

Originalarbeit

Pädagogisches Wissen von berufserfahrenen Mathematiklehrkräften – Validierung der Konstruktrepräsentation im TEDS- M-Instrument

Zusammenfassung: Das TEDS-M-Instrument zur Erfassung pädagogischen Wissens am Ende der Lehrerausbildung hat sich in diversen Untersuchungen von (angehenden) Lehrkräften in unterschiedlichen Ausbildungsstadien und -kontexten als zuverlässiges Messinstrument erwiesen, und es liegt eine Reihe von Ergebnissen zu seiner Validität vor. Ein wesentlicher Validierungsschritt steht jedoch noch aus: Die Überprüfung im Sinne der Konstruktrepräsentation (Embretson, 1983), die kontrolliert, ob das Instrument konstrukt-relevante, kognitive Bearbeitungsprozesse auch bei berufserfahrenen Mathematiklehrkräften abbildet. Im Rahmen des Projektes TEDS-Validierung wurde an 113 berufserfahrenen Mathematiklehrkräften (mit durchschnittlich 10 Jahren Unterrichtserfahrung) geprüft, ob sich die Item-Schwierigkeiten auf die von König (2009) und Klemenz und König (2019) modellierten kognitiven Prozesse zurückführen lassen. Die Ergebnisse bestätigen, dass das Instrument auch bei berufserfahrenen Mathematiklehrkräften ein reliables Messinstrument darstellt. Des Weiteren unterstreichen die Analysen, dass das Komplexitätsniveau von Aufgabenlösungen einen bedeutsamen Anteil der Schwierigkeitsvarianz aufklärt, und liefern somit einen ersten Hinweis für die Konstruktrepräsentation. Abschließende Varianzanalysen unterstreichen die Vorhersagekraft der kognitiven Komplexitätsniveaus für situationsspezifische Fähigkeiten und stützen die kriteriale Validität des Modells.

Schlüsselwörter: Pädagogisches Wissen, berufserfahrene Mathematiklehrkräfte, Lehrerkompetenzen, Validität, Konstruktrepräsentation, Leistungstest, Item-Response-Theorie

Pedagogical knowledge – Validating the construct representation in the TEDS-M instrument among professionally experienced teachers of Mathematics

Abstract: The TEDS-M test for pedagogical knowledge has proven reliable for measuring (future) teachers' knowledge in several studies, and broad evidence of its validity across professional education phases and contexts has been provided. Yet,

one essential validation step has remained untaken so far: the examination of the test's construct representation (Embretson, 1983), i.e., whether the test also requires and represents construct-relevant cognitive processes of professionally experienced mathematics teachers. In the frame of a pilot study for the project TEDS-Validate, 113 mathematics teachers (with 10 years of teaching experience on average) were tested in order to determine whether task difficulty could be attributed to cognitive processes as modeled by König (2009) and Klemenz und König (2019). The results confirm that the instrument also works reliably with in-service teachers. The analysis further underlines that the complexity level of task solutions explains a significant proportion of the variance in item difficulty, thus providing first evidence for construct representation. Finally, variance analyses show that the cognitive complexity levels predict teachers' situation-specific skills, thus underlining the criterion validity of our model.

Keywords: Pedagogical knowledge, professionally experienced mathematics teachers, teacher competence, validity, construct representation, item-response theory

Einleitung

Im Rahmen der internationalen Vergleichsstudie *Teacher Education and Development Study: Learning to Teach Mathematics* (TEDS-M) wurde ein Test entwickelt, um das pädagogische Wissen angehender Lehrkräfte aus Deutschland, den USA und Taiwan als kognitive Komponente von Lehrkompetenz zu erfassen und zu modellieren (König, Blömeke, Paine, Schmidt & Hsieh, 2011). Das TEDS-M-Instrument hat sich in Studien bewährt, um die pädagogische Wissensstruktur von Lehrenden in verschiedenen Ländern (König et al., 2011; König & Pflanzl, 2016) und Ausbildungsstadien zu untersuchen (für eine Übersicht, siehe König, 2014b). So lieferten Studien zu Lehramtsstudierenden in der universitären Ausbildungsphase (König & Seifert, 2012; König & Klemenz, 2015) und im Referendariat (König, 2013) wiederholt Hinweise für die curriculare, faktorielle, konvergente und kriteriale Validität des Instruments (für eine Übersicht, siehe Voss, Kunina-Habenicht, Hoehne & Kunter, 2015). Zugleich konnten erste Validitätsbelege zu Lehrenden in der Berufseinstiegsphase vorgelegt werden (König, Blömeke, Klein, Suhl & Busse, 2014).

Ein wesentlicher Schritt zur Validierung des TEDS-M-Instruments steht jedoch aus: Die Prüfung der Konstruktrepräsentation bei berufserfahrenen Mathematiklehrkräften. Der Test wurde ursprünglich konzipiert, um das pädagogische Wissen zu erfassen, das angehende Lehrkräfte während ihrer Ausbildung erwerben (König et al., 2011). Doch ergibt sich in Anbetracht der Debatte um die Wirksamkeit der Lehrerbildung (vgl. Terhart, 2012) die Chance, das Instrument auch bei berufserfahrenen Lehrkräften einzusetzen und empirisch zu untersuchen, welche praktische Bedeutung dem Ausbildungswissen für das Unterrichtshandeln von Lehrkräften zukommt (König & Pflanzl, 2016). Erste Befunde zeigen, dass das Instrument auch bei der Zielgruppe berufserfahrener Lehrkräfte funktioniert (König et al., 2014; König, 2015; König & Pflanzl, 2016; Lauermaun & König, 2016). Doch gilt es weiterführend zu prüfen, ob das Instrument speziell bei der Zielgruppe von Mathematiklehrkräften eingesetzt werden kann und ob dabei Konstrukt-relevante, kognitive Bearbeitungsprozesse abgebildet werden.

Bei der Konstruktrepräsentation handelt es sich um einen Validitätsaspekt, der von Embretson (1983) eingeführt wurde, um faktorielle und korrelative Validitätsanalysen im Sinne von Cronbach und Meehl (1955) zu ergänzen. Diese liefern für Validität zwar „circumstantial evidence“ (Borsboom, Mellenbergh & van Heerden, 2004, S. 1062) im Sinne von „construct significance“ (Daniel & Embretson, 2010, S. 349), indem sie die Relevanz unterstreichen, die ein Konstrukt für die Erklärung individueller Unterschiede in einem übergeordneten Netz an Zusammenhängen hat. Sie sind jedoch kein hinreichender Nachweis für Konstruktvalidität (Hartig, Frey & Jude, 2012). Vielmehr gilt

es, auch Hinweise auf „construct meaning“ (Daniel & Embretson, 2010, S. 349) zu finden, d.h. aufzuzeigen, dass das Konstrukt in Form von kognitiven Prozessen im Test selbst repräsentiert wird, die bei Testpersonen angenommen werden und sich *während* der Testbearbeitung auf das Zustandekommen der Antworten auswirken. Diese Konstruktrepräsentation stellt nach Messick (1994) einen fundamentalen Bestandteil der Konstruktvalidität dar und sollte gemäß Borsboom, Mellenbergh und van Heerden (2004, S. 1067) *primär* im Fokus von Konstruktvalidierungen stehen. Erst wenn die Konstruktrepräsentation sichergestellt ist, kann die zuverlässige Interpretation von Testwerten und ihrer Zusammenhänge mit externen Variablen gewährleistet werden.

Neben der Stützung der Konstruktvalidität ist die Analyse des Antwortverhaltens von praktischer Bedeutung für die Untersuchung kognitiver Fähigkeiten. Stützt ein Modell die Annahmen über kognitive Lösungsprozesse, erlauben diese es, kontinuierliche Kompetenzskalen in ein Kompetenzniveaumodell zu überführen, d.h. in qualitativ beschreibbare Abschnitte zu unterteilen (Hartig, 2007). Anhand eines Kompetenzniveaumodells lassen sich die Testwerte von Personen schließlich dahingehend interpretieren, zu welcher Art von Operationen die Personen jeweils in der Lage sind. Damit bieten Niveaumodelle einen Vorteil gegenüber kontinuierlichen Skalen, deren Interpretation weitestgehend auf normorientierte Aussagen und Vergleiche von Personengruppen beschränkt ist: Sie ermöglichen eine Festlegung und Überprüfung von qualitativen, kriteriumsorientierten Kompetenzstandards (Rost, 2004) und bieten somit konkrete Ansatzpunkte für pädagogisches Handeln (Hartig & Frey, 2012).

Theoretischer Bezugsrahmen

Pädagogisches Wissen

Pädagogisches Wissen gilt als zentraler kognitiver Bestandteil der professionellen Kompetenz von Lehrkräften (Baumert & Kunter, 2006). Es wird als notwendige Voraussetzung für die erfolgreiche Bewältigung lehrberufsspezifischer Anforderungen erachtet, die sich in qualitativem Unterricht und einer effektiven Leistungsförderung bei Schülerinnen und Schülern niederschlagen soll (Terhart, 2012). Pädagogisches Wissen wird allgemein als Kenntnis über fächerübergreifende, allgemeine Prinzipien des Klassenmanagements verstanden sowie als Wissen über Lernende und ihre Lernprozesse, Lehr-Lern-Methoden, Diagnostikprinzipien, Bildungskontexte und -ziele (Gindele & Voss, 2017; König, 2014a; Shulman, 1987). Bisherige Untersuchungen stützen über verschiedene Ausbildungskontexte und Testinstrumente hinweg die Annahme, dass es sich beim pädagogischen Wissen um ein mehrdimensionales Konstrukt handelt, das sich in inhaltliche und kognitive Anforderungsfacetten unterteilt (Baumert & Kunter, 2006; König & Blömeke, 2009; Voss et al., 2015).

Kognitive Prozesse beim Lösen unterrichtsspezifischer Aufgaben

Bis heute ist nur wenig darüber bekannt, wie sich welche Prozesse bei Lehrenden „im Kopf“ abspielen, wenn sie berufsspezifische Aufgaben meistern. Embretson (2016, S. 7) stellt für den Bildungssektor im Allgemeinen fest, dass die Prüfung von Bearbeitungsprozessen bei Leistungs- und Kompetenztests typischerweise unberücksichtigt bleibt. Obwohl kognitive Prozesse häufig in Leitlinien zur Testkonstruktion skizziert werden, werden diese in der Regel nicht im Sinne einer Theorie der Bearbeitungsprozesse hergeleitet und empirisch geprüft (Daniel & Embretson, 2010). Im deutschsprachigen Raum existieren einzelne Studien, die sich der Analyse von kognitiven Prozessen und Aufgabenmerkmalen widmen, jedoch meist auf Kompetenztests für Schülerinnen und Schüler ausgerichtet sind

(bspw. Beck & Klieme, 2007; Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010; Prenzel, Häußler, Rost & Senkbeil, 2002; Viering, Fischer & Neumann, 2010). Studien zu Lehrkräften hingegen sind rar und fokussieren in der Regel nicht auf fächerübergreifendes, pädagogisches Wissen (ein Beispiel zum Fachwissen von Mathematiklehrkräften findet sich in Blömeke et al., 2008).

Eine Ausnahme bilden die Arbeiten von König (2009) und Klemenz und König (2019). Sie liefern erste Hinweise für die Konstruktrepräsentation des pädagogischen Wissens im TEDS-M-Test und ermöglichen so eine qualitative Beschreibung der Fähigkeiten von angehenden Lehrenden. In ihrem Modell gehen sie davon aus, dass für die erfolgreiche Bewältigung unterrichtsbezogener, pädagogischer Aufgaben einerseits das differenzierte Explizieren von Wissen und andererseits das Verbinden mehrerer Wissens Elemente als kognitive Prozesse zentral sind. Das Explizieren durch die Anwendung differenzierter, unterrichtsbezogener Sprache ist notwendig, um Handlungen zu analysieren und kritisch-reflektiert zu beurteilen. Erst durch seine Expliziertheit wird das Wissen analysier- und kommunizierbar gemacht (Altrichter & Posch, 2007). Das Verbinden mehrerer Wissens Elemente zu komplexeren, bedeutungstragenden Einheiten (*chunks*) fördert das Einnehmen multipler Perspektiven und das schnellere Verarbeiten unterrichtsbezogener Situationen hin zur Entwicklung von Handlungsvorschlägen (Wipperfürth, 2015). Es stellt somit eine Voraussetzung für das situationsspezifische Handeln von Lehrenden dar (Bromme, 1992).

Fragestellungen

Anliegen der vorliegenden Arbeit ist es, dem aufgezeigten Validierungsbedarf für das TEDS-M-Instrument nachzukommen. Es stellt sich die Frage, ob der Test kognitive Prozesse abbildet, die berufserfahrene Mathematiklehrkräfte bei der Bearbeitung pädagogischer, unterrichtsbezogener Aufgaben anwenden. Hierzu soll geprüft werden, ob sich das Modell und die Befunde von König (2009) und Klemenz und König (2019) replizieren lassen. Darauf aufbauend soll zudem der Frage nachgegangen werden, ob sich das Modell als kriterial valide erweist und Unterschiede in weiteren Personenmerkmalen erklären kann.

Untersuchungsdesign

Stichprobe und Datenerhebung

Zur Untersuchung der Fragestellung verwenden wir Daten aus der Studie TEDS-Unterricht. Die Stichprobe umfasst 113 Mathematiklehrkräfte an Hamburger Schulen (49% weiblich). Alle Teilnehmenden verfügten zum Testzeitpunkt über Mathematikunterrichtserfahrung in den Jahrgangsstufen 7 bis 9. Mehr als 92% von ihnen wurden speziell für die Sekundarstufe I ausgebildet. Das durchschnittliche Alter der Lehrenden lag zum Zeitpunkt der Erhebung bei 39 Jahren ($SD = 10.48$; Min = 23; Max = 71), ihre durchschnittliche Unterrichtserfahrung an Schulen bei 10 Jahren ($SD = 9.87$; Min = 0.5; Max = 40).

Die Datenerhebung fand im Schuljahr 2015/16 statt und wurde mit Hilfe der Online-Umfragesoftware *Unipark* durchgeführt. Die Teilnehmenden konnten individuell über Zeitpunkt und Ort der etwa 3-stündigen Testung entscheiden und erhielten nach Abschluss eine monetäre Aufwandsentschädigung. Die Testung umfasste fünf umfangreiche Instrumente zu mathematischem, mathematikdidaktischem und pädagogischem Wissen, die klassische *paper-and-pencil*-Formate in digitalisierter Form ebenso einschlossen wie videobasierte Vignettentests. Zwischen

den einzelnen Testblöcken konnten die Teilnehmenden individuell Unterbrechungen vornehmen. Jeder Testabschnitt unterlag jedoch einer zeitlichen Begrenzung, sodass Teilnehmende nach Ablauf der vorgesehenen Zeit automatisch zum nächsten Test weitergeleitet wurden, selbst wenn die Aufgabenbearbeitung noch nicht abgeschlossen war. Nicht abgeschlossene Aufgaben wurden als fehlende Werte kodiert.

TEDS-M-Test pädagogischen Wissens

Um eine Überbeanspruchung der Lehrkräfte zu vermeiden, wurde für die Testung des pädagogischen Wissens eine Kurzfassung des TEDS-M-Instruments eingesetzt, deren Bearbeitung auf 20 Minuten begrenzt wurde und in etwa im letzten Drittel der Erhebung stattfand. Die Kurzfassung kam bereits in mehreren Studien zum Einsatz (bspw. König, Blömeke & Kaiser, 2015) und setzt sich aus 15 Aufgaben zusammen, aus denen insgesamt 47 Items resultieren. Zehn der Aufgaben (28 Items) liegen in offenem und fünf Aufgaben (19 Items) in geschlossenem Format vor. Die Aufgabeninhalte decken berufliche Anforderungsdimensionen ab, die unter Berücksichtigung von Erkenntnissen der Lehr-Lern-Forschung sowie der Allgemeinen Didaktik entwickelt wurden: Unterrichtsstrukturierung, Motivierung, Umgang mit Heterogenität, Klassenführung und Leistungsbeurteilung (vgl. Tabelle 1). Zugleich repräsentieren die 15 Aufgaben qualitativ unterschiedliche kognitive Anforderungen nach Anderson und Krathwohl (2001): Sie erfordern Prozesse des Erinnerns, Prozesse des Verstehens bzw. Analysierens von Sachverhalten, Konzepten und Zusammenhängen sowie Prozesse des Kreierens von Handlungsoptionen (siehe Beispiel in Abbildung 1).

»Tabelle 1 hier einfügen«

»Abbildung 1 hier einfügen«

Alle Aufgaben mit offenem Antwortformat wurden auf Basis einer umfassenden Anleitung von zwei geschulten Studierenden mit mehrjähriger Kodiererfahrung kodiert. Für 43 der 113 Fälle wurden die offenen Antworten von beiden Kodiererinnen unabhängig kodiert. Diese Doppelkodierungen weisen mit einem Cohen's Kappa-Wert von $\kappa > .79$ ein hohes Übereinstimmungsmaß auf und unterstreichen die Interrater-Reliabilität des Instruments.

Methode

Zur Prüfung der Konstruktrepräsentation wird im Sinne von Embretsons „*item difficulty modeling*“ (Hervorhebung im Original; 2016, S. 8) ein 4-stufiges Verfahren angewendet, das sich in vergangenen Untersuchungen bereits bewährt hat (vgl. Blömeke et al., 2010). In einem ersten Schritt werden die (1a) Bearbeitungsprozesse in Form von Aufgabenmerkmalen operationalisiert, die konstruktrelevante kognitive Prozesse erfordern. Ihnen werden jeweils (1b) Ausprägungsgrade zugeordnet, von denen erwartet wird, dass sie einen Einfluss auf die Schwierigkeit der Aufgaben haben. Auf Basis des daraus entstehenden Kategoriensystems werden (2) die Aufgaben hinsichtlich der Merkmale und Ausprägungen klassifiziert. Schließlich wird mittels (3) IRT-Skalierung die Schwierigkeit der Items geschätzt, um den Einfluss der Merkmale auf die Testbearbeitung (4) in einem linearen Regressionsmodell zu prüfen. Stützen die Ergebnisse der Regressionsanalyse unsere Annahmen zur Konstruktrepräsentation, werden sie anschließend zu einem Kompetenzniveauomodell für pädagogisches Wissen zusammengetragen, das in einem letzten Schritt mittels Varianzanalysen kriterial validiert wird.

Operationalisierung und Klassifizierung der Items

Zur Operationalisierung der kognitiven Prozesse wird auf das von König (2009) und Klemenz und König (2019) entwickelte Modell zurückgegriffen. Sie operationalisieren den Prozess des differenzierten Explizierens von Wissen, indem sie auf Forschung zur Verknüpfung von Lehrersprache und -wissen (Terhart, 1993) zurückgreifen. Im vorliegenden Fall wird der Prozess des Verbindens mehrerer Wissensselemente untersucht. Dieser wird operationalisiert, indem die Komplexität der Aufgabenlösung erfasst wird. Bezugnehmend auf das Konzept der kognitiven Komplexität (Eye, 1999, S. 82) wird unter der Komplexität der Aufgabenlösung die Zahl der Wissensselemente verstanden, die für die „Strukturierung und Beurteilung eines Objektes oder Sachverhalts“ notwendig sind. Werden also für die vollständig korrekte Lösung der Aufgabe mehrere Wissensselemente benötigt, wird die Komplexität der Aufgabenlösung als hoch eingestuft. Muss lediglich ein einzelnes Wissensselement abgerufen werden, gilt sie als niedrig. Wie viele Wissensselemente den Erwartungshorizont der Testaufgaben widerspiegeln, wurde theoretisch vorweg in der umfassenden Kodieranleitung festgelegt (König & Blömeke, 2010).

Um die Komplexität abzubilden wurden offene Items, die zum selben Aufgabenstamm gehören und daher zusammenhängende Lösungsaspekte einer Aufgabe repräsentieren, zu einem Partial-Credit-Item zusammengefasst. Abbildung 2 zeigt PK39, eine Aufgabe zum Konzept der strukturierten Unterrichtsanalyse, die als hoch komplex klassifiziert wurde, mit beispielhaften offenen Antworten.

»Abbildung 2 hier einfügen«

Die Aufgabe wurde nach dem in Tabelle 2 dargestellten Schema kodiert. Die zehn Fragen, die die Testperson zur Beantwortung der Aufgabe entwickelt hat, wurden anhand des Kodiermanuals dahingehend kodiert, ob sie bestimmte Inhalte ansprachen (bspw. Voraussetzungen bei der Lerngruppe, methodisch-didaktische Prinzipien, kognitive Aktivierung). Insgesamt waren zwölf Kodierungskategorien gebildet worden (König & Blömeke, 2010, S. 27f.), die vier unterschiedliche Wertungskriterien (PK39S01-04) abbilden, um inhaltlich nahe Antworten nicht doppelt zu werten (ebd., S. 45). Diese Wertungs-Items wurden schließlich in einem Partial-Credit-Item zusammengefasst (PK39S). Formuliert eine Testperson also wie in Abbildung 2 Fragen zum Kontext (bspw. Vorwissen), Input (bspw. Ziele), Prozess (bspw. Lehrmethoden) und Output (bspw. Lernzuwachs) der Unterrichtsstunde, erfüllte sie vier von vier Kriterien und erhielt für PK39S den Wert „4“. Formuliert eine Person ausschließlich Fragen zum Vorwissen und Prozess, erfüllte sie lediglich zwei von vier Kriterien (PK39S01 und PK39S03), sodass sie für PK39S den Wert „2“ erhielt.

»Tabelle 2 hier einfügen«

Nach Klemenz und König (2019) gilt die Komplexität einer Aufgabe als hoch, wenn ein Konzept zur korrekten Lösung in seiner Gänze angewendet wurde, also alle erforderten Wissensselemente genannt wurden. Für das Beispiel aus Abbildung 2 bedeutet dies, dass die Person mit PK39S = „4“ das hoch komplexe Item korrekt beantwortet und somit das Konzept der strukturierten Unterrichtsanalyse vollständig angewendet hat. Um nicht nur zwischen niedriger Komplexität (einzelnes Wissensselement) und hoher (vollständig verknüpftes Konzept) zu unterscheiden, sondern auch partielle Verknüpfungsprozesse von Personen mit Teilscores abzubilden, wurde das Modell von Klemenz und König (2019) ausdifferenzierend weiterentwickelt. Einzelne Verknüpfungen, die sich aus zwei oder mehr Wissensselementen zusammensetzten, jedoch noch kein vollständiges Konzept darstellten, wurden einer mittleren Komplexität zugeordnet.

Auf Items geschlossener Aufgaben wurde diese Partial-Credit-Modellierung nicht angewendet, da diese nicht als komplexitätsabbildend interpretiert wurden. Ihre Lösungen konnten zwar mehrere Wissens Elemente erfordern, jedoch musste diese Multiperspektivität nicht von den Testteilnehmenden selbst entwickelt werden, sondern wurde durch die Distraktoren aktiviert. Insgesamt ergaben die geschlossenen Items zusammen mit der Bildung der Partial-Credit-Items eine Summe von 32 Items, von denen 21 dichotom (niedrig komplex) waren und 11 als mehrstufig modelliert wurden, also Teilprozesse niedriger, mittlerer und hoher Komplexität einschlossen.

Skalierung und Bestimmung der Item-Schwierigkeiten

Zur Bestimmung der Item-Schwierigkeiten wurden die Daten mittels des in *ConQuest* (Wu, Adams & Wilson, 1997) implementierten einparametrischen Partial-Credit-Modells skaliert, um auf Basis der 32 Items ein eindimensionales Modell zu schätzen. Das Modell zeigt akzeptable Reliabilitätswerte (WLE-Reliabilität .68; EAP/PV-Reliabilität .74). Die Items weisen mit Weighted-MNSQ-Werten zwischen 0.80 und 1.20 einen zufriedenstellenden Fit auf (vgl. Tabelle 3). Auch die Trennschärfe kann mit Werten $> .20$ bei den meisten Items als akzeptabel eingestuft werden. Vier Items liegen mit $.10 < r_{i(x-i)} < .18$ darunter, werden jedoch beibehalten, da sie zur inhaltlichen Differenzierung beitragen und so theoretisch gesehen eine bessere Konstruktrepräsentation gewährleisten. Die Streuung der Item-Schwierigkeitswerte umfasst mehr als 3 Logits (-2.801 bis 0.889). Die mittlere Personenfähigkeit liegt bei 1.25 und somit über der mittleren Aufgabenschwierigkeit.

»Tabelle 3 hier einfügen«

Jedes der 11 Partial-Credit-Items deckt mehrere Komplexitätsstufen ab, sodass für jede Komplexitätsstufe ein eigener Schwierigkeitsparameter geschätzt werden muss. Hierzu wurden mithilfe des R-Softwarepakets *WrightMap* (Torres Irribarra & Freund, 2016; R Core Team, 2018) für jedes Partial-Credit-Item die Thurstonian-Threshold-Parameter geschätzt. Diese geben diejenige Stelle auf der Skala an, bei der die Wahrscheinlichkeit 50% beträgt eine bestimmte Kategorie oder eine höhere zu erreichen (Wu, Tam & Jen, 2016). Diese Kategorienschwierigkeiten wurden anschliessend mithilfe einer Dummy-Kodierung jeweils als eigenständige Items behandelt und nach ihrem Komplexitätsgrad klassifiziert (vgl. Tabelle 4). Aus dem vierstufigen Partial-Credit-Item PK39S zu Abbildung 2 wurden so bspw. vier Dummy-Items generiert. Die Schwierigkeiten dieser Dummy-Variablen entsprachen den Kategorienschwierigkeiten der jeweiligen Stufe des Partial-Credit-Items. Beispielsweise entstand so aus der Kategorie „4“ von PK39S das Dummy-Item PK39S_04 mit der Schwierigkeit 2.605 und dem Komplexitätsniveau „hoch“, das von 20% der Lehrkräfte korrekt gelöst wurde

»Tabelle 4 hier einfügen«

Aus dieser Dummy-Kodierung ergab sich für die Folgeanalyse eine erweiterte Item-Basis von insgesamt 57 Items. Von diesen wurden 32 Items einem niedrigen, 14 Items einem mittleren und 11 Items einem hohen Komplexitätsniveau zugeordnet.

Vorhersage der Item-Schwierigkeiten auf Basis des Aufgabenmerkmals

Auf dieser erweiterten Item-Basis wurde anhand einer multiplen Regressionsanalyse untersucht, inwiefern sich das Merkmal der kognitiven Lösungskomplexität einer Aufgabe zur Vorhersage der Item- bzw.

Kategorienschwierigkeiten eignet. Das Ergebnis (Tabelle 5) zeigt erwartungskonform, dass die Komplexität einen Einfluss auf die Schwierigkeit hat ($F(2, 54) = 26.72, p < .001$). Die Effekte der Komplexitätsabstufungen erweisen sich als praktisch bedeutsam. Mit einem $adj. R^2 = .48$ kann das Modell etwa die Hälfte der Gesamtstreuung der Item-Schwierigkeiten durch das Merkmal Komplexität erklären, was nach Cohen (1988) einer hohen Anpassungsgüte entspricht.

»Tabelle 5 hier einfügen«

Bildung eines Kompetenzniveauomodells und kriteriale Validierung

Um die Testwerte der untersuchten Lehrkräfte qualitativ interpretierbar zu machen, werden die Regressionskoeffizienten zu Kompetenzniveaus aufaddiert (vgl. Hartig, 2007). Die Regressionskonstante stellt hierbei die Schwierigkeit einer Aufgabe mit der niedrigsten Merkmalsausprägung dar. In unserem Fall liegt die Schwelle für Niveau I, auf dem Aufgaben von niedriger Komplexität mit hinreichender Wahrscheinlichkeit gelöst werden, bei -0.463 Logits. Durch die Addition von Konstante und den jeweiligen Regressionskoeffizienten ergeben sich die Schwellenwerte für Niveaus II und III, auf denen Aufgaben von mittlerer bzw. hoher Komplexität gelöst werden können (s. Tabelle 6). Eine Zuordnung der Versuchspersonen zu den einzelnen Niveaus zeigt, dass sich der Großteil der Lehrkräfte (84%) auf Niveau II befinden und 4% Niveau III erreichen.

»Tabelle 6 hier einfügen«

Um die weiterführende Interpretierbarkeit des Kompetenzniveauomodells sicherzustellen, wurde abschließend geprüft, inwiefern sich das Erreichen bestimmter Kompetenzniveaus in externen Kriterien widerspiegelt. Für diese Analyse wurden proximale Maße herangezogen, die in TEDS-Unterricht erhoben wurden und von denen ein Zusammenhang mit dem pädagogischen Wissen von Lehrpersonen angenommen werden kann. Dies sind beispielsweise situationsspezifische Fähigkeiten, d.h. Fähigkeiten, erfolgreich Unterrichtssituationen wahrzunehmen und zu interpretieren sowie unterrichtliche Entscheidungen zu treffen (Blömeke, Gustafsson & Shavelson, 2015). Durch ihre Situietheit bieten sie einen inhaltlich passgenauen Anknüpfungspunkt an das Kompetenzniveauomodell: Um im Unterricht erfolgreich zu handeln, muss eine Lehrperson in der Lage sein, informationsreiche Situationen schnell und erfolgreich zu bearbeiten (Weinert, 1996), also kognitiv komplexe Aufgaben zu bewältigen. Es kann daher angenommen werden, dass ein höheres kognitives Komplexitätsniveau im pädagogischen Wissen der Lehrpersonen auch mit ausgeprägteren situationsspezifischen Fähigkeiten einhergeht. Erste Forschungsarbeiten stützen diese Annahme und zeigen positive Korrelationen zwischen den globalen Testwerten aus dem TEDS-M-Instrument zum pädagogischen Wissen und Ergebnissen aus video-basierten Tests wie dem TEDS-Follow Up-Videoinstrument (pädagogische Skala P_PID ; König et al., 2014; Blömeke, Hoth et al., 2015; Kaiser, Busse, Hoth, König & Blömeke, 2015) oder dem Classroom Management Expertise-Test (CME; König, 2015; König & Kramer, 2016).

Zur Prüfung unserer Annahme werden zwei einfaktorielle Varianzanalysen durchgeführt, die das Kompetenzniveau als Faktor und die WLE-Scores der Testpersonen im TEDS-FU $_{P_PID}$ - bzw. CME-Video-Test als abhängige Variable berücksichtigen. Die Ergebnisse zeigen in Tabelle 7 für beide Instrumente, dass die pädagogischen Kompetenzniveaugruppen sich signifikant in den Mittelwerten der situationsspezifischen Tests

unterscheiden und es sich nach Cohen (1988) um Effekte mittlerer Stärke handelt (TEDS-FU_{P_PID}: $F(3,109) = 9,165$, $p < .001$, $\eta^2 = .20$; CME: $F(3,109) = 4,981$, $p = .003$, $\eta^2 = .12$).¹

» Tabelle 7 hier einfügen«

Die anschließend angewandten Post-hoc-Tests mit Bonferroni-Korrektur unterstreichen für beide Tests, dass Lehrpersonen mit mittlerer und hoher kognitiver Komplexität über ausgeprägtere situationsspezifische Fähigkeiten verfügen als Lehrpersonen mit niedriger Komplexität (vgl. Tabelle 7). Die Mittelwerte der Personen auf Kompetenzniveau I fallen in beiden Tests signifikant niedriger aus als die der Personen auf Kompetenzniveau II (TEDS-FU_{P_PID}: $p < .001$; CME: $p = .05$) und auf Kompetenzniveau III (TEDS-FU_{P_PID}: $p < .001$; CME: $p = .02$). Auch zwischen dem untersten und dem höchsten Niveau zeigen sich jeweils signifikante Unterschiede (TEDS-FU_{P_PID}: $p = .08$; CME: $p = .04$). Die Unterschiede zwischen dem untersten Kompetenzniveau und den Niveaus I bzw. II fallen nicht signifikant, jedoch in ihrer Richtung erwartungskonform aus. Ähnliches gilt für die Mittelwertunterschiede zwischen den zwei obersten Niveaus. Einzig beim CME-Test zeigt sich für Personen auf Niveau III ein erwartungswidrig niedrigeres Ergebnis für Personen auf Niveau II. Zusammenfassend betrachtet, unterstreichen die Ergebnisse der Varianzanalysen die prädiktive Kriteriumsvalidität (Messick, 1990) des Kompetenzniveauomodells.

Diskussion

Ziel dieses Beitrags war es, das Instrument zur Messung des pädagogischen Wissens von TEDS-M hinsichtlich seiner Konstruktrepräsentation zu untersuchen. Es wurde der Frage nachgegangen, ob das Instrument die erwarteten kognitive Prozesse abbildet, die berufserfahrene Mathematiklehrkräfte bei der Bearbeitung pädagogischer, unterrichtsbezogener Aufgaben anwenden. Zunächst kann festgehalten werden, dass sich das Instrument auch in dieser spezifischen Lehrpersonengruppe als geeignet erweist, um pädagogisches Wissen zu erfassen: Die Ergebnisse der IRT-Skalierung weisen zufriedenstellende Reliabilitäts- und Item-Fit-Werte auf.

Darüber hinaus konnten Hinweise für die Validität des Instruments im Sinne der Konstruktrepräsentation erbracht werden. Die Ergebnisse der Regressionsanalysen bestätigen, dass sich das Modell von König (2009) und Klemenz und König (2019) in dem hier angewandten Teil replizieren lässt: Die Varianz in den Aufgabenschwierigkeiten lässt sich bedeutsam durch die Komplexität der Aufgabenlösung aufklären. Dies stützt die Annahme, dass während der Testbearbeitung der angenommene Prozess des Verbindens mehrerer Wissens Elemente zum Tragen kommt, und kann daher als positives Indiz für die Konstruktrepräsentation interpretiert werden. Die kognitive Leistung, einzelne Wissens Elemente partiell zu verknüpfen, konnten wir in unserer Stichprobe für den Großteil der Lehrkräfte belegen; diese zu einem ganzheitlichen Konzept zusammenzuführen, ließ sich allerdings nur für einen kleinen Anteil zeigen

¹ Aufgrund der niedrigen Gruppengrößen wurden zudem zwei Kruskal-Wallis-Tests durchgeführt, die die Ergebnisse der parametrischen Verfahren bestätigen (TEDS-FU_{P_PID}: Chi-Quadrat(3) = 19.952, $p < .001$; CME: Chi-Quadrat(3) = 11.552, $p = .009$).

(vgl. Tabelle 6). Angesichts einer vermutlich positiv selektierten Stichprobe² werten wir dies als Hinweis, dass das aufgestellte Kompetenzmodell für berufserfahrene Lehrkräfte gut erreichbar, aber nicht zu einfach gehalten ist – jedenfalls liegen keine starken Deckeneffekte mit Blick auf die modellierte kognitive Komplexität im pädagogischen Wissen vor.

Methodisch betrachtet konnte die Operationalisierung kognitiver Komplexität so weiterentwickelt werden, dass im Gegensatz zu bisherigen Ansätzen keine Dichotomisierung der Items mehr notwendig war. Somit konnten zwei wesentliche Nachteile der Dichotomisierung behoben werden: Zum einen wurde durch die Partial-Credit-Modellierung komplexer Items bei der Skalierung und Ermittlung der Itemschwierigkeiten ein Informationsverlust vermieden. Zum anderen ermöglichte die anschließende Dummy-Kodierung der Partial-Credit-Items es, kognitive Komplexität in mehr als zwei Niveaus abzubilden.

Dass diese differenziertere Modellierung geeignet ist unterschiedliche kognitive Komplexitätsgrade abzubilden, unterstreichen auch die Ergebnisse der Varianzanalysen zwischen Kompetenzniveau und situationsspezifischen Fähigkeiten. Diese zeigten nicht nur, dass situationsspezifische Fähigkeiten in pädagogischen Bereichen durch das Kompetenzniveau im pädagogischen Wissen vorhergesagt werden konnten. In den Post-hoc-Tests wurde auch ersichtlich, dass niedrige kognitive Komplexität sich signifikant von mittlerer Komplexität unterscheidet. Es scheint also empfehlenswert, Prozesse mittlerer kognitiver Komplexität (d.h. partielle Verknüpfungen mehrerer Wissens Elemente) nicht mit Prozessen niedriger Komplexität (d.h. Anwenden einzelner Wissens Elemente) gleichzusetzen, wie es im Rahmen der Dichotomisierung in bisherigen Arbeiten umgesetzt wurde. Stattdessen legen die Ergebnisse nahe, beide Teilprozesse als eigenständige Komplexitätsniveaus zu konzipieren. Gleichzeitig ist allerdings zu bemerken, dass die Unterschiede zwischen mittlerer und hoher Komplexität nicht signifikant ausfallen. Dies lässt sich womöglich durch die äußerst kleine Randgruppe auf Niveau III ($n = 5$) erklären, sollte dennoch in zukünftigen Studien an größeren Stichproben überprüft werden.

Insgesamt ist die Größe der Stichprobe als eine Limitierung unserer Studie zu benennen. In zukünftigen Analysen wird es darum gehen, das weiterentwickelte Modell auf größere Gruppen anzuwenden und zu validieren. Ähnliches gilt für die Anzahl der Items: Das Modell konnte hier nur auf die reduzierte Item-Basis angewendet werden, die sich aus der Kürze der eingesetzten TEDS-M-Testversion ergibt. Dadurch musste die Anwendung unterrichtsbezogener Terminologie als kognitiver Prozess aus den Analysen ausgeschlossen werden. Es bleibt also noch offen, wie generalisierbar die Befunde sind und wie hoch der Aufklärungsanteil des gesamten Kompetenzniveau Modells für die hier untersuchte Gruppe ausfällt. Für zukünftige Untersuchungen empfiehlt es sich daher, analog zu den Analysen von Klemenz und König (2019) die Langversion des TEDS-M-Instruments zu verwenden, entsprechend des hier durchgeführten Vorgehens bei berufserfahrenen Lehrkräften einzusetzen und ebenfalls hinsichtlich der Konstruktrepräsentation zu prüfen.

² Im Rahmen der Teilnehmendenrekrutierung wurde einer positiven Verzerrung mithilfe breiter Öffentlichkeitsarbeit und monetärer Aufwandsentschädigungen entgegenwirkt. Dennoch macht die Mobilisierung privater Netzwerke des Projektteams eine Selektion überdurchschnittlich motivierter Testpersonen wahrscheinlich.

Eine breitere Testpersonen- und Item-Basis böte überdies die Möglichkeit weitere kognitive Teilprozesse und Aufgabenmerkmale zu berücksichtigen, die für die Konstruktrepräsentation des pädagogischen Wissens bedeutsam sind. Die hier vorgenommene Konzeptualisierung der kognitiven Komplexität könnte um den Aspekt hierarchischer Komplexität (Commons et al., 2008) ergänzt werden, sodass neben der Anzahl gleichwertiger Wissens Elemente auch die Anzahl der unterschiedlich gearteten, einander übergeordneten Operationen berücksichtigt würden, die zur Bearbeitung konsekutiv durchlaufen werden müssen. Hierzu könnten beispielsweise die von Kauertz et al. (2010) modellierten Prozesse des Reproduzierens, Selektierens, Organisierens und Integrierens oder die von Embretson (2016) identifizierten Prozesse des Dekodierens, Integrierens und Entscheidens aufgegriffen werden. Weiterführend wäre schließlich interessant zu prüfen, inwiefern sich kognitive Komplexität auch eignet, um die Konstruktrepräsentation in anderen pädagogischen Kompetenztests zu stützen, und möglicherweise als Basis bei zukünftigen Test- und Itementwicklungen zu dienen.

Literatur

- Altrichter, H. & Posch, P. (2007). *Lehrerinnen und Lehrer erforschen ihren Unterricht. Unterrichtsentwicklung und Unterrichtsevaluation durch Aktionsforschung* (4., überarbeitete und erweiterte Auflage). Bad Heilbrunn: Julius Klinkhardt.
- Anderson, L. & Krathwohl, D. (Hrsg.). (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- 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>
- Beck, B. & Klieme, E. (Hrsg.). (2007). *Sprachliche Kompetenzen. Konzepte und Messung: DESI-Studie (Deutsch Englisch Schülerleistungen International)* (Beltz-Pädagogik). Weinheim: Beltz Verlag.
- 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., Hoth, J., Döhrmann, M., Busse, A., Kaiser, G. & König, J. (2015). Teacher Change During Induction. Development of Beginning Primary Teachers' Knowledge, Beliefs and Performance. *International Journal of Science and Mathematics Education*, 13 (2), 287–308. <https://doi.org/10.1007/s10763-015-9619-4>
- Blömeke, S., Lehmann, R., Seeber, S., Schwarz, B., Kaiser, G., Felbrich, A. et al. (2008). Niveau- und institutionenbezogene Modellierungen des fachbezogenen Wissens. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *Professionelle Kompetenz angehender Lehrerinnen und Lehrer. Wissen, Überzeugungen und Lerngelegenheiten deutscher Mathematikstudierender und -referendare; erste Ergebnisse zur Wirksamkeit der Lehrerausbildung* (S. 105–134). Münster u.a.: Waxmann.
- Blömeke, S., Lehmann, R., Seeber, S., Schwarz, B., Kaiser, G., Felbrich, A. et al. (2010). Niveau- und institutionenbezogene Modellierungen des fachbezogenen Wissens. In S. Blömeke, G. Kaiser & R. Lehmann (Hrsg.), *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich* (S. 105–134). Münster: Waxmann.
- Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111 (4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Bromme, R. (1992). *Der Lehrer als Experte. Zur Psychologie des professionellen Wissens* (Huber-Psychologie-Forschung, 1. Aufl.). Bern: Huber.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). Hillsdale, NJ u.a.: Erlbaum.
- Cronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52 (4), 281–302.
- Daniel, R. C. & Embretson, S. (2010). Designing Cognitive Complexity in Mathematical Problem-Solving Items. *Applied Psychological Measurement*, 34 (5), 348–364. <https://doi.org/10.1177/0146621609349801>

- Embretson, S. (1983). Construct validity. Construct representation versus the nomothetic span. *Psychological Bulletin*, 93 (1), 179–197.
- Embretson, S. (2016). Understanding Examinees' Responses to Items. Implications for Measurement. *Educational Measurement: Issues and Practice*, 35 (3), 6–22. <https://doi.org/10.1111/emip.12117>
- Eye, A. von. (1999). Kognitive Komplexität - Messung und Validität. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 20, 81–96. Verfügbar unter <http://econtent.hogrefe.com/doi/full/10.1024//0170-1789.20.2.81>
- Gindele, V. & Voss, T. (2017). Pädagogisch-psychologisches Wissen. Zusammenhänge mit Indikatoren des beruflichen Erfolgs angehender Lehrkräfte. *Zeitschrift für Bildungsforschung*, 108 (6), 800. <https://doi.org/10.1007/s35834-017-0192-5>
- Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In B. Beck & E. Klieme (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung: DESI-Studie (Deutsch Englisch Schülerleistungen International)* (Beltz-Pädagogik, S. 83–99). Weinheim: Beltz Verlag.
- Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, 63 (1), 43–49. <https://doi.org/10.1026/0033-3042/a000109>
- Hartig, J., Frey, A. & Jude, N. (2012). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, S. 143–171). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20072-4_7
- Kaiser, G., Busse, A., Hoth, J., König, J. & Blömeke, S. (2015). About the complexities of video-based assessments. Theoretical and methodological approaches to overcoming shortcomings of research on teachers' competence. *International Journal of Science and Mathematics Education*, 13 (2), 369–387. <https://doi.org/10.1007/s10763-015-9616-7>
- Kauertz, A., Fischer, H., Mayer, J., Sumfleth, E. & Walpuski, M. (2010). Standardbezogene Kompetenzmodellierung in den Naturwissenschaften der Sekundarstufe I. *Zeitschrift für Didaktik der Naturwissenschaften*, 19, 135–153. Verfügbar unter http://archiv.ipn.uni-kiel.de/zfdn/pdf/16_Kauertz.pdf
- Klemenz, S. & König, J. (2019). Modellierung von Kompetenzniveaus im pädagogischen Wissen bei angehenden Lehrkräften. Zur kriterialen Beschreibung von Lernergebnissen der fächerübergreifenden Lehramtsausbildung. *Zeitschrift für Pädagogik*, 65 (3).
- 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.
- König, J. (2013). First comes the theory, then the practice? On the acquisition of general pedagogical knowledge during initial teacher education. *International Journal of Science and Mathematics Education*, 11 (4), 999–1028. <https://doi.org/10.1007/s10763-013-9420-1>

- König, J. (2014a). *Designing an International Instrument to Assess Teachers' General Pedagogical Knowledge (GPK): Review of Studies, Considerations, and Recommendations*. Technical paper prepared for the OECD Innovative Teaching for Effective Learning (ITEL) - Phase II Project: A Survey to Profile the Pedagogical Knowledge in the Teaching Profession (ITEL Teacher Knowledge Survey). Paris: OECD. Verfügbar unter <http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/CERI/CD/RD%282014%293/REV1&doclanguage=en>
- König, J. (2014b). *Designing an International Instrument to Assess Teachers' General Pedagogical Knowledge (GPK): Review of Studies, Considerations, and Recommendations*. Technical paper prepared for the OECD Innovative Teaching for Effective Learning (ITEL) - Phase II Project: A Survey to Profile the Pedagogical Knowledge in the Teaching Profession (ITEL Teacher Knowledge Survey). Paris: OECD. Verfügbar unter <http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=EDU/CERI/CD/RD%282014%293/REV1&doclanguage=en>
- König, J. (2015). Measuring classroom management expertise (CME) of teachers. A video-based assessment approach and statistical results. *Cogent Education*, 2 (1), 1–15. <https://doi.org/10.1080/2331186X.2014.991178>
- König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften. Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. *Zeitschrift für Erziehungswissenschaft*, 12 (3), 499–527. <https://doi.org/10.1007/s11618-009-0085-z>
- König, J. & Blömeke, S. (2010). *Pädagogisches Unterrichtswissen (PUW). Dokumentation der Kurzfassung des TEDS-M-Testinstruments zur Kompetenzmessung in der ersten Phase der Lehrerausbildung*. Berlin: Humboldt-Universität.
- König, J., Blömeke, S. & Kaiser, G. (2015). Early career mathematics teachers' general pedagogical knowledge and skills. Do teacher education, teaching experience, and working conditions make a difference? *International Journal of Science and Mathematics Education*, 13 (2), 331–350. <https://doi.org/10.1007/s10763-015-9618-5>
- König, J., Blömeke, S., Klein, P., Suhl, U. & Busse, A. (2014). Is teachers' general pedagogical knowledge a premise for noticing and interpreting classroom situations? A video-based assessment approach. *Teaching and Teacher Education*, 38, 76–88. <https://doi.org/10.1016/j.tate.2013.11.004>
- König, J., Blömeke, S., Paine, L., Schmidt, W. H. & Hsieh, F.-J. (2011). General pedagogical knowledge of future middle school teachers. On the complex ecology of teacher education in the United States, Germany, and Taiwan. *Journal of Teacher Education*, 62 (2), 188–201. <https://doi.org/10.1177/0022487110388664>
- König, J. & Klemenz, S. (2015). Der Erwerb von pädagogischem Wissen bei angehenden Lehrkräften in unterschiedlichen Ausbildungskontexten. Zur Wirksamkeit der Lehrerausbildung in Deutschland und Österreich. *Zeitschrift für Erziehungswissenschaft*, 18 (2), 247–277. <https://doi.org/10.1007/s11618-015-0623-9>
- König, J. & Kramer, C. (2016). Teacher professional knowledge and classroom management. On the relation of general pedagogical knowledge (GPK) and classroom management expertise (CME). *ZDM*, 48 (1-2), 139–151. <https://doi.org/10.1007/s11858-015-0705-4>

- König, J. & Pflanzl, B. (2016). Is teacher knowledge associated with performance? On the relationship between teachers' general pedagogical knowledge and instructional quality. *European Journal of Teacher Education*, 39 (4), 419–436. <https://doi.org/10.1080/02619768.2016.1214128>
- König, J. & Seifert, A. (Hrsg.). (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerbildung*. Münster u.a.: Waxmann.
- Laueremann, F. & König, J. (2016). Teachers' professional competence and wellbeing. Understanding the links between general pedagogical knowledge, self-efficacy and burnout. *Learning and Instruction*, 45, 9–19. <https://doi.org/10.1016/j.learninstruc.2016.06.006>
- Messick, S. (1990). *Validity of test interpretation and use* (ETS Research Report RR-90-11). Princeton, U.S.A.: Educational Testing Service. Verfügbar unter <https://files.eric.ed.gov/fulltext/ED395031.pdf>
<https://doi.org/10.1002/j.2333-8504.1990.tb01343.x>
- Messick, S. (1994). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *ETS Research Report Series* (2), i-28. <https://doi.org/10.1002/j.2333-8504.1994.tb01618.x>
- Prenzel, M., Häußler, P., Rost, J. & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, 30 (2), 120–135. Verfügbar unter https://www.pedocs.de/volltexte/2013/7682/pdf/UnterWiss_2002_1_Prenzel_et_al_PISA_Naturwissenschaftstest.pdf
- R Core Team. (2018). R. A Language and Environment for Statistical Computing (Version 3.5.0) [Computer software]. Wien: R Foundation for Statistical Computing. Verfügbar unter <http://www.R-project.org>
- Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik*, 50 (5), 662–678. Verfügbar unter https://www.pedocs.de/volltexte/2011/4834/pdf/ZfPaed_2004_5_Rost_Psychometrische_Modelle_zur_Ueberpruefung_D_A.pdf
- Shulman, L. S. (1987). Knowledge and teaching. Foundations of the new reform. *Harvard Educational Review*, 57 (1), 1–21.
- Terhart, E. (1993). Pädagogisches Wissen. Überlegungen zu seiner Vielfalt, Funktion und sprachlichen Form am Beispiel des Lehrerwissens. In J. Oelkers & H.-E. Tenorth (Hrsg.), *Pädagogisches Wissen* (Zeitschrift für Pädagogik Beiheft, Bd. 27, S. 129–141). Weinheim: Beltz.
- 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>
- Torres Iribarra, D. & Freund, R. (2016). Wright Map. IRT item-person map with ConQuest integration (Version 1.2.1) [Computer software]. Verfügbar unter <http://github.com/david-ti/wrightmap>

- Viering, T., Fischer, H. & Neumann, K. (2010). Die Entwicklung physikalischer Kompetenz in der Sekundarstufe I. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes*. 56. Beiheft (1. Auflage, S. 92–104). Weinheim: Beltz.
- Voss, T., Kunina-Habenicht, O., Hoehne, V. & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften. Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft*, 18 (2), 187–223.
<https://doi.org/10.1007/s11618-015-0626-6>
- Weinert, F. E. (1996). „Der gute Lehrer“, „die gute Lehrerin“ im Spiegel der Wissenschaft. Was macht Lehrende wirksam und was führt zu ihrer Wirksamkeit. *Beiträge zur Lehrerinnen- und Lehrerbildung*, 14 (2), 141–151. Verfügbar unter <https://core.ac.uk/download/pdf/83642611.pdf>
- Wipperfürth, M. (2015). *Professional vision in Lehrernetzwerken. Berufssprache als ein Weg und ein Ziel von Lehrerprofessionalisierung* (Münchener Arbeiten zur Fremdsprachen-Forschung, Bd. 32, 1. Aufl.). Münster: Waxmann Verlag GmbH.
- Wu, M., Adams, R. & Wilson, M. (1997). ConQuest. Multi-aspect test software [Computer software]. Camberwell: Australian Council for Educational Research.
- Wu, M., Tam, H. P. & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Singapore: Springer.
<https://doi.org/10.1007/978-981-10-3302-5>

Tabellen:

Tabelle 1. Design-Matrix zum TEDS-M-Test des pädagogischen Wissens (König & Blömeke, 2009) mit Angaben zur Itemanzahl je Subdimension in der Test-Kurzversion

Berufliche Anforderungen von Lehrpersonen beim Unterrichten		Kognitive Anforderung		
Subdimensionen	operationalisierte Themen	Erinnern	Verstehen / Analysieren	Kreieren
Strukturierung von Unterricht	<ul style="list-style-type: none"> ▪ Komponenten- und prozessbezogene Planung ▪ Analyse von Unterricht ▪ Curriculare Strukturierung von Unterricht 	4	10	4
Motivierung	<ul style="list-style-type: none"> ▪ Leistungsmotivation ▪ Motivierungsstrategien im Unterricht 	-	6	3
Umgang mit Heterogenität	<ul style="list-style-type: none"> ▪ Differenzierungsmaßnahmen ▪ Methodenvielfalt 	4	8	-
Klassenführung	<ul style="list-style-type: none"> ▪ Störungspräventive Unterrichtsführung ▪ effektive Nutzung der Unterrichtszeit 	-	-	3
Leistungsbeurteilung	<ul style="list-style-type: none"> ▪ Funktionen und Formen ▪ zentrale Kriterien ▪ Urteilsfehler 	5	-	-

Tabelle 2. Kodierungsverlauf bei offenen Items am Beispiel von Aufgabe PK39

	Kodierungsitems	Wertungsitems	Partial-Credit-Item	Komplexitätsniveau
Kategorien		0, 1	0, 1, 2, 3, 4	niedrig, mittel, hoch
Itembezeichnung	PK39K1	PK39S01 (Kontext)	PK39S	niedrig für PK39S=1 mittel für $1 < PK39S < 4$ hoch für PK39S=4
	PK39K2	PK39S02 (Input)		
	PK39K5			
	PK39K6			
	PK39K7			
	PK39K8			
	PK39K3	PK39S03 (Prozess)		
	PK39K4			
	PK39K9			
	PK39K10			
	PK39K11			
	PK39K12	PK39S04 (Output)		

Tabelle 3. Itemschwierigkeitsparameter der 1-dimensionalen Skalierung sowie Angaben zum Weighted Item Fit (Mean Square, Konfidenzintervall, T-Wert, Trennschärfe, erforderliche Wissensselemente)

Item	Itemschwierigkeitsparameter	Standardfehler	MNSQ	Konfidenzintervall	T-Wert	Trennschärfe	Anzahl erforderlicher Wissensselemente
PK01S01	-2.070	0.232	0.99	(0.10, 1.90)	0.1	.30	1
PK01S02	-1.453	0.218	0.97	(0.36, 1.64)	0.0	.39	1
PK01S03	-2.801	0.244	0.95	(0.00, 2.34)	0.1	.48	1
PK01S04	-0.626	0.196	1.02	(0.61, 1.39)	0.2	.28	1
PK01S05	1.401	0.168	1.03	(0.88, 1.12)	0.5	.31	1
PK07S+	1.356	0.088	1.13	(0.78, 1.22)	1.1	.55	3
PK09S	2.943	0.191	0.92	(0.66, 1.34)	-0.4	.35	1
PK12	0.412	0.183	0.98	(0.79, 1.21)	-0.1	.37	1
PK15S+	1.659	0.147	1.03	(0.71, 1.29)	0.3	.38	3
PK20B	0.183	0.186	1.02	(0.75, 1.25)	0.2	.28	1
PK20CS01	-1.250	0.218	1.01	(0.39, 1.61)	0.1	.18	1
PK20CS02	-2.607	0.244	1.00	(0.00, 2.31)	0.2	.10	1
PK20CS03	-0.603	0.202	1.00	(0.59, 1.41)	0.1	.24	1
PK20CS04	-1.095	0.214	1.01	(0.44, 1.56)	0.1	.22	1
PK22AS+	-0.154	0.132	0.77	(0.42, 1.58)	-0.8	.61	4
PK26S01	-0.021	0.180	0.91	(0.74, 1.26)	-0.7	.49	1
PK26S02	-0.084	0.181	0.88	(0.73, 1.27)	-0.8	.51	1
PK26S03	0.271	0.174	0.91	(0.79, 1.21)	-0.8	.47	1
PK32S1+	1.637	0.129	0.86	(0.71, 1.29)	-0.9	.60	3
PK32S2+	1.386	0.120	1.09	(0.73, 1.27)	0.7	.50	3
PK32S3+	0.865	0.122	0.99	(0.71, 1.29)	-0.0	.54	3
PK32S4+	0.889	0.130	0.98	(0.70, 1.30)	-0.1	.52	3
PK33S+	0.193	0.119	1.06	(0.59, 1.41)	0.4	.46	4
PK39S+*	0.297	0.133	1.11	(0.60, 1.40)	0.6	.41	4
PK40AS+	0.761	0.155	1.05	(0.71, 1.29)	0.3	.28	3
PK67S+	0.266	0.141	1.03	(0.66, 1.34)	0.2	.51	3
PK82S01	-1.462	0.218	0.98	(0.36, 1.64)	0.0	.30	1

PK82S02	0.033	0.180	1.03	(0.75, 1.25)	0.2	.30	1
PK82S03	-0.309	0.188	1.11	(0.68, 1.32)	0.7	.15	1
PK82S04	-0.465	0.192	0.99	(0.65, 1.35)	-0.0	.32	1
PK82S05	0.546	0.172	0.98	(0.83, 1.17)	-0.3	.37	1
PK82S06	-0.097	0.994	1.07	(0.72, 1.28)	0.6	.13	1

Anmerkungen: Bei den mit + markierten Items handelt es sich um Partial-Credit-Items, deren Stufen unterschiedliche Komplexitätsniveaus abbilden. Das mit * markierte Item entspricht der in Abbildung 2 dargestellten Beispielaufgabe.

Tabelle 4. Lösungshäufigkeiten und Thurstonian Thresholds der Partial-Credit-Items sowie Komplexitätsniveaus der resultierenden Dummy-Variablen

Partial-Credit-Item	Kategorie	Lösungshäufigkeit (gültige %)	Thurstonian Threshold	Dummy-Items	Komplexität
PK07S	1	10,3	1.029	PK07_01	niedrig
	2	18,7	1.263	PK07_02	mittel
	3	32,0	1.717	PK07_03	hoch
PK15S	1	45,7	-0.949	PK15S_01	niedrig
	2	44,4	1.391	PK15S_02	mittel
	3	2,3	4.534	PK15S_03	hoch
PK22AS	1	0,0	-0.374	PK22A_01	niedrig
	2	2,0	-0.371	PK22A_02	mittel
	3	15,8	-0.260	PK22A_03	mittel
	4	79,2	0.162	PK22A_04	hoch
PK32S1	1	40,3	0.175	PK32S1_01	niedrig
	2	29,2	1.644	PK32S1_02	mittel
	3	8,3	3.093	PK32S1_03	hoch
PK32S2	1	26,8	0.325	PK32S2_01	niedrig
	2	35,2	1.224	PK32S2_02	mittel
	3	15,5	2.590	PK32S2_03	hoch
PK32S3	1	22,2	-0.190	PK32S3_01	niedrig
	2	37,5	0.766	PK32S3_02	mittel
	3	27,8	2.006	PK32S3_03	hoch
PK32S4	1	24,0	-0.436	PK32S4_01	niedrig
	2	45,1	0.707	PK32S4_02	mittel
	3	21,1	2.385	PK32S4_03	hoch
PK33S	1	2,2	-0.462	PK33_01	niedrig
	2	5,4	-0.225	PK33_02	mittel
	3	42,4	-0.003	PK33_03	mittel
	4	46,7	1.360	PK33_04	hoch
PK39S*	1	1,1	-0.846	PK39_01	niedrig
	2	14,4	-0.687	PK39_02	mittel
	3	62,2	-0.012	PK39_03	mittel
	4	20,0	2.605	PK39_04	hoch
PK40AS	1	55,0	-2.930	PK40A_01	niedrig
	2	38,8	1.503	PK40A_02	mittel
	3	5,0	3.709	PK40A_03	hoch
PK67S	1	12,5	-0.987	PK67_01	niedrig
	2	50,0	-0.066	PK67_02	mittel
	3	33,3	1.832	PK67_03	hoch

Anmerkungen: Bei dem mit * markierten Item handelt es sich um die in Abbildung 2 dargestellte Beispielaufgabe.

Tabelle 5. Nicht standardisierte Regressionsgewichte als Prädiktoren der Schwierigkeitsparameter

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten	<i>T</i>	<i>p</i>
	Regressionskoeffizient	Standardfehler	Beta		
<i>Adj. R</i> ² = 0.48					
(Konstante)	-0.463	.196		-2.356	.022
mittlere Komplexität	0.954	.356	.269	2.680	.010
hohe Komplexität	2.826	.388	.731	7.279	<.001

Tabelle 6. Schwellenwerte, Kurzbeschreibungen der Kompetenzniveaus und Häufigkeitsverteilung der Testteilnehmenden

Kompetenzniveau	Logit	Kurzbeschreibung	Häufigkeit	Prozent
Unterhalb von I	<-0.463	Bearbeitungsprozesse von niedriger Komplexität können <i>nicht</i> mit hinreichender Wahrscheinlichkeit gelöst werden	3	2.7
I	-0.463	Bearbeitungsprozesse von niedriger Komplexität können mit hinreichender Wahrscheinlichkeit gelöst werden	10	8.8
II	0.491	Bearbeitungsprozesse von mittlerer Komplexität können mit hinreichender Wahrscheinlichkeit gelöst werden	95	84.1
III	2.363	Bearbeitungsprozesse von hoher Komplexität können mit hinreichender Wahrscheinlichkeit gelöst werden	5	4.4

Tabelle 7. Mittelwerte und Standardabweichungen der WLE-Scores zu den situationsspezifischen TEDS-FU_{P_PID}- und CME-Tests, nach Kompetenzniveau

Kompetenzniveau	TEDS-FU _{P_PID} -Videotest			CME-Test	
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
< I	3	-0,179 ³	0,288	-1,276 ^{III}	0,874
I	10	-0,455 ^{II, III}	0,760	-0,688 ^{III, 2}	0,924
II	95	0,503 ^I	0,584	0,402 ^I	0,936
III	5	1,070 ^{I, <1}	0,741	0,236 ^{<1, I}	0,494

Anmerkungen: Die Indices ^{<1, I, II, III} geben diejenigen Gruppen an, deren Mittelwerte sich gemäß der einfaktoriellen Varianzanalysen und Post-hoc-Tests mit Bonferroni-Korrektur signifikant von den jeweils angegebenen Mittelwerten unterscheiden ($p < .05$). Die mit ^{<1, 1, 2, 3} markierten Mittelwerte weisen einen statistischen Trend zu Mittelwertunterschieden auf ($p < .10$).

Abbildungen:

Abbildung 1. Offene Aufgabe mit beispielhaften Antworten (kursiv) zur Erfassung des pädagogischen Wissens zu den Anforderungen *Strukturierung von Unterricht und Erinnern* (a) bzw. *Verstehen/Analysieren* (b).

Phasenmodelle von Unterricht stellen ein Grundgerüst dar, nach dem Unterricht strukturiert werden kann.

- a) Nennen Sie die zentralen Phasen eines üblichen Unterrichtsverlaufs.
 b) Nennen Sie die Funktion der jeweiligen Phase.

a) Name der Phase: **b) Funktion der Phase:**

<i>Einstieg</i>	<i>Motivation Themenpräsentation</i>
<i>Problemstellung</i>	<i>SuS verdeutlichen sich das Problem, sodass jeder es versteht</i>
<i>Erarbeitungsphase</i>	<i>SuS gehen dem Problem „auf die Spur“. Hier kann ganz differenziert gearbeitet werden.</i>
<i>Sicherungsphase</i>	<i>Die Lösung wird präsentiert. Jeder kann die Lösung übernehmen – mögliche Diskussion nötig</i>
<i>Anwendung/Transfer</i>	<i>Die Lösung wird bei weiteren Aufgaben benötigt, Relevanz der Lösung transparent</i>

Abbildung 2. Offene Aufgabe mit beispielhaften Antworten (kursiv) zur Erfassung des pädagogischen Wissens zu den Anforderungen *Strukturierung von Unterricht und Kreieren*.

Stellen Sie sich vor, Sie helfen einer angehenden Lehrperson bei der Auswertung ihres Unterrichts, weil sie dies noch nie gemacht hat.

Welche Fragen würden Sie ihr stellen, damit sie ihren Unterricht angemessen analysiert?

Nennen Sie zehn zentrale Fragen und formulieren Sie diese bitte aus.

- 1) *Verfügen Ihre Schülerinnen und Schüler über Vorwissen zum Thema?*
- 2) *Was sind Ihre Ziele?*
- 3) *Arbeiten die Lernenden einzeln oder in Gruppen?*
- ...
- 10) *Haben die Lernenden in der Stunde neues Wissen erworben?*