Unterrichtswissenschaft
https://doi.org/10.1007/s42010-024-00200-w

ALLGEMEINER TEIL

Empirisch-kriterienorientierte Analyse des fachdidaktischen Wissens angehender Physiklehrkräfte

Welche inhaltlichen Strukturen zeigen sich über unterschiedliche Projekte hinweg?

Jannis Zeller · Dustin Schiering · Christoph Kulgemeyer · Knut Neumann · Josef Riese · Stefan Sorge

Eingegangen: 28. Juni 2023 / Überarbeitet: 5. Februar 2024 / Angenommen: 1. März 2024

Zusammenfassung In den letzten Jahren wurde das Professionswissen (angehender) Lehrkräfte intensiv untersucht. Neben Aussagen zur inneren Struktur liegen auch Ergebnisse über den Zusammenhang zwischen Professionswissen, Performanz in prototypischen Handlungssituationen sowie Unterrichtserfolg vor. In diesen Analysen hat sich gezeigt, dass insbesondere dem fachdidaktischen Wissen eine zentrale Rolle zukommt. Es mangelt bisher jedoch an empirisch fundierten Beschreibungen von Niveaustufen des fachdidaktischen Wissens. Zwar liegen einzelne Vorschläge vor, diese sind jedoch entweder empirisch nicht fundiert oder post hoc generiert, so dass unklar ist, inwieweit die Beschreibung der Ausprägungen auch außerhalb der jeweiligen Projektkontexte anwendbar ist. Der vorliegende Artikel stellt eine projektübergreifende Analyse des fachdidaktischen Wissens mithilfe zweier Ansätze zur Bildung von Niveaustufen vor. Dazu werden Niveaumodelle mit Daten zum fachdidaktischen Wissen aus zwei Projekten ($N = 427$ und $N = 779$) mithilfe des Scale-Anchoring-Verfahrens sowie eines regressionsanalytischen Ansatzes auf Basis eines Modells hierarchischer Komplexität erstellt. Das Scale-Anchoring-Verfahren liefert Niveaubeschreibungen, die sich zwar bezüglich fachlicher und fachdidaktischer Inhalte unterschieden, aber Parallelen bezüglich lernpsychologisch interpretierbarer Operatoren zeigten. Projektübergreifend deuteten die Ergebnisse daraufhin, dass sich das fachdidaktische Wissen in niedrigen Ausprägungen auf reproduktive

✉ Jannis Zeller · Josef Riese
Didaktik der Physik, Department Physik, Universität Paderborn, Warburger Str. 100, 33098 Paderborn, Deutschland
E-Mail: jannis.zeller@uni-paderborn.de

Dustin Schiering · Knut Neumann · Stefan Sorge
Didaktik der Physik, IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Olshausenstraße 62, 24118 Kiel, Deutschland

Christoph Kulgemeyer
Institut für Didaktik der Naturwissenschaften, Abteilung Physikdidaktik, Universität Bremen, Otto-Hahn-Allee 1, 28359 Bremen, Deutschland

Published online: 15 April 2024

Aspekte beschränkt, in höheren Ausprägungen aber kreative und evaluierende Elemente hinzukommen. Das Modell hierarchischer Komplexität zeigte sich nur für einen der Datensätze als geeignet, um ein Niveaumodell abzuleiten und konnte daher für projektübergreifende Analysen nicht weiter genutzt werden. Nichtsdestotrotz lieferte die projektübergreifende Analyse mithilfe des Scale-Anchoring-Verfahrens kontextunabhängige Beschreibungen von Ausprägungen des fachdidaktischen Wissens und ermöglicht so erste Schritte in Richtung eines empirisch fundierten, inhaltlich reichhaltigen Assessments, welches über eine Einordnung mittels eines Scores hinaus geht.

Schlüsselwörter Fachdidaktisches Wissen · Niveaumodell · Projektübergreifende Analyse · Physik

Cross-project empirical and criteria-oriented analysis of pre-service physics teachers' pedagogical content knowledge

What content structures emerge in the context of different models?

Abstract In recent years, the professional knowledge of (pre-service) teachers has been intensively investigated. In addition to statements regarding its internal structure, there are also findings on the relationship between professional knowledge, performance in prototypical action situations, and teaching effectiveness. These analyses have shown that pedagogical content knowledge plays a central role. However, there is still a lack of an empirically grounded description of competency levels of pedagogical content knowledge. There have been some individual proposals, though they are either not empirically grounded or post hoc generated, leaving the extent to which the descriptions of such levels are applicable outside the specific project contexts unclear. This article presents a cross-project analysis of pedagogical content knowledge using two approaches to establish levels of proficiency. Therefore, level models were constructed based on data regarding pedagogical content knowledge from two projects ($N = 427$ and $N = 779$) using the Scale-Anchoring procedure and a regression-analytical approach based on a model of hierarchical complexity. The Scale-Anchoring procedure provided level descriptions that, despite differences in subject matter and pedagogical content, exhibited parallels in terms of operators that are interpretable in terms of learning psychology. Across projects, the results indicated that pedagogical content knowledge in low levels is limited to reproductive aspects but incorporates creative and evaluative elements in higher levels. The model of hierarchical complexity turned out to be properly applicable only for one of the datasets and thus could not be further utilized for cross-project analyses. Nevertheless, the cross-project analysis using the Scale-Anchoring procedure provided context-independent descriptions of levels of pedagogical content knowledge, thus enabling initial steps towards an empirically grounded, conceptually rich assessment that goes beyond solely preparing a quantitative score.

Keywords Pedagogical content knowledge · Competency level model · Cross-project analysis · Physics

1 Einleitung

Die professionelle Kompetenz (angehender) Lehrkräfte steht seit langem im Fokus der fachdidaktischen Forschung zur Professionalisierung von Lehrkräften (Baumert und Kunter 2006; Gess-Newsome 1999; Shulman 1986; Terhart 2012). Die professionelle Kompetenz wird dabei in unterschiedlichen Konzeptualisierungen als wesentlich für die Handlungsqualität im Unterricht oder für den Unterrichtserfolg aufgefasst (Ball et al. 2001; Harms und Riese 2018; Terhart 2012). Eine zunehmende Anzahl an Studien belegt diese Annahme (z. B. Keller et al. 2017; Kunter et al. 2013; Blömeke et al. 2022; Kulgemeyer und Riese 2018). Speziell in den Naturwissenschaften wurden in den vergangenen Jahren insbesondere die innere Struktur und die globale Entwicklung des Professionswissens sowie die Abhängigkeit dieser Entwicklung von anderen Konstrukten untersucht (Neumann et al. 2019; Riese et al. 2017; Sorge et al. 2018, 2019). Darüber hinaus liegen Ergebnisse zur Bedeutung des Professionswissens für die Performanz in prototypischen Handlungssituationen vor (z. B. Förtsch et al. 2016; Kulgemeyer und Riese 2018; Kulgemeyer et al. 2020; Riese et al. 2022).

Im Rahmen von Projekten wie den genannten werden üblicherweise ausgehend von gängigen Operationalisierungen des Professionswissens Testinstrumente erstellt, die häufig konkrete Aspekte wie das thematisierte Fachwissen oder spezielle Professionswissensdimensionen fokussieren. Dadurch wird ein direkter Vergleich der vorliegenden Ergebnisse erschwert, da unklar ist, inwieweit die durch diese Testinstrumente abgebildeten Konstrukte deckungsgleich sind. Gleichzeitig stellt die möglichst allgemeingültige, theoretisch begründete und empirisch fundierte Beschreibung von Ausprägungen oder sogar Entwicklungsstufen des Professionswissens und der Professionswissensdimensionen bereits länger ein Forschungsdesiderat dar (z. B. Kaiser et al. 2020), denn die Möglichkeit zur Einordnung von Personen oder Lerngruppen in ein entsprechendes Niveaumodell ist für eine inhaltlich nützliche Diagnose und die Identifikation von Entwicklungspotenzialen notwendig.

Das fachdidaktische Wissen (FDW) stellt in den meisten theoretischen Modellen eine Kerndomäne des Professionswissens von Lehrkräften dar und eine Vielzahl empirischer Ergebnisse belegt seine praktische Relevanz (z. B. Kulgemeyer und Riese 2018). Gerade für das FDW als „special amalgam" (Shulman 1987; siehe auch Neumann et al. 2019), d. h. als spezielle, für die Lehrprofession einzigartige „Mischung" von fachlichem und pädagogischem Wissen, gestaltet sich jedoch eine projektunabhängige Beschreibung von Ausprägungen als herausfordernd, denn auch aufgrund dieses Mischungscharakters fokussieren die in unterschiedlichen Studien verwendeten Testinstrumente häufig einzelne Aspekte wie z. B. konkretes Fachwissen und Subskalen (siehe z. B. Hume et al. 2019)[1]. Daher können bisherige Untersuchun-

[1] Den Autoren ist bewusst, dass gewisse Unterschiede zwischen den international üblichen, auf Shulman (1986, 1987) zurückgehenden Konzeptualisierungen des „Pedagogical Content Knowledge" (PCK) und dem im deutschsprachigen Raum verwendeten Konstrukt des FDW gibt (z. B. Gramzow et al. 2013; Vollmer und Klette 2023). Da sich die Analyse auf empirisch-inhaltliche Ergebnisse stützt, wird auf eine genaue Beschreibung der hier zugrundeliegenden theoretischen Modellierungen verzichtet. Ergebnisse zum Forschungsstand werden hier unter dem FDW gelabelt, auch wenn teilweise eher PCK untersucht wurde.

gen des FDW und deren Ergebnisse bisher meist nur eingeschränkt miteinander verglichen werden.

Aussagen über das FDW, die auf Analysen mithilfe quantitativer Globalscores von Bearbeitungen der Testinstrumente basieren, bleiben also inhaltlich recht allgemein und die Gültigkeit über die konkreten Projektkontexte hinaus ist trotz gemeinsamer theoretischer Fundierung ungeklärt, was zusammenfassende Betrachtungen und Implikationen über mehrere Projekte hinweg schwierig macht. Dass Operationalisierung des FDW entsprechend der Natur des Konstrukts in der Regel in (unterschiedliche) fachliche Kontexte/Inhaltsbereiche eingebettet sind[2] erschwert eine Analyse zusätzlich. Die vorliegende Arbeit macht sich daher ein regressionsanalytisches Verfahren (z. B. Woitkowski und Riese 2017) sowie das Scale-Anchroing-Verfahren (Beaton und Allen 1992; OECD 2018) zur Bildung von Niveaumodellen zunutze, um die nicht unmittelbar vergleichbaren quantitativen Aussagen unter Nutzung des vorhandenen Datenschatzes in inhaltlich-kriterienorientierte Beschreibungen zu überführen. Einerseits kann mithilfe solcher Beschreibungen die Vergleichbarkeit der tatsächlich abgebildeten Konstrukte, die durch die in den Projekten jeweils verwendeten Testinstrumente erfasst werden, durch eine Gegenüberstellung eingeschätzt werden. Andererseits können mithilfe der inhaltich-kriterienorientierten Beschreibungen auch inhaltliche Aussagen über Ausprägungen oder sogar Entwicklungsstufen des FDW empirisch fundiert abgeleitet werden, die wiederum differenziertere Einschätzungen der Kenntnisstände von Proband:innen oder Lerngruppen über die bloße Angabe eines Scores hinaus ermöglichen. Solche Einschätzungen würden beispielsweise in einem (Self-) Assessment für Studierende eine Möglichkeit bieten, neben quantitativen Einordnungen auch inhaltliche Lücken wie beispielweise Nachholbedarfe bezüglich konkreter fachdidaktischer Inhalte oder im Kontext konkreter Anforderungssituationen zu ermitteln. Sowohl die Gültigkeit empirischer Ergebnisse über die konkreten Projektkontexte hinaus als auch eine inhaltliche Einschätzung von Proband:innen sind grundlegend für einen effektiven und nützlichen Transfer der wissenschaftlichen Ergebnisse in die Praxis der Lehramtsausbildung.

Im Kontext des Professionswissens von Lehramtsstudierenden wurden entsprechende Verfahren zur Niveaubildung bereits mit Erfolg angewendet (König 2009; Schiering et al. 2023; Woitkowski 2020; Zeller et al. 2022). Hier werden erstmals im deutschsprachigen Raum solche Niveaumodelle genutzt, um die Ergebnisse zur empirisch-inhaltlichen Beschreibung des FDW zweier Projekte vergleichend zu analysieren. Dazu werden hier die Projekte ProfiLe-P+[3] (z. B. Vogelsang et al. 2019) und

[2] In der hier vorliegenden Analyse wurde dabei im ProfiLe-P-Projekt der fachphysikalische Inhalt auf „Mechanik“ fokussiert, während in den Projekten KiL/KeiLa mehrere Fachinhalte (Mechanik, Elektrizitätslehre, Optik, Thermodynamik, Atom- und Kernphysik, spezielle Relativitätstheorie, Festkörperphysik & Quantenmechanik) abgedeckt wurden.

[3] Akronym ProfiLe-P: „**Prof**essionskompetenz im **Le**hramtsstudium **P**hysik“, gefördert durch das Bundesministerium für Bildung und Forschung. In der ersten Projektphase (ProfiLe-P, siehe z. B. Riese und Reinhold 2012) wurde auf die Modellierung und Operationalisierung der Domänen des Professionswissens für das Fach Physik fokussiert. In der zweiten Projektphase (ProfiLe-P+ siehe z. B. Vogelsang et al. 2019) wurde die längsschnittliche Entwicklung sowie der Zusammenhang der Domänen des Professionswissens zur Performanz in prototypischen Handlungssituationen in den Blick genommen. Für die hier vorgestellte Analyse sind primär die Daten aus dem in ProfiLe-P entwickelten und in ProfiLe-P+ verwendeten FDW-Testinstruments (Gramzow 2015) relevant.

KiL[4] (z. B. Kleickmann et al. 2014) bzw. dessen Folgeprojekt KeiLa[5] (z. B. Schiering et al. 2023) gemeinschaftlich in den Blick genommen. In beiden Projekten waren Physik-Lehramtsstudierende die primäre Zielpopulation der Untersuchung. Insgesamt werden 1206 Testbearbeitungen (779 aus dem ProfiLe-P+-Projekt und 427 aus den Projekten KiL/KeiLa) von Physik-Lehramtsstudierenden zum FDW genutzt, um Niveaumodelle mithilfe des Scale-Anchoring-Verfahrens (z. B. Mullis und Fishbein 2020) und eines regressionsanalytischen Ansatzes (z. B. Nold et al. 2008; Woitkowski und Riese 2017) auf Basis hierarchischer Komplexität (Commons et al. 1998) zu entwickeln, welche anschließend zu projektübergreifenden, vergleichenden Betrachtungen auf inhaltlicher Ebene genutzt werden.

Diese projektübergreifende Betrachtung soll, wie oben bereits angedeutet, die Verallgemeinerbarkeit bzw. Allgemeingültigkeit etwaiger inhaltlicher Beschreibungen untersuchen. Durch die bisher isoliert stehenden Modellierungen können beispielsweise Untersuchungen der Entwicklung des FDW mithilfe der projektspezifischen Testinstrumente, wie etwa zur Evaluation einer Lehrveranstaltung, keine allgemeingültigen inhaltlichen Aussagen über den Wissenszuwachs der Proband:innen treffen. Es bleibt unklar, ob oder inwieweit ein über beide Projekte äquivalenter Wissenszuwachs auf Basis quantitativer Scores auch ähnliche Zuwächse in der Fähigkeit konkrete Anforderungen zu bewältigen beschreibt. Unter Umständen kann auch aus methodischer Sicht die Vorgehensweise selbst als Vorlage für projektübergreifende Analysen in Fällen dienen, in denen eine direkte gemeinsame quantitative Analyse nicht möglich ist, da sich Testinstrumente und Stichproben unterscheiden bzw. sogar beide disjunkt sind.

Abschließend werden Limitationen und Anwendungsmöglichkeiten der erhaltenen inhaltlichen Beschreibungen von Ausprägungen des FDW diskutiert. Darüber hinaus werden Optionen für weiterführende Forschung erörtert.

2 Theoretischer Hintergrund

Das Professionswissen von Lehrkräften wird in der Tradition Shulmans (1986, 1987) üblicherweise in Fachwissen (FW), Pädagogisches Wissen (PW) und FDW gegliedert (Baumert und Kunter 2006; speziell für das Fach Physik vgl. Riese 2009). Das FDW wird demnach als dasjenige Wissen aufgefasst, welches zur adressatengerechten Aufbereitung des FW notwendig ist und stellt somit eine zentrale Komponente des Professionswissens dar (Shulman 1987). Nachfolgend wird das in diesem Bei-

[4] Akronym KiL: „Messung professioneller **K**ompetenzen **i**n mathematischen und naturwissenschaftlichen **L**ehramtsstudiengängen“, gefördert durch Leibniz Gemeinschaft. In diesem Projekt wurde das Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer gemeinsam modelliert. Für die Physik wurde dabei ein FDW-Testinstrument durch Kröger (2019) entwickelt und in KiL sowie KeiLa eingesetzt.

[5] Akronym KeiLa: „**K**ompetenz**e**ntwicklung **i**n mathematischen und naturwissenschaftlichen **L**ehr**a**mtsstudiengängen“, gefördert durch Leibniz Gemeinschaft. Aufbauend auf den Modellierungen aus KiL wurde in KeiLa die Entwicklung des Professionswissens im Zusammenhang mit Lerngelegenheiten und individuellen Merkmalen der Proband:innen untersucht (z. B. Sorge et al. 2019). Auch hier wurde das FDW-Testinstrument nach Kröger (2019) wieder eingesetzt. In der hier vorgestellten Analyse werden die FDW-Daten aus beiden Projektphasen genutzt.

trag fokussiert betrachtete Konstrukt des FDW aus der Perspektive der Naturwissenschaftsdidaktik präzisiert und in relevante theoretische Rahmungen eingebettet.

2.1 Fachdidaktisches Wissen

Die Modellierungen des FDW (im englischsprachigen und internationalen Raum auch „Pedagogical Content Knowledge", kurz PCK, genannt[2]) unterscheiden sich zwar häufig im Detail (Gess-Newsome und Lederman 1999; Hume et al. 2019), gemein ist jedoch allen theoretischen Grundmodellen die o. g. Auffassung von FDW als spezifisches Wissen von Lehrkräften, welches zur adressatengerechten Aufbereitung von Fachwissen notwendig ist und mit den anderen Domänen des Professionswissens (FW & PW) in Beziehung steht (z. B. Shulman 1986; Baumert und Kunter 2006; Riese 2009). Dabei gibt es unterschiedliche strukturelle Ansätze, das FDW in der Bandbreite von eher deklarativem Wissen bis hin zu gezeigten Handlungen zu positionieren.

Einen prominenten Ansatz stellt hier das häufig als „Kontinuumsmodell" bezeichnete Konzept von Blömeke et al. (2015) dar, das Kompetenz als Kontinuum zwischen latenten kognitiven Dispositionen und gezeigter Performanz in für die Profession spezifischen Handlungssituationen beschreibt. Das in Testinstrumenten abrufbare FDW im hier beschriebenen Sinne lässt sich in diesem Modell eher auf Seite der kognitiven Dispositionen verorten, die wiederum eine Grundlage für situationsspezifische Fähigkeiten und Fertigkeiten darstellen (Blömeke et al. 2015). International speziell im Bereich der Naturwissenschaftsdidaktik etabliert ist darüber hinaus auch das sog. „Refined Consensus Model of PCK" (kurz RCM, Carlson und Daehler 2019), welches das FDW in die Bereiche *collective* PCK (cPCK), *personal* PCK (pPCK) und *enacted* PCK (ePCK) gliedert (siehe auch Alonzo et al. 2019). Dabei stellt cPCK die kollektive Wissensbasis der fachdidaktischen Community dar, pPCK das explizite Wissen einzelner Akteur:innen und ePCK das internalisierte Wissen, welches sich durch Performanz in spezifischen Situationen äußert. Eine knappe Gegenüberstellung der beiden theoretischen Ansätze des Kontinuumsmodells und des RCMs ist z. B. bei Kulgemeyer et al. (2020, S. 4–7) zu finden. Beide Modelle nehmen dabei an, dass das FDW bzw. PCK eine wichtige Voraussetzung für späteres professionelles Handeln im Klassenzimmer ist.

Hierzulande ist eine Gliederung des FDW in drei Dimensionen üblich (z. B. Gramzow 2015; Kröger 2019; Tepner et al. 2012). Dabei wird das FDW grundsätzlich als abhängig vom konkret betrachteten Fachinhalt (Dimension 1) aufgefasst. Im Falle der Physik sind dabei konkrete Inhaltsgebiete wie beispielsweise „Mechanik", „Optik" oder „Elektrizitätslehre" und nicht übergeordnete fachliche Dimensionen wie „Erkenntnisgewinnung" gemeint. Weiterhin umfassen die Modellierungen meist eine Dimension, die unterschiedliche fachdidaktische Inhalte/Facetten (Dimension 2) wie beispielsweise Schülerkognition oder Instruktionsstrategien abbildet. Es existieren zahlreiche Kataloge relevanter Facetten, die u. a. Kirschner (2013) in einer Übersicht gegenübergestellt hat. Dabei ist auffällig, dass die Facetten *Schüler und Schülerkognition*[6] sowie *Instruktions- und Vermittlungsstrategien* fast allen Model-

[6] Die Facette wird hier wie im Original benannt und daher nicht geschlechtsneutral umformuliert.

lierungen gemein ist. Diese und die weiteren genutzten Facetten werden primär aus den ursprünglichen theoretischen Modellierungen des FDW (z. B. Shulman 1986; Carlson und Daehler 2019), Analysen der Curricula der Lehrerbildung bzw. Literatur-Reviews (z. B. Kröger 2019; Gramzow et al. 2013) sowie Expertenbefragungen zu Sicherstellung der curricularen Validität entsprechender Items (z. B. Gramzow 2015) abgeleitet. Auch die Items zu den o. g. Facetten *Schüler und Schülerkognition* und *Instruktions- und Vermittlungsstrategien* wurden in den entsprechenden Befragungen als curricular passend eingeschätzt (Gramzow 2015, S. 166–168). Aus Gründen der Testökonomie und Zumutbarkeit wird bei der Entwicklung konkreter Testinstrumente meist eine Auswahl entsprechender Facetten getroffen. Die dritte Dimension der Itementwicklungsmodelle dient üblicherweise zur Anreicherung der Anforderungsbereiche der Testinstrumente (Klieme et al. 2003). So findet sich bei Tepner et al. (2012) sowie Kröger (2019) eine Dimension „Wissensarten" (S. 50) und bei Gramzow (2015) eine Dimension „Kognitive Aktivität" (S. 104).

Für die Physik sind hier die Modelle des FDW, die den Testinstrumenten von Kröger (2019) und Gramzow (2015) (zur Itementwicklung) zugrunde liegen, exemplarisch dargestellt (Abb. 1). Auffällig ist auch hier, dass in beiden Modellen jeweils eine Facette zu Schülerkognition und eine Facette zu Instruktionsstrategien enthalten ist. Auch Tepner et al. (2012) schließen in ihrer Dimensionierung, die weitgehend Analog zu der von Kröger (2019) aufgebaut ist, die Facette der Schülervorstellungen explizit mit ein. Die anderen beiden Facetten weichen jedoch voneinander ab. Für die Begründung der Auswahl der entsprechenden Facetten sei auf die Originalquellen (z. B. Gramzow 2015, S. 96–105; Kröger 2019, S. 46–47; Tepner et al. 2012, S. 13–16) verwiesen.

Speziell für das Fach Physik belegen konkrete Forschungsergebnisse aus Quer- und Längsschnitten signifikante Zuwächse des FDW im Studium und Vorbereitungsdienst (Kirschner 2013; Kröger 2019; Riese und Reinhold 2012). Weiterhin zeigen sich im naturwissenschaftlichen Bereich signifikante Zusammenhänge zwischen FDW und FW bzw. PW (Riese und Reinhold 2012; Sorge et al. 2019) und Zusammenhänge zwischen FDW und Performanz in prototypischen Anforderungssituationen, wie beispielsweise (1) der Unterrichtsplanung (Behling et al. 2022b; Riese et al. 2022; Schröder et al. 2020), (2) dem Erklären physikalischer Phänomene (Kulgemeyer und Riese 2018; Kulgemeyer et al. 2020), (3) dem Reflektieren über Unterricht (Kulgemeyer et al. 2021), (4) der kognitiven Aktivierung (Förtsch et al. 2016), (5) der Nutzung von physischen Modellen (Förtsch et al. 2018) sowie (6) diagnostischen Handlungen (Kramer et al. 2021). Für den MINT-Bereich wurden zudem (häufig mediative) Einflüsse des FDW auf Aspekte der Unterrichtsqualität bzw. des Unterrichtserfolgs (Behling et al. 2022a, b; Blömeke et al. 2022; Keller et al. 2017) festgestellt. Diese Ergebnisse sind konform zu den theoretischen Annahmen, beispielsweise der angenommenen Notwendigkeit von FDW zur Aufbereitung fachlicher Inhalte bei Shulman (1986). Auch die angenommene Wirkkette der schulischen Bildung nach Terhart (2012) macht diese Ergebnisse plausibel. Somit ist das besondere Augenmerk auf das FDW als wichtige Dimension des Professionswissens sowohl empirisch als auch theoretisch zu rechtfertigen.

Statistische Zusammenhangs- und Mediationsanalysen in der Art der genannten Studien zielen dabei naturgemäß im Wesentlichen auf Schlussfolgerungen auf Basis

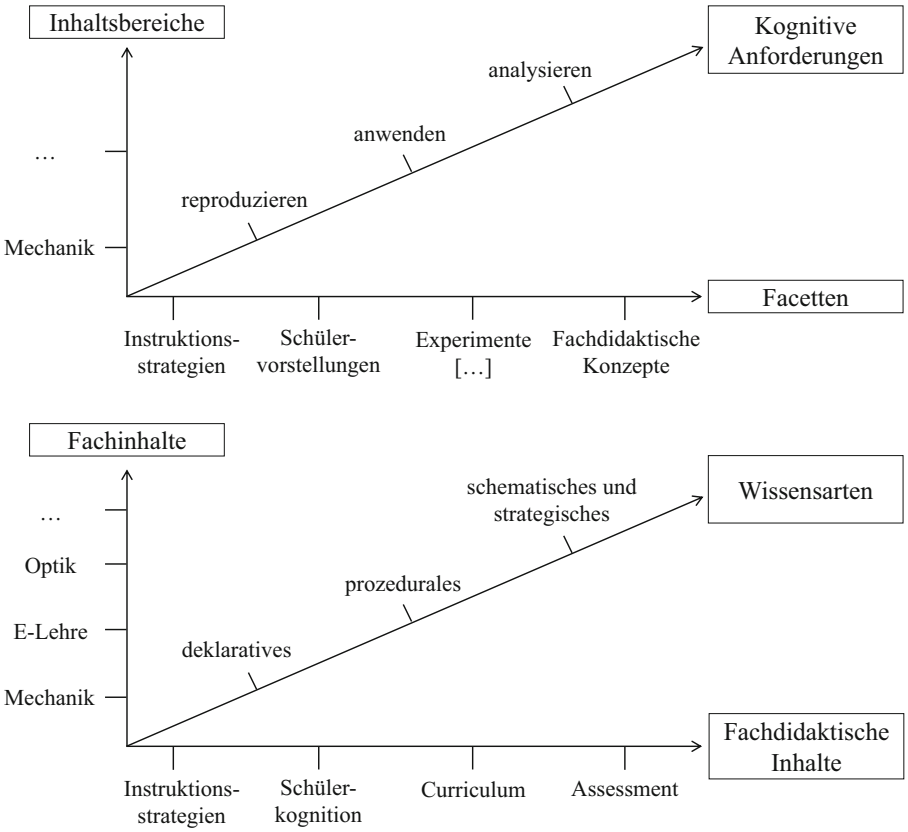


Abb. 1 Itementwicklungsmodelle zu den Testinstrumenten nach Kröger (2019, S. 50) *oben* und Gramzow (2015, S. 104) *unten*

quantitativer Ausprägungen ab (Reinhold et al. 2017) und treffen dabei keine Aussagen über die (inhaltliche) Art dieser Ausprägungen. In der Folge stellen Mientus et al. (2022) im Rahmen eines systematischen Reviews fest, dass in bisheriger internationaler Forschung zur inhaltlichen Charakterisierung des FDW im MINT-Bereich primär qualitative Untersuchungsmethoden genutzt wurden. Weiterhin beobachten sie, dass quantitative empirische Analysen, wenn auch zur Beantwortung unterschiedlicher Forschungsfragen und Untersuchung unterschiedlicher Zusammenhänge, weitestgehend auf Globaleinschätzungen abzielen.

2.2 Kompetenzniveaumodelle

Kompetenzniveaumodelle werden allgemein als geeignetes Mittel zur inhaltlichen Beschreibung von hierarchischen Ausprägungen unterschiedlicher Konstrukte aufgefasst (Beaton und Allen 1992; Lok et al. 2016) und wurden beispielsweise in den Large-Scale Schulleistungsstudien wie PISA und TIMSS zur inhaltlichen Beschreibung von Fähigkeitsniveaus verwendet (z. B. Mullis et al. 2016; OECD 2018). Die

inhaltliche Beschreibung entsprechender Ausprägungen auf Basis quantitativer Daten bietet dabei die Chance, quantitative Ergebnisse und qualitative Beschreibungen zu verbinden. Die Nutzung der Testdaten validierter Testinstrumente stellt hierbei auch ein Validitätsargument für die erhaltenen Niveaumodelle dar. Es existieren unterschiedliche Möglichkeiten, aus Testscores inhaltliche Niveaumodelle abzuleiten, die sich deutlich unterscheiden. Woitkowski (2020) stellt im Rahmen seiner Adaption eines dieser Verfahren eine Übersicht u. a. des Scale-Anchoring-Verfahrens und regressionsanalytischer Ansätze vor. Beide Verfahren nutzen ein IRT[7]-Modell als Ausgangspunkt, mit dem eine gemeinsame Abbildung von Personenfähigkeiten und Aufgabenschwierigkeiten auf eine Skala mit inhärenter Hierarchie ausgenutzt wird, so dass Aufgaben und Personen direkt miteinander in Beziehung gesetzt werden können (siehe z. B. Moosbrugger und Kelava 2020; Neumann 2014).

Im Scale-Anchoring-Verfahren wird über mehrere Schritte aus einem IRT-Modell ein inhaltliches Niveaumodell gebildet (Mullis und Fishbein 2020; OECD 2018). Dabei werden zunächst Personengruppen mithilfe der Fähigkeits-Verteilungen gebildet (beispielsweise eine Gruppe mit niedriger, eine mit mittlerer und eine mit hoher Fähigkeit). Anschließend werden die Aufgaben gemäß ihrer Lösungshäufigkeit in den unterschiedlichen Personengruppen wiederum in Gruppen eingeteilt. Die mittleren Schwierigkeitsparameter der Aufgabengruppen dienen dann zur Bildung der Niveaugrenzen, da sie sich durch die Nutzung des IRT-Modells direkt auf die Personenfähigkeiten beziehen lassen. Die inhaltlichen Beschreibungen der Niveaus werden anschließend durch die Aufgaben, deren Schwierigkeitsparameter sich nahe an den Niveaugrenzen befinden, erstellt. Der genaue Ablauf des Verfahrens wird auch in Abschn. 4.2 noch einmal bei der konkreten Anwendung deutlich. Die Niveaustruktur und die inhaltlichen Niveaucharakterisierungen werden somit vollständig induktiv aus dem Modell abgeleitet, wodurch der qualitative Aufwand sich auf die inhaltliche (Re-)Analyse weniger Aufgaben reduziert. Das Verfahren zeichnet sich dadurch durch vergleichsweise hohe Objektivität und Effizienz aus. Allerdings ist eine möglichst große Anzahl an Aufgaben an den jeweiligen Niveaugrenzen für eine reliable Niveaucharakterisierungen hier optimal. Das Scale-Anchoring-Verfahren wurde bereits mehrfach zur Analyse des FDW im deutschsprachigen Raum eingesetzt (Schiering et al. 2019, 2023; Zeller et al. 2022). In Niveauanalysen im Kontext anderer Domänen des Professionswissens werden anstelle des Scale-Anchoring-Verfahrens meist stärker theoriegeleitete Ansätze genutzt.

Eine Alternative zum Scale-Anchoring-Verfahren bietet beispielsweise ein regressionsanalytischer Ansatz (Blömeke et al. 2008; Nold et al. 2008; Woitkowski 2020). Dazu werden schwierigkeitserzeugende Merkmale aus theoretischen Überlegungen abgeleitet (z. B. sprachliche Terminologie und Komplexität kognitiver Bearbeitungsprozesse bei König 2009) und die Aufgaben bezüglich dieser Merkmale gruppiert. Anschließend wird mithilfe einer linearen Regression die Varianzaufklärung dieser Gruppierung bzgl. der Aufgabenschwierigkeit bestimmt und somit die Eignung des Modells geprüft. Zeigt das Modell eine ausreichende Passung, können wiederum die mittleren Aufgabenschwierigkeiten durch das IRT-Modell als Niveaugrenzen aufgefasst werden (analog zu den Aufgabengruppen aus dem Scale-

[7] IRT wird als Abkürzung für Item Response Theorie verwendet.

Anchoring-Verfahren). Die Niveaucharakterisierungen ergeben sich dann implizit durch die Beschreibung der schwierigkeitserzeugenden Merkmale. Da der regressionsanalytische Ansatz die Entwicklung eines Modells für schwierigkeitserzeugende Merkmale und eine (Re-)Analyse aller Aufgaben bzgl. dieser Merkmale erfordert, ist er aufwändiger als das Scale-Anchoring-Verfahren. Auf der anderen Seite können mithilfe des regressionsanalytischen Ansatzes (nach entsprechender theoretischer Vorarbeit) Informationen aus allen Aufgaben und Expertenwissen bzgl. aller Aufgaben zur inhaltlichen Charakterisierung mit herangezogen werden, weshalb dieser Ansatz gerade bei einer geringen Anzahl verfügbarer Aufgaben attraktiv ist. Besonders für eine projektübergreifende Analyse sollte das theoretisch zugrunde gelegte Modell schwierigkeitserzeugender Merkmale unabhängig vom konkreten Testinstrument sein. Im naturwissenschaftsdidaktischen Kontext wurde der regressionsanalytische Ansatz bereits mehrfach bei Fachwissenstests eingesetzt (Bernholt 2010; Woitkowski 2019; Woitkowski und Riese 2017).

2.3 Hierarchische Komplexität des FDW

Bei den in Abschn. 2.2 genannten regressionsanalytischen Ansätzen zur Kompetenzniveauermittlung wurde als „schwierigkeitserzeugendes Merkmal“ mehrfach ein Modell hierarchischer Komplexität der Aufgabenanforderungen angelehnt an das „Model of hierarchical Complexity“ nach Commons et al. (1998) (siehe auch Commons et al. 2014) entwickelt bzw. für das jeweils fokussierte Konstrukt adaptiert. Die hierarchische Komplexität stellt dabei ein Schema dar, nach dem die Qualität von Wissen als propositionales Netzwerk im lernpsychologischen Sinne (z. B. Schnotz 1994) eingeschätzt werden kann. Der grundlegende Ansatz ist, dass höhere Qualität des Wissens nicht durch bloße Breite, sondern durch den Grad der Vernetzung des Wissensnetzwerks entsteht. Höhere Komplexitätsstufen bauen dabei auf niedrigeren auf, indem sie die Wissensstrukturen dieser niedrigeren Stufen reorganisieren. Es stellt somit einen etablierten, vereinheitlichten Ansatz dar, um die Qualität von Wissensstrukturen in unterschiedlichen Bereichen zu beschreiben (siehe Woitkowski und Riese 2017).

Das Modell hierarchischer Komplexität wurde also bereits in unterschiedlichen Kontexten erfolgreich genutzt. Es umfasst allgemeine kognitive Prozesse und ist insofern auch für das FDW ein aussichtsreicher Kandidat zur vereinheitlichten Beschreibung schwierigkeitserzeugender Merkmale. Da für das physikalische Fachwissen bereits ein Komplexitätsmodell existiert, welches mit Erfolg zur Modellierung von Niveaustufen genutzt wurde (Woitkowski und Riese 2017) wäre es zudem wünschenswert die Adaptierbarkeit dieses Modells für das FDW zu überprüfen (siehe Abschn. 4.4).

3 Ziele der Analyse

Die empirisch fundierte inhaltliche Beschreibung von Ausprägungen des FDW z. B. in Form von Niveaumodellen stellt nach wie vor ein Desiderat fachdidaktischer Forschung dar. Eine Möglichkeit der Beschreibung solcher Ausprägungen von Stu-

dierenden und Lerngruppen, ist sowohl für individual- als auch systemdiagnostische Zwecke und die Entwicklung oder Auswahl passender Fördermöglichkeiten notwendig. Bisher liegen jedoch von empirischer Seite im deutschsprachigen Raum hauptsächlich quantitative, globale Analysen und Ergebnisse zum FDW vor, in welchen die inhaltliche Komponente weniger fokussiert wurde. Erste entsprechend inhaltlich angereicherte, kriterienorientierte Ergebnisse sind Projekt- bzw. Testinstrument-spezifisch und stehen dadurch zunächst isoliert. Prinzipiell bieten IRT-Modellierungen die Möglichkeit, auch Datensätze zu unterschiedlichen Testinstrumenten zu verbinden, indem Stichproben von Proband:innen die mehrere Testinstrumente bearbeiten haben, gebildet werden oder indem identische Ankeritems in beiden Tests verwendet werden (siehe z. B. Lee und Lee 2018). Die nachträgliche Erhebung von entsprechenden Normstichproben gestaltet sich aber in der Fachdidaktik aufgrund kleiner Populationsgrößen und schwierigem Zugriff auf geeignete Stichproben meist nicht praktikabel. Eine projektübergreifende inhaltliche Beschreibung von Ausprägungen des FDW ist aber sowohl zur Vergleichbarkeit von gefundenen quantitativen Ausprägungen des FDW unter der Nutzung unterschiedlicher Testinstrumente als auch zur Validierung von Einordnungen von Proband:innen vor dem Hintergrund einzelner Modellierungen notwendig.

Erst seit kurzem wird auch die inhaltliche Beschreibung von Ausprägungen des FDW auf Basis quantitativer empirischer Ergebnisse in den Blick genommen. Dazu wurden erste datenbasierte kriterienorientierte/inhaltliche Beschreibungen von Ausprägungen des FDW im Rahmen von IRT-Modellierungen entwickelt. Dabei wurde das Scale-Anchoring-Verfahren (z. B. Mullis et al. 2016) auf die Daten aus dem KiL-Projekt (Schiering et al. 2019) sowie vorläufigen Daten ($N < 150$) zu einer geschlossenen Version des in ProfiLe-P konzipierten und verwendeten Testinstruments (Kulgemeyer et al. 2023) angewandt (Zeller et al. 2022). Die Ergebnisse dieser Analysen deuteten in beiden Projekten auf übergeordnete Parallelen bzgl. der erhaltenen Niveaustufen hin: In niedrigen Ausprägungen schien sich das FDW vor allem auf reproduktive Aspekte zu beschränken, während in höheren Ausprägungen auch kreative und evaluierende Elemente hinzukamen (Schiering et al. 2019, S. 224; Zeller et al. 2022, S. 770). Um diese Beobachtung weiter zu systematisieren und ggf. zu bestätigen, soll in diesem Beitrag eine erweiterte Niveauanalyse der Daten aus den KiL/KeiLa-Projekten von Schiering et al. (2023) mit einer Re-Analyse des ProfiLe-P+-Datensatzes im Rahmen von Niveaumodellierungen inhaltlich verglichen werden. Dieses Vorgehen kann sich unter Umständen als Vorlage für ähnliche projektübergreifende Betrachtungen in anderen verwandten Felder erweisen.

Ziel dieses Beitrags ist also erstens die datengestützte kriterienorientiert-inhaltliche Beschreibung von Ausprägungen des FDW, um damit zweitens die Verknüpfung der Ergebnisse zweier unabhängiger Large-Scale Studien (für fachdidaktische Größenordnungen) auf Basis entsprechender inhaltlicher Ergebnisse zu ermöglichen. Dazu werden die folgenden Forschungsfragen formuliert:

- *FF1:* Inwieweit lassen sich mithilfe des Scale-Anchoring-Verfahrens projektübergreifend inhaltliche Strukturen des FDW identifizieren und inhaltlich charakterisieren?
- *FF2:* Inwieweit lassen sich Stufen hierarchischer Komplexität des FDW projektübergreifend identifizieren und inhaltlich charakterisieren?

Zunächst wird dazu analog zum Vorgehen von Schiering et al. (2023) das Scale-Anchoring-Verfahren auf den ProfiLe-P+-Datensatz angewendet. Der inhaltliche Vergleich der Ergebnisse findet dann durch eine Gegenüberstellung der erhaltenen Niveaubeschreibungen statt. Anschließend wird ein Modell hierarchischer Komplexität für das FDW zur Niveaubildung mithilfe eines regressionsanalytischen Ansatzes ausgehend vom ProfiLe-P+-Datensatz vorgeschlagen und die Übertragbarkeit auf die KiL/KeiLa-Daten untersucht. Es wird dabei in den Blick genommen, ob mit den Scale-Anchoring-Analysen erhaltene inhaltliche Parallelen sich durch ein solches Modell hierarchischer Komplexität unterstützen, erweitern oder erklären lassen. Etwaige projektübergreifende Strukturen bieten einerseits Potenziale für die Nutzung als Grundlage für Feedback im Rahmen der Lehrpraxis, andererseits erweitern sie den Forschungsstand um allgemein zutreffende Aussagen über Ausprägungen des FDW.

4 Methoden

Zur Beantwortung der Forschungsfragen werden das Scale-Anchoring Verfahren und ein regressionsanalytischer Ansatz zur Niveaubildung synchron auf die Daten der beiden Projekte angewandt. Im Falle des Scale-Anchroing Verfahrens findet die projektübergreifende Analyse durch die gemeinsame vergleichende Betrachtung der erhaltenen Niveauformulierungen statt. Die regressionsanalytische Betrachtung fußt auf einem zu diesem Zweck entwickelten Modell hierarchischer Komplexität für das FDW. Die projektübergreifende Analyse findet hierbei durch die Überprüfung der Anwendbarkeit des Komplexitätsmodells auf beide Datensätze statt. Beide in dieser Analyse verwendete Operationalisierungen lassen sich vor dem Hintergrund des RCM im Rahmen des pPCK, d. h. dem „testbaren“ persönlichen FDW der Proband:innen, interpretieren (siehe Riese et al. 2022 für ProfiLe-P sowie Schiering et al. 2023 für KiL/KeiLa).

Sowohl das Scale-Anchoring-Verfahren als auch der regressionsanalytische Ansatz basieren auf einem IRT-Modell des jeweiligen Datensatzes. Für die KiL/KeiLa-Daten wurde dasselbe IRT-Modell wie bei Schiering et al. (2023) verwendet. Für die ProfiLe-P+-Daten wurde nach einer Bereinigung des Datensatzes ein neues IRT-Modell erstellt. In beiden Fällen wurde dabei das Paket „Test Analysis Modules“ (TAM, Robitzsch et al. 2022) auf Basis der Statistik-Software R (R Core Team 2023) verwendet. Der Workflow der Analysen ist in Abb. 2 dargestellt.

4.1 Testinstrumente und Stichproben

Der Datensatz des ProfiLe-P+-Projekts (Vogelsang et al. 2019) beinhaltet 846 Bearbeitungen des FDW-Testinstruments nach Gramzow (2015), das FDW in den Facetten *Schülervorstellungen, Fachdidaktische Konzepte, Experimente und Vermittlung eines angemessenen Wissenschaftsbegriffs* sowie *Instruktionsstrategien* abbildet. Beschreibungen des inhaltlichen Verständnisses dieser Facetten haben Riese et al. (2017, S. 103–104) knapp zusammengefasst. Bezüglich des fachphysikalischen Inhalts wurde sich im ProfiLe-P-Projektverbund übergreifend auf die Mechanik fest-

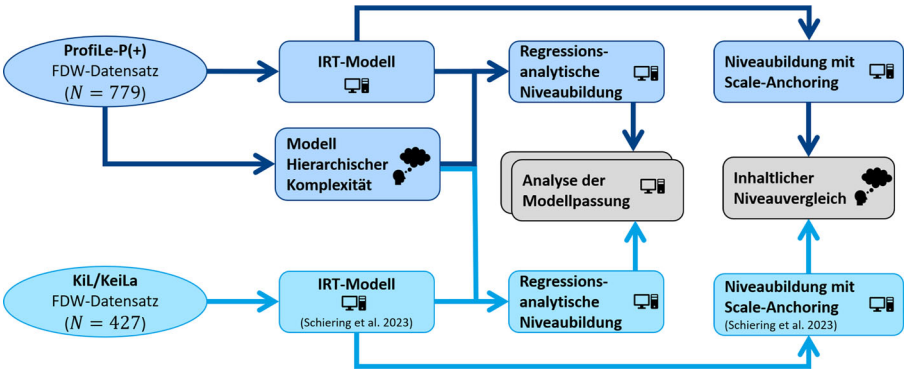


Abb. 2 Analyse-Workflow der vorgestellten Untersuchung

gelegt, um zu diesem Bereich empirisch trennbare Teilskalen auf Facettenebene erfassen zu können (Riese et al. 2015). Insgesamt besteht das Testinstrument aus 20 offenen und 4 geschlossenen (Multiple-Choice) Aufgaben und wurde im Rahmen des ProfiLe-P+-Projekts in den Jahren 2016 bis 2019 von Bachelor- und Masterstudierenden des Physik-Lehramts aus 12 deutschsprachigen Universitäten bearbeitet. Ein Beispielitem aus diesem Testinstrument ist in Abb. 3 dargestellt. Aus diesen Erhebungen blieben nach einer intensiven Bereinigung der Daten und dem Ausschluss von unvollständigen Bearbeitungen 779 Bearbeitungen (34 % weiblich, Studienjahr $M = 2{,}11$, $SD = 1{,}75$) für die hier verwendete Modellierung.

In den Projekten KiL und KeiLa wurde ein FDW-Testinstrument (Kröger 2019; Sorge et al. 2019) eingesetzt, welches FDW im Rahmen der fachdidaktischen Inhalte (analoge Dimension zu den „Facetten“ in ProfiLe-P+) *Schülerkognition, Instruktionsstrategien, Curriculum* und *Assessment* abbildet. Das inhaltliche Verständnis dieser Aspekte führt Kröger (2019, S. 46–47) genauer aus. Es wurde darauf abgezielt, das FDW bzgl. der fachlichen Inhalte breit zu untersuchen und somit die fachphysikalischen Inhalte Mechanik, Elektrizitätslehre, Optik, Thermodynamik, Atom- und Kernphysik, spezielle Relativitätstheorie, Festkörperphysik sowie Quantenmechanik eingeschlossen. Das Testinstrument besteht insgesamt aus 18 offenen und 21 geschlossenen Aufgaben. Ein Beispielitem aus diesem Testinstrument ist in Abb. 4 dargestellt. Der Datensatz des KiL/KeiLa-IRT-Modells besteht insgesamt aus 200 Bearbeitungen dieses Testinstruments aus der Querschnitterhebung des KiL-Projekts (2013, 12 Universitäten) und 227 Bearbeitungen aus den Längsschnitterhebungen des KeiLa-Projekts (2014 bis 2017, 20 Universitäten)[8].

4.2 Item-Response-Modellierungen

Um möglichst vergleichbare Niveaumodelle zu konstruieren, wurde bereits bei der IRT-Modellierung ein analoges Vorgehen zu der bereits bestehenden Analyse von Schiering et al. (2023) gewählt. Aufgrund der für die Anwendung von Niveaubil-

[8] Eine ausführlichere Beschreibung der Stichproben der Projekte KiL und KeiLa kann in Schiering et al. (2023, S. 8) gefunden werden.

Aufgabe 10 [27d]

Im Physikunterricht der Klasse 10 möchten Sie als Ziel Ihrer Unterrichtsstunde den Zusammenhang zwischen Weg und Zeit ($s \sim t^2$) beim freien Fall im Schülerversuch erarbeiten lassen.

Im Klassengespräch wurden Vermutungen über denkbare Zusammenhänge von Weg und Zeit formuliert und an der Tafel zur Prüfung durch Schülerversuche festgehalten. Von den Schülern wurden ein linearer und ein nicht-linearer Zusammenhang vermutet.

Im Schülerversuch lassen Schülergruppen jeweils eine kleine Stahlkugel im Treppenhaus der Schule aus verschiedenen Höhen fallen und messen die Zeit vom Loslassen bis zum Aufschlagen mit einer Stoppuhr.

Anschließend tragen sie ihre Messergebnisse jeweils in ein Zeit-Weg-Diagramm ein und stellen die von ihnen daraus gezogenen Schlussfolgerungen bei der abschließenden Präsentation auf Folien dar.

Sie bemerken, dass die Gruppen zu keinem eindeutigen Ergebnis gekommen sind. Einige präsentieren einen quadratischen, andere einen linearen, wieder andere einen nicht linearen Zusammenhang.

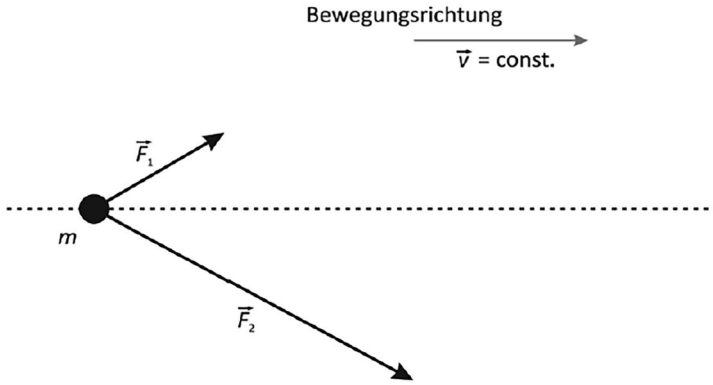
Formulieren Sie eine angemessene Reaktion: Skizzieren Sie dazu stichwortartig Ihr mögliches Vorgehen im weiteren Unterrichtsverlauf, um ausgehend von der gegebenen Situation den Zusammenhang $s \sim t^2$ zu erarbeiten.

Abb. 3 Beispielitem aus dem FDW-Testinstrument des ProfiLe-P-Projekts (Gramzow 2015, S. 235)

dungsverfahren vergleichsweise geringen Aufgabenanzahl wurde ein eindimensionales Partial-Credit-Modell (Masters 1982) verwendet, wobei Thurstone-Thresholds zur Schätzung der Itemschwierigkeiten bei polytomen Aufgaben verwendet wurden (Linacre 1998). Zur gemeinsamen Modellierung wurden Datensätze, die derselben Person zugeordnet sind, im Rahmen der Methode virtueller Proband:innen (von Davier et al. 2008) als unabhängige Datensätze modelliert, d. h. jede Bearbeitung fließt in die Modellierung als eigene „Datenzeile“ ein, ohne dass weiter beachtet wird, dass es sich um dieselbe Person handelt. Das erhaltene Modell für die ProfiLe-P+-Daten wies mit einer EAP-Reliabilität von 0,71 und Item-Outfits im Bereich von 0,8 bis 1,2 hinreichende Fit-Qualität für die weitere Analyse auf.

Für die Daten der KiL/KeiLa-Projekte wurde das bereits bestehende IRT-Modell von Schiering et al. (2023) basierend auf 427 Bearbeitungen herangezogen. Auch hier waren die Fit-Gütekriterien mit einer EAP-Reliabilität von 0,72 und Item-Outfits ebenfalls im Bereich von 0,8 bis 1,2 zufriedenstellend.

Schülerinnen und Schülern fällt es oft schwer, die Newtonschen Axiome zur Lösung konkreter Aufgaben anzuwenden.
Betrachten Sie die folgende Situation: Ein kleiner Körper der Masse m bewegt sich reibungsfrei und mit konstanter Geschwindigkeit $\vec{v}$ nach rechts. Auf den Körper wirken dabei drei Kräfte. Zwei davon sind eingezeichnet. Sie bitten die Schülerinnen und Schüler, die dritte Kraft einzuzeichnen.



Welche physikalisch falsche Antwort würden Sie von den Schülerinnen und Schülern erwarten?

Mögliche korrekte Antworten:
[Kraftpfeil in Bewegungsrichtung] – Es muss eine Kraft für die Bewegung verantwortlich sein.
[Kraftpfeil als Summe von F_1 und F_2] – Reflexartiges Zeichnen eines Kräfteparallelogramms.
[Kraftpeil als Verlängerung von F_1] – Summe aller Kräfte muss in Bewegungsrichtung zeigen.

Abb. 4 Beispielitem aus dem FDW-Testinstrument des KiL-Projekts (Schiering et al. 2019, S. 225)

4.3 Scale-Anchoring-Verfahren

Zur Beantwortung der ersten Forschungsfrage wurde das Scale-Anchoring-Verfahren (z. B. Mullis et al. 2016) auf das IRT-Modell der ProfiLe-P+-Daten angewendet. Im ersten Schritt wurden dazu die Item- und Personenparameter gemeinsam auf eine praktikablere Skala mit Mittelwert 500 und Standardabweichung 100 transformiert. Anschließend wurden drei Probandengruppen durch eine äquidistante Zerlegung der Fähigkeitsskala gebildet (Abb. 5). Zur absichernden Kontrolle, dass die so gefundenen Gruppen ausreichend unterschiedlich (Woitkowski und Riese 2017) waren, wurden inferenzstatistische Betrachtung mithilfe verteilungsfreier Tests (Kruskal-Wallis und Mann-Whitney *U* Tests) nach dem Vorbild von Schiering et al. (2023) durchgeführt, die eine ausreichende Differenzierung der Gruppen bestätigten (Tab. 1).

Auf Basis dieser Probandengruppen wurden die Aufgaben analog zum von Schiering et al. (2023; adaptiert nach Mullis und Fishbein 2020) genutzten Schema in Aufgabengruppen eingeteilt:

1. Aufgabengruppe 1: Mehr als 55 % der Personen aus Personengruppe 1 haben die Aufgabe gelöst.
2. Aufgabengruppe 2: Mehr als 55 % der Personen aus Personengruppe 2 und weniger als 50 % der Personen aus Personengruppe 1 haben die Aufgabe gelöst.

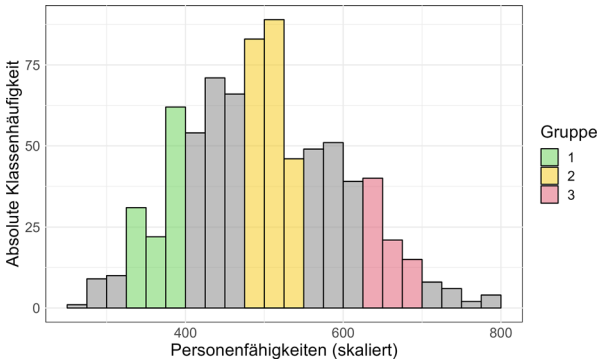


Abb. 5 Personengruppen aus dem ersten Schritt des Scale-Anchoring-Verfahrens (ProfiLe-P+-Daten). Die Personengruppen wurden als äquidistante Abschnitte der (skalierten) Fähigkeitsparameter gewählt. Das Scale-Anchoring Verfahren erwies sich als robust gegenüber leichter Verschiebungen dieser Abschnitte

Tab. 1 Beschreibung der Personengruppen aus dem ersten Schritt des Scale-Anchroing-Verfahrens (ProfiLe-P+-Daten). Ein Kruskal-Wallis Test bestätigte signifikante Gruppenunterschiede ($\chi^2(2) = 335$, $p < 0{,}001$). In der Tabelle sind anschließend paarweise Post-Hoc Mann-Whitney *U* Tests berichtet

Gruppe	Fähigkeitsspanne	*N*	*M*	SD	Differenz und *p*-Wert
1	325–400	115	370	23	140 ($W = 0$, $p < 0{,}001$)
2	475–550	218	510	19	143 ($W = 0$, $p < 0{,}001$)
3	625–700	76	653	22	–

3. Aufgabengruppe 3: Mehr als 55 % der Personen aus Personengruppe 3 und weniger als 50 % der Personen aus Personengruppe 2 haben die Aufgabe gelöst.
4. Aufgabengruppe 3+: Weniger als 50 % der Personen aus Personengruppe 3 haben die Aufgabe gelöst.

Die Mittelwerte der Schwierigkeitsparameter der Aufgabengruppen dienten dann als Schätzungen für die empirischen Niveaugrenzen. Auch hier wurden, um eine Vergleichbarkeit zu Schiering et al. (2023) beizubehalten, anschließend an die Zuordnung der Aufgaben verteilungsfreie statistische Tests zur Überprüfung der Unterscheidbarkeit der Aufgabengruppen durchgeführt (Tab. 2). Dabei wurde zudem das Abstandskriterium überprüft, d. h. es wurde getestet, ob eine Person mit einem Fähigkeitsparameter, der der Niveaugrenze des Niveaus *n* entspricht, einer Aufgabe

Tab. 2 Beschreibung der Aufgabengruppen aus dem zweiten Schritt des Scale-Anchroing-Verfahrens (ProfiLe-P+-Daten). Ein Kruskal-Wallis Test bestätigte signifikante Gruppenunterschiede ($\chi 2(3) = 27,9$, $p < 0{,}001$). In der Tabelle sind anschließend paarweise Post-Hoc Mann-Whitney *U* Tests berichtet. Dabei ist der Vergleichstest für die Aufgabengruppen 1 und 2 hier nur der Vollständigkeit halber angegeben, da er aufgrund der geringen Aufgabenanzahl in Aufgabengruppe 1 nicht sinnvoll interpretierbar ist – hier ist $p = 0{,}096$ bereits der „minimal erreichbare" *p*-Wert beim Vergleich zweier Gruppen mit 2 und 5 Elementen

Aufgaben-gruppe	*N*	*M*	SD	Differenz und *p*-Wert	*P* Abstandskriterium
1	2	–1,57	0,24	1,06 ($W = 0$, $p = 0{,}096$)	0,26
2	5	–0,51	0,24	0,84 ($W = 2$, $p < 0{,}001$)	0,30
3	13	0,32	0,41	1,52 ($W = 2$, $p < 0{,}001$)	0,18
3+	14	1,85	0,78	–	–

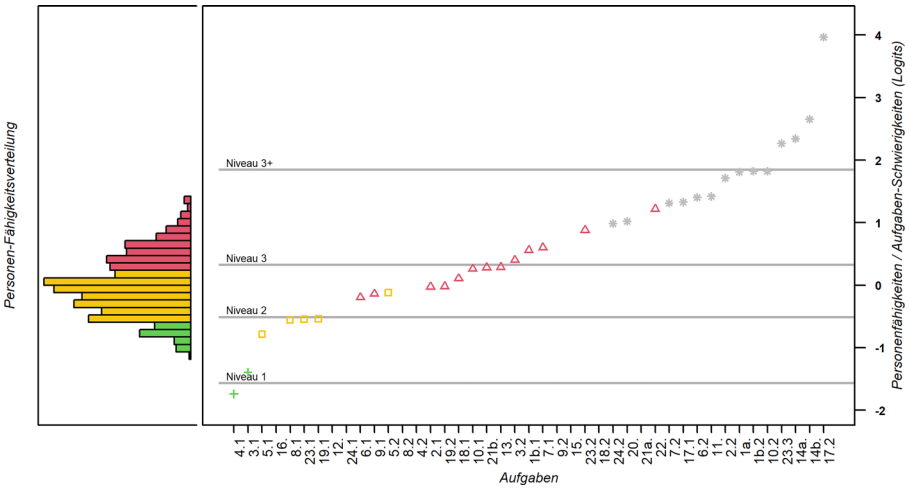


Abb. 6 Finale Wright-Map mit Ergebnissen des Scale-Anchoring-Verfahrens (ProfiLe-P+-Daten)

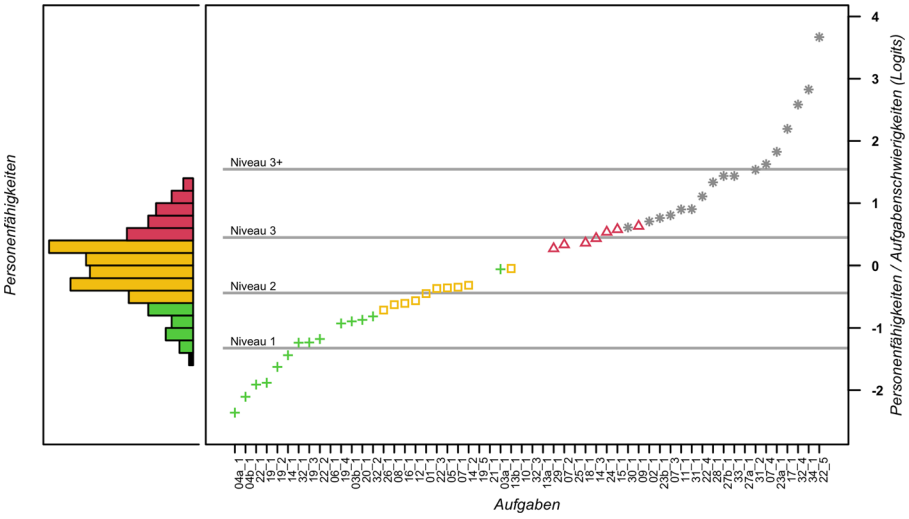


Abb. 7 Finale Wright-Map mit Ergebnissen des Scale-Anchoring-Verfahrens (KiL/KeiLa) nach Schiering et al. (2023, S. 15)

an der Niveaugrenze des Niveaus $n + 1$ mit einer Wahrscheinlichkeit von maximal 30 % (Beaton und Allen 1992) löst. Zur inhaltlichen Charakterisierung der Niveaus wurden diejenigen Aufgaben herangezogen, die sich nahe bei den Niveaugrenzen befinden.

Die Ergebnisse der Anwendung des Scale-Anchoring-Verfahrens beider Projekte sind in den Abb. 6 und 7 dargestellt. Die sich aus diesen Ergebnissen ergebenden inhaltlichen Niveaubeschreibungen und deren Gegenüberstellung werden in Abschn. 5.1 vorgestellt.

4.4 Regressionsanalytisches Verfahren auf Basis eines Modells hierarchischer Komplexität des FDW

In der Naturwissenschaftsdidaktik zeigen Ansätze wie die bereits genannten Analysen von Bernholt (2010) sowie Woitkowski und Riese (2017), dass das Modell der hierarchischen Komplexität nach Commons et al. (1998) geeignet sein kann, Niveaustufen im Fachwissen auf Basis theoretischer Überlegungen zu definieren und erklären. In einem weiteren Analyseschritt wurde daher überprüft, ob und inwieweit sich die gefundenen Gemeinsamkeiten in den Niveaumodellen des FDW mithilfe eines Modells hierarchischer Komplexität untermauern, erklären und ggf. erweitern lassen.

Zu diesem Zweck wurde zunächst ein Modell hierarchischer Komplexität für das FDW entwickelt. Dazu wurden die bereits genannten Arbeiten zur Entwicklung von hierarchischen Komplexitätsmodellen für das Fachwissen von Woitkowski (2015) bzw. Woitkowski und Riese (2017) auf das FDW übertragen. Über mehrere Iterationen hinweg wurde das in Tab. 3 beschriebene 3-stufige Modell ausgearbeitet. Die Stufen *„(I) Fakten“* und *„(II) Einstufige Kausalität“* (Tab. 3) umfassen die bloße Reproduktion sowie die Verknüpfung einzelner Wissenselemente und sind weitgehend analog zu den Stufen *„(I) Fakten“* und *„(III) Lineare Kausalität“* des Komplexitätsmodells nach Woitkowski und Riese (2017, S. 41) angelegt. Die Stufe *„(II) Prozessbeschreibungen“* von Woitkowski und Riese (2017) ließ sich auf das FDW in der operationalisierten Form nicht übertragen, da für das FDW weniger „Prozesse“ im Sinne eines zeitlichen Ablaufs als vielmehr Ursache-Wirkungs-Argumentationen im Zentrum stehen. Daher wird die Stufe der Prozessbeschreibungen in die Einstufige Kausalität integriert (siehe Tab. 3). Die höchste hier betrachtete Komplexitätsstufe stellt somit die Stufe *„(III) Mehrstufige Kausalität“* dar. Sie tritt an die Stelle der Stufe *„(IV) Multivariate Interdependenz“* des Fachwissensmodells und umfasst mehrstufige Argumentationsstränge. Wir argumentieren, dass es sich bei mehrstufigen Argumentationen um eine substanziell höhere Anforderungsstufe im Sinne des Modells hierarchischer Komplexität handelt, als bei einstufigen Argumentationen, da hier mehrere mentale Schemata miteinander in Beziehung gesetzt werden müssen und diese Beziehungen wiederum voneinander abhängig sind.

Um die Passung dieses Komplexitätsmodells zu den empirischen Daten zu testen, wurden die Aufgaben der jeweiligen Testinstrumente zunächst disjunkt zu den Komplexitätsstufen zugeordnet. Dies geschah durch die Analyse der jeweiligen Aufgabe vor dem Hintergrund der in Tab. 3 beschriebenen Komplexitätsstufen. Leitfragen der Zuordnung waren:

1. Erfordert die Aufgabe lediglich die Reproduktion von Fakten? (→ Fakten)
2. Erfordert die Aufgabe die Analyse eines komplexeren Elements (z. B. beschriebene Unterrichtssituation, Dialog, Zeichnung)? (→ einstufige Kausalität)
3. Erfordert die Aufgabe die Kreation eines komplexeren Elements (z. B. Beschreibung eines Experiments, Beschreibung einer Handlungsoption)? (→ einstufige Kausalität)
4. Erfordert die Aufgabe mehrere Schritte im Sinne der Frage 2 und/oder Frage 3? (→ mehrstufige Kausalität)

Tab. 3 Dreistufiges Modell hierarchischer Komplexität für das FDW. Die Charakterisierung diente als Grundlage für die Einordnung der Testaufgaben in das Komplexitätsmodell und wurde an die jeweiligen Rater gegeben

(I) Fakten
Reproduktion einzelner, unverbundener Informationen
Keine oder kaum Bezugnahme auf Situation oder sonstige Beschreibung
Keine oder kaum Verknüpfung der genannten Informationen
Beispiel: Nennen von Fakten zu einem Fachdidaktischen Konzept
(II) Einstufige Kausalität
Verknüpfung von zwei oder mehr Fakten, Informationen oder Äußerungen zu einem Produkt (z. B. Schlussfolgerungen, Argumentationen)
Begründungen, Analysen und Argumentationen mir nur einer Argumentations-/Analysestufe
Beispiel: (einstufige) Analyse oder Evaluation einer Situation
(III) Mehrstufige Kausalität
Begründungen, Argumentationen, Evaluationen mit mehr als einer Argumentations-/Analysestufe
Alle Anforderungen, die komplexere Analysen/Argumentation verlangen als II
Beispiel: Analyse und Evaluation einer Situation

Tab. 4 Anzahl an Aufgaben in den Komplexitätsstufen nach Projekt getrennt. Die Gesamtaufgabenanzahl weicht hier für beide Testinstrumente von den in Abschn. 4.1 ab, da Punkteschwellen (z. B. 1 vs. 2 Punkte) im Rahmen der Partial-Credit Modellierung getrennt wurden

Komplexitätsstufe	*N* ProfiLe-P(+)	*N* KiL/KeiLa
I – Fakten	13	12
II – Einstufige Kausalität	23	34
III – Mehrstufige Kausalität	7	10

Beide dargestellten Beispielaufgaben (Abb. 3 und 4) werden somit der mehrstufigen Kausalität zugeordnet. In der ProfiLe-P-Aufgabe muss zunächst eine beschriebene Unterrichtssituation analysiert werden, um auftretende Problemstellen zu identifizieren und anschließend müssen darauf aufbauend geeignete Handlungsoptionen generiert werden, um diese Probleme zu bewältigen[9]. In der KiL/KeiLa-Aufgabe muss im ersten Schritt eine komplexe Schüleraufgabe analysiert (und dabei mutmaßlich auch selbst gedanklich korrekt gelöst) werden und im zweiten Schritt davon ausgehend eine typische falsche Lösung mithilfe des Wissens über Schülervorstellungen generiert werden[10].

Diese Zuordnung wurde pro Testinstrument durch zwei Personen durchgeführt. Die Beurteilerübereinstimmung betrug beim ProfiLe-P-Testinstrument $\kappa = 0{,}86$ und beim KiL/KeiLa-Testinstrument $\kappa = 0{,}82$. Uneinigkeiten wurden durch eine kommunikative Validierung (Steinke 1999) geklärt, sodass für beide Testinstrumente eine Konsens-Aufgabenzuordnung vorlag. Tab. 4 zeigt die Anzahl an Aufgaben pro

[9] Eine „analoge" Aufgabe in der einstufigen Kausalität wäre beispielsweise die reine Kreation eines Unterrichtsverlaufs zum Fallgesetz.

[10] Eine „analoge" Aufgabe in der einstufigen Kausalität wäre dies beispielsweise dann, wenn eine typisch falsche Lösung aufgrund von Schülervorstellungen bereits eingezeichnet wäre und lediglich die zugehörige Schülervorstellung identifiziert werden müsste.

Komplexitätsstufe nach Projekt getrennt. Diese Zuordnung wurde anschließend genutzt, um mithilfe einer linearen Regression der Aufgaben-Schwierigkeitsparameter gegen die Aufgabenzuordnung zum Komplexitätsmodell die Passung auf die jeweiligen Datensätze und somit die „Gültigkeit" des Komplexitätsmodells für die jeweils abgebildeten Konstrukte einzuschätzen (Abschn. 5.2).

5 Ergebnisse

5.1 Scale-Anchoring-Verfahren: Niveauformulierungen und Vergleich

Der zentrale Gegenstand des Scale-Anchoring-Verfahrens ist die erhaltene Wright-Map mit den entsprechenden Zuordnungen und Werten (Abb. 6 und 7). Für beide Datensätze zeigt sich hier ein vergleichsweise homogenes Bild, d. h. die Aufgabengruppen zerfasern nicht stark über die Schwierigkeitsspanne hinweg. Gleichzeitig zeigen die statistischen Betrachtungen (Tab. 1 und 2 & Schiering et al. 2023, S. 14–15) die empirische Trennbarkeit der Stufen. Im Falle des ProfiLe-P+-Modells erkennt man, dass das Testinstrument vergleichsweise schwierig für die Zielgruppe ist. Dementsprechend stehen für die Charakterisierung der unteren Niveaus nur wenige Aufgaben zur Verfügung, was die spätere Interpretation erschwert. Die Niveauformulierungen auf Basis der Aufgaben nahe der entsprechenden Niveaugrenzen sind in Tab. 5 zusammengefasst, wobei eine Loslösung vom fachlichen Inhalt der jeweiligen Aufgabe hier vorerst nicht forciert wurde, da allgemein eine Abhängigkeit des FDW vom jeweils nötigen FW angenommen wird.

Für die projektübergreifende Analyse werden die erhaltenen Niveaustufen aus beiden Datensätzen verglichen. Es zeigen sich keine auffälligen Parallelen in den fachlichen und fachdidaktischen Inhalten. Demgegenüber sind allerdings Gemeinsamkeiten der Niveaubeschreibungen bzgl. der auftretenden lernpsychologisch interpretierbaren Operatoren (Tab. 6) auffällig. In den niedrigen Niveaus 1 und 2 treten primär Operatoren, welche reproduktive Aspekte beschreiben (*kursiv* in Tab. 6), auf. In den höheren Niveaus kommen Operatoren, die kreative (**fett** in Tab. 6) und bewertende (***fettkursiv*** in Tab. 6) Aspekte beschreiben, hinzu. Es zeigt sich eine deutliche Parallele bezüglich des Auftretens dieser Operatoren auf den jeweiligen Niveaus.

5.2 Passung eines Modells hierarchischer Komplexität des FDW zu den Testdaten

Zur Einschätzung der Passung des Modells hierarchischer Komplexität bzw. der Nutzbarkeit von Stufen hierarchischer Komplexität als schwierigkeitserzeugendes Merkmal des FDW wurden Regressionsanalysen für beide Testinstrumente bzw. beide Datensätze durchgeführt. Die Zuordnungen zu den Komplexitätsniveaus werden dabei als 3 Dummy-Variablen kodiert (Woitkowski und Riese 2017). Die Ergebnisse der Regressionsanalysen sind in Tab. 7 zusammengefasst und Abb. 8 illustriert diese mithilfe von Violinplots.

Sowohl Abb. 8 als auch die Varianzaufklärung von $R^2 = 0{,}39$ (multiples R^2) im Regressionsmodell ($F(2, 40) = 12{,}77$, $p < 0{,}001$) zeigen, dass das Komplexitäts-

Tab. 5 Gegenüberstellung der Scale-Anchoring Niveauformulierungen der ProfiLe-P+- und KiL/KeiLa-Modelle. Die jeweiligen Aufgaben, auf die sich der Aspekt bezieht, sind in Klammern mit angegeben

	ProfiLe-P+	KiL/KeiLa (Übers. nach Schiering et al. 2023, S. 15)
Niveau 1:	*Schülervorstellungen:* Studierende können einzelne Ursachen für die Entstehung von Schülervorstellungen nennen. (A4.1) *Experimente:* Studierende können einzelne Ziele des Experimentierens im Physikunterricht nennen. (A3)	*Schülervorstellungen:* Studierende unterscheiden in ihrer Charakterisierung wissenschaftliche Modelle von der gängigen Schülervorstellung, weil sie ein wissenschaftliches Modell nicht als richtig oder falsch, sondern als geeignet für die Erklärung eines Phänomens charakterisieren. (A32.1) *Instruktionsstrategien:* Studierende kennen typische Merkmale des entdeckenden Physikunterrichts. (A14.1) *Curriculum:* Studierende kennen Bedeutungsdimensionen der Wissenschaftsgeschichte für den Physikunterricht. (A19.2, A19.3) *Curriculum:* Studierende können zwischen zwei der drei Leistungsniveaus von Aufgaben unterscheiden. (A22.2)
Niveau 2:	*Schülervorstellungen:* Studierende können einzelne problematische Äußerungen, die durch Schülervorstellungen zum Thema Kraft und Reibung entstehen, erkennen. (A8.1) *Fachdidaktische Konzepte:* Studierende können einzelne Aspekte Didaktischer Rekonstruktion erkennen und nennen. (A19.1, A23.1)	*Schülervorstellungen:* Studierende kennen typische und untypische Schülervorstellungen im Bereich des Elektromagnetismus. (A1.1) *Schülervorstellungen:* Studierende können einfache Experimente planen, um zu demonstrieren, dass die menschliche Haut keine Temperatur misst. (A5.1) *Schülervorstellungen:* Studierende können das Verständnis der Schüler:innen für wissenschaftliche Methoden durch Experimente fördern. (A7.1) *Assessment:* Studierende können zwischen allen drei Leistungsniveaus für Aufgaben unterscheiden. (A22.3)
Niveau 3:	*Experimente:* Studierende können erste Planungselemente in Bezug auf eine situationsspezifische Unterrichtssituation zum Thema gleichmäßig beschleunigte Bewegung entwickeln. (A10.1) Studierende können mehrere Ziele des Experimentierens im Physikunterricht nennen. (A3.1) *Schülervorstellungen:* Studierende können manche Schülervorstellungen aus Schüleräußerungen zum Thema Kraft und Reibung rekonstruieren. (A21b) *Instruktionsstrategien:* Studierende können die Missverständlichkeit eines Diagramms im Kontext der Kinematik evaluieren. (A13)	*Instruktionsstrategien:* Studierende kennen typische Merkmale verschiedener Unterrichtsmethoden. (A14.3) *Curriculum:* Studierende können Themen (z. B. zur Elektrizität) gemäß dem Spiralansatz anordnen. (A18.1) *Assessment:* Studierende können, Multiple-Choice-Aufgaben hinsichtlich des Stammes und der Distraktoren bewerten. (A24.1)

Tab. 5 (Fortsetzung)

	ProfiLe-P+	KiL/KeiLa (Übers. nach Schiering et al. 2023, S. 15)
Niveau 3+:	*Experimente:* Studierende können vollständige Reaktionen in Bezug auf eine situationsspezifische Unterrichtssituation zum Thema gleichmäßig beschleunigte Bewegung entwickeln. (A10.2) *Schülervorstellungen:* Studierende können mehrere Schülervorstellungen aus einem Schülerdialog zum 3. Newtonsches Axiom rekonstruieren. (A1b.2) *Instruktionsstrategien:* Studierende können das Vorgehen einer Lehrkraft zum Erklären des 3. Newtonschen Axiom evaluieren. (A1a.)	*Schülervorstellungen:* Studierende können mögliche Quellen von Missverständnissen in wissenschaftlichen Darstellungen identifizieren. (A31.1) *Schülervorstellungen:* Studierende können die Vorstellungen der Schüler:innen zu wissenschaftlichen Experimenten (z. B. zum Verständnis der Natur der Wissenschaft) durch Experimente zu fördern. (A7.4) *Instruktionsstrategien:* Studierende können Anweisungen auf der Grundlage des Verständnisses der Schüler erstellen, die ihnen helfen, ihre wissenschaftlichen Konzepte zu ändern. (A33.1) *Curriculum:* Studierende können außerschulische Aktivitäten im Hinblick auf das Lernen der Schüler zu begründen. (A23a.1) *Assessment:* Studierende können Validität hinsichtlich eines Physiktests definieren. (A27b.1) *Assessment:* Studierende können Aspekte der Kompetenz der Schüler zu identifizieren, die durch Aufgaben bewertet werden können. (A28.1)

Tab. 6 Gegenüberstellung der Scale-Anchoring Niveauformulierungen der Projekte. Die Operatoren der KiL/KeiLa-Ergebnisse wurden aus Schiering et al. (2023) übersetzt

Niveau	ProfiLe-P+	KiL/KeiLa
1	*Nennen, erkennen*	*Unterscheiden (× 2), kennen (× 2), charakterisieren*
2	*Nennen, erkennen (× 2)*	*Unterscheiden, kennen,* **planen, fördern**
3	*Nennen,* **entwickeln, rekonstruieren,** ***evaluieren***	*Kennen* **anordnen** ***bewerten***
3+	**Entwickeln, rekonstruieren,** ***evaluieren***	*Definieren* **identifizieren (× 2), erstellen, fördern,** ***begründen***

modell für den Datensatz aus ProfiLe-P+ einen substanziellen Anteil der Varianz der Aufgabenschwierigkeit aufklärt. Hier wäre es durchaus geeignet, als Niveaustufenmodell für das FDW herangezogen zu werden. Allerdings ist dies für den Datensatz aus KiL/KeiLa nicht in gleicher Form möglich. In Abb. 8 zeigt sich nur ein leichter tendenzieller Anstieg der Aufgabenschwierigkeiten mit zunehmendem Komplexitätsniveau. Das Regressionsmodell selbst wird nicht signifikant ($F(2, 53) = 1{,}13$, $p = 0{,}33$) und klärt weniger als 5 % ($R^2 = 0{,}041$) der Varianz der Aufgabenschwierigkeit auf.

Die Komplexitätsstufen scheinen also nicht geeignet, um eine vom Testinstrument unabhängige Beschreibung von inhaltlichen Ausprägungen des FDW liefern

Tab. 7 Ergebnisse der Regressionsanalysen zur Passung des Komplexitätsmodells an die Daten. Signifikanzniveaus $p < 0{,}05 : *$, $p < 0{,}001 : ***$. Das Regressionsmodell ist so konfiguriert, dass die Regressionskonstante den Mittelwert der Schwierigkeiten der Komplexitätsstufe I-Aufgaben beschreibt. Die Mittelwerte der anderen Stufen ergeben sich durch Addition ihrer jeweiligen Regressionsparameter zur Konstanten. Die Signifikanzniveaus geben an, ob die jeweiligen Schätzer signifikant von 0 verschieden sind. Auch wenn diese Frage hier zweitrangig ist, sind die Signifikanzniveaus der Vollständigkeit halber hier mit angegeben

Komplexitätsstufe	Regr. – Parameter b_i ProfiLe-P+	Regr. – Parameter b_i KiL/KeiLa
Konstante ($\simeq$ I – Fakten)	–0,11 (n. s.)	–0,11 (n. s.)
II – Einstufige Kausalität	$0{,}71^{*}$	0,18 (n. s.)
III – Mehrstufige Kausalität	$2{,}15^{***}$	0,77 (n. s.)

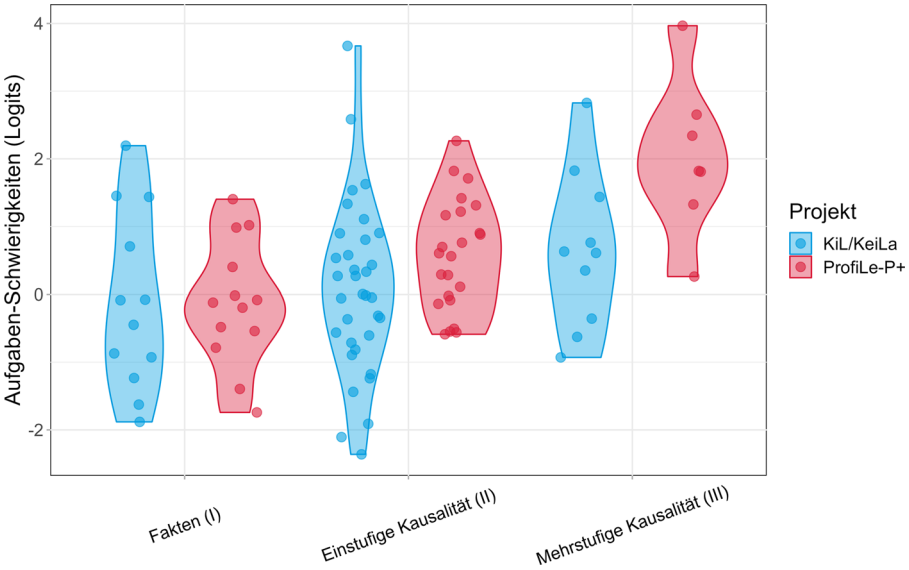


Abb. 8 Violinplots der Item-Schwierigkeiten beider Projekte mit Einordnung in die Stufen hierarchischer Komplexität. Die *Formen* stellen die Wahrscheinlichkeitsverteilung der Datenpunkte dar; die *Punkte* sind die tatsächlichen Schwierigkeiten der Aufgaben

zu können. Es wird daher hier darauf verzichtet, mögliche Wright-Maps mit Personenzuordnungen in die Niveaus abzubilden.

6 Diskussion

Ziel dieses Beitrags war es, zu überprüfen, inwieweit sich projektübergreifend inhaltliche Ausprägungen des FDW mithilfe des Scale-Anchoring-Verfahrens sowie eines regressionsanalytischen Ansatzes zur Bildung von Niveaumodellen finden lassen. Solche inhaltlichen Beschreibungen von Ausprägungen stellen eine notwendige Voraussetzung für die gewinnbringende Übertragung der Forschungsergebnisse in die Lehrpraxis dar und sind darüber hinaus von übergeordnetem Interesse für das

Forschungsfeld. Die projektübergreifende Analyse stellt zudem einen Forschungsansatz in Richtung einer vereinheitlichten Beschreibung des FDW nicht nur auf theoretischer, sondern auch auf empirischer Ebene dar. Die verwendete Methode der Überführung quantitativer Ergebnisse in Niveaumodelle mithilfe von IRT-Analysen kann ggf. als Vorlage für andere verwandte Felder dienen.

Zunächst wurden die durch das Scale-Anchoring-Verfahren erhaltenen Niveaubeschreibungen der Projekte gegenübergestellt. Es zeigten sich dabei keine Ähnlichkeiten bzgl. fachlicher oder fachdidaktischer Inhalte, aber bzgl. des Auftretens von Handlungsoperatoren, die sich auf einer lernpsychologischen Ebene interpretieren lassen. Dabei fällt die Limitation der beschränkten Anzahl an Aufgaben für die Beschreibung des ersten Niveaus in ProfiLe-P(+)-Daten weniger ins Gewicht, da die beobachtete Systematik bzgl. des Auftretens der Operatoren hier für Niveau 1 *und* Niveau 2 gilt. Die so erhaltenen Abstufungen sind insgesamt konform mit Ergebnissen der Kognitionspsychologie zum Wissenserwerbsprozess (z. B. Gagné und White 1978) und lassen sich mit Standard-Taxonomien, wie beispielsweise der auf Lehr-Lernprozesse angepassten Bloom'schen Taxonomie nach Anderson und Krathwohl (2001; Erinnern, Verstehen, Anwenden, Analysieren, Bewerten, Kreieren) in Verbindung setzen. Insgesamt lässt sich somit auch die unsystematische Beobachtung zu Ähnlichkeiten in den Niveaumodellen der beiden Projekte (Abschn. 3) im Sinne der FF1 bekräftigen:

> FDW beschränkt sich unabhängig von der konkret zugrundeliegenden Operationalisierung in niedrigen Ausprägungen auf reproduktive Aspekte und erweitert sich in höheren Ausprägungen hin zu evaluierenden und kreierenden Elementen.

Bemerkenswert ist hierbei, dass sich diese Parallele trotz einem deutlich größeren Anteil an Anfängerstudierenden im ProfiLe-P+-Datensatz (vgl. Abschn. 4.1 und Schiering et al. 2023, S. 8) zeigt.

Für den Transfer der Niveaumodelle in die Lehrpraxis zeigt sich, dass die durch das Scale-Anchoring-Verfahren erhaltenen Niveaus für die Einordnung von Lernenden in Niveaus und damit als Grundlage für das Erstellen entsprechenden Feedbacks geeignet sind. Die Niveaus und somit entsprechendes Feedback sind aber bzgl. des fachdidaktischen Inhalts abhängig vom jeweils verwendeten Testinstrument bzw. zugrundeliegender Modellierung. Das ist nicht direkt überraschend, da die beiden Testinstrumente nur in zwei von vier fachdidaktischen Facetten übereinstimmen und zudem im KiL/KeiLa-Testinstrument zusätzliche physikalisch-fachliche Inhalte thematisiert werden.

Es konnte gezeigt werden, dass die projektunabhängigen Systematiken entsprechender Niveaus primär eher allgemeine lernpsychologische Abstufungen darstellen, bzgl. derer dann auch projektunabhängige Aussagen unter Verwendung eines einzelnen Testinstruments getroffen werden können. Eine Einordnung von einzelnen Lernenden oder Lerngruppen in die Scale-Anchoring-Niveaus würde projektunabhängig bislang also beispielsweise eine Entscheidungshilfe für Lehrende bzgl. des Wechsels von eher theoretischen Lerninhalten (z. B. Vermittlung von Elementen entdeckenden Unterrichts) hin zu praxisorientierteren Elementen (z. B. Evaluation von Unterrichtsbeobachtungen) bieten. Auch bezüglich dieser lernpsychologischen

Stufung kann eine Niveau-Einordnung allerdings noch keine differenziertere Empfehlung für eher kreative oder eher evaluierende Lerninhalte für Lernende auf den höheren Niveaus unterstützen.

Aus theoriebildender Perspektive zeigen die Ergebnisse des Scale-Anchoring-Verfahrens, dass bei Austausch des fachlichen Inhalts sowie der fachdidaktischen Facetten bei ansonsten nahezu identischen theoretischen Annahmen in der Operationalisierung im Wesentlichen allgemeine kognitive Anforderungen als gemeinsame Systematiken einer hierarchischen Modellierung des FDW verbleiben. Es stellt sich also die Frage, ob aus Datenanalysen der Erhebungen mit entsprechenden Testinstrumenten abgeleitete Aussagen nicht grundsätzlich enger an die einbezogenen fachlichen (hier: physikalischen) Inhalte und fachdidaktischen Facetten gekoppelt sein müssten. Andererseits kann man die Ergebnisse des Scale-Anchoring-Verfahrens in folgendem Sinne auch als (Konstrukt-) Validitätsargument für die verwendeten Testinstrumente auffassen: In den beiden Testinstrumenten weichen die fokussierten Inhalte bzgl. der ersten zwei Dimensionen (1. fachphysikalische Inhalte und 2. fachdidaktische Inhalte/Facetten) der äußerst ähnlichen Itementwicklungsmodelle voneinander ab. Die sich zeigende übergeordnete Niveaustruktur lässt sich anschließend gerade durch die vergleichbare übrige Facette der „kognitiven Aktivierung“ (Gramzow 2015) bzw. „Wissensarten“ (Tepner et al. 2012; Kröger 2019) interpretieren. Dadurch werden die Annahmen der Operationalisierungen bzgl. einer entsprechenden Dimensionierbarkeit des FDW unterstützt.

Um die Vergleichbarkeit unterschiedlicher Operationalisierungen darüber hinaus weiter zu untersuchen, wären Studien wünschenswert, in welchen Proband:innen Testinstrumente aus unterschiedlichen Projekten bearbeiten. Korrelations- und Faktorenanalysen entsprechender Datensätze können ggf. weitere Aufschlüsse über Gemeinsamkeiten und Unterschiede der entsprechenden abgebildeten Konstrukte liefern. Für die Anwendung des Scale-Anchoring Verfahrens wären solche Datensätze auch interessant, da dann mehr Aufgaben in einem gemeinsamen Datensatz vorliegen würden, sodass die Niveaus detaillierter beschrieben werden und ggf. bisher unerkannte Systematiken zu Tage treten können.

Um die Ergebnisse der durch das Scale-Anchoring-Verfahren erhaltenen Stufen weiter auszuschärfen, wurde anschließend versucht, mithilfe der projektunabhängigen, lernpsychologisch begründeten Stufen hierarchischer Komplexität die Varianz der Aufgabenschwierigkeiten im FDW zu erklären. Während das entwickelte Modell hierarchischer Komplexität sich als sehr passend für die Daten aus ProfiLe-P+ erwiesen, zeigten sich trotz guter Übereinstimmung der Aufgabeneinordnung in das Komplexitätsmodell für beide Testinstrumente deutliche Limitationen in Bezug auf die Übertragbarkeit auf die Daten der KiL/KeiLa-Projekte. Da das Komplexitätsmodell aus dem ProfiLe-P+-Team heraus vorgeschlagen wurde, ist nicht auszuschließen, dass es sich bei der mangelnden Übertragbarkeit auf KiL/KeiLa-Daten um ein Artefakt der Modellentwicklung handelt. Eine Konfundierung des Komplexitätsmodells durch bestimmte Überzeugungen und Blickwinkel auf das Konstrukt des FDW oder durch die Art der verwendeten Aufgabentypen des ProfiLe-P+-Testinstruments konnte hier eventuell nicht vollständig vermieden werden. Das FDW scheint als „amalgam“ (Shulman 1987) im Vergleich zum FW eine weniger stark kumulative Struktur aufzuweisen, was die Konstruktion eines projektunabhängigen

theoretischen Modells schwierigkeitserzeugender Merkmale erschwert. (Physikalisches) FW ist auch aufgrund der starken Mathematisierung und damit verbundenen sehr klaren Beschreibbarkeit von Begriffen und Konzepten stark hierarchisch geprägt. Begriffe und Konzepte aus der Fachdidaktik sind oft schwieriger exakt zu beschreiben und werden erst durch die gegenseitigen Beziehungen greifbar (z. B. „Didaktische Rekonstruktion“, „Elementarisierung“ und „Schülervorstellungen“).

Das hier vorgeschlagene Modell hierarchischer Komplexität allein stellt somit kein geeignetes Modell zur projektübergreifenden Aufklärung der Aufgabenschwierigkeit dar. Weitere mögliche Einflussfaktoren im Sinne eines „amalgams“ sind z. B. der thematisierte Fachinhalt, der sich in den beiden Projekten unterschied, das auftretende Fachvokabular oder auch die theoretische Thematisierung unterschiedlicher didaktischer Inhalte zu unterschiedlichen Zeitpunkten im Studium, d. h. die vorhandene Studienstruktur (Schiering et al. 2021). Letzteres kann auch einen Ansatzpunkt bieten, um zu erklären, weshalb auch auf hohen Niveaustufen offenbar teilweise noch neue deklarative Aspekte hinzukommen (siehe Tab. 5 und 6). Die Interaktion der genannten und weiterer möglicher Einflussfaktoren, scheint die hierarchische Struktur des FDW deutlich komplexer werden zu lassen, als mit einem stark verdichteten Modell hierarchischer Komplexität fassbar ist. Für eine umfassendere regressionsanalytische Niveaubildung mit einer größeren Anzahl an möglichen schwierigkeitserzeugenden Merkmalen wären allerdings Testinstrumente mit einer deutlich größeren Anzahl an Testitems notwendig, damit entsprechenden multivariaten Regressionsmodellen eine ausreichende Datengrundlage geboten wird.

Insgesamt konnten in diesem Beitrag vor allem mithilfe des Scale-Anchoring-Verfahrens trotz Unterschieden in der Testinstrument-Konzeption besonders hinsichtlich fachlicher und fachdidaktischer Inhalte projektübergreifende kriterienorientierte Systematiken von Ausprägungen des FDW ermittelt werden. Limitiert werden diese Beschreibungen vor allem durch die aus Gründen der Testökonomie und Zumutbarkeit vergleichsweise kleinen Aufgabenanzahl der FDW-Testinstrumente. So kann etwa in den höheren Niveaustufen keine Hierarchie zwischen kreierenden und evaluieren Elementen festgestellt werden. Es ist also noch weitere Forschung zu Vergleichen und zur Vereinheitlichung der empirischen Ergebnisse notwendig.

Da für die oben vorgeschlagene Erhebung neuer Datensätze mit Proband:innen, die mehrere Testinstrumente bearbeiten, große organisatorische Hürden überwunden werden müssten, wäre es dafür auch denkbar, ein gemeinsames IRT-Modell durch eine Normierung über die mittlere Personenfähigkeit einer hinsichtlich relevanter demographischer Merkmale ununterscheidbaren jeweiligen Unterstichprobe und anschließender konditionierter Schätzung der Item-Schwierigkeiten aufzustellen. In einer neuerlichen Anwendung des Scale-Anchoring Verfahrens könnten dann die Aufgabenschwierigkeiten auf Basis der fixen gemeinsam normierten Personenparameter geschätzt werden und es stände unmittelbar ein deutlich vergrößerter Aufgabenpool für die Charakterisierung der Niveaustufen zur Verfügung. Dafür müssten sowohl die Stichproben noch einmal im Detail auf eine Vergleichbarkeit geprüft werden als auch eine andere Software genutzt oder selbst entwickelt werden, da das hier genutzte R-Paket TAM (Robitzsch et al. 2022) keine direkte Schätzung von Aufgabenschwierigkeiten unter fixierten Personenfähigkeiten ermöglicht.

Die Betrachtung der Systematiken bzgl. lernpsychologisch interpretierbarer Operatoren als Teil der inhaltlich kriterienorientierten Niveaubeschreibungen weisen auf eine praktikable Anwendbarkeit von lernpsychologischen Taxonomien auf das FDW hin. Gleichzeitig scheinen hierarchische Modelle evaluierende und kreative Elemente, die ab einer mittleren FDW-Ausprägung auftreten, nicht trennen zu können. Eine Alternative zu hierarchischen Modellen bieten Clusteranalysen (z. B. Duda et al. 2001) oder auch eng verwandte Latente Profil- oder Klassenanalysen (z. B. Spurk et al. 2020), die im naturwissenschaftsdidaktischen Kontext bisher nur wenig eingesetzt wurden (Zhai et al. 2020a, 2020b). Daher bestehen in diesem Kontext noch keine prototypischen Vorgehensweisen, die synchron auf Datensätze unterschiedlicher Projekte angewendet werden könnten; die Entwicklung entsprechender Vorgehensweisen ist hier also zunächst das Ziel weiterer Forschung. Für die Daten aus dem ProfiLe-P+-Projekt werden in diesem Kontext aktuell Vorgehensweisen erprobt, welche Clusteranalysen der Scores (Zeller und Riese 2023) mit Methoden zur Machine-Learning-basierten Sprachanalyse der Sprachproduktionen der Proband:innen verbinden. Im Gegensatz zu IRT-Modellen können solche Ansätze auch nicht-hierarchische Strukturen aufdecken und hier womöglich zur Unterscheidung der Einflüsse von kreativen und evaluierenden Aspekten dienen.

Danksagung Der Erstautor bedankt sich bei seinen Koautoren, diese Veröffentlichung im Rahmen seiner kumulativen Promotion nutzen zu können und daher besonders für die Zusammenarbeit und das erhaltene Feedback.

Förderung Das Projekt „Messung professioneller Kompetenzen in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen“ (Akronym *KiL*) wurde durch die Leibniz Gemeinschaft unter dem Kennzeichen SAW-2011-IPN-2 gefördert. Das Projekt „Kompetenzentwicklung in mathematischen und naturwissenschaftlichen Lehramtsstudiengängen“ (Akronym *KeiLa*) wurde durch die Leibniz Gemeinschaft unter dem Kennzeichen SAW-2014-IPN-1 gefördert. Das Projekt „Professionskompetenz im Lehramtsstudium Physik“ (Akronym *ProfiLe-P+*) wurde vom Bundesministerium für Bildung und Forschung im Rahmen des BMBF-Rahmenprogramms „Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen“ (Akronym *KoKoHs*) unter dem Kennzeichen 01PK15005A-D gefördert. Die hier verwendeten Daten stammen aus den o. g. Projekten. Das Manuskript ist im Rahmen einer kumulativen Promotion entstanden, die mit einem Promotionsstipendium der Studienstiftung des deutschen Volkes gefördert wurde.

Author Contribution *Konzeptualisierung:* Jannis Zeller, Dustin Schiering, Josef Riese. *Datenerhebung:* Christoph Kulgemeyer, Knut Neumann, Stefan Sorge, Josef Riese (sowie andere an den Projekten beteiligte Personen, die hier nicht alle als Autor:innen fungieren). *Datenpflege:* Jannis Zeller (ProfiLe-P+-Daten), Dustin Schiering & Stefan Sorge (KiL/KeiLa-Daten). *Methodik und Formale Analyse:* Jannis Zeller (Scale-Anchoring-Verfahren: ProfiLe-P+-Daten, Regressionsanalytischer Ansatz: beide Datensätze), Dustin Schiering (Scale-Anchoring-Verfahren: KiL/KeiLa-Daten). *Ergebnisinterpretation:* Jannis Zeller, Dustin Schiering, Christoph Kulgemeyer, Knut Neumann, Stefan Sorge, Josef Riese. *Ursprünglicher Entwurf:* Jannis Zeller, Josef Riese. *Review und Überarbeitung:* Jannis Zeller, Dustin Schiering, Christoph Kulgemeyer, Knut Neumann, Josef Riese, Stefan Sorge. *Fördermittelbeschaffung:* Jannis Zeller, Christoph Kulgemeyer, Knut Neumann, Josef Riese. Alle Autoren haben der veröffentlichten Version des Manuskripts zugestimmt.

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt J. Zeller, D. Schiering, C. Kulgemeyer, K. Neumann, J. Riese und S. Sorge geben an, dass kein Interessenkonflikt besteht.

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. 93–116). Singapore: Springer. https://doi.org/10.1007/978-981-13-5898-2_12.

Anderson, L. W., & Krathwohl, D. R. (Hrsg.). (2001). *A taxonomy for learning, teaching, and assessing A revision of Bloom's taxonomy of educational objectives* (4. Aufl.). New York: Longman.

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: the unsolved problem of teachers' mathematical knowledge. In V. Richardson (Hrsg.), *Handbook of research on teaching* (4. Aufl. S. 433–456). Washington: American Educational Research Association.

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

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*(2), 191–204. https://doi.org/10.2307/1165169.

Behling, F., Förtsch, C., & Neuhaus, B. J. (2022a). The refined consensus model of pedagogical content knowledge (PCK): detecting filters between the realms of PCK. *Education Sciences*, *12*(9), 592. https://doi.org/10.3390/educsci12090592.

Behling, F., Förtsch, C., & Neuhaus, B. J. (2022b). Using the plan-teach-reflect cycle of the refined consensus model of PCK to improve pre-service biology teachers' personal PCK as well as their motivational orientations. *Education Sciences*, *12*(10), 654. https://doi.org/10.3390/educsci12100654.

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

Blömeke, S., Felbrich, A., Müller, C., Kaiser, G., & Lehmann, R. (2008). Effectiveness of teacher education. *ZDM Mathematics Education*, *40*, 719–734. https://doi.org/10.1007/s11858-008-0096-x.

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

Blömeke, S., Jentsch, A., Ross, N., Kaiser, G., & König, J. (2022). Opening up the black box: Teacher competence, instructional quality, and students' learning progress. *Learning and Instruction*, *79*, 101600. https://doi.org/10.1016/j.learninstruc.2022.101600.

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* (S. 77–94). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-13-5898-2_2.

Commons, M. L., Trudeau, E. J., Stein, S. A., Richards, F. A., & Krause, S. R. (1998). Hierarchical complexity of tasks shows the existence of developmental stages. *Developmental Review*, *18*(3), 237–278. https://doi.org/10.1006/drev.1998.0467.

Commons, M. L., Crone-Todd, D., & Chen, S. J. (2014). Using SAFMEDS and direct instruction to teach the model of hierarchical complexity. *The Behavior Analyst Today*, *14*(1–2), 31–45. https://doi.org/10.1037/h0101284.

von Davier, A. A., Carstensen, C. H., & von Davier, M. (2008). Linking competencies in horizontal, vertical, and longitudinal settings and measuring growth. In J. Hartig, E. Klieme & D. Leutner (Hrsg.), *Assessment of competencies in educational contexts* (S. 121–149). Göttingen: Hogrefe.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2. Aufl.). New York: Wiley.

Förtsch, C., Werner, S., von Kotzebue, L., & Neuhaus, B. J. (2016). Effects of biology teachers' professional knowledge and cognitive activation on students' achievement. *International Journal of Science Education*, *38*(17), 2642–2666. https://doi.org/10.1080/09500693.2016.1257170.

Förtsch, S., Förtsch, C., Von Kotzebue, L., & Neuhaus, B. J. (2018). Effects of teachers' professional knowledge and their use of three-dimensional physical models in biology lessons on students' achievement. *Education Sciences*, *8*(3), 118. https://doi.org/10.3390/educsci8030118.

Gagné, R. M., & White, R. T. (1978). Memory structures and learning outcomes. *Review of Educational Research*, *48*(2), 187–222. https://doi.org/10.3102/00346543048002187.

Gess-Newsome, J. (1999). Pedagogical content knowledge: an introduction and orientation. In J. Gess-Newsome & N. G. Lederman (Hrsg.), *Examining pedagogical content knowledge* (S. 3–17). Dordrecht: Springer. https://doi.org/10.1007/0-306-47217-1_1.

Gess-Newsome, J., & Lederman, N. G. (Hrsg.). (1999). *Examining pedagogical content knowledge*. Dordrecht: Springer. https://doi.org/10.1007/0-306-47217-1.

Gramzow, Y. (2015). Fachdidaktisches Wissen von Lehramtsstudierenden im Fach Physik: Modellierung und Testkonstruktion. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* Bd. 181. Berlin: Logos.

Gramzow, Y., Riese, J., & Reinhold, P. (2013). Modellierung fachdidaktischen Wissens angehender Physiklehrkräfte. *Zeitschrift für Didaktik der Naturwissenschaften*, *19*, 7–30.

Harms, U., & Riese, J. (2018). Professionelle Kompetenz und Professionswissen. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Theorien in der naturwissenschaftsdidaktischen Forschung* (S. 283–298). Berlin Heidelberg: Springer Spektrum. https://doi.org/10.1007/978-3-662-56320-5_17.

Hume, A., Cooper, R., & Borowski, A. (Hrsg.). (2019). *Repositioning pedagogical content knowledge in teachers' knowledge for teaching science*. Singapore: Springer Nature. https://doi.org/10.1007/978-981-13-5898-2.

Kaiser, G., Bremerich-Vos, A., & König, J. (2020). Professionswissen. In C. Cramer, J. König, M. Rothland & S. Blömeke (Hrsg.), *Handbuch Lehrerinnen- und Lehrerbildung* (S. 811–818). Bad Heilbrunn: Klinkhardt. https://doi.org/10.35468/hblb2020-100.

Keller, M. M., Neumann, K., & Fischer, H. E. (2017). The impact of physics teachers' pedagogical content knowledge and motivation on students' achievement and interest. *Journal of Research in Science Teaching*, *54*(5), 586–614. https://doi.org/10.1002/tea.21378.

Kirschner, S. (2013). *Modellierung und Analyse des Professionswissens von Physiklehrkräften*. Duisburg, Essen: Universitätsbibliothek Duisburg-Essen.

Kleickmann, T., Großschedl, J., Harms, U., Heinze, A., Herzog, S., Hohenstein, F., Köller, O., Kröger, J., Lindmeier, A., Loch, C., Mahler, D., Möller, J., Neumann, K., Parchmann, I., Steffensky, M., Taskin, V., & Zimmermann, F. (2014). Professionswissen von Lehramtsstudierenden der mathematisch-naturwissenschaftlichen Fächer – Testentwicklung im Rahmen des Projekts KiL. *Unterrichtswissenschaft*, *42*(3), 280–288.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., Reiss, K., Riquarts, K., Rost, J., Tenorth, H.-E., & Vollmer, H. J. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise*. BMBF. https://doi.org/10.25656/01:20901.

König, J. (2009). Zur Bildung von Kompetenzniveaus im Pädagogischen Wissen von Lehramtsstudierenden: Terminologie und Komplexität kognitiver Bearbeitungsprozesse als Anforderungsmerkmale von Testaufgaben? *Lehrerbildung auf dem Prüfstand*, *2*(2), 244–262. https://doi.org/10.25656/01:14703.

Kramer, M., Förtsch, C., Boone, W. J., Seidel, T., & Neuhaus, B. J. (2021). Investigating pre-service biology teachers' diagnostic competences: relationships between professional knowledge, diagnostic activities, and diagnostic accuracy. *Education Sciences*, *11*(3), 89. https://doi.org/10.3390/educsci11030089.

Kröger, J. (2019). *Struktur und Entwicklung des Professionswissens angehender Physiklehrkräfte*. Diss., Christian-Albrechts Universität Kiel

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.

Kulgemeyer, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Reinhold, P., Riese, J., Schecker, H., Schröder, J., & Vogelsang, C. (2020). Professional knowledge affects action-related skills: The development of preservice physics teachers' explaining skills during a field experience. *Journal of Research in Science Teaching*, *52*(10), 1554–1582. https://doi.org/10.1002/tea.21632.

Kulgemeyer, C., Kempin, M., Weißbach, A., Borowski, A., Buschhüter, D., Enkrott, P., Reinhold, P., Riese, J., Schecker, H., Schröder, J., & Vogelsang, C. (2021). Exploring the impact of pre-service science teachers' reflection skills on the development of professional knowledge during a field experience.

International Journal of Science Education, *43*(18), 3035–3057. https://doi.org/10.1080/09500693.2021.2006820.

Kulgemeyer, C., Riese, J., Vogelsang, C., Buschhüter, D., Borowski, A., Weißbach, A., Jordans, M., Reinhold, P., & Schecker, H. (2023). How authenticity impacts validity: Developing a model of teacher education assessment and exploring the effects of the digitisation of assessment methods. *Zeitschrift für Erziehungswissenschaft*, *26*, 601–625. https://doi.org/10.1007/s11618-023-01154-y.

Kunter, M., Klusmann, U., Baumert, J., Richter, D., Voss, T., & Hachfeld, A. (2013). Professional competence of teachers: effects on instructional quality and student development. *Journal of Educational Psychology*, *105*, 805–820. https://doi.org/10.1037/a0032583.

Lee, W.-C., & Lee, G. (2018). Linking and equating. In P. Irwing, T. Booth & D.J. Hughes (Hrsg.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (S. 639–673). John Wiley & Sons. https://doi.org/10.1002/9781118489772.ch21.

Linacre, J.M. (1998). Thurstone thresholds and the Rasch model. *Rasch Measurement Transactions*, *12*(2), 634–635.

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: compatibility and complementarity. *Assessment & Evaluation in Higher Education*, *41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136.

Masters, G.N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272.

Mientus, L., Hume, A., Wulff, P., Meiners, A., & Borowski, A. (2022). Modelling STEM teachers' pedagogical content knowledge in the framework of the refined consensus model: a systematic literature review. *Education Sciences*, *12*(6), 385. https://doi.org/10.3390/educsci12060385.

Moosbrugger, H., & Kelava, A. (Hrsg.). (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Berlin: Springer.

Mullis, I.V.S., & Fishbein, B.G. (2020). Using scale anchoring to interpret the TIMSS 2019 achievement scales. In M.O. Martin, M. von Davier & I.V.S. Mullis (Hrsg.), *Methods and procedures: TIMSS 2019 technical report* (S. 15.1–15.60). TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-15.html.

Mullis, I.V.S., Cotter, K.E., Centurino, V.A.S., Fishbein, B.G., & Liu, J. (2016). Using scale anchoring to interpret the TIMSS 2015 achievement scales. In M.O. Martin, I.V.S. Mullis & M. Hooper (Hrsg.), *Methods and procedures in TIMSS 2015* (S. 14.1–14.47). Lynch School of Education.

Neumann, K. (2014). Rasch-Analyse naturwissenschaftsbezogener Leistungstests. In D. Krüger, I. Parchmann & H. Schecker (Hrsg.), *Methoden in der naturwissenschaftsdidaktischen Forschung* (S. 355–369). Berlin, Heidelberg: Springer Spektrum. https://doi.org/10.1007/978-3-642-37827-0_28.

Neumann, K., Kind, V., & Harms, U. (2019). Probing the amalgam: the relationship between science teachers' content, pedagogical and pedagogical content knowledge. *International Journal of Science Education*, *41*(7), 847–861. https://doi.org/10.1080/09500693.2018.1497217.

Nold, G., Rossa, H., & Hartig, J. (2008). Proficiency scaling in DESI listening and reading EFL tests: Task characteristics, item difficulty and cut-off points. In L. Taylor & C.J. Weir (Hrsg.), *Multilingualism and assessment. Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin conference* (S. 94–116). Cambridge: Cambridge University Press.

OECD (Hrsg.). (2018). PISA 2018 technical report. https://www.oecd.org/pisa/data/pisa2018technicalreport/

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/. Zugegriffen: 21. März 2024.

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

Riese, J. (2009). Professionelles Wissen und professionelle Handlungskompetenz von (angehenden) Physiklehrkräften. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* Bd. 97. Berlin: Logos.

Riese, J., & Reinhold, P. (2012). Die professionelle Kompetenz angehender Physiklehrkräfte in verschiedenen Ausbildungsformen. *Zeitschrift für Erziehungswissenschaften*, *15*, 111–143. https://doi.org/10.1007/s11618-012-0259-y.

Riese, J., Kulgemeyer, C., Zander, S., Borowski, A., Fischer, H.E., Gramzow, Y., Reinhold, P., Schecker, H., & Tomczyszyn, E. (2015). Modellierung und Messung des Professionswissens in der Lehramtsausbildung Physik. *Zeitschrift für Pädagogik*, *61*, 55–79.

Riese, J., Gramzow, Y., & Reinhold, P. (2017). Die Messung fachdidaktischen Wissens bei Anfängern und Fortgeschrittenen im Lehramtsstudiengang Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*, 99–112. https://doi.org/10.1007/s40573-017-0059-2.

Riese, J., Vogelsang, C., Schröder, J., Borowski, A., Kulgemeyer, C., Reinhold, P., & Schecker, H. (2022). Entwicklung von Unterrichtsplanungsfähigkeit im Fach Physik: Welchen Einfluss hat Professionswissen? *Zeitschrift für Erziehungswissenschaft*, *25*, 843–867. https://doi.org/10.1007/s11618-022-01112-0.

Robitzsch, A., Kiefer, T., & Wu, M. (2022). TAM: Test Analysis Modules [R package version 4.1-4. https://CRAN.R-project.org/package=TAM. Zugegriffen: 31. Okt. 2022.

Schiering, D., Sorge, S., Petersen, S., & Neumann, K. (2019). Konstruktion eines qualitativen Niveaumodells im fachdidaktischen Wissen von angehenden Physiklehrkräften. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*, 211–229. https://doi.org/10.1007/s40573-019-00100-y.

Schiering, D., Sorge, S., & Neumann, K. (2021). Hilft viel viel? Der Einfluss von Studienstrukturen auf das Professionswissen angehender Physiklehrkräfte. *Zeitschrift für Erziehungswissenschaft*, *24*, 545–570. https://doi.org/10.1007/s11618-021-01003-w.

Schiering, D., Sorge, S., Keller, M. M., & Neumann, K. (2023). A proficiency model for pre-service physics teachers' pedagogical content knowledge (PCK)—What constitutes high-level PCK? *Journal of Research in Science Teaching*, *60*(1), 136–163. https://doi.org/10.1002/tea.21793.

Schnotz, W. (1994). *Aufbau von Wissensstrukturen. Untersuchungen zur Kohärenzbildung beim Wissenserwerb mit Texten*. Weinheim: Beltz.

Schröder, J., Riese, J., Vogelsang, C., Borowski, A., Buschhüter, D., Enkrott, P., Kempin, M., Kulgemeyer, C., Reinhold, P., & Schecker, H. (2020). Die Messung der Fähigkeit zur Unterrichtsplanung im Fach Physik mit Hilfe eines standardisierten Performanztests. *Zeitschrift für Didaktik der Naturwissenschaften*, *26*(1), 103–122. https://doi.org/10.1007/s40573-020-00115-w.

Shulman, L. S. (1986). Those who understand: knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14. https://doi.org/10.3102/0013189X015002004.

Shulman, L. S. (1987). Knowledge and teaching: foundations of the new reform. *Harvard Educational Review*, *57*(1), 1–23. https://doi.org/10.17763/haer.57.1.j463w79r56455411.

Sorge, S., Keller, M., Petersen, S., & Neumann, K. (2018). Die Entwicklung des Professionswissens angehender Physiklehrkräfte. In C. Maurer (Hrsg.), *Qualitätsvoller Chemie- und Physikunterricht – Normativ und empirische Dimensionen. Tagungsband der GDCP Jahrestagung 2017* (S. 114–117). Regensburg: Universität Regensburg.

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. https://doi.org/10.1080/09500693.2017.1346326.

Spurk, D., Hirschi, A., Wang, M., Valero, D., & Kauffeld, S. (2020). Latent profile analysis: A review and "how to" guide of its application within vocational behavior research. *Journal of Vocational Behavior*, *120*, 103445. https://doi.org/10.1016/j.jvb.2020.103445.

Steinke, I. (1999). *Kriterien qualitativer Forschung: Ansätze zur Bewertung qualitativ-empirischer Sozialforschung*. München, Weinheim: Juventa.

Tepner, O., Borowski, A., Dollny, S., Fischer, H. E., Jüttner, M., Kirschner, S., Leutner, D., Neuhaus, B. J., 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.

Terhart, E. (2012). Wie wirkt Lehrerbildung? Forschungsprobleme und Gestaltungsfragen. *Zeitschrift für Bildungsforschung*, *2*(1), 3–21. https://doi.org/10.1007/s35834-012-0027-3.

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. https://doi.org/10.25656/01:23990.

Vollmer, H. J., & Klette, K. (2023). Pedagogical content knowledge and subject didactics—An intercontinental dialogue? In F. Ligozat, K. Klette & J. Almqvist (Hrsg.), *Didactics in a changing world: European perspectives on teaching, learning and the curriculum* (S. 17–33). Cham: Springer. https://doi.org/10.1007/978-3-031-20810-2_2.

Woitkowski, D. (2015). Fachliches Wissen Physik in der Hochschulausbildung: Konzeptualisierung, Messung, Niveaubildung. In H. Niedderer, H. Fischler & E. Sumfleth (Hrsg.), *Studien zum Physik- und Chemielernen* Bd. 185. Berlin: Logos.

Woitkowski, D. (2019). Erfolgreicher Wissenserwerb im ersten Semester Physik. *Zeitschrift für Didaktik der Naturwissenschaften*, *25*, 97–114. https://doi.org/10.1007/s40573-019-00094-7.

Woitkowski, D. (2020). Tracing physics content knowledge gains using content complexity levels. *International Journal of Science Education*, *42*(10), 1585–1608. https://doi.org/10.1080/09500693.2020.1772520.

Woitkowski, D., & Riese, J. (2017). Kriterienorientierte Konstruktion eines Kompetenzniveaumodells im physikalischen Fachwissen. *Zeitschrift für Didaktik der Naturwissenschaften*, *23*, 39–52. https://doi.org/10.1007/s40573-016-0054-z.

Zeller, J., & Riese, J. (2023). Datenbasierte Fähigkeitsprofile im Physikdidaktischen Wissen. In H. van Vorst (Hrsg.), *Lernen, Lehren und Forschen in einer digital geprägten Welt, Tagungsband der GDCP Jahrestagung 2022*. Essen: Universität Duisburg-Essen.

Zeller, J., Jordans, M., & Riese, J. (2022). Ansätze zur Ermittlung von Kompetenzniveaus im Fachdidaktischen Wissen. In S. Habig (Hrsg.), *Unsicherheit als Element von naturwissenschaftsbezogenen Bildungsprozessen, Tagungsband der GDCP Jahrestagung 2021, virtuell*. Essen: Gesellschaft für Didaktik der Chemie und Physik.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*, 1430–1459. https://doi.org/10.1002/tea.21658.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: a systematic review. *Studies in Science Education*, *56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757.