. B. Strukturierung des Unterrichts im Klassenverband und Anpassung des Unterrichts an heterogene Voraussetzungen). Unterrichtsplanung hat daher theoretisch auch Ähnlichkeiten zu Designaufgaben (vgl. Aprea 2014). Die Planung von Unterricht verläuft auch hierbei nicht linear, sondern zirkulär mit vielen Interdependenzen und Rückbezügen. Sie verbindet dabei die zwei übergeordneten Zielklassen der *Legitimation* und der *Kreation* (Vogelsang und Riese 2017). Es müssen zunächst mögliche Unterrichtsaktivitäten vorausgedacht, also kreiert werden. Darüber hinaus sollte auch eine Legitimation der erdachten Handlungsmöglichkeiten erfolgen, um die Berücksichtigung von curricularen Vorgaben oder didaktischen Prinzipien zu belegen, was insbesondere in Prüfungssituationen der zweiten Ausbildungsphase bedeutsam ist.

Unterrichtsplanung soll ferner der Lehrkraft flexibles Unterrichtshandeln ermöglichen (Klafki 2007), da der finale Verlauf des Unterrichts aufgrund vieler Unwägbarkeiten nicht exakt vorausgedacht werden kann. Daher sollten auch alternative Handlungsoptionen antizipiert werden, um geeignet auf unterrichtliche Bedingungen reagieren zu können.

Zur Frage, welche Elemente auf theoretischer Ebene Bestandteil von Unterrichtsplanungen sein sollen, existieren in der Praxis viele Modelle, Ratgeber und Lehrbücher, wobei

insbesondere der bildungstheoretische Ansatz der *Didaktischen Analyse* (Klafki 1969) und das lerntheoretische *Strukturmodell der Berliner Schule* (Heimann et al. 1969) häufig rezipiert werden. Neben diesen beiden Ansätzen wird im naturwissenschaftlichen Bereich hinsichtlich einer qualitativ hochwertigen Unterrichts- bzw. Reihenplanung häufig auf das Modell der *Didaktischen Rekonstruktion* (Kattmann et al. 1997) verwiesen, das beide Modelle zusammenführt bzw. erweitert und zur Rekonstruktion eines Themengebietes verwendet werden kann. Grundpfeiler der Unterrichtsplanung in Anlehnung an die Didaktische Rekonstruktion sind die fachliche Klärung des Inhaltes sowie die Beachtung von Schülervorstellungen, welche durch ein iteratives Vorgehen didaktisch zu Unterrichtselementen strukturiert werden. Aufgrund der Komplexität der fachlichen Inhalte wird die ursprüngliche Sachstruktur für den Unterricht elementarisiert, es müssen also im Sinne Klafkis die elementaren Ideen herausgestellt und komplexitätsreduziert in unterrichtstaugliche Sinneinheiten gebracht werden. Die im Berliner Modell postulierte Interdependenz der zu beachtenden Variablen „Thematik“, „Methoden“, „Medien“ und „Intention“ wird innerhalb der Didaktischen Rekonstruktion weiterhin beibehalten, allerdings das „Primat der Didaktik“ (Klafki 1979) anerkannt, weshalb Entscheidungen zu Bildungsinhalten denen über Medien und Methoden vorangestellt werden (Kattmann et al. 1997).

Erfassung von Fähigkeiten zur Unterrichtsplanung

Die Fähigkeit von Lehrkräften im Bereich der Unterrichtsplanung ist bislang vergleichsweise wenig empirisch untersucht worden (z. B. Jacobs et al. 2008). Nach Wernke und Zierer (2017) ist sie sogar „ein in Vergessenheit geratener Kompetenzbereich“.

Einige Studien untersuchen das Vorgehen von Lehrkräften bei der Unterrichtsplanung mit Hilfe der Methode des Lauten Denkens (z. B. Tebrügge 2001; Haas 1998; Bromme 1981). Die Ergebnisse zeigen, dass der Prozess der Unterrichtsplanung fachabhängig und individuell unterschiedlich verläuft. Dennoch werden übergreifende Planungsmuster beschrieben, die den angenommenen zirkulären Verlauf von Planungen bestätigen (vgl. Gassmann 2013). Darüber hinaus stellte Bromme (1981) für Mathematiklehrkräfte bspw. heraus, dass sich der Planungsprozess inhaltlich an der „Auswahl von mathematischen Aufgaben und als Antizipation ihrer Bearbeitung im Zuge des Unterrichts orientiert“ (S. 192).

Ergebnisse zur Planungsdokumentation, d. h. zum Planungsprodukt, das von Lehrkräften erstellt wird, zeigen, dass Planungen selten und wenn dann eher in kurzer Form schriftlich fixiert werden. Ausführliche schriftliche Unterrichtsplanungen werden daher auch als Ausbildungsartefakte betrachtet (Seel 2011). In Befragungen oder Studien auf

Basis von Arbeitstagebüchern wurde festgestellt, dass erfahrene Lehrkräfte für die Planung einer Unterrichtsstunde durchschnittlich zwischen 20–30 min aufwenden, wobei die Daten sich auf unterschiedliche Fächer und Schulformen beziehen (z. B. Hübner und Werle 1997; Wengert 1989). Insgesamt untersuchen diese Arbeiten Unterrichtsplanung eher deskriptiv, machen selten Aussagen zur Planungsqualität und fokussieren daher nicht die Planungsfähigkeit der Lehrkräfte. Dieses Ziel wird eher in Studien verfolgt, die sich auf angehende Lehrkräfte in Ausbildungskontexten beziehen.

Die Messung der Fähigkeit erfolgt dabei meist durch zwei verschiedene Ansätze. Ein testökonomischer Ansatz besteht in der Verwendung von schriftlichen Vignetten, bei denen die Testpersonen ihr Wissen bzw. ihre Kompetenzen im Bereich der Unterrichtsplanung demonstrieren sollen (z. B. Blömeke et al. 2008; Baer et al. 2011). Solche Vignetten erlauben einen hohen Grad an Standardisierung und die Antworten der Testpersonen lassen sich anhand eines konkreten und gut objektivierbaren Erwartungshorizontes bewerten. Darüber hinaus bieten sie einen hohen Grad an Inhaltsvalidität, da das zu überprüfende Wissen meist normativ vorgegeben ist und durch das Testformat direkt überprüft werden kann, ob dieses vorhanden ist. Blömeke et al. (2008) nutzten beispielsweise eine Vignette, in der eine angehende Mathematik Lehrkraft zunächst Kriterien zur Beurteilung einer Unterrichtsplanung entwickeln soll und diese anschließend zur Beurteilung anwenden muss. Die querschnittliche Vergleichsstudie konnte verschiedene Schwerpunkte innerhalb der Bearbeitung der Vignette zwischen Studierenden verschiedener Länder bzw. aus unterschiedlichen Ausbildungssystemen diagnostizieren, so wurde beispielsweise besonders häufig über Lernziele oder den mathematischen Inhalt gesprochen. Darüber hinaus konnten vergleichende Analysen der Antworten von Studierenden zu Beginn und Ende der Ausbildung zeigen, dass die Verwendung von Fachvokabular und die Anzahl verwendeter Kriterien zur Bewertung des Unterrichts zunimmt, woraus sich ein Effekt der Ausbildungsphase schließen lässt. Einen positiven Zusammenhang mit der universitären Ausbildung konnten ebenfalls Baer et al. (2011) für schweizerische Studierende der Primarstufe feststellen. Über drei Messzeitpunkte hinweg nimmt das Planungswissen, erfasst über Vignetten, in den Dimensionen „Didaktik“ und „Diagnostik“ bei Studierenden deutlich zu. Stender et al. (2017) erfassten mit einem adaptiven Online-Test auf Basis von drei Vignetten die Handlungsskripte von Physik Lehrkräften. Die offenen Antworten der Befragten bzgl. verschiedener Planungsbereiche wurden dahingehend ausgewertet, inwiefern sie potentiell eine höhere Unterrichtsqualität ermöglichen. Dabei konnten signifikante Zusammenhänge zwischen der Qualität der Skripte und dem professionellen Wissen der Lehrkräfte gefunden werden.

Trotz der ökonomischen Vorteile solch schriftlicher, objektiver Vignettestests ist zu bedenken, dass kein bzw. nur in Ansätzen reales und authentisches Planungshandeln erforderlich ist, also kein vollständiger Unterricht eigenständig geplant werden muss, um einen hohen Testscore zu erreichen. Daher besteht ein anderer Ansatz in der Analyse schriftlicher „Realplanungen“ (z. B. Hasenkamp et al. 2016; Gassmann 2013), die während des Vorbereitungsdienstes oder im Studium angefertigt werden. König (2019) fokussiert bspw. bei Lehrkräften des Faches Deutsch den spezifischen Aspekt der didaktischen Adaptivität und kann entlang des Referendariats einen signifikanten Zuwachs ebenjenes Aspektes und auch Zusammenhänge zu relevanten Teilen des Professionswissens feststellen.

Bei einer allgemeinen, umfassenden Betrachtung der Planungsfähigkeit mittels realer Planungsprodukte zeigen sich jedoch Probleme mit der Messgenauigkeit und der Vergleichbarkeit der Scores, die aus der fehlenden Standardisierung resultieren, wodurch die Messung einer möglichen Veränderung der Planungsfähigkeit im Laufe der Ausbildung deutlich erschwert wird (Hasenkamp et al. 2016). Die Herausforderung besteht dabei in der Komplexität von ganzheitlich geplantem Unterricht, da viele Interdependenzen einzelner Aspekte zu betrachten sind und die Vergleichbarkeit zwischen den Planungen aufgrund einer notwendigen Adaption an unterschiedliche Lerngruppen im Sinne der Didaktischen Rekonstruktion (Kattmann et al. 1997), kaum vergleichbaren schulischen Randbedingungen und unterschiedlichen zu planenden Fachinhalten erschwert wird. Hinzu kommen uneinheitliche Vorgaben der jeweiligen Ausbildungsinstitution, nach denen die Unterrichtsentwürfe anzufertigen sind. Die Analyse und Bewertung realer Unterrichtsentwürfe ist daher zwar ein deutlich authentischeres und ökologisch valideres Verfahren zur Einschätzung der Planungsfähigkeit. Das Bilden eines objektiven und reliablen Maßes ist methodisch jedoch erheblich anspruchsvoller im Vergleich zur Auswertung klassischer Wissens- bzw. Kompetenztests. In diesem Beitrag wird daher als ein dritter Ansatz die Verwendung von Performanztests betrachtet, die das Potenzial haben, die jeweiligen Vorteile der Standardisierung bzw. der Authentizität bei der Erfassung der Fähigkeit zur Unterrichtsplanung zusammenzuführen.

Performanztests

Miller (1990) unterscheidet verschiedene Formen der Erfassung von Kompetenz bzw. Kompetenzaspekten. Neben der schriftlichen Abfrage von Wissen („test for knowledge“) und der Bewertung des realen Berufshandeln („test for action“) existieren noch weitere Methoden, Fähigkeiten zu testen. Er beschreibt einen traditionell in der medizinischen Ausbildung verwendeten *Performanztest* („test for perfor-

mance“), bei dem Testpersonen reale Handlungssituationen unter kontrollierten und standardisierten Rahmenbedingungen bewältigen müssen. So kann z. B. mithilfe eines trainierten Schauspielers ein „standardisierter Patient“ simuliert werden, an dem Testpersonen z. B. ein Anamnesegespräch oder Routineuntersuchungen demonstrieren können. Durch die Standardisierung der Handlungssituation resultiert gegenüber der Beobachtung realer Berufssituationen eine erhöhte Vergleichbarkeit und aufgrund des authentischen Handelns eine höhere Validität als bei klassischen Wissenstests. Obwohl die eigentliche Durchführung eines Performanztests durch die standardisierten Patienten relativ einfach erscheint, besteht die Schwierigkeit bei der Bewertung der demonstrierten Leistungen. Hierzu wird oft auf einen Bewertungsbogen zurückgegriffen, der einer Checkliste oder einem Erwartungshorizont ähnelt und somit eine möglichst objektive und reproduzierbare Bewertung ermöglicht (z. B. Harden et al. 1975; Sturpe et al. 2013). Kulgemeyer und Riese (2018) nutzten einen Performanztest in der Domäne der Lehrerbildung und entwickelten ein Instrument zur Standardsituation des dialogischen Erklärens von physikalischen Sachverhalten. Dabei konnten sie positive Einflüsse des Professionswissens auf die gemessene Erklärfähigkeit feststellen. Somit scheinen Performanztests grundsätzlich geeignet zu sein, eine Fähigkeit in Standardsituationen des Lehrberufs zu messen und diese mit Ergebnissen aus klassischen Wissenstests zu verbinden. Neben üblichen Validitätsbetrachtungen resümieren Bartels et al. (2019) für Performanztests dabei drei gesonderte Qualitätsaspekte. Um als Simulation einer realen Handlungssituation zu dienen, muss (1) die Performanzsituation möglichst authentisch wirken und (2) für die Testpersonen eine reale Interaktion mit der Testsituation ermöglichen. Zudem sollte (3) die Relevanz des Performanztests für die Testpersonen ersichtlich sein, weshalb die Testsituation einen wichtigen Aspekt des professionellen Alltags unter standardisierten Bedingungen darstellen sollte.

Bezogen auf die oben beschriebenen Methoden zur Erfassung der Planungsfähigkeit von angehenden Lehrkräften würde man diese in der Kategorisierung von Miller (1990) wie folgt einordnen: Verfahren zur Erfassung des Planungswissens mit Hilfe von Vignetten sind eher als „test for knowledge“ anzusehen, während Analysen von realen Planungsdokumenten eher „tests for action“ entsprechen. Ansätze zur Erfassung der Planungsfähigkeit mit Hilfe eines Performanztests im obigen Sinne liegen bisher nicht vor. Diese Strukturierung möglicher Verfahren korrespondiert mit gängigen Modellen zur Beschreibung des professionellen Wissens und der professionellen Kompetenz von Lehrkräften. Bspw. kann die Fähigkeit zur Planung von Unterricht auch als „enacted pedagogical content knowledge (ePCK)“ im Sinne des „Refined Consensus Model of PCK“ verstanden werden (Alonzo et al. 2019), während das über

Vignetten erfasste Planungswissen eher dem „personal PCK (pPCK)“ oder – je nach Test – „collective PCK“ entspricht. Im Kontinuumsmodell professioneller Kompetenz (Blömeke et al. 2015) entspricht erfasstes Planungswissen dem „cognition“-Anteil von Kompetenz, während Planungsfähigkeit, erfasst durch einen Performanztest, eher dem „performance“-Anteil entspricht.

Forschungsfragen

Ausgehend von den in Abschn. 2.2 diskutierten Problemen bzw. Limitationen, die einerseits aus der Analyse realer Unterrichtsplanungen und andererseits aus der Durchführung schriftlicher Vignettentests resultieren, wird in der vorgestellten Studie ein alternativer, standardisierter Ansatz zur Erfassung der Planungsfähigkeit im naturwissenschaftsdidaktischen Bereich vorgeschlagen. In dem neuen Testverfahren soll eine möglichst ganzheitliche Unterrichtsplanung und damit ein reales Planungsprodukt in einem fiktiven, standardisierten Planungssetting erstellt werden. Dabei werden zunächst im Sinne einer Weiterentwicklung bisheriger Erhebungsmethoden ein Performanztest sowie ein Verfahren zur Generierung eines Fähigkeitsmaßes gemäß der in Abschn. 2.3 dargestellten Grundidee und den diskutierten Qualitätskriterien entwickelt. Darüber hinaus wird der Frage nachgegangen, inwieweit das entwickelte Testverfahren die Fähigkeit zur Unterrichtsplanung von Physiklehramtsstudierenden objektiv, reliabel und valide erfassen kann und eine authentische Handlungssituation darstellt. Dazu werden im Sinne der argumentbasierten Validierung (Kane 2013) Argumente gesammelt, die getroffene Interpretationen, beispielsweise im Hinblick auf die kausale Erklärung und Bewertung der gezeigten Ergebnisse (vgl. Hartig et al. 2012), stützen sollen. Das Validitätskonzept beruht also auf der Sammlung von zusammenhängenden Annahmen und Aussagen, die die angestrebten Testwertinterpretationen stützen (sog. Interpretations-Nutzungs-Argumente), und Validitätsargumenten, die die Plausibilität, Klarheit und Kohärenz dieser Annahmen überprüfen (Kane 2013). Eingedenk bisher genutzter Verfahren zur Messung der Planungsfähigkeit soll der entwickelte Performanztest so hinsichtlich bestimmter Aspekte untersucht werden, bei denen sich das Testformat (nachteilig) auf die Testgüte auswirken könnte.

Die Betrachtung realer Unterrichtsplanungen stellt ein ökologisch valides Verfahren dar, da Einblicke in reale Planungsprodukte gewährt werden. Durch die standardisierte Ausgangssituation, die durch den Ansatz des Performanztests geschaffen wird, werden einschränkende Rahmenbedingungen geschaffen (z.B. durch zeitliche Vorgaben bei der Testung, Präsentation einer fiktiven Lerngruppe statt ei-

ner realen). Daher stellt sich die Frage, ob die ökologische Validität hinreichend bestehen bleibt:

Forschungsfrage 1

Inwieweit führt der durch die Standardisierung bedingte fiktive Charakter der Planungssituation zu einem deutlich anderen Planungsprodukt, als es die Testpersonen bei der Planung einer realen Unterrichtsstunde anfertigen würden, die tatsächlich nach der Planung auch durchgeführt wird?

Sofern das Testverfahren eine authentische, also ökologisch valide Messung ermöglicht und das Planungsprodukt der standardisierten Situation demnach hinreichend dem Produkt bei realen Unterrichtsplanungen ähnelt, stellt sich die Frage nach der Aussagekraft des generierten Testscores und ob ein Zusammenhang zwischen einer augenscheinlichen Qualität der Unterrichtsplanung und dem generierten Testscore besteht. Bei dem Einsatz von Vignetten-Tests zum Planungswissen geht ein hoher Testscore per Konstruktion mit einem höheren Planungswissen einher. Für das Format des Planungsperformanztests ist jedoch zu klären, ob ein hoher Gesamtscore, der im Detail aus ganz unterschiedlichen Einzelindikatoren gebildet wird, auch mit dem Gesamteindruck einer qualitativ hochwertigen Planung einhergeht. Die folgende Forschungsfrage widmet sich damit der Augenscheinvalidität des generierten Testscores:

Forschungsfrage 2

Inwieweit sagt der Testscore für Planungsqualität das Ergebnis einer ganzheitlichen Bewertung der Unterrichtsplanungen durch projektexterne Expertinnen und Experten voraus?

Eine dritte Forschungsfrage widmet sich der Instruktionssensitivität, also einem besonderen Aspekt der instruktionalen Validität (Polikoff 2010) des entwickelten Testverfahrens. Die Instruktionssensitivität beschreibt die Eigenschaft eines Tests, Effekte von Instruktionen (Polikoff 2010), im weiteren Sinne also auch von mutmaßlichen Lerngelegenheiten, zu erfassen. Da während eines mehrmonatigen, universitär betreuten Praxissemesters die Planung und Durchführung eigenen Unterrichts unter Anleitung curricular verankert ist und somit Lerngelegenheiten bezüglich der (schriftlichen) Planung von Unterricht anzunehmen sind, erscheint eine Zunahme der Fähigkeit zur Unterrichtsplanung plausibel. Vor diesem Hintergrund wird die folgende Forschungsfrage zur Überprüfung der instruktionalen Validität formuliert:

Forschungsfrage 3

Inwieweit ist der entwickelte Planungsperformanztest instruktionssensitiv?

Diese drei Forschungsfragen lassen sich durch eine Zielformulierung der Testentwicklung zusammenfassend ausdrücken: Der Test soll aufgrund einer realitätsnahen Anforderungssituation die Planungsfähigkeit messen. Er soll sich an einem Qualitätsurteil orientieren, das auch außerhalb von Experten geteilt wird. Und er soll Veränderungen aufgrund von Lerngelegenheiten messen können, um später bspw. Ausbildungsprogramme zu vergleichen. Dem schließt sich eine vierte Forschungsfrage an, die Zusammenhänge zwischen Wissen und Performanz in den Blick nimmt und damit die Wirkung von Lehrerbildung in Physik. Darüber hinaus kann so an bisherige Forschungen angeknüpft werden, wenngleich hier nicht direkt Zusammenhänge des Professionswissens zur Unterrichtsqualität gemessen werden, sondern vielmehr das vorhandene Professionswissen mit der Performanz in berufsrelevanten Standardsituationen im Sinne der Konstruktvalidität überprüft werden soll.

Forschungsfrage 4

Welche Zusammenhänge bestehen zwischen der gemessenen Planungsfähigkeit und dem Professionswissen (Fachwissen (FW), fachdidaktisches Wissen (FDW), erziehungswissenschaftliches (EW)) bei Lehramtsstudierenden, die sich in einem Masterstudiengang befinden?

Methoden

Entwicklung des Testinstruments

Ziel bei der Entwicklung des Testinstruments ist es, einen Performanztest im Sinne Millers (1990) zu entwickeln und dabei eine Planung einzufordern, bei der Aspekte der Didaktischen Rekonstruktion (vgl. Abschn. 2.1) zu bearbeiten sind. Es stellt sich die Frage, inwieweit und in welchen Bereichen eine Standardisierung vorgenommen werden kann, ohne dass ein Eingreifen der Testleitung während der Erstellung der Planung erforderlich wird. Da bei der Planung von realem Unterricht gemäß Didaktischer Rekonstruktion z. B. Informationen über die Lerngruppe, Lernvoraussetzungen (wie fachliches Vorwissen/Präkonzepte) sowie antizipierte Lernziele zu beachten sind, kann hier durch explizite Vorgaben eine Standardisierung erreicht werden (Standardisierung der inhaltlichen und organisatorischen Voraussetzungen und der vorgegebenen Stundenziele). Darüber hinaus kann auch eine Standardisierung durch konkrete Vorgaben zu Mindestanforderungen der schriftlichen Planung erreicht werden.

Die Authentizität des Planungsperformanztests kann allerdings durch zu viele äußere Vorgaben oder Rahmenbedingungen eingeschränkt werden, weshalb hier ein Mittelweg gefunden werden muss, der dennoch eine vergleich-

bare Auswertung ermöglicht. In diesem Sinne werden die Testpersonen in eine Situation versetzt, wie sie während eines Praxissemesters auftreten könnte. Der fiktive betreuende Physiklehrer bietet der Testperson kurzfristig (für den nächsten Tag) an, eine 45-minütige Physikunterrichtsstunde selbst zu planen und anschließend durchzuführen. Aufgrund seiner Reihenplanung stehen für die zu planende Stunde bereits das dritte Newtonsche Axiom als Thema und zwei inhaltsbezogene, von den Schülerinnen und Schülern zu erreichende Lernziele fest, außerdem soll ein Experiment in der Stunde durchgeführt werden. Das Experiment stellt zwar im Allgemeinen keine zwingende Anforderung für den Physikunterricht dar, ist jedoch in einer prototypischen Stunde meist enthalten (vgl. Tesch und Duit 2004). Im Rahmen des Einleitungstexts wird eine kurze Beschreibung der fiktiven Lerngruppe gegeben, wie sie in einem informellen Gespräch denkbar wäre, außerdem wird eine Schülervorstellung aus dem Themenbereich Mechanik („nur aktive Körper können Kräfte ausüben, passive leisten Widerstand“, vgl. Schecker und Wilhelm 2018, S. 78) angedeutet, die in der vorherigen Stunde beobachtet wurde. Hierdurch wird sichergestellt, dass ein Experiment geplant wird und dass auf standardisierte Lernvoraussetzungen eingegangen werden kann.

Neben der Aufgabenstellung sind auch die zugelassenen Hilfsmittel standardisiert, indem Auszüge aus zwei schulischen Standardlehrwerken sowie ein Ausdruck aus einer online verfügbaren Quelle zur Verfügung stehen (insgesamt sieben Seiten). Schließlich erfolgt die Dokumentation auf einem vorstrukturierten Planungspapier, welches Prompts zu bestimmten zu explizierenden Teilbereichen der Unterrichtsplanung enthält und so eine erhöhte Vergleichbarkeit zwischen den Planungen der Teilnehmer ermöglichen soll. In Anlehnung an den Grundgedanken der Didaktischen Rekonstruktion, typische Planungsratgeber sowie Kategorien bestehender Studien, werden die Testpersonen durch die Prompts des Planungstests zunächst dazu angeregt, sich (a) fachlich mit dem zu unterrichtenden Gegenstand auseinanderzusetzen, indem zentrale Aspekte wie Fachbegriffe, fachliche notwendige Voraussetzungen, mögliche Elementarisierungen oder sinnvolle Anwendungsbeispiele zu dokumentieren sind. Auch sollen (b) weitere mögliche, relevante Schülervorstellungen beschrieben werden. Den Vorgaben aus der Aufgabenstellung folgend sollen weiterhin (c) geplante Experimente sowie die dabei erwarteten Ergebnisse beschrieben werden. Ebenfalls ist (d) ein zentraler Arbeitsauftrag oder eine zentrale Fragestellung, welche von den Schülerinnen und Schülern bearbeitet werden und der Erreichung des vorgegebenen Lernziels dienen soll, zusammen mit einer Musterlösung zu formulieren. Ein Schwerpunkt, der so auf ein Experiment und auf die konkrete Formulierung einer Aufgabe oder Fragestellung gelegt wird, lehnt sich an die Erkenntnisse Brommes (1997) an, nach

denen Lehrkräfte beim Unterrichten mit der Herstellung einer fachtypischen Aufgabenstruktur konfrontiert sind. Es wird angenommen, dass bei der Planung von typischem Physikunterricht ein Fokus auf die Auswahl und Planung von Experimenten gelegt wird, was z. B. anhand der für Experimente im Realunterricht aufgewendete Lernzeit von bis zu 70 % (z. B. Tesch und Duit 2004) plausibel erscheint.

Neben der Aufgabe bzw. Fragestellung, die direkt die Schülerinnen und Schüler adressiert, soll ebenfalls (e) ein schülergerechtes Tafelbild (Bsp. s. Abb. 1), welches in der Stunde als Sicherung genutzt werden und den Schülerinnen und Schülern zum Lernen und Wiederholen dienen kann, skizziert werden. Wie für Unterrichtsplanungen im Rahmen des Ausbildungsunterrichts üblich (MSB NRW, S. 19), sollen zudem (f) getroffene Entscheidungen begründet werden (Legitimation, vgl. Vogelsang und Riese 2017) und die gesamte Stunde soll nachvollziehbar in einem (g) Verlaufsplan dargestellt werden. Somit werden durch die Prompts des Planungstests möglichst viele Bereiche der Unterrichtsplanung angesprochen, die auch in gängigen Planungsratgebern auftreten, sodass der Tests Aufschluss über die Planungsqualität als Ganzes geben können sollte. Die einzelnen Elemente des Planungspapiers wurden ausgewählt, da sie wichtige und häufig auftretende Teile der Unterrichtsplanung repräsentieren (z. B. Mühlhausen 2007) und sowohl Elemente enthalten, die direkt und z. T. wörtlich an die Schülerinnen und Schüler adressiert sind als auch solche, die an den Fachlehrer adressiert sind und somit im Rahmen eines Ausbildungsgesprächs auf einem anderen fachlichen Niveau besprochen werden können. Die einzelnen Planungselemente stellen die üblichen Anforderungen an Unterrichtsplanungen dar, die an allen beteiligten Universitäten gestellt werden und innerhalb der Curricula verankert sind, ebenso wie sie in allgemein- und fachdidaktischen Lehrwerken und Ratgebern (vgl. auch Abschn. 4.2) verankert sind. Die Bearbeitungszeit für den vollständigen

Planungsperformanztest beträgt 60 min, da für Planungen in alltäglichen Situationen ebenfalls nur begrenzt Zeit zur Verfügung steht (z. B. aufgrund hoher Arbeitsbelastung im Praxissemester oder aufgrund einer kurzfristigen Übernahme einer Stunde) und die durchschnittliche Vorbereitungszeit bei erfahrenen Lehrkräften noch geringer ist (vgl. Hübner und Werle 1997). Darüber hinaus stellt die begrenzte Bearbeitungszeit im Sinne des „Handelns unter Druck“ (Wahl 1991) eine authentische Abbildung des beruflichen Handelns dar.

Der fachliche Inhalt der zu planenden Unterrichtsstunde wurde ausgehend von einer Lehrplananalyse ausgewählt. Dazu wurden die Kern-, Rahmen- bzw. Bildungslehrpläne der für den Einsatz ausgewählten Bundesländer analysiert und auf gemeinsame Unterrichtsinhalte überprüft. Aufgrund der unterschiedlichen Bildungssysteme und den damit einhergehend verschiedenen Lehramtsstudiengängen wurde als gemeinsame Schulform die Gesamtschule ausgewählt und das dritte Newtonsche Axiom als gemeinsames Thema des Mechanikunterrichts in der zehnten Jahrgangsstufe identifiziert.

Entwicklung eines Fähigkeitsmaßes zur Unterrichtsplanung

Zur Bewertung der demonstrierten Planungsfähigkeit wurde zunächst ein Kategoriensystem mit konkreten Kodieritems entwickelt, anhand derer die Qualität der Unterrichtsplanungen eingeschätzt werden soll. Die Grundstruktur des Kategoriensystems orientiert sich dabei am Modell der Didaktischen Rekonstruktion (Kattmann et al. 1997). Die fachliche Klärung des zu vermittelnden Inhalts, das Beachten der Lernvoraussetzungen, die daran anknüpfende adressatengerechte Elementarisierung und die Passung der Planungselemente zu den zugrundeliegenden Lernzielen stellen dabei auf erster Strukturierungsebene vier un-

Abb. 1 Beispiellösung einer Testperson zum Arbeitsauftrag (e) Tafelbild

Skizzieren Sie für eine Kreidetafel ein Tafelbild (oder alternativ einen Sicherungstext) mit Skizze, das

- 1.) gemeinsam mit den Schülerinnen und Schülern als Sicherung entwickelt werden könnte und
- 2.) das den Schülerinnen und Schülern zum Lernen und Wiederholen dienen kann.

<p>(1. Demo - Exp)</p> <p>Skizze:</p> <p>Beob:</p> <ul style="list-style-type: none"> - S1 & 2 rollt weg - übt Kraft auf S2 aus - dieser übt ebenfalls Kraft auf S1 aus → Federkraftmessung ausgeklübelt - Kräfte gleich groß 	<p>3. Newton'sches Axiom:</p> <p>Das Wechselwirkungsprinzip</p> <p>▷ actio = reactio</p> <p>übt Körper A eine Kraft $\vec{F}_{A \rightarrow B}$ auf Körper B aus, so wirkt auch dieser eine gleichgroße entgegengesetzt gerichtete Kraft auf Körper A aus.</p> <p>Anm.: - Wechselwirkungskräfte greifen an zwei Körper an</p>	<p>Beispiele</p> <ul style="list-style-type: none"> - Mensch auf Erde - Trampolin mit Mensch - weitere Bsp. <p>(auf zugedruckte RS)</p>
--	--	--

terschiedliche Oberkategorien dar. Um die Bewertung der konkreten, kreierte Planungselemente der geplanten Unterrichtsstunde auszuschärfen, wurden zur Erweiterung und Ausdifferenzierung des Kategoriensystems Planungsratgeber sowie (physik-)didaktische Lehrwerke analysiert und berücksichtigt (z.B. Sandfuchs 2009; Becker 2007; Tulodziecki et al. 2004; Mühlhausen 2007; Kircher et al. 2015). Dabei wurden die Oberkategorien „Darstellung und Einsatz von Experimenten“, „Einsatz von Kontexten und Alltagsbeispielen“ sowie die „Darstellung und Verwendung von Aufgaben und Fragestellungen“ identifiziert und in das Modell mit aufgenommen. Darüber hinaus ist bei der Planung von Unterricht die Auswahl geeigneter Medien, Sozial- und Organisationsformen (z.B. Tulodziecki et al. 2004), die Beachtung von Heterogenität (z.B. Becker 2007) sowie die Erstellung konkreter Arbeitsmaterialien bzw. das Entwerfen einer schriftlichen Sicherung (z.B. Mühlhausen 2007) relevant. Aufgrund des angestrebten Testformates und der damit einhergehenden zeitlichen Limitation bei der Testbearbeitung und der Vorgabe einer fiktiven Lerngruppe wurden die Beachtung von Heterogenität und das Entwerfen von Unterrichtsmaterialien im Kriterienkatalog jedoch nicht durch eine eigenständige Oberkategorie repräsentiert. Die Auswahl geeigneter Medien (über Experimente hinaus) bzw. Sozialformen wurde in einer ersten Version des Modells integriert, jedoch nach einer Pilotierung mit $N=17$ Testpersonen verworfen, da sich innerhalb der betrachteten Planungen keine Varianz innerhalb der Items zur Darstellung und Nutzung der Medien und Sozialformen ergab und Deckeneffekte auftraten. Das finale Kategoriensystem umfasst auf erster Strukturierungsebene somit die folgenden sieben Oberkategorien, anhand derer nachfolgend konkrete Kodieritems zur Bewertung der Planungsfähigkeit erstellt wurden:

- Fachlicher Inhalt und fachliche Korrektheit
- Elementarisierungen
- Darstellung von Lernvoraussetzungen und Eingehen auf diese
- Erreichbarkeit der Lernziele bzw. Möglichkeiten zum Kompetenzerwerb
- Darstellung und Einsatz von Experimenten
- Darstellung und Einsatz von Fragestellungen bzw. Aufgaben
- Einsatz von Kontexten bzw. Alltagsbeispielen

Zur Genese einzelner Kodieritems wurden literaturbasiert Qualitätskriterien abgeleitet, z.B. zum erfolgreichen Einsatz von Experimenten im Physikunterricht (z.B. Muth 2018) oder zur angemessenen Verwendung von Kontexten und Alltagsbeispielen (z.B. Muckenfuß 2004). Da die Unterrichtsplanung neben kreativen Elementen auch legitimierende Elemente enthalten soll, wurde zwischen Kodieritems zur Bewertung von Begründungen und solchen,

die der Bewertung des Planungsprodukts (als Kreativeleistung) dienen, unterschieden. Die Kodieritems zur Bewertung der Qualität des Planungsprodukts lassen sich wiederum nach ihrer Funktion für die Gesamtplanung im Material unterscheiden. So muss z.B. innerhalb der Kategorie „Darstellung und Einsatz von Experimenten“ einerseits bewertet werden, ob ein Experiment lokal, also punktuell nachvollziehbar und vollständig beschrieben wurde. Andererseits muss für eine hohe Planungsqualität auch die Interdependenz der einzelnen Planungselemente global beachtet werden, also z.B. die Einbettung des Experiments in die gesamte Stunde bzw. die Passung zu anderen relevanten Aspekten (z.B. Lernziele). Beide Aspekte werden durch entsprechende Kodieritems repräsentiert.

Das vollständige Kodiermanual enthielt in einer ersten Fassung insgesamt 26 zwei- oder dreistufige Kodieritems, jedoch zeigten sich bei der Erprobung mit $N=32$ Planungsdokumenten Schwierigkeiten in der Beurteilerübereinstimmung (lediglich 75 %), insbesondere für Kodieritems, die Qualitätseinschätzungen zu bestimmten Elementen des Planungsprodukts vornahmen. Daher wurde in einem folgenden Schritt das Kodiermanual weiter ausdifferenziert. Bei dieser Weiterentwicklung wurden die teils mehrstufigen Items ausnahmslos in kleinschrittigere, weniger komplexe und damit leichter einschätzbare dichotome Items überführt und um einige Items ergänzt, sodass insgesamt $N=59$ Kodieritems (Bsp. s. Abb. 2) zu den sieben oben genannten Kategorien entstanden (in obiger Reihung: 8, 8, 9, 5, 18, 6, 5 Items). Alle vorliegenden Planungen wurden schließlich mit diesem 59 Kodieritems umfassenden Kategoriensystem analysiert. Bei Kodieritems, die auf lokale Items zurückgreifen, aufgrund fehlender Darstellungen aber nicht bewertet werden können (z.B. kein Experiment beschrieben) wurden aufbauende Items zunächst als „Missing“ kodiert und anschließend mit 0 Punkten bewertet.

Design

Der beschriebene Performanztest zur Erfassung der Planungsfähigkeit wurde an vier deutschen Hochschulen in drei Bundesländern im Pre-Post-Design bei $N=96$ (Pre-Test) bzw. $N=78$ (Post-Test) Masterstudierenden des Physiklehramts eingesetzt. Die erste Befragung fand unmittelbar zu Beginn des Praxissemesters statt, die zweite Befragung in den letzten Wochen des Semesters, sodass zwischen den beiden Messzeitpunkten im Mittel 4,5 Monate lagen ($M=139$ d, $SD=16$ d).

Im Rahmen des Gesamtprojekts wurden zu beiden Messzeitpunkten neben dem Performanztest zur Unterrichtsplanung zusätzlich drei bereits erprobte und validierte paper-pencil-Tests zum Professionswissen (aus Platzgründen hier nicht ausführlich dargestellt, vgl. für FW: Enkrott et al. 2020; FDW: Riese et al. 2017; EW: Riese und Reinhold

Abb. 2 Beispiel eines dichotomen Kodieritems zur Nennung von Schülervorstellungen passend zum Dritten Newtonschen Axiom

LVNennWWP	
Ziel: Überprüfen, ob passende Schülervorstellungen (SV) zum Inhalt formuliert werden konnten.	
Erwartungshorizont korrekt (1P)	Erwartungshorizont inkorrekt (0P)
<p>Schülervorstellung explizit zum Wechselwirkungsprinzip (WWP) bzw. zur Wechselwirkung (WW) wird genannt:</p> <ul style="list-style-type: none"> • Es wirkt (immer) nur eine Kraft • Kraft und Gegenkraft greifen immer am gleichen Körper an → Umdeutung Kräftegleichgewicht • WWP nur bei vorhandenem Kräftegleichgewicht → Gleichsetzen von WWP und Kräftegleichgewicht • Gegenkraft wird als zu überwindende Kraft verstanden → Bewegung resultiert nur, wenn <i>actio</i> größer als <i>reactio</i> ist • Gegenkraft wird bei Gewichtskraft (Erde auf Körper) nicht wahrgenommen • „Der Boden/die Erde wirkt keine Gegenkraft aus“ • „Trifft ein bewegtes Objekt auf ein Ruhendes, so übt nur das bewegte Objekt eine Kraft aus.“ • ... 	<p>Keine SV genannt</p> <p>Die genannten Schülervorstellungen passen nicht zum WWP</p> <p>Es sich z.B. um falsche Vorhersagen zum Experiment handelt statt um allgemein formulierte /generalisierte SV:</p> <ul style="list-style-type: none"> • „Der stärkere gewinnt beim Experiment“ (keine Schülervorstellung) • „Nur ein Schüler wird sich bewegen“ • „Beide Kraftmesser zeigen etwas Anderes an“ • ...

2012), zwei weitere Performanztests zur Fähigkeit der Reflexion von Physikunterricht und zum Erklären eines physikalischen Phänomens (Kulgemeyer und Riese 2018, Kempin et al. 2020) sowie ein Fragebogenteil zur Erfassung demographischer Daten eingesetzt. Die gesamte Testzeit je Messzeitpunkt und Testperson betrug bei Bearbeitung aller sechs Testinstrumente insgesamt 305 min, wobei aufgrund der hohen Testbelastung 20 % der Testpersonen zum zweiten Messzeitpunkt den Planungstest nicht bearbeiteten. Die einzelnen Messzeitpunkte variieren um wenige Wochen, da die Probanden parallel ihren regulären Lehrveranstaltungen nachgehen mussten und eine gebündelte Erhebung daher nicht möglich war. Zur Überprüfung der Motivation bei der Testbearbeitung (was aufgrund der großen Testbelastung geboten schien) wurde bei den Performanztests zum Ende der Testung zudem eine Kurzsкала eingesetzt (3 Items, adaptiert aus Thelk et al. (2009)), in der die Probanden auf einer 5-stufigen Likert-Skala (1: „stimme gar nicht zu“, 5: „stimme völlig zu“) z.B. folgendes Item beantworten sollten: „Ich habe bei der Planung mein Bestes gegeben“. Darüber hinaus wurde zum zweiten Messzeitpunkt ein Fragebogen zum Praxissemester eingesetzt, um Informationen zu (Unterrichts-)Erfahrungen, Betreuungsverhältnissen und Ausbildungsgesprächen während des Praxissemesters zu gewinnen.

Um die ökologische Validität des Testverfahrens einschätzen zu können (vgl. Forschungsfrage 1), wurde für eine Teilstichprobe untersucht, inwieweit sich Unterrichtsplanungen, die aus der Bearbeitung des standardisierten Planungsperformanztests resultieren, von Unterrichtsplanun-

gen unterscheiden, die unter Nutzung analoger Prompts für eine real durchzuführende Stunde im Praxissemester angefertigt wurden („Realplanungen“). Ziel war es dabei, mögliche Abweichungen im Planungsprodukt des Planungsperformanztest gegenüber dem „normalen“ Planungsprodukt, wie es innerhalb des Praxissemesters angefertigt wird, zu identifizieren, welche sich einerseits aus dem fiktiven Charakter der unbekannten Lerngruppe und andererseits dadurch ergeben, dass die im Planungstest vorbereitete Stunde nicht in der Realität durchgeführt werden muss. Für $N=17$ Testpersonen, welche von zwei der vier beteiligten universitären Standorte stammen, wurde ein bzw. zwei Monate vor dem Post-Test zusätzlich eine Unterrichtsplanung erfasst, die unter Nutzung einer Vorlage mit den Prompts des standardisierten Tests für eine reale Lerngruppe angefertigt und anschließend im Praxissemester durchgeführt wurde. Darüber hinaus mussten die Studierenden die eigene Unterrichtsstunde selbst in die Unterrichtsreihe einordnen, die Situation der Lerngruppe sowie deren Lernvoraussetzungen beschreiben und eigenständig Lernziele festlegen.

Um die Planungsdokumente, die während des Praxissemesters von den Studierenden für reale Lerngruppen angefertigt wurden, vergleichen zu können, wurde mittels strukturierender qualitativer Inhaltsanalyse (vgl. Mayring 2015) ein zusätzliches, separates Kategoriensystem entwickelt, das in seiner finalen Version insgesamt $N=66$ Untersuchungsaspekte in zehn verschiedenen Kategorien enthielt (fachlicher Inhalt, Kompetenzen, Lernvoraussetzungen, Kontext, Experimente, Aufgaben, Methoden, Kreativität, Verlaufsplan, Tafelbild). Mithilfe dieser Kategorien wurden

sowohl die standardisierten als auch die realen Planungsdokumente im Hinblick auf Auftreten, Umfang und Struktur bestimmter Elemente analysiert und verglichen. Das Kategoriensystem, welches zuvor für die Bewertung der Planungsperformanztests entwickelt wurde (vgl. Abschn. 4.2) konnte hierfür nicht verwendet werden, da es sich aufgrund der Passung zum vorgegebenen Inhalt (3. Newtonsche Axiom) und der Rahmenbedingungen des Performanztests als zu spezifisch für die Analyse freier Planungen erwies.

Um eine Qualitätseinschätzung für bearbeitete Planungsperformanztests zu erhalten und Forschungsfrage 2 beantworten zu können wurden Fachleitungen, die angehende Physiklehrkräfte während des Praxissemesters sowie innerhalb der zweiten Ausbildungsphase betreuen, als mutmaßliche, außeruniversitäre Experten für Unterrichtsplanung und deren Bewertung herangezogen. Diese wurden in einem leitfadengestützten Interview darum gebeten, exemplarisch ausgewählte Planungstests anhand ihrer eigenen Kriterien zu bewerten. Sie erhielten nach der Einführung in das Interview jeweils drei von Studierenden ausgefüllte Produkte des Planungsperformanztests. Die Planungsprodukte waren so ausgewählt, dass die korrespondierenden Testscores eine möglichst hohe Varianz aufwiesen, ohne dass die Unterschiede im Testscore auf unvollständige Bearbeitungen zurückzuführen waren. Nach erster Durchsicht der drei Unterrichtsplanungen sollten die Fachleiter zunächst zu Beginn des Interviews eine Gesamtbewertung der Planungen im Schulnotensystem abgeben und ihre Entscheidung begründen. Die Bewertung der Planungen erfolgte anhand der individuellen Kriterien des Fachleiters; weder das Kodiermanual noch der damit berechnete Testscore waren ihnen bekannt. Im Verlauf des Interviews wurden die Fachleiter gebeten, die Planungstests etwas differenzierter hinsichtlich sechs der sieben Kategorien (vgl. Abschn. 4.2) zu analysieren. Zum fachlichen Inhalt bzw. zur fachlichen Korrektheit wurden keine Nachfragen gestellt, da die fachliche Korrektheit auch durch andere Personen mit physikalischer Ausbildung eingeschätzt werden kann und somit das Experteninterview nicht übermäßig verlängert wurde. Nach der ausführlichen Diskussion der Planungen sollten sie anschließend am Ende des Interviews eine erneute Gesamtbewertung über die Unterrichtsplanungen abgeben. Darüber hinaus wurden sie zu möglichen Gewichtungen befragt, die sie bei der Bewertung vorgenommen hatten.

Zur Beantwortung von Forschungsfrage 3 werden die Daten der ausgefüllten Planungsperformanztests, die jeweils zu Beginn und Ende des Praxissemesters angefertigt wurden, herangezogen. Das Praxissemester stellt bezüglich der Fähigkeit zur Unterrichtsplanung eine Lerngelegenheit dar, da eigenständig geplant wird und durch Hospitationen vertiefte Einblicke in Fremdunterricht und ggf. auch in dessen Planung ermöglicht werden. Auch solche Hospitationen können das eigene Lehrrepertoire erweitern und dazu füh-

ren, sich differenziertere Unterrichtsmethoden anzueignen. Darüber hinaus ist durch nachbereitende Gespräche mit betreuenden Lehrkräften und deren Feedback zu Unterrichtsplanungen auch die Möglichkeit zur Reflexion des eigenen Handelns gegeben, was einen wichtigen Schritt im zirkularen Prozess der Planung darstellt (Shavelson und Stern 1981). Es ist davon auszugehen, dass ein großer Anteil der selbst gehaltenen Physikstunden von den Studierenden auch weitgehend eigenständig vorbereitet wurde (z. B. MSB NRW 2019), weshalb hier zumindest moderate Zuwächse erwartbar sind.

Für Forschungsfrage 4 werden je Messzeitpunkt bivariate Korrelationen zwischen den Ergebnissen der drei eingesetzten Instrumente zum Professionswissen und den Testscores zur Planungsfähigkeit betrachtet.

Beschreibung der Stichprobe

Die folgenden Angaben zur Stichprobe beziehen sich im Weiteren auf $N=62$ Testpersonen (davon 25 weiblich), die zu beiden Messzeitpunkten am Planungstest teilnahmen und für die der Praxissemesterbogen sowie demographische Angaben vorliegen. Für Forschungsfrage 3 wurden darüber hinaus die Planungen von sechs weiteren Probanden verwendet, die regulär am Planungstest teilnahmen, aber für die keine demographischen Daten vorlagen. Bei den Subgruppen zur Beantwortung von Forschungsfrage 4 wurden jeweils die größtmögliche Probandenzahl je Messzeitpunkt, die den Planungstest und den jeweiligen Test zum Professionswissen bearbeitet haben, verwendet. Dieses Vorgehen wurde gewählt, da aufgrund der separat erhobenen Tests zum Professionswissen viele Fälle hätten ausgeschlossen werden müssen, wenn nur vollständige aufgenommen worden wären.

Das durchschnittliche Alter der $N=62$ Testpersonen mit demographischen Daten betrug $M=25,2$ Jahre ($SD=3,8$ Jahre). Die Teilnehmer gaben ihr Fachsemester im Lehramt Physik zum zweiten Messzeitpunkt im Mittel mit 9,6 an (Median=9; $SD=3,1$) und die Gesamtnote ihrer Hochschulzugangsberechtigung mit $M=2,02$ ($SD=0,56$). Trotz der relativ guten durchschnittlichen Hochschulzugangsberechtigungsnote kann die Stichprobe als repräsentativ angenommen werden, da an den teilnehmenden Hochschulen über zwei Jahre hinweg nahezu Vollerhebungen stattfanden. Die Daten zu Umfang und Erfahrungen im Praxissemester sind Tab. 1 zu entnehmen. Während des Praxissemesters hospitierten die Studierenden Unterrichtsstunden im Fach Physik sowie im Zweitfach und führten eigenständig geplante Physikstunden durch, welche üblicherweise mit den betreuenden Lehrkräften vor- und nachbereitet wurden (Median=10 min pro Gespräch). Nur 25 der Studierenden unterrichteten während des Praxissemesters eigenständig den Inhaltsbereich Mechanik,

Tab. 1 Kenndaten zu Umfang und Erfahrungen während des Praxissemesters

	<i>M</i>	<i>SD</i>	<i>N</i>
Hospitierte Physikstunden	66	34	61
Vorbereitungszeit je Einzelstunde (Minuten)	171	74	39
Vorbereitungszeit je Doppelstunde (Minuten)	293	204	46
Vollständig gehaltene Physikstunden	20	11	61
Davon gehaltene Stunden im Bereich Mechanik	9	6	25

der dem Performanztest zugrunde lag. Darüber hinaus gaben 22 Studierende an, nur Doppelstunden unterrichtet zu haben, hingegen zwölf Studierende ausschließlich Einzelstunden unterrichtet zu haben. Die selbst angegebene Vorbereitungszeit je Einzel- oder Doppelstunden beträgt im Praxissemester knapp 3 bzw. 5 h (vgl. Tab. 1), wobei hier nicht genauer aufgeschlüsselt werden kann, wie viel Zeit für die explizite Planung des Unterrichts und wie viel für die weitere Vorbereitung, z. B. den Aufbau von Experimenten sowie das konkrete Designen von Arbeitsmaterialien, aufgewendet wird.

Die drei Fachleiter, die zur Beantwortung von Forschungsfrage 2 interviewt wurden, übten ihre Rolle als Ausbilder im Referendariat zum Zeitpunkt der Interviews zwischen sieben und zwölf Jahren im gleichen Bundesland aus.

Ergebnisse der Validierungsstudien

Im folgenden Abschnitt werden zunächst Analysen zu Testgütekriterien des entwickelten Planungsperformanztests durchgeführt. Anschließend werden Ergebnisse aus verschiedenen Validierungsstudien vorgestellt, wobei insbesondere die zuvor formulierten Forschungsfragen diskutiert werden.

Reliabilität und Objektivität

Bei der Analyse von allen $N=174$ vorliegenden Planungsdokumenten mithilfe des 59 Kodieritems umfassenden Kategoriensystems zeigte sich zunächst, dass 14 Kodieritems eine sehr hohe oder sehr geringe relative Lösungshäufigkeit ($p > 0,9$ bzw. $p < 0,1$) oder eine sehr geringe Trennschärfe ($r < 0,1$) aufwiesen, sodass diese für die weiteren Analysen in einem ersten Schritt ausgeschlossen wurden. Die resultierende Gesamtskala, auf der alle folgenden Ergebnisse zum Testscore beruhen, weist 45 Kodieritems (in Reihung der Oberkategorien: 7, 6, 7, 4, 10, 6, 5) auf, wobei die Punkteverteilung eine zufriedenstellende Varianz des Gesamtscores als Summe der dichotomen Kodieritems auf-

weist ($M=50\%$, $SD=15\%$, $\text{Min}=11\%$, $\text{Max}=87\%$). Es lässt sich als klassischer Wert für die interne Konsistenz $\alpha_{\text{Cronbach}} = 0,81$ berechnen, jedoch muss dieser Wert mit Vorsicht interpretiert werden, da aufgrund der Konstruktion der Kodieritems diese nicht alle voneinander stochastisch unabhängig sind. Beispielsweise kann die Passung des Experiments zum physikalischen Inhalt nur dann beurteilt werden, wenn zuvor ein Experiment hinreichend erläutert wurde. Daher impliziert ein nicht erläutertes Experiment für folgende Kodieritems, die über die Qualität urteilen sollen, bereits null Punkte. Die Beurteilerübereinstimmung wurde anhand einer Doppelkodierung von $N=52$ Unterrichtsplanungen durch zwei geschulte Rater überprüft, wobei sich über alle 45 Kodieritems hinweg bezüglich der vergebenen Punktzahl eine Übereinstimmung von 87,6% bzw. ein Gwet's AC_1 von 0,83 (Cohens $\kappa = 0,75$) ergab, was einer guten Übereinstimmung entspricht (Gwet 2014). Zur Bewertung der Beurteilerübereinstimmung eignet sich Gwet's AC_1 besonders, da Anomalien von Cohens κ kaum auftreten: Der Maximalwert von κ ist bei asymmetrischer Randverteilung nicht immer exakt 1 und kann auch bei hoher absoluter Übereinstimmung sehr geringe Werte produzieren, wofür neben der Randverteilung auch die angenommene Wahrscheinlichkeit einer zufälligen Übereinstimmung verantwortlich ist.

Die sieben Oberkategorien (vgl. Abschn. 4.2) enthielten nach der Reduktion der Kodieritemzahl noch zwischen vier und zehn Items, die allerdings empirisch nicht voneinander getrennt werden können, weshalb ein Gesamtscore als Maß für die Fähigkeit zur Unterrichtsplanung verwendet wird. Vor der Grundannahme, dass bei der Unterrichtsplanung unterschiedliche Bereiche (z. B. Abstimmung von Experiment/Aufgabe/Frage auf bestehende Schülervorstellungen) ineinandergreifen und miteinander vernetzt werden müssen, erscheint dies aus inhaltlicher Sicht unproblematisch.

Während der Testungen konnten anhand der verwendeten Kurzskaala zur Testmotivation keine Motivationsprobleme trotz der hohen Testbelastung beobachtet werden. Es konnten keine signifikanten Unterschiede zwischen Pre- und Post-Test festgestellt werden ($t(66) = 1,2$; $p = 0,2$), die Motivation befand sich zu beiden Messzeitpunkten im Mittel im obersten Drittel der Skala, deren Maximum 15 Punkte beträgt ($M_{\text{Pre}} = 12,4$; $M_{\text{Post}} = 11,3$).

Forschungsfrage 1: Vergleich des Planungsprodukts aus realen und standardisierten Unterrichtsplanungen

Zur Beantwortung von Forschungsfrage (FF) 1 wurden Realplanungen aus dem Praxissemester und bearbeitete Planungsperformanztests für $N=17$ Probanden hinsichtlich des Auftretens, Umfang und Struktur analysiert. Beispielsweise wurde überprüft, ob in den geplanten Unterrichtsstunden je-

Tab. 2 Übersicht über alle Kategorien, die zum paarweisen Vergleich ($N=17$) von Realplanung und standardisierter Planung verwendet wurden sowie zugehörige Übereinstimmung jeweils bei Betrachtung aller Items und bei Reduktion auf solche Items, bei denen ein Vergleich prinzipiell möglich ist, da entsprechende Angaben gemacht wurden

Kategorie (Items in Kategorie)	Übereinstimmung bei allen Items in % (Anzahl Items)	Übereinstimmung bei einschätzbaren Items in % (Anzahl Items)
Tafelbild (7)	56 (119)	58 (115)
Verlaufsplan (7)	69 (119)	73 (112)
Kreativität (4)	72 (68)	78 (63)
Methoden (2)	76 (34)	79 (33)
Aufgaben (7)	53 (119)	91 (69)
Experimente (20)	46 (340)	74 (210)
Kontext (5)	71 (85)	86 (70)
Lernvoraussetzungen (6)	65 (102)	70 (94)
Kompetenzen (2)	62 (34)	72 (29)
Fachlicher Inhalt (5)	73 (85)	92 (67)

Die absolute Zahl der Items, bei denen ein Vergleich nicht möglich war, ergibt sich als Differenz der Anzahl der Items, z. B. zu $68-63=5$ für die Kategorie Kreativität

weils Differenzierungsmöglichkeiten vorgesehen sind oder ob und in welcher Form zu einer geplanten Aufgabenstellung auch eine Musterlösung formuliert wurde. Anschließend wurde überprüft, ob in beiden Planungsdokumenten einer Testperson für jeden Untersuchungsaspekt identische Ausprägungen festgestellt wurden, was im Falle einer Übereinstimmung als Hinweis auf ein ähnliches Planungsverhalten interpretiert wird. Das hierzu entwickelte Kategoriensystem mit $N=66$ Kodieritems dient ausschließlich dem strukturellen Vergleich von standardisierten Planungen zum 3. Newtonschen Axiom und Realplanungen variierenden Inhaltes und wird nicht zur Qualitätseinschätzung im Sinne des Testscores verwendet.

Bei den Kategorien „Experimente“ und „Aufgaben“ stellte sich heraus, dass in den Realplanungen nicht immer ein Experiment geplant oder Aufgaben formuliert wurden, sodass eine Einschätzung zu diesen beiden Elementen der Unterrichtsplanungen nicht immer möglich war. Werden nur die echten Nicht-Übereinstimmungen zwischen Realplanung und standardisierter Planung betrachtet, und nicht solche, die durch Besonderheiten der Unterrichtsthemen oder andere nicht einschätzbare Aspekte bedingt sind, erhöht sich die Übereinstimmung im Planungsverhalten im Aspekt Experimente von 46 auf 74 % und in der Kategorie Aufgaben von 53 auf 91 %. Somit lässt sich schlussfolgern, dass falls in Realplanung sowie standardisierter Planung Experimente oder Aufgaben vorbereitet werden, diese in ähnlichem Umfang und ähnlich strukturiert geplant bzw. in den Unterricht eingebunden werden. Deutlich stärkere Abweichungen gibt es bezüglich der Kategorie „Tafelbild“, hier zeigt sich auch nach Einschränkung auf prinzipiell einschätzbare Kriterien nur eine Übereinstimmung von 58 %. Die Übereinstimmungen der verbleibenden sieben Kategorien liegen zwischen 62 und 76 % bzw. zwischen 73 und 92 % bei Einschränkung auf prinzipiell einschätzbare

Untersuchungsaspekte, eine vollständige Darstellung aller Werte findet sich in Tab. 2. Die Anzahl der Items, die prinzipiell nicht einschätzbar waren, ergibt sich als Differenz der absoluten Werte in den beiden Spalten. Da nicht jede reale Unterrichtsstunde ein Experiment oder eine zentrale Aufgabe bzw. Frage enthält, was z. B. in der Stellung der geplanten Stunde innerhalb der gesamten Unterrichtsreihe begründet sein kann, sind bei diesen Kategorien besonders viele Items bzw. Itempaare nicht bewertbar (130 von 340 für Experimente und 50 von 119 für Aufgaben).

Insgesamt lässt sich aus dieser Analyse schließen, dass unter Beachtung der äußeren Bedingungen, die zu anderen Schwerpunkten bei der Planung führen können (z. B. Unterrichtsstunden ohne Experiment oder Sicherung ohne Tafelbild), das Planungsprodukt für etwa 76 % der untersuchten Aspekte übereinstimmt, wobei die Kategorie „Tafelbild“ am stärksten abweicht. Allerdings gilt es einige Einschränkungen zu betrachten. So wurden Realplanungen zu verschiedenen Inhaltsbereichen der Physik für verschiedene Jahrgangsstufen betrachtet und die Vorbereitungszeit für die Realplanungen konnte nicht überprüft werden. Aufgrund dieser thematischen und strukturellen Varianz lag der Fokus vor allem auf Aspekten der Vollständigkeit und weniger auf Qualitätsbetrachtungen bzw. Aspekten der Vernetzung.

Um zu überprüfen, in wie weit das Planungsprodukt eines gänzlich offenen Planungsauftrags mit analogem Setting vom standardisierten Planungsprodukt abweicht, sind weitere Analysen in Vorbereitung. Bei einer ersten Erkundung mit $N=10$ Testpersonen deutete sich zunächst an, dass das Planungsverhalten im offenen Format weniger ausführlich und breiter gestreut ist. So wird z. B. Schülervorstellungen ohne entsprechende Prompts deutlich weniger Aufmerksamkeit gewidmet, wodurch bei der Nutzung eines solchen Formats als Performanztest unklar bliebe, ob die Test-

Abb. 3 Vergleich der Bewertungen ausgewählter Planungsdokumente (A, B, C) von drei Fachleitern (FL) und den mithilfe des Kodiermanuals generierten Scores. Die Bewertung durch die Fachleiter erfolgte jeweils einmalig zu Beginn des Interviews und einmalig zum Ende des Interviews, nachdem über Teilaspekte der Planung diskutiert wurde

personen potentiell zur Adaption an Schülervorstellungen in der Lage wären.

Eine Aussage über Qualitätsunterschiede zwischen standardisierter und realer Unterrichtsplanung ist nur sehr begrenzt möglich. Dennoch lässt sich feststellen, dass das Planungsprodukt der Studierenden viele Ähnlichkeiten bei einer analogen Struktur des Planungspapiers aufweist und dies somit als Hinweis auf ökologische Validität aufgefasst werden kann. Das Fehlen eines persönlichen Eindrucks von der Lerngruppe und der fiktive Charakter der standardisierten Situation scheinen daher kein gravierendes Problem bzgl. der Qualitätskriterien an einen Performanztest (insbesondere die Authentizität, vgl. Abschn. 2.3) und der ökologischen Validität des Planungstests darzustellen.

Forschungsfrage 2: Vorhersage von Expertenurteilen zur Planungsqualität

Zur Überprüfung von FF2 wurden Interviews mit drei Fachleitern durchgeführt, in denen die Urteile der Experten zu drei ausgewählten Planungsdokumenten erfasst wurden (vgl. Abschn. 4.3). Die verwendeten Planungsdokumente A, B, und C erhielten Testscores von 53 %, 38 % und 73 %. Die zwei Gesamtnoten jedes Fachleiters, die zu Beginn und zum Ende des Interviews pro Unterrichtsplanung verteilt wurden, wurden anschließend mithilfe des Notenschemas der gymnasialen Oberstufe in prozentuale Werte umgerechnet (vgl. KMK 2016), um sie mit dem Testscore vergleichen zu können. Bezogen auf die Rangfolge der Unterrichtsplanungen bestand zu Beginn wie auch zum Ende des Interviews Einigkeit bei den Fachleitern bezüglich der Qualität der Planungen (vgl. Abb. 3). Lediglich bei der

Qualitätseinschätzung von Planungsdokument B weicht Fachleiter 1 stark von der Meinung der Fachkollegen ab und bewertet die Planung mit der Note 6, was aufgrund der ab der Note 4,0 nicht mehr linearen Skala zu größeren Abweichungen bei der Umrechnung in Prozentwerte führt (vgl. KMK 2016).

Dennoch bleibt zu beiden Zeitpunkten die Rangfolge aus den Fachleiterurteilen identisch mit der Rangfolge, die sich auch durch die Kodierung mithilfe des Kodiermanuals ergibt. Lediglich Fachleiter 2 bewertet Planungsdokument A und B vor der detaillierten Analyse sehr ähnlich, nach der Analyse jedoch unterschiedlich. Bei Verwendung der Gesamtpunktzahl, die sich auf Grundlage der Summe der verbliebenen 45 Kodieritems ergibt, kann also für drei exemplarisch ausgewählte Planungsprodukte die qualitätsbezogene Rangfolge, die durch die Fachleiterurteile vorgegeben wird, reproduziert werden. Darüber hinaus stimmt die relative Punktzahl bzw. Bewertung durch das Kodiermanual für zwei der drei Planungsdokumente mit den Fachleiterbewertungen gut überein. Bei Betrachtung der Mittelwerte, die sich je Planungsdokument aus den sechs Noten der Fachleiter ergibt, erhalten die Planungen A, B und C Bewertungen von $(51 \pm 4)\%$, $(28 \pm 16)\%$ und $(78 \pm 6)\%$. Im Rahmen der Standardabweichungen, die sich aus den sechs Urteilen ergibt, passen die Testscores für alle drei Dokumente zu den Urteilen der Experten, jedoch fällt die große Varianz bei Planung B auf. Die Bewertung anhand des Planungsscores scheint jedoch insgesamt ein ähnliches Verständnis von Planungsqualität widerzuspiegeln, wie es den Urteilen der Fachleiter zugrunde liegt.

Bezüglich der Gewichtungen der Oberkategorien konnten sich die Experten nicht auf eine einheitliche Rangfolge

Tab. 3 Mittlere erreichte Punktzahl (Mean) und Standardabweichung (SD) im Planungstest je Messzeitpunkt (MZP) und Ergebnisse aus der längsschnittlichen Betrachtung

<i>N</i> =68 je MZP	MZP1		MZP2		Gepaarter t-Test, <i>t</i> (67) = -2,8	
	Mean	SD	Mean	SD	<i>p</i>	Cohens <i>d</i>
Score	48 %	14 %	53 %	15 %	0,008	0,33

der sechs betrachteten Oberkategorien festlegen. Zwar konnten die Kategorie „Elementarisierung“ als eine der wichtigeren identifiziert werden, die Qualität der Elementarisierung sei jedoch ohne Einsicht in die Reihenplanung schwer zu beurteilen. Die Lernvoraussetzungen wurden nur von zwei Fachleitern als eine der wichtigsten Kategorien beurteilt, da sie bei der Planung richtig eingeschätzt werden müssen, um den Verlauf der Stunde sowie die Schülerreaktionen antizipieren zu können. Das mittlere Feld bilden die Kategorien „Experimente“ und „Aufgaben/Fragestellungen“, die Kategorie „Kontexte“ wurde eher untergeordnet betrachtet.

Keiner der Experten wollte sich insgesamt auf eine Gewichtung festlegen, sondern maximal auf eine Rangfolge der Aspekte. Bei der Bildung des betrachteten Gesamtscores wird deshalb bisher keine Gewichtung einzelner Oberkategorien oder Kodieritems vorgenommen, jedoch sind (vgl. Abschn. 4.2 und 5.1) die Oberkategorien durch vier (Kompetenzen) bis zu zehn (Experimente) Kodieritems vertreten, was indirekt zu einer Gewichtung der Kategorien mit mehr Kodieritems führt.

Forschungsfrage 3: Sensitivität des Testverfahrens für den Zuwachs an Planungsfähigkeit

Wie bereits in Abschn. 4.3 und 4.4 dargestellt, bietet das Praxissemester den getesteten Studierenden durch die 4,5-monatige Teilnahme am Schulbetrieb die Möglichkeit, eigene Unterrichtserfahrungen zu sammeln.

Bei Betrachtung der standardisierten Planungen von *N*=68 Testpersonen, die zu beiden Messzeitpunkten am Planungsperformanztest teilnahmen, zeigt sich ein signifikanter Zuwachs im Gesamtscore, wie in Tab. 3 zusammengefasst. Der Planungsperformanztest scheint im Sinne der dritten Forschungsfrage somit instruktionssensitiv zu sein, also dazu in der Lage, einen mutmaßlichen Zuwachs der Planungsfähigkeit abbilden zu können, wenngleich die Effektstärke mit *d*=0,33 eher gering ausfällt. Es bleibt allerdings ungeklärt, welcher Anteil des Zuwachses auf die Testwiederholung zurückzuführen ist und welcher auf eine Weiterentwicklung der Fähigkeit zur Unterrichtsplanung. Um zu überprüfen, ob sich die Probanden zum zweiten Messzeitpunkt aufgrund der Testwiederholung leichter in die Situation einfinden können oder sich an bestimmte Elemente des Planungstests erinnern, wurde überprüft, ob sich die Anzahl von Missings bei der Kodierung zwischen den

Messzeitpunkten unterscheidet. Zum ersten Messzeitpunkt wurden für die *N*=68 Probanden insgesamt 341 fehlende Werte von insgesamt 3060 Kodierungen, (45 Items je Probanden und MZP) kodiert, zum zweiten Messzeitpunkt 347. Bei probandenweisem Vergleich mittels *t*-test für verbundene Stichproben konnte kein signifikanter Unterschied bezüglich der Anzahl fehlender Werte zwischen Pre- und Post-Test festgestellt werden (*t* (67) = -0,14; *p* = 0,89). Somit scheinen die Planungen zum zweiten Messzeitpunkt zumindest nicht umfangreicher bzw. vollständiger zu sein.

Forschungsfrage 4: Zusammenhänge zwischen Testscore und Professionswissen

Da im Rahmen des Gesamtprojekts zu beiden Messzeitpunkten neben dem Planungsperformanztest auch Instrumente zur Erhebung des Professionswissens eingesetzt wurden, können auch mögliche Zusammenhänge zwischen den Testwerten bzw. zwischen der erhobenen Planungsfähigkeit und den drei Bereichen des Professionswissen im Sinne einer Untersuchung der Konstruktvalidität und zur Untersuchung der Wirksamkeit der Lehrerbildung betrachtet werden. Hierzu wurden die manifesten Testwerte für FW, FDW und EW verwendet und getrennt nach Messzeitpunkt mit dem Testscore aus dem Planungstest bivariat korreliert. Die Ergebnisse sind in Tab. 4 zusammengefasst. Die unterschiedlichen Fallzahlen sind darin begründet, dass nicht alle Probanden an allen Erhebungen teilnehmen konnten.

Dabei zeigte sich für beide Messzeitpunkte kein Zusammenhang (*r* < 0,1) zwischen der Planungsfähigkeit und dem Fachwissen. Zu Beginn des Praxissemesters zeigen sich jeweils mittlere signifikante Korrelationen zwischen Planungsfähigkeit und FDW und EW (Cohen 1988). Zum zweiten Messzeitpunkt ist die Korrelation zum FDW nahezu unverändert, zum EW ist sie um etwa ein Drittel geringer als zuvor. Zwischen den Messzeitpunkten fand jeweils eine signifikante Steigerung der FDW bzw. EW-Testwerte statt (Cohens *d*_{FDW} = 0,43; Cohens *d*_{EW} = 0,75).

Zusammenfassung und Diskussion

Die Fähigkeit, Unterricht adäquat planen zu können, stellt eine zentrale Anforderung für Lehrkräfte dar. Obgleich die Relevanz schriftlicher Unterrichtsplanungen für erfahrene Lehrkräfte abnimmt (Livingston und Borko 1989), so stellt

Tab. 4 Übersicht über die bivariaten Korrelationen (Pearson's r) zwischen den manifesten Gesamtscores der Tests zum Professionswissen (vgl. für FW: Enkrott et al. 2020; FDW: Riese et al. 2017; EW: Riese und Reinhold 2012) und dem Testscore aus dem Planungsperformanztest, getrennt nach Messzeitpunkt

		FW		FDW		EW	
		Pearson's r	N	Pearson's r	N	Pearson's r	N
Planung	Beginn PS	0,01 (n. s.)	59	0,32**	74	0,36**	55
	Ende PS	–0,06 (n. s.)	38	0,33**	71	0,25*	71

sie doch während der Ausbildung und für Berufsanfänger eine unverzichtbare Tätigkeit dar, da sie u. a. eine Grundlage für anschließende Reflexionen über das eigene Berufshandeln darstellt. Trotzdem ist die Fähigkeit zur Unterrichtsplanung empirisch in standardisierter Form bislang eher selten untersucht worden. Daher wurde im ProfiLe-P+ ein Performanztest entwickelt, der die Fähigkeit zur Unterrichtsplanung standardisiert in einem vorgegebenen Szenario erfasst. Das Testinstrument scheint dabei trotz des speziellen Settings eine authentische, also ökologisch valide Situation hinreichend widerzuspiegeln, da keine bedeutsamen Unterschiede zwischen dem Planungsprodukt der Testsituation und dem Produkt der Planung einer realen Stunde unter ähnlichen Rahmenbedingungen (Stunde mit Experiment und Aufgabe) gefunden werden konnten (FF1), wenn auf solche Aspekte fokussiert wird, die gleichermaßen sowohl in den Realplanungen als auch in den standardisierten Planungen des Performanztests bewertet werden können.

Die geringe Vergleichbarkeit der Kategorie „Tafelbild“ ist vermutlich dadurch begründet, dass im Planungsperformanztest ein „Tafelbild für eine Kreidetafel (oder alternativ einen Sicherungstext) mit Skizze“ eingefordert wird, wohingegen die Studierenden in den Realplanungen die Art der Sicherung den örtlichen Gegebenheiten anpassen mussten, woraus z. B. die Sicherung auf einem Arbeitsblatt oder mithilfe eines interaktiven Whiteboards resultiert. Eine Sicherung durch andere Medien erfordert aber auch eine schriftliche Formulierung des Lerngegenstands, wenngleich diese möglicherweise anders in den Unterricht eingebunden würde. Insgesamt bildet der Performanztest damit solche Planungen authentisch ab, in denen (a) ein Experiment genutzt wird und dazu (b) eine zentrale Aufgabe gestellt wird. Somit ist die Generalisierbarkeit auf die Fähigkeit zur Unterrichtsplanung auf diese Art der Stunde eingeschränkt, da keine Informationen darüber gewonnen können, ob z. B. ein qualitativ hochwertiges Lehrangebot, welches den Fokus auf das Üben setzt, geplant werden kann.

Das Format des Performanztests erwies sich gemessen an der Beurteilerübereinstimmung als hinreichend objektives Testinstrument. Darüber hinaus konnte ein Gesamtscore generiert werden, für den mittels Expertenbeurteilungen Argumente für externe Validität gewonnen werden konnten (FF2). Gemessen an der mittleren Bewertung der Experten zeigten sich auch absolut betrachtet gute Übereinstimmungen mit den erhobenen Testwerten, jedoch liegen

für die Planung mit dem geringsten Testscore sehr unterschiedliche Urteile der Experten vor. Eine mögliche Ursache hierfür könnte sein, dass bei der Beurteilung des Performanztests auch Punkte für das bloße Vorhandensein von bestimmten Planungselementen vergeben werden, die Qualität hierfür aber noch keinen Einfluss auf die Punktzahl hat. Experten hingegen könnten unpassende Planungselemente genauso wie nicht vorhandene Elemente bewerten oder sogar Punktabzüge dafür vorsehen, bspw. bei einem irreführenden Tafelbild oder überkomplexen Experimenten, die womöglich zum Scheitern der geplanten Stunde führen könnten. Somit sind möglicherweise Unterrichtsplanungen an den Rändern der Bewertungsskala weniger gut abgebildet.

Da die Kodieritems, die zu den sieben Oberkategorien gehören, nicht empirisch voneinander trennbar sind, wurde bisher keine Gewichtung der Items bzw. der Oberkategorien eingeführt, um z. B. den Einfluss der Kategorie Experimente für den Gesamtscore zu reduzieren, wie durch die Experteninterviews angestoßen, auch weil manche Items per Konstruktion nicht ausschließlich einer Kategorie zugeordnet werden könnten. Bspw. kann die vorhandene Verbindung von Experiment und Kontext dahingehend interpretiert werden, dass das Experiment passend zu Inhalt und Kontext ausgewählt wurde oder aber als konsequente Umsetzung des angedachten Kontextes, was über die Zuordnung des Items zu einer Oberkategorie entscheiden könnte. Der Einfluss der Kategorie Experimente ist gegenüber den Urteilen der Fachleiter mit zehn von 45 Items überdurchschnittlich stark, jedoch bei einer Unterrichtsstunde, die nach Vorgabe auch ein Experiment enthält, inhaltlich vertretbar. Aufgrund der Charakteristik des Instruments, dass einzelne Elemente des Planungsprodukts – wie z. B. das Experiment – unter unterschiedlichen Gesichtspunkten – z. B. dessen Passung zum Lernziel oder zur Aufgabenstellung – analysiert werden, ist eine stochastische Unabhängigkeit einzelner Items nicht in allen Fällen gegeben, so dass die Reliabilität des Instruments bei bloßer Betrachtung von Cronbachs α möglicherweise überschätzt wird.

Bei einem Testeinsatz vor und nach dem Praxissemester, das als Lern Gelegenheit für Planungsfähigkeit angenommen werden kann, reagiert das Testinstrument offenbar sensitiv auf Veränderungen der Planungsfähigkeit (FF3), da eine Steigerung (Cohens $d=0,33$) der gemessenen Planungsfähigkeit für $N=68$ Probanden gemessen werden konnte.

Zwar konnte aufgrund fehlender Kontrollgruppen nicht untersucht werden, in wie weit es sich um Testwiederholungseffekte handelt, jedoch liegen zwischen den beiden Messzeitpunkten etwa 4,5 Monate, sodass keine detaillierten Erinnerungen befürchtet werden müssen. Perspektivisch wäre ein follow-up Test nach weiteren 4,5 Monaten denkbar.

Zu beachten ist, dass durch das Testformat vorgegeben wird, welche Aspekte der Unterrichtsplanung dokumentiert werden sollen. Bei diesen handelt es sich zwar um Aspekte, die an den beteiligten Universitäten in den fachdidaktischen Veranstaltungen behandelt werden und die in Unterrichtsplanungen typischerweise eingefordert werden, jedoch lassen sich diese Anforderungen nicht uneingeschränkt auf andere Standorte oder für inhaltlich abweichende Unterrichtsplanungen übertragen. Ebenfalls handelt es sich um Aspekte, die innerhalb der universitären Ausbildungsphase eingefordert werden. Während des Vorbereitungsdienstes oder danach können auch andere Anforderungen an eine schriftliche Planung gestellt werden, weshalb die Vorstrukturierung und die zugehörige Bewertung zunächst nur für diesen speziellen Ausbildungsabschnitt verwendet werden sollte. Daher kann mithilfe des Instrumentes nur die Qualität von Planungen, die den Anforderungen während des Studiums entsprechen, erfasst werden. Aus diesem Grund wurden teils restriktive Prompts in den Test integriert, da z. B. das „Beachten von Schülervorstellungen“ eine Anforderung an Unterrichtsplanungen während dieses Stadiums darstellt. Darüber hinaus ist durch das Format des Performanztests ein Zeitlimit von 60 min für die Unterrichtsplanung gesetzt, was für schriftliche Planungen zu diesem Zeitpunkt der Ausbildung zwar kurz erscheint, jedoch findet die alltägliche Planung von Lehrkräften deutlich kürzer statt (vgl. Abschn. 2.2) und auch während der Ausbildungsphase sind durch kurzfristige Übernahmen von Unterrichtsstunden kürzere Planungsdauern möglich, weshalb durch den zeitlichen Umfang des Performanztests durchaus authentische Rahmenbedingungen im Sinne eines „Handelns unter Druck“ dargestellt werden (vgl. Wahl 1991).

Im Hinblick auf die Aussagekraft der mit dem Planungsperformanztest gewonnenen Daten ist einschränkend anzumerken, dass die Fähigkeit zur Unterrichtsplanung nur zu einem speziellen fachlichen Inhalt aus dem Bereich der Mechanik untersucht wird. Aus diesem Grund wird aktuell erkundet, inwieweit sich das Kodiermanual öffnen lässt, um auch weniger standardisierte, beliebige Unterrichtsplanungen zu anderen Inhalten bewerten zu können, was für die inhaltsunabhängigen Kodieritems weitgehend möglich ist. Bei der Bewertung der vorliegenden, standardisierten Planungen mit dem geöffneten Manual deuten erste Ergebnisse darauf hin, dass sich eine identische Rangfolge des jeweils ermittelten Fähigkeitsscores wie bei der Bewertung mit dem eigentlichen Kodiermanual zeigt, wenngleich insgesamt natürlich ein Informationsverlust gegenüber ei-

ner inhaltspezifischen Analyse zu verzeichnen ist. Somit scheint die Bewertung von beliebigen schriftlichen Planungen mit dem geöffneten Kodiermanual prinzipiell möglich zu sein, was perspektivisch auch einen Einsatz des entwickelten Instruments zu Lehr- bzw. Diagnosezwecken (z. B. im Rahmen von Praxisphasen) möglich erscheinen lässt.

Des Weiteren ist zu bedenken, dass unklar ist, welche Bedeutung der mit dem entwickelten Performanztest ermittelte Fähigkeitsscore für die Qualität von real durchgeführtem Unterricht hat. Dies konnte im Rahmen des Projekts nicht untersucht werden, da ohnehin bereits eine hohe Testbelastung für die Testpersonen bestand. Vielmehr dient der entwickelte Performanztest dem Zweck, die Lücke bei der Klärung des Zusammenhanges von Professionswissen und Unterrichtsqualität zu verkleinern, indem zunächst der Zusammenhang zwischen Professionswissen und einer speziellen Standardanforderung des Lehrberufs untersucht wird. Die hier gewonnenen Erkenntnisse zeigen mittlere Korrelationen zwischen der Planungsfähigkeit und dem fachdidaktischen Wissen (FDW) zu beiden Messzeitpunkten sowie zum erziehungswissenschaftlichen Wissen (EW) zum ersten Messzeitpunkt. Trotz einer großen Zunahme des EW über das Praxissemesters ist die Korrelation zwischen EW und der Planungsfähigkeit zum zweiten Messzeitpunkt geringer als zu Beginn. Das Fachwissen hingegen korrelierte zu keinem Messzeitpunkt mit der Planungsfähigkeit. Die Korrelationen lassen sich im Sinne der Konstruktvalidität dahingehend deuten, dass ein erhöhtes FDW und EW auch mit einer leicht erhöhten Planungsfähigkeit einhergeht, was mit Blick auf die Konzeption der Lehramtsausbildung wünschenswert ist. Zwar lassen sich im Rahmen dieser Studien nur Korrelationen und keine kausalen Zusammenhänge betrachten, dies ist jedoch perspektivisch im Rahmen von Cross-lagged-panel Analysen mit dem Datensatz des Gesamtprojektes möglich, auch unter Einbezug der Fähigkeiten bezüglich der Reflexions und des Erklärens von Physik. Die gemessenen Zusammenhänge zum Professionswissen zeigen eine ähnliche Tendenz wie bisherige Studien (z. B. Stender et al. 2017), jedoch sind die Studien aufgrund der unterschiedlichen Testformate nur bedingt vergleichbar.

Beim Vergleich mit anderen Studien zur Erfassung der Planungsfähigkeit bzw. des Planungswissens können bei längsschnittlicher Betrachtung, ähnlich wie z. B. bei König (2019) oder Baer et al. (2011), Zuwächse erkannt werden, jedoch sind die hier gemessenen Effekte kleiner (bei König: $d=1,7$). Eine denkbare Ursache ist, dass das Praxissemester zwar für das allgemeine Unterrichten und Aspekte wie das Classroom-Management eine wichtige Rolle spielt, jedoch die fachdidaktisch fundierte und ausführliche Planung nicht so sehr im Vordergrund steht wie während des Referendariats. Ebenfalls befinden sich hier nur wenige Monate zwischen den beiden Messzeitpunkten, wohingegen Baer

et al. (2011) das Wissen entlang des gesamten Studiums erfassen.

Die Ergebnisse zu FF4 können im Rahmen des Refined Consensus Models of PCK so interpretiert werden, dass ein erhöhtes pPCK auch mit einem erhöhten ePCK im Bereich der Unterrichtsplanung einhergeht. Jedoch lässt sich auch hier die Einschränkung des Instrumentes auf eine bestimmte Art der Unterrichtsplanung erkennen, da das ePCK aufgrund seiner Spezifität keine vollständige Abbildung des gesamten PCK, aber dennoch einen relevanten Teil, darstellt.

Förderung Das Projekt *ProfiLe-P+* wird vom Bundesministerium für Bildung und Forschung im Rahmen des BMBF-Rahmenprogramms *X* unter dem Kennzeichen *x* gefördert.

Funding Open Access funding provided by Projekt DEAL.

Open Access Dieser Artikel wird unter der Creative Commons Namensnennung 4.0 International Lizenz veröffentlicht, welche die Nutzung, Vervielfältigung, Bearbeitung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden.

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

Weitere Details zur Lizenz entnehmen Sie bitte der Lizenzinformation auf <http://creativecommons.org/licenses/by/4.0/deed.de>.

Literatur

- Alonzo, A., Berry, A., Nilsson, P. (2019). Unpacking the Complexity of science Teachers' PCK in Action: Enacted and Personal PCK. In A. Hume, R. Cooper, A. Borowski (Hrsg.), *Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science* (S. 271–288). Singapore: Springer.
- Apra, C. (2014). Unterrichtsplanung als Designaufgabe. *Journal für Lehrerinnenbildung*, 2014(4), 47–50.
- Baer, M., Kocher, M., Wyss, C., Guldemann, T., Larcher, S., & Dörr, G. (2011). Lehrerbildung und Praxiserfahrung im ersten Berufsjahr und ihre Wirkung auf die Unterrichtskompetenz von Studierenden und jungen Lehrpersonen im Berufseinstieg. *Zeitschrift für Erziehungswissenschaft*, 14, 85–117.
- Bartels, H., Geelan, D., Kulgemeyer, C. (2019). Developing an approach to the performance-oriented testing of science teachers' action-related competencies. *International Journal of Science Education*, 41(14). <https://doi.org/10.1080/09500693.2019.1658241>.
- Becker, G. (2007). *Unterricht planen – Handlungsorientierte Didaktik Teil I*. Weinheim: Beltz.
- Blömeke, S., Gustafsson, J. & Shavelson, R. (2015). Beyond Dichotomies: Competence viewed as a continuum. *Zeitschrift für Psychologie* 223(1), 3–13.
- Blömeke, S., Paine, L., Houang, R., Hsieh, F., Schmidt, W., Tatto, M., Bankov, K., Cedillo, T., Coga, L., Han, S., Santillan, M., & Schwille, J. (2008). Future teachers, competence to plan a lesson: first results of a six-country study on the efficiency of teacher education. *ZDM Mathematics Education*, 40(5), 749–762.
- Bromme, R. (1981). *Das Denken von Lehrern bei der Unterrichtsvorbereitung. Eine empirische Untersuchung zu kognitiven Prozessen von Mathematiklehrern*. Weinheim: Beltz.
- Bromme, R. (1997). Kompetenzen, Funktionen und unterrichtliches Handeln des Lehrers. In F.E. Weinert (Hrsg.), *Psychologie des Unterrichts und der Schule* (S. 177–212). Göttingen: Hogrefe.
- Carlson, J., & Daehler, K.R. (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*. Singapore: Springer.
- Cauet, E., Liepertz, S., Borowski, A., & Fischer, H.E. (2015). Does it matter what we measure? Domain-specific professional knowledge of physics teachers. *Schweizerische Zeitschrift für Bildungswissenschaften*, 37, 463–480.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ: Erlbaum.
- Enkrott, P., Buschhüter, D., Spoden, C., Fischer, H. E., & Borowski, A. (2020). Entwicklung des fachlichen- und fachdidaktischen Wissens angehender Physik Lehrkräfte. In S. Habig (Hrsg.), *Naturwissenschaftliche Kompetenzen in der Gesellschaft von morgen. Jahrestagung der GDGP 2019 in Wien* (S. 431–434). Duisburg: Universität Duisburg-Essen.
- Gassmann, C. (2013). *Erlebte Aufgabenschwierigkeit bei der Unterrichtsplanung – Eine qualitativ-inhaltsanalytische Studie zu den Praktikumsphasen der universitären Lehrerbildung*. Wiesbaden: Springer.
- Großschädl, J., Mahler, D., Kleickmann, T., & Harms, U. (2014). Content-related knowledge of biology teachers from secondary schools: structure and learning opportunities. *International Journal of Science Education*, 36(14), 2335–2366.
- Gwet, K.L. (2014). *Handbook of inter-rater reliability* (4. Aufl.). Gaithersburg, MD: Advanced Analytics, LLC.
- Haas, A. (1998). *Unterrichtsplanung im Alltag: Eine empirische Untersuchung zum Planungshandeln von Hauptschul-, Realschul- und Gymnasiallehrern*. Regensburg: Roderer.
- Harden, R., Stevenson, M., Downie, W., & Wilson, G. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, 1, 447–451.
- Hardwig, T., & Mußmann, F. (2018). Zeiterfassungsstudien zur Arbeitszeit von Lehrkräften in Deutschland. Konzepte, Methoden und Ergebnisse von Studien zu Arbeitszeiten und Arbeitsverteilung im historischen Vergleich. Expertise im Auftrag der Max-Träger-Stiftung. Göttingen. https://kooperationsstelle.uni-goettingen.de/fileadmin/user_upload/Hardwig_Mussmann_MTS-Expertise_-_Zeiterfassungsstudien_zur_Arbeitszeit_von_Lehrkraefte_in_Deutschland.pdf. Zugegriffen: 03.04.2020.
- Hartig, J., Frey, A., & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 143–173). Berlin: Springer.
- Hasenkamp, A., Windt, A., & Rumann, S. (2016). Qualität der Sachunterrichtsplanung im Vorbereitungsdienst. In C. Maurer (Hrsg.), *Authentizität und Lernen – das Fach in der Fachdidaktik* (S. 278–280). Regensburg: GDGP.
- Heimann, P., Otto, G., & Schulz, W. (1969). *Unterricht, Analyse und Planung* (4. Aufl.). Hannover: Schroedel.
- Hübner, R., & Werle, M. (1997). Arbeitszeit und Arbeitsbelastung Berliner Lehrerinnen und Lehrer. In S. Buchen (Hrsg.), *Jahrbuch der Lehrerforschung* (Bd. 1, S. 203–226). Weinheim: Juventa.
- Jacobs, C.L., Martin, S.N., & Otieno, T.C. (2008). A science lesson plan analysis instrument for formative and summative program evaluation of a teacher education program. *Science Education*, 92(6), 1096–1126.

- Kane, M. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>.
- Kattmann, U., Duit, R., Gropengießer, H., & Komorek, M. (1997). Das Modell der Didaktischen Rekonstruktion – Ein Rahmen für naturwissenschaftsdidaktische Forschung und Entwicklung. *Zeitschrift für Didaktik der Naturwissenschaften*, 3(3), 3–18.
- Keller, M., Neumann, K., & Fischer, H. (2016). The impact of teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching*, 54(5), 568–614.
- 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. Jahrestagung der GDGP 2019 in Wien* (S. 439–442). Duisburg: Universität Duisburg-Essen.
- Kircher, E., Girwidz, R., & Häußler, P. (Hrsg.). (2015). *Physikdidaktik – Theorie und Praxis* (3. Aufl.). Berlin: Springer.
- Klafki, W. (1969). Didaktische Analyse als Kern der Unterrichtsvorbereitung. In H. Roth & A. Blumental (Hrsg.), *Didaktische Analyse. Auswahl Reihe A 10*. Aufl. Hannover: Schroedel.
- Klafki, W. (2007). *Neue Studien zur Bildungstheorie und Didaktik* (6. Aufl.). Weinheim und Basel: Beltz.
- Klafki, W. (1979). Der Begriff Didaktik und der Satz vom Primat der Didaktik im Verhältnis zur Methodik. In W. Klafki, et al. (Hrsg.), *Funkkolleg Erziehungswissenschaften* (S. 55–92). Frankfurt a.M.: Fischer.
- KMK (2016). Schema und Beispiele zur Notenbildung im Zentralabitur. Vorgabe der Kultusministerkonferenz. https://www.standardsicherung.schulministerium.nrw.de/cms/upload/abiturgost/verfuegungen/2020/Anlage_4_Schema_und_Beiispiele_zur_Notenbildung.pdf. Zugegriffen: 15.01.2020.
- KMK (2019). *Ländergemeinsame inhaltliche Anforderungen für die Fachwissenschaften und Fachdidaktiken in der Lehrerbildung*. Beschluss der Kultusministerkonferenz vom 16.10.2008 i.d. F. vom 16.05.2019
- König, J. (2019). PlanvoLL-D: Planungskompetenz von angehenden Lehrerinnen und Lehrern im Fach Deutsch. In N. McElvany, W. Bos, H. Holtappels & A. Ohle-Peters (Hrsg.), *Bedingungen und Effekte von Lehrerbildung, Lehrkraftkompetenzen und Lehrkraft Handeln* (S. 67–86). Münster: Waxmann.
- Kulgemeyer, C., & Riese, J. (2018). From professional knowledge to professional performance: The impact of CK and PCK on teaching quality in explaining situations. *Journal of Research in Science Teaching*, 55(10), 1393–1418. <https://doi.org/10.1002/tea.21457>.
- Lederman, N. G., & Abell, S. K. (Hrsg.). (2014). *Handbook of research on science education*. Bd. 2. New York, NY: Routledge.
- Liepert, S., & Borowski, A. (2019). Testing the consensus model: relationships among physics teachers' professional knowledge, interconnectedness of content structure and student achievement. *International Journal of Science Education*, 41(7), 890–910.
- Livingston, C., & Borko, H. (1989). Expert-novice differences in teaching: a cognitive analysis and implications for teacher education. *Journal of Teacher Education*, 40(4), 36–42. <https://doi.org/10.1177/002248718904000407>.
- Mayring, P. (2015). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (12. Aufl.). Weinheim: Beltz.
- Miller, G.E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine*, 65(5), 63–67.
- Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (MSB NRW) (2019). Rahmenkonzeption zur strukturellen und inhaltlichen Ausgestaltung des Praxissemesters im lehramtsbezogenen Masterstudiengang. https://www.lbz.rwth-aachen.de/global/show_document.asp?id=aaaaaaaaaabcwp. Zugegriffen: 15.01.2020.
- Muckenfuß, H. (2004). Themen und Kontexte als Strukturelement des naturwissenschaftlichen Unterrichts. *PhyDid A – Physik und Didaktik in Schule und Hochschule*, 3, 57–66.
- Mühlhausen, U. (2007). Unterrichtsvorbereitung – wie am besten? In U. Daschner & U. Drews (Hrsg.), *Kursbuch Referendariat* (S. 40–58). Weinheim: Beltz.
- Muth, L. (2018). *Einfluss der Auswertephase von Experimenten im Physikunterricht*. Berlin: Logos.
- Plöger, W. (2008). *Unterrichtsplanung. Ein Lehr- und Arbeitsbuch für Studium und Seminar*. Köln: Kölner Studien Verlag.
- Polikoff, M.S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, 29, 3–14.
- Riese, J. & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen: Empirische Hinweise für eine Verbesserung des Lehramtsstudiums. *Zeitschrift für Erziehungswissenschaft*, 15(1), 111–143.
- Riese, J., Gramzow, Y., & Reinhold, P. (2017). Die Messung fachdidaktischen Wissens bei Anfängern und Fortgeschrittenen im Lehramtsstudiengang Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, 23(1), 99–112.
- Sandfuchs, U. (2009). Grundfragen der Unterrichtsplanung. In K. H. Arnold, U. Sandfuchs & J. Wiechmann (Hrsg.), *Handbuch Unterricht* (2. Aufl. S. 512–519). Bad Heilbrunn: Klinkhardt.
- Schecker, H., & Wilhelm, T. (2018). Schülervorstellungen in der Mechanik. In H. Schecker, T. Wilhelm, M. Hopf & R. Duit (Hrsg.), *Schülervorstellungen und Physikunterricht* (S. 63–88). Berlin: Springer.
- Seel, A. (2011). Wie angehende Lehrer/innen das Planen lernen – Empirische Befunde zur ausbildungsbezogenen Unterrichtsplanung. In K. Zierer, et al. (Hrsg.), *Jahrbuch für Allgemeine Didaktik 2011* (S. 304–318). Baltmannsweiler: Schneider Hohengehren.
- Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgements, decisions and behavior. *Review of Educational Research*, 51(4), 455–498.
- Shulman, L. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher*, 15(2), 4–14.
- Sorge, S., Kröger, J., Petersen, S., & Neumann, K. (2019). Structure and development of pre-service physics teachers' professional knowledge. *International Journal of Science Education*, 41(7), 862–889.
- Stender, A. (2014). *Unterrichtsplanung: Vom Wissen zum Handeln*. Berlin: Logos.
- Stender, A., Brückmann, M., & Neumann, K. (2017). Transformation of topic-specific professional knowledge into personal pedagogical content knowledge through lesson planning. *International Journal of Science Education*, 39(12), 1690–1714.
- Sturpe, D.A., Layson-Wolf, C., McPherson, M.L., & Haines, S.T. (2013). An objective structured teaching examination (OSTE) for faculty development. *Currents in Pharmacy Teaching and Learning*, 5(6), 616–619.
- Tebrügge, A. (2001). *Unterrichtsplanung zwischen didaktischen Ansprüchen und alltäglicher Berufs Anforderung – Eine empirische Studie zum Planungs Handeln von Lehrerinnen und Lehrern in den Fächern Deutsch Mathematik und Chemie*. Frankfurt am Main: Peter Lang.
- Tepner, O., Borowski, A., Dollny, S., Fischer, H., Jüttner, M., Kirschner, S., Leutner, D., Neuhaus, B., Sandmann, A., Sumfleth, E., Thillmann, H., & Wirth, J. (2012). Modell zur Entwicklung von Testitems zur Erfassung des Professionswissens von Lehrkräften in den Naturwissenschaften. *Zeitschrift für Didaktik der Naturwissenschaften*, 18, 7–28.
- Tesch, M., & Duit, R. (2004). Experimentieren im Physikunterricht – Ergebnisse einer Videostudie. *Zeitschrift für Didaktik der Naturwissenschaften*, 10, 51–69.
- Thelk, A.D., Sundre, D.L., Jeanne Horst, S., & Finney, S.J. (2009). Motivation matters: Using the Student Opinion Scale to make va-

- lid inferences about student performance. *Journal of general education*, 58(3), 129–151.
- Tulodziecki, G., Herzig, B., & Blömeke, S. (2004). *Gestaltung von Unterricht. Eine Einführung in die Didaktik*. Bad Heilbrunn: Klinkhardt.
- Vogelsang, C. (2014). *Validierung eines Instruments zur Erfassung der professionellen Handlungskompetenz von (angehenden) Physiklehrkräften*. Berlin: Logos.
- Vogelsang, C. & Riese, J. (2017). Wann ist eine Unterrichtsplanung „gut“? Planungsperformanz in Praxisratgebern zur Unterrichtsplanung. In S. Wernke & K. Zierer (Hrsg.), *Die Unterrichtsplanung - Ein in Vergessenheit geratener Kompetenzbereich?! (S. 47–61)*. Bad Heilbrunn: Klinkhardt.
- Vogelsang, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., Riese, J., Schecker, H. & Schröder, J. (2019). Entwicklung von Professionswissen und Unterrichtsperformanz im Lehramtsstudium Physik. Analysen zu valider Testwertinterpretation. *Zeitschrift Für Pädagogik*, 65(4), 473–491.
- Wahl, D. (1991). *Handeln unter Druck. Der weite Weg vom Wissen zum Handeln bei Lehrern, Hochschullehrern und Erwachsenenbildnern*. Weinheim: Deutscher Studien Verlag.
- Wengert, H.G. (1989). *Untersuchungen zur alltäglichen Unterrichtsplanung von Mathematiklehrern*. Frankfurt: Peter Lang.
- Wernke, S., & Zierer, K. (Hrsg.). (2017). *Die Unterrichtsplanung: Ein in Vergessenheit geratener Kompetenzbereich?! Status Quo und Perspektiven aus Sicht der empirischen Forschung*. Bad Heilbrunn: Klinkhardt.