Originalarbeit

Pädagogisches Wissen von berufstätigen Mathematiklehrkräften

Validierung der Konstruktrepräsentation im TEDS-M-Test

Caroline Felske[1], Johannes König[1], Gabriele Kaiser[2,3], Stefan Klemenz[1], Natalie Ross[2] und Sigrid Blömeke[4]

[1]Humanwissenschaftliche Fakultät, Department Erziehungs- und Sozialwissenschaften, Universität zu Köln
[2]Fakultät für Erziehungswissenschaften, Universität Hamburg
[3]Institute for Learning Sciences & Teacher Education, Australian Catholic University, Brisbane, Australien
[4]Centre for Educational Measurement, University of Oslo, Norwegen

Zusammenfassung: Der in der TEDS-M-Studie (Teacher Education and Development Study: Learning to Teach Mathematics) entwickelte Test zur Erfassung pädagogischen Wissens am Ende der Lehramtsausbildung hat sich in diversen Untersuchungen als zuverlässiges Messinstrument erwiesen, für das eine Reihe von Ergebnissen vorliegt, die die Validität der Testwertinterpretationen bei (angehenden) Lehrkräften in unterschiedlichen Ausbildungsstadien und -kontexten unterstreichen. Ein wesentlicher Validierungsschritt steht jedoch noch aus: Die Überprüfung, ob sich der Test eignet, um quantitative und qualitative Aussagen zum pädagogischen Wissen von berufstätigen Mathematiklehrkräften zu treffen. Im Rahmen des Projekts TEDS-Validierung wurde an 113 Mathematiklehrkräften geprüft, ob der Test das Wissen der Lehrkräfte reliabel und differenziert erfasst. Darauf aufbauend wurde im Sinne der Konstruktrepräsentation (Embretson, 1983) untersucht, ob er konstruktrelevante, kognitive Bearbeitungsprozesse erfordert, wie sie von König (2009) und Klemenz und König (2019) modelliert wurden. Die Analysen bestätigen, dass der Test auch bei berufstätigen Mathematiklehrkräften ein reliables Messinstrument darstellt und unterstreichen, dass die kognitive Komplexität der erforderlichen Bearbeitungsprozesse einen bedeutsamen Anteil der Schwierigkeitsvarianz aufklärt. Sie liefern somit einen ersten Hinweis für die Konstruktrepräsentation und die Grundlage für eine qualitative Interpretation der Testwerte. Diese Interpretation wird durch Varianzanalysen validiert, die zeigen, dass Personen, die kognitiv komplexere Bearbeitungsprozesse im pädagogischen Wissenstest vollziehen können, auch ausgeprägtere situationsspezifische, pädagogische Fähigkeiten aufweisen als Vergleichspersonen.

Schlüsselwörter: Pädagogisches Wissen, berufstätige Mathematiklehrkräfte, Lehrkompetenzen, Konstruktrepräsentation, situationsspezifische Fähigkeiten

Pedagogical Knowledge – Validating the Construct Representation in the TEDS-M Test Among In-Service Mathematics Teachers

Abstract: The test for pedagogical knowledge that was developed within TEDS-M (Teacher Education and Development Study: Learning to Teach Mathematics) has proven reliable in several studies for measuring future teachers' knowledge. Across professional education phases and contexts, broad evidence has been presented for valid interpretations of the measures it provides about (future) teachers. Yet, one essential validation step has not been carried out to date: examination of the suitability of the test for quantitative and qualitative inferences about the pedagogical knowledge of in-service mathematics teachers. In the frame of a study within the project TEDS-Validate, 113 mathematics teachers were tested in order to determine whether the test allows for a reliable and differentiated measurement of their pedagogical knowledge. Following the concept of construct representation (Embretson, 1983), the study examined whether the test also requires construct-relevant cognitive processes as modeled by König (2009) and Klemenz and König (2019). The results confirm that the test works reliably with in-service mathematics teachers. They further underline that the variance in item difficulties can be explained by the complexity of the cognitive processes they require, thus providing first evidence for construct representation and the basis for qualitative interpretations of teachers' test results. Finally, these qualitative interpretations were validated via variance analyses showing that teachers who apply more complex cognitive processes in the pedagogical knowledge test also show higher situation-specific pedagogical skills.

Keywords: pedagogical knowledge, in-service mathematics teachers, teacher competence, construct representation, situation-specific skills

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 als kognitive Komponente von Lehrkompetenz zu erfassen (König, Blömeke, Paine, Schmidt & Hsieh, 2011). Der TEDS-M-

Test 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, 2014). So lieferten Studien zu Lehramtsstudierenden in der universitären Ausbildungsphase (König & Seifert, 2012; König & Klemenz, 2015) und im Referendariat (König, 2013) wiederholt inhalts-, struktur- und kriteriumsbezogene Validitätshinweise (für eine Übersicht, siehe Voss, Kunina-Habenicht, Hoehne & Kunter, 2015). Zugleich konnten erste Belege für eine valide Verwendung bei Lehrenden in der Berufseinstiegsphase vorgelegt werden (König, Blömeke, Klein, Suhl & Busse, 2014).

Der Test wurde konzipiert, um das pädagogische Wissen zu erfassen, das angehende Lehrkräfte in ihrer Ausbildung erwerben (König et al., 2011) und das sich an dem Wissen orientiert, welches sie für die Bewältigung beruflicher Anforderungen im Unterricht benötigen. Die Diskussion um die Wirksamkeit der Lehramtsausbildung (vgl. Terhart, 2012) macht es notwendig, den Test auch bei berufstätigen Lehrkräften einzusetzen, um empirisch zu prüfen, welche praktische Bedeutung dem Ausbildungswissen für das Unterrichtshandeln zukommt (König & Pflanzl, 2016). Erste Befunde zeigen, dass er auch bei berufstätigen Lehrkräften einsetzbar ist (König et al., 2014; König, 2015; König & Pflanzl, 2016; Lauermann & König, 2016). Jedoch ist weiterführend zu prüfen, ob der Test speziell bei der Zielgruppe von Mathematiklehrkräften eingesetzt werden kann und ob im Sinne der Konstruktrepräsentation konstruktrelevante, kognitive Bearbeitungsprozesse erforderlich sind.

Bei der Konstruktrepräsentation handelt es sich um einen Validitätsaspekt, der von Embretson (1983) eingeführt wurde, um die Prüfung struktureller und korrelativer Validitätsaspekte im Sinne von Cronbach und Meehl (1955) zu ergänzen. Diese liefern für Validität zwar „circumstancial 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 und aufzuzeigen, dass das Konstrukt in Form von kognitiven Prozessen im Test selbst repräsentiert ist, 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 et al. (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. Zu diesen werden in der Regel hierarchisch gestufte Modelle aufgestellt. Stützt eine Validitätsanalyse die Annahmen zu kognitiven Bearbeitungsprozessen, erlauben es diese, kontinuierliche Kompetenzskalen in ein Kompetenzniveaumodell mit qualitativ beschreibbaren Abschnitten zu überführen (Hartig, 2007). So lassen sich Testwerte von Personen dahingehend interpretieren, zu welcher Art von Operationen diese jeweils in der Lage sind. Damit bieten Niveaumodelle trotz Informationsverlustes 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 kognitiver Bestandteil der professionellen Kompetenz von Lehrkräften (Baumert & Kunter, 2006). Es wird als notwendige Voraussetzung für die erfolgreiche Bewältigung berufsspezifischer Anforderungen erachtet, die sich in qualitätsvollem Unterricht und effektiver 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, 2014; Shulman, 1987). Bisherige Untersuchungen stützen die Annahme, dass es sich beim pädagogischen Wissen um ein mehrdimensionales Konstrukt mit inhaltlichen und kognitiven Anforderungsfacetten handelt (Baumert & Kunter, 2006; König & Blömeke, 2009; Voss, Kunina-Habenicht, Hoehne & Kunter, 2015).

Kognitive Prozesse beim Lösen unterrichtsspezifischer Aufgaben

Nur wenig ist darüber bekannt, welche kognitiven Prozesse bei Lehrenden zum Tragen kommen, wenn sie berufsspezifische Aufgaben meistern. Embretson (2016) stellt für den Bildungssektor fest, dass die Prüfung von Bearbeitungsprozessen in Leistungs- und Kompetenztests typischerweise unberücksichtigt bleibt. Obwohl kognitive Prozesse häufig in Leitlinien zur Testkonstruktion skizziert werden, werden diese selten im Sinne einer Theorie der Bearbeitungsprozesse hergeleitet und empirisch geprüft (Daniel & Embretson, 2010). Im deutschsprachigen Raum widmen sich einzelne Studien kognitiven Prozessen in Kompetenztests für Schülerinnen und Schüler (bspw. Beck & Klieme, 2007; Kauertz, Fischer, Mayer, Sumfleth & Walpuski, 2010; Prenzel, Häußler, Rost & Senkbeil, 2002). Studien zu Lehrkräften hingegen sind rar und fokussieren in der Regel nicht auf pädagogisches Wissen. So untersuchen Blömeke et al. (2008) die kognitiven Auseinandersetzungsformen von Mathematiklehrkräften, konzentrieren sich dabei jedoch auf Aufgaben des Fachwissenstests, der in der Studie MT21 (Mathematics Teaching in the 21st Century) eingesetzt wurde.

Eine Ausnahme, die sich pädagogischen Aufgaben von Lehrkräften widmet, bilden die Arbeiten von König (2009) und Klemenz und König (2019). Sie liefern erste Hinweise auf eine angemessene Konstruktrepräsentation des pädagogischen Wissens im TEDS-M-Test und ermöglichen so eine qualitative Beschreibung der Fähigkeiten von angehenden Lehrkräften. In ihrem Modell gehen sie davon aus, dass für die erfolgreiche Bewältigung unterrichtsbezogener, pädagogischer Aufgaben einerseits das Explizieren von Wissen und andererseits das Verbinden von Wissenselementen als kognitive Prozesse zentral sind.

Das Explizieren durch die Anwendung differenzierter, unterrichtsbezogener Sprache ist notwendig, um Handlungen zu analysieren und reflektiert zu beurteilen. Erst durch seine Explizitheit wird das Wissen analysier- und kommunizierbar gemacht (Altrichter & Posch, 2007). Das Verbinden mehrerer Wissenselemente zu komplexeren, bedeutungstragenden Einheiten (*chunking*) fördert das Einnehmen multipler Perspektiven und das schnelle Verarbeiten unterrichtsbezogener Situationen (Wipperfürth, 2015): Es ermöglicht eine Automatisierung, mit der Wissenselemente unter verringerter kognitiver Belastung abgerufen und verarbeitet werden können, sodass Gedächtnisleistung für zusätzliche Verarbeitungsprozesse verfügbar wird (Yinger, 1987).

Situationsspezifische, pädagogische Fähigkeiten

Die Fähigkeit zur Multiperspektivität und schnellen Verarbeitung ist insbesondere in informationsreichen Kontexten wie im Unterricht relevant, in denen Lehrkräfte unter Zeitdruck agieren und situationsspezifische Fähigkeiten anwenden müssen. Unter diesen werden die Fähigkeiten verstanden, Unterrichtssituationen professionell wahrzunehmen, zu interpretieren und unterrichtliche Entscheidungen zu treffen (Blömeke, Gustafsson & Shavelson, 2015; van Es & Sherin, 2002). Zum erfolgreichen Agieren in Schlüsselsituationen müssen Lehrkräfte ihr professionelles Wissen abrufen, um relevante Aspekte zu identifizieren und in einen ganzheitlichen Kontext zu bringen. Diese setzen sie auf Basis ihres Wissens zueinander in Bezug, um sie zu analysieren und interpretieren, Folgeaktivitäten zu antizipieren, Handlungsoptionen abzuwägen und Entscheidungen zu treffen (Kaiser, Busse, Hoth, König & Blömeke, 2015). Empirische Indizien für den relativ engen Zusammenhang zwischen dem Wissen von Lehrkräften und ihren situationsspezifischen Fähigkeiten konnten bereits in verschiedenen Untersuchungen im mathematikdidaktischen (Blömeke et al., 2014) und deutschdidaktischen (König, Bremerich-Vos, Buchholtz, Fladung & Glutsch, 2020), aber auch im pädagogischen Bereich erbracht werden (König et al., 2014; König & Kramer, 2016).

Zielsetzung und Annahmen

Ziel der vorliegenden Arbeit ist es, zu prüfen, ob die aus dem im Rahmen der TEDS-M-Studie entwickelten Test zum pädagogischen Wissen resultierenden Testwerte geeignet sind, um valide Aussagen über das Ausmaß und die Qualität des unterrichtsbezogenen, pädagogischen Wissens berufstätiger Mathematiklehrkräfte zu treffen. Hierzu werden folgende Annahmen untersucht:

1. Der TEDS-M-Test für pädagogisches Wissen ermöglicht eine trennscharfe und reliable Erfassung des pädagogischen Wissens von berufstätigen Mathematiklehrkräften.
2. Der Test erfordert konstruktrelevante, kognitive Prozesse, die berufstätige Mathematiklehrkräfte bei der Bearbeitung pädagogischer, unterrichtsbezogener Testaufgaben anwenden.
3. Basierend auf den zugrundeliegenden kognitiven Prozessen lässt sich ein Kompetenzniveaumodell ableiten, das eine qualitative Interpretation der pädagogischen Wissenstestwerte von berufstätigen Mathematiklehrkräften erlaubt.

4. Die auf diese Weise qualitativ interpretierten Unterschiede im pädagogischen Wissen der Lehrkräfte erklären Unterschiede in ihren situationsspezifischen, pädagogischen Fähigkeiten.

Untersuchungsdesign

Stichprobe und Datenerhebung

Die Untersuchung der Annahmen basiert auf Daten von TEDS-Unterricht, einer Studie, die im Kontext des BMBF-geförderten Projekts TEDS-Validierung durchgeführt wurde. Die Stichprobe umfasst 113 Mathematiklehrkräfte an Hamburger Schulen (49 % weiblich). Alle Teilnehmenden verfügten über Mathematikunterrichtserfahrung in den Jahrgangsstufen 7 bis 9. Über 92 % von ihnen wurden speziell für die Sekundarstufe I ausgebildet. Ihr durchschnittliches Alter lag zum Testzeitpunkt bei 39 Jahren (*SD* = 10.48; Altersspanne: 23–71 Jahre), ihre durchschnittliche Unterrichtserfahrung an Schulen bei 10 Jahren (*SD* = 9.87; Erfahrungsspanne: 0.5–40 Jahre).

Die Datenerhebung fand im Schuljahr 2015/16 statt. Die Lehrkräfte wurden in Veranstaltungen, über Schulleitungen und Fachleitungen an den Schulen sowie Mailsendungen rekrutiert und konnten online an der Befragung teilnehmen, die mit Hilfe der Umfragesoftware *Unipark* umgesetzt wurde. Sie konnten individuell über Zeitpunkt und Ort der etwa dreistündigen Testung entscheiden und erhielten nach Abschluss eine monetäre Aufwandsentschädigung. Die Testung umfasste fünf Instrumente zu mathematischem, mathematikdidaktischem und pädagogischem Wissen, die digitalisierte Paper-and-Pencil-Formate ebenso einschlossen wie Videovignettentests. Zwischen den einzelnen Testblöcken konnten die Teilnehmenden Unterbrechungen vornehmen. Jeder Testabschnitt war zeitlich begrenzt, 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 für pädagogisches Wissen

Um eine Überbeanspruchung der Lehrkräfte zu vermeiden, wurde für die Testung des pädagogischen Wissens eine Kurzfassung des im Rahmen der TEDS-M-Studie entwickelten Tests eingesetzt, dessen Bearbeitung auf 20 Minuten begrenzt wurde und im letzten Drittel der Erhebung stattfand. Die Kurzfassung wurde bereits in mehreren Studien eingesetzt (bspw. König, Blömeke & Kaiser, 2015) und setzt sich aus 15 Aufgaben zusammen, die im Rahmen der Studie TEDS-Follow-Up (TEDS-FU) so aus der Langfassung ausgewählt wurden, dass sie die Dimensionen sowie das Schwierigkeits- und Fähigkeitsspektrum des Originals möglichst ausgeglichen repräsentierten. Die Lösungen dieser 15 Aufgaben werden in insgesamt 47 Antwort- bzw. Teilantwortitems abgebildet, wobei zehn der Aufgaben bzw. 28 (Teil-)Antwortitems in offenem und fünf Aufgaben bzw. 19 (Teil-)Antwortitems in geschlossenem Format vorliegen. 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 (s. elektronisches Supplement 1). Zugleich repräsentieren die 15 Aufgaben unterschiedliche kognitive Anforderungen nach Anderson, Krathwohl, Airasian, Cruikshank, Mayer und Pintrich (2001): Sie erfordern Prozesse des Erinnerns, des Verstehens bzw. Analysierens von Sachverhalten, Konzepten und Zusammenhängen sowie des Kreierens von Handlungsoptionen (für ein Beispiel s. elektronisches Supplement 2).

Alle Aufgaben mit offenem Antwortformat wurden auf Basis einer umfassenden Anleitung von zwei geschulten studentischen Hilfskräften mit mehrjähriger Kodiererfahrung kodiert. Für 43 der 113 Fälle wurden die offenen Antworten von beiden Kodiererinnen unabhängig kodiert. Diese Doppelkodierungen unterstreichen mit einem Cohen's $\kappa > .79$ die hohe Interrater-Reliabilität des Tests.

Videovignettentests für situationsspezifische, pädagogische Fähigkeiten

Die situationsspezifischen, pädagogischen Fähigkeiten der Lehrkräfte wurden mit zwei videobasierten Tests erfasst: Der in der TEDS-FU-Studie entwickelte Test für situationsspezifische Fähigkeiten der professionellen Unterrichtswahrnehmung (Blömeke, Hoth et al., 2015; Kaiser et al., 2015) und der in der Studie Classroom Management Expertise (CME) entwickelte Test (König, 2015; König & Kramer, 2016) umfassen jeweils eine Reihe von zwei- bis fünfminütigen Videosequenzen, in denen pädagogische Schlüsselsituationen im Unterricht gezeigt werden. Jedes Video wird einmalig gezeigt und gefolgt von offenen und geschlossenen Testfragen.

Der Classroom Management Expertise-Test

Der CME-Test legt einen Fokus auf Klassenführung und zeigt in vier Videos Situationen, in denen Lehrkräfte Übergänge gestalten, eine effektive Lernzeitnutzung sicherstel-

len, das Klassenverhalten regeln oder Schülerinnen und Schülern Feedback geben. Insgesamt 24 Items (19 offen, 5 geschlossen) erfordern Prozesse der Wahrnehmung sowie der Interpretation der gezeigten Situationen. Offene Items wurden auf Basis eines Kodiermanuals von geschulten Hilfskräften kodiert (Cohen's $\kappa > .77$; für ein Beispiel, s. elektronisches Supplement 3). Die WLE-Skalenreliabilität (Weighted Maximum Likelihood Estimation) liegt bei WLE = .73.

Das TEDS-FU-Videoinstrument

Der im Rahmen der TEDS-FU-Studie entwickelte Test für situationsspezifische, pädagogische Fähigkeiten umfasst neben situationsspezifischer Wahrnehmung und Interpretation auch Entscheidungsaufgaben. Er umfasst Fragen im mathematikdidaktischen und pädagogischen Bereich in 22 geschlossenen sowie 18 offenen Items (für ein Beispiel, s. elektronisches Supplement 4). Im Folgenden wird nur die pädagogische Subskala P_PID (Pedagogy: Perception, Interpretation and Decision-Making) verwendet. Sowohl die Interrater-Reliabilität als auch die WLE-Skalenreliabilität (Weighted Maximum Likelihood Estimation) sind zufriedenstellend (Cohen's $\kappa > .76$; König et al., 2014; WLE = .70).

Methodisches Vorgehen

Zur Prüfung der Annahmen 1 und 2 wird eine IRT-Skalierung (Item-Response-Theory) der Daten vorgenommen, die Itemschwierigkeiten und Personenfähigkeitswerte auf einer gemeinsamen Skala abbildet. Dies ermöglicht es, von einem Itemschwierigkeitswert auf den Fähigkeitswert zu schließen, den eine Testperson mindestens erreichen sollte, um das Item mit hinreichender Wahrscheinlichkeit korrekt zu beantworten. Diese Eigenschaft der IRT-Skalierung kann genutzt werden, um von Bearbeitungsprozessen, die Items eines bestimmten Schwierigkeitsgrads erfordern, auf die Fähigkeiten von Personen zu schließen, deren Testwert dem Schwierigkeitsgrad entspricht.

Zur Prüfung der Konstruktrepräsentation wird das „item difficulty modeling" (Embretson, 2016, S. 8) angewendet, das sich in vergangenen Untersuchungen bewährt hat (vgl. Blömeke, Lehmann et al., 2010). Hierzu werden konstruktrelevante, kognitive Bearbeitungsprozesse definiert, die in Form von Itemmerkmalen operationalisiert werden, die solche Prozesse erfordern. Diese Operationalisierung berücksichtigt Ausprägungsgrade, von denen ein Einfluss auf die Itemschwierigkeit erwartet wird, und bildet so ein Kategoriensystem, mit dem Testitems hinsichtlich der Merkmale und Ausprägungen klassifiziert werden. Mittels IRT-Skalierung werden die Itemschwierigkeits- und die Personenfähigkeitswerte auf einer Skala abgebildet. Fallen die Reliabilitäts- und Item-Fit-Werte des Skalierungsmodells positiv aus, kann dies als Evidenz für Annahme 1 zur differenzierten Erfassung des pädagogischen Wissens der Lehrkräfte gewertet werden. Der Einfluss der Itemmerkmale auf die Schwierigkeiten wird in einem linearen Regressionsmodell geprüft. Lässt sich anhand der Merkmale mindestens ein Drittel der Gesamtstreuung der Itemschwierigkeiten aufklären und den Merkmalen eine moderate, praktische Bedeutsamkeit zuschreiben (Cohen, 1988), stützt dies Annahme 2 zur Konstruktrepräsentation.

Ferner erlaubt dies eine Einteilung der Itemschwierigkeits- und Fähigkeitsskala in Abschnitte, die anhand der kognitiven Prozesse beschreibbar werden. Verteilen sich die Lehrkräfte so auf die Niveaus pädagogischen Wissens, dass keine starken Boden- oder Deckeneffekte entstehen, kann das als Hinweis für die qualitative Interpretierbarkeit der Testwerte von berufstätigen Lehrkräften gewertet werden (Annahme 3). Diese Interpretation wird validiert, indem mittels Varianzanalysen geprüft wird, ob sich durch die Kompetenzniveaus der Lehrkräfte auch Unterschiede in den situationsspezifischen Fähigkeitswerten erklären lassen, wie sie im TEDS-FU_{P_PID}- und dem CME-Videoinstrument erfasst werden (Annahme 4).

Operationalisierung und Klassifizierung der Items

Zur Operationalisierung der kognitiven Prozesse wird auf das Modell von König (2009) und Klemenz und König (2019) zurückgegriffen. Sie operationalisieren den Prozess des differenzierten Explizierens von Wissen, indem sie auf eine Diskussion zur Verknüpfung von Lehrersprache und -wissen (Terhart, 1993) zurückgreifen und drei sprachliche Niveaus unterschieden: umgangssprachlich, fachsprachlich und wissenschaftssprachlich. Da die hier eingesetzte Kurzversion des TEDS-M-Tests diese Merkmalsausprägungen (im Gegensatz zu seiner Langversion) nur unzureichend abbildet, muss das Merkmal und der damit verbundene Prozess des differenzierten Explizierens jedoch aus den Analysen ausgeschlossen werden.

Stattdessen wird nur derjenige Teil des Modells angewendet und weiterentwickelt, der den Prozess des Verbindens mehrerer Wissenselemente untersucht. Dieser wird operationalisiert, indem die Komplexität der zur Aufgabenbearbeitung erforderlichen Prozesse erfasst wird. Bezugnehmend auf die Konzeptualisierung nach Eye (1999, S. 8) wird unter kognitiver Komplexität die Zahl der Wissenselemente verstanden, die für die „Strukturierung und Beurteilung eines Objektes oder Sachverhalts" notwendig sind. Werden also für die vollständig korrekte Lösung der

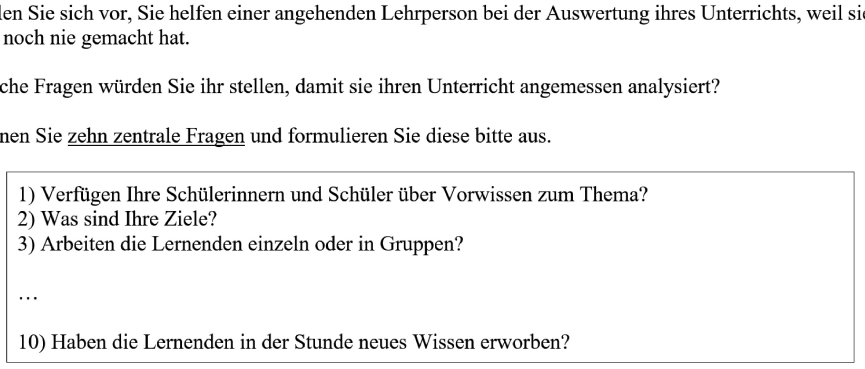

Abbildung 1. Offene Aufgabe mit beispielhaften Antworten zur Erfassung des pädagogischen Wissens zu den Anforderungen „Strukturierung von Unterricht" und „Kreieren" (König & Blömeke, 2009).

Aufgabe mehrere Wissenselemente benötigt und daher in mehreren Teilantworten abgefragt, gilt die Komplexität der erforderlichen Prozesse als hoch. Muss nur ein einzelnes Wissenselement abgerufen werden, gilt sie als niedrig. Wie viele Wissenselemente den Erwartungshorizont der Testaufgaben widerspiegeln, wurde vorweg in der Kodieranleitung festgelegt (König & Blömeke, 2010).

Um die kognitive Komplexität abzubilden, wurden Teilantwortitems, die zum selben offenen Aufgabenstamm gehören und daher zusammenhängende Lösungsaspekte repräsentieren, zu einem Antwortitem zusammengefasst, indem sie zu einem Partial-Credit-Score aufsummiert wurden. Abbildung 1 zeigt die Aufgabe PK39 (Pedagogical Knowledge) zur Unterrichtsanalyse, deren Bearbeitungsprozesse als hoch komplex klassifiziert wurden.

Die Aufgabe wurde nach dem folgenden Schema kodiert (vgl. elektronisches Supplement 5). Die 10 Fragen, die die Testperson zur Aufgabenlösung entwickelt hat, wurden dahingehend kodiert, ob sie bestimmte Inhalte ansprachen (bspw. methodisch-didaktische Prinzipien, kognitive Aktivierung). Insgesamt wurden 12 Kodierungskategorien gebildet (König & Blömeke, 2010, S. 27f.), die vier unterschiedliche Wertungskategorien für Teilantwortitems (PK39S01–04) abbilden, um inhaltlich nahe Antworten nicht doppelt zu werten. Diese Teilantwortitems wurden zu einem Partial-Credit-Antwortitem aufsummiert (PK39S). Formulierte eine Testperson also Fragen zum Kontext, Input, Prozess und Output der Unterrichtsstunde, erfüllte sie vier von vier Wertungskategorien und erhielt für das Partial-Credit-Antwortitem PK39S den Score „4". Formulierte eine Person ausschließlich Fragen zum Kontext und Prozess, erfüllte sie nur zwei Wertungskategorien und erhielt den Score „2".

Nach Klemenz und König (2019) gilt die Komplexität der Bearbeitungsprozesse als hoch, wenn ein Konzept in seiner Gänze angewendet wurde, also alle erforderten Wissenselemente verknüpft wurden. Für das Beispiel aus Abbildung 1 bedeutet dies, dass die Person mit PK39S = „4" das Konzept der strukturierten Unterrichtsanalyse vollständig angewendet und somit das hoch komplexe Antwortitem korrekt beantwortet hat. Bei einer Person, die nur drei Aspekte nennt, würde die Aufgabe als nicht gelöst gelten und ihr Wissen als nicht umfassend verknüpft interpretiert. Dies würde jedoch eine Gleichsetzung mit einer Person bedeuten, die keinen Lösungsaspekt genannt hat, und zu einer Unterschätzung ihrer Fähigkeiten führen.

Um nicht nur dichotom zwischen „hoch komplexen" und „nicht hoch komplexen" Fähigkeiten zu unterscheiden, sondern auch partielle Verknüpfungsprozesse von Personen mit Teilscores abzubilden, wurde das Modell von Klemenz und König (2019) ausdifferenziert. Jeder Stufe der Partial-Credit-Antwortitems wurde ein eigenes Komplexitätsniveau zugewiesen, sodass einzelne Verknüpfungen, die sich aus zwei oder mehr Wissenselementen zusammensetzten, jedoch noch kein vollständiges Konzept darstellten, einer mittleren Komplexität zugeordnet wurden. Die erste Stufe erforderte das Nennen eines einzelnen Wissenselements und wurde als niedrig komplex kategorisiert. Die höchstmögliche Stufe, die das Verknüpfen zu einem vollständigen Konzept erforderte, galt weiterhin als hoch komplex.

Auf (Teil-)Antwortitems zu geschlossenen Aufgaben wurde diese Partial-Credit-Modellierung nicht angewendet, da die Komplexität der für sie erforderlichen Prozesse als niedrig interpretiert wurde. Ihre Lösungen konnten zwar mehrere Wissenselemente erfordern, jedoch musste diese Multiperspektivität nicht von den Testteilnehmenden selbst entwickelt werden, sondern wurde durch die Distraktoren aktiviert. Sie gingen als einstufige Antwortitems in die weiteren Analysen ein, ebenso wie Antwortitems aus offenen Aufgaben, deren Lösung nur ein einzelnes Wissenselement erfordern. Insgesamt ergab die Modellierung eine Summe von 32 Antwortitems, von denen 21 dichotom (niedrig komplex) und 11 mehrstufig (Partial-Credit) waren, also Teilprozesse niedriger, mittlerer und hoher Komplexität einschlossen.

Skalierung und Bestimmung der Itemschwierigkeiten

Zur Bestimmung der Schwierigkeiten der dichotomen und Partial-Credit-Antwortitems wurden die Daten mittels des in ConQuest 2.0 (Wu, Adams & Wilson, 2007) implementierten einparametrischen Partial-Credit-Modells skaliert, um ein eindimensionales Modell zu schätzen. Das Modell zeigt für die Parameterschätzungen nach den Methoden der Weighted Maximum Likelihood Estimation (WLE) und der Expected A-Posteriori Estimation basierend auf Plausible Values (EAP / PV) akzeptable Reliabilitätswerte (WLE = .68; EAP / PV = .74). Die Items weisen mit einer Ausnahme Weighted-MNSQ-Werte (Weighted Mean Square) zwischen .80 und 1.20 auf und können somit weitgehend als zufriedenstellend eingestuft werden (s. elektronisches Supplement 6). Auch die Trennschärfe ist mit Werten > .20 bei den meisten Items akzeptabel. 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 Spannweite der Itemschwierigkeitswerte umfasst mehr als 5 Logits (-2.80 bis 2.94), während ihr Mittelwert im Rahmen der Raschskalierung auf null fixiert wurde, um eine Identifizierbarkeit des Modells zu gewährleisten. Die mittlere Personenfähigkeit liegt bei 1.25 (SD = 0.75) und somit über der mittleren Aufgabenschwierigkeit.

Jedes der 11 Partial-Credit-Antwortitems deckt mehrere Komplexitätsstufen ab, sodass für jede Stufe ein eigener Schwierigkeitsparameter geschätzt werden muss. Mit der Statistiksoftware R 3.5.0 (R Core Team, 2018) und dem R-Paket WrightMap 1.2.1 (Torres Irribarra & Freund, 2016) wurden hierzu 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).

Die Kategorieschwierigkeiten ermöglichten es, jeder Stufe eines Antwortitems eine eigene Schwierigkeit und ein eigenes Komplexitätsniveau zuzuordnen. Hierzu wurden aus jedem Antwortitem so viele virtuelle Fälle des Items abgeleitet wie das Item Stufen beinhaltete. Aus dem vierstufigen Antwortitem PK39S wurden so vier virtuelle (Antwortitem-)Fälle, denen jeweils ein Komplexitätsniveau zugeordnet wurde. Die Schwierigkeiten dieser virtuellen (Antwortitem-)Fälle entsprachen den Kategorieschwierigkeiten der jeweiligen Stufe des Partial-Credit-Antwortitems. So entstand etwa aus der Kategorie „4“ von PK39S der virtuelle (Antwortitem-)Fall PK39S_04 mit der Schwierigkeit 2.61 und dem Komplexitätsniveau „hoch“, der von 20 % der Lehrkräfte korrekt gelöst wurde (vgl. Tabelle 1; für ein Auswertungsbeispiel s. elektronisches Supplement 7).

Diese Umkodierung ergab für die Folgeanalyse eine Basis von 57 virtuellen (Antwortitem-)Fällen, von denen 32 Bearbeitungsprozesse niedriger Komplexität, 14 solche mittlerer und 11 solche hoher Komplexität erforderten. Einen zusammenfassenden Überblick des Kodierungsverlaufs gibt das elektronische Supplement 8.

Vorhersage der Itemschwierigkeiten und Bildung eines Kompetenzniveaumodells

Auf Basis der virtuellen (Antwortitem-)Fälle wurde untersucht, ob sich die Komplexität der erforderlichen Bearbeitungsprozesse zur Vorhersage der Item- bzw. Kategorieschwierigkeiten eignet. Hierzu wurde in IBM SPSS Statistics 25 eine multiple Regressionsanalyse durchgeführt, in der die Komplexitätsniveaus jeweils als Dummy-Variablen einbezogen wurden. Das Ergebnis (s. Tabelle 2) zeigt erwartungskonform, dass die Komplexität die Schwierigkeit beeinflusst ($F_{2, 54} = 26.72$, $p < .001$). Die aufsteigenden Regressionskoeffizienten bestätigen die angenommene Rangfolge der Komplexitätsausprägungen und ihre Effekte erweisen sich als praktisch bedeutsam. Mit einem *adj.* $R^2 = .48$ kann das Modell etwa die Hälfte der Gesamtstreuung der Kategorieschwierigkeiten durch die Komplexität der Bearbeitungsprozesse erklären, was nach Cohen (1988) einer hohen Anpassungsgüte entspricht.

Um die Schwierigkeitswerte der Antwortitems qualitativ zu beschreiben, wurden die Regressionskoeffizienten zu Kompetenzniveaus aufsummiert (vgl. Hartig, 2007). Die Regressionskonstante stellt die Schwierigkeit eines Antwortitems mit der niedrigsten Merkmalsausprägung dar. Die Schwelle für Niveau I, auf dem Antwortitems das Abrufen eines einzelnen Wissenselements erfordern, liegt somit bei -0.463 Logits. Durch die Addition von Konstante und den jeweiligen Regressionskoeffizienten ergeben sich die Schwellenwerte für Niveaus II und III (s. Tabelle 3), auf denen Antwortitems verortet sind, die partielle Verknüpfungen (mittlere Komplexität) oder Verknüpfungen zu einem vollständigen Konzept (hohe Komplexität) erfordern. Dank der IRT-Skalierung ist die qualitative Beschreibung der Itemschwierigkeiten auf die Fähigkeitswerte der Testpersonen übertragbar. Beispielhaft können Personen mit einem Testwert von 0.491 Logits oder höher somit als Personen angesehen werden, die Items mit hinreichender Wahrscheinlichkeit lösen, für die Wissenselemente partiell verknüpft werden müssen. Eine Zuordnung der Testpersonen zu den Niveaus zeigt, dass sich der Großteil der Lehrkräfte (84 %) auf Niveau II befindet und lediglich 4 % Niveau III erreichen.

Tabelle 1. Lösungshäufigkeiten und Thurstonian Thresholds der Partial-Credit Antwortitems sowie Komplexitätsniveaus der virtuellen (Antwortitem-)Fälle

Partial-Credit Antwortitem	Kategorie	Lösungshäufigkeit (gültige %)	Thurstonian Threshold	virtueller (Antwortitem-)Fall	Komplexität
PK07S	1	10	1.03	PK07_01	niedrig
	2	19	1.26	PK07_02	mittel
	3	32	1.72	PK07_03	hoch
PK15S	1	46	-0.95	PK15S_01	niedrig
	2	44	1.39	PK15S_02	mittel
	3	2	4.53	PK15S_03	hoch
PK22AS	1	0	-0.37	PK22 A_01	niedrig
	2	2	-0.37	PK22 A_02	mittel
	3	16	-0.26	PK22 A_03	mittel
	4	79	0.16	PK22 A_04	hoch
PK32S1	1	40	0.18	PK32S1_01	niedrig
	2	29	1.64	PK32S1_02	mittel
	3	8	3.09	PK32S1_03	hoch
PK32S2	1	27	0.33	PK32S2_01	niedrig
	2	35	1.22	PK32S2_02	mittel
	3	16	2.59	PK32S2_03	hoch
PK32S3	1	22	-0.19	PK32S3_01	niedrig
	2	38	0.77	PK32S3_02	mittel
	3	28	2.01	PK32S3_03	hoch
PK32S4	1	24	-0.44	PK32S4_01	niedrig
	2	45	0.71	PK32S4_02	mittel
	3	21	2.39	PK32S4_03	hoch
PK33S	1	2	-0.46	PK33_01	niedrig
	2	5	-0.23	PK33_02	mittel
	3	42	-0.00	PK33_03	mittel
	4	47	1.36	PK33_04	hoch
PK39S*	1	1	-0.85	PK39_01	niedrig
	2	14	-0.69	PK39_02	mittel
	3	62	-0.01	PK39_03	mittel
	4	20	2.61	PK39_04	hoch
PK40AS	1	55	-2.93	PK40 A_01	niedrig
	2	39	1.50	PK40 A_02	mittel
	3	5	3.71	PK40 A_03	hoch
PK67S	1	13	-0.99	PK67_01	niedrig
	2	50	-0.07	PK67_02	mittel
	3	33	1.83	PK67_03	hoch

Anmerkungen: Bei dem mit * markierten Partial-Credit Antwortitem handelt es sich um die in Abbildung 1 dargestellte Beispielaufgabe.

Tabelle 2. Nicht standardisierte Regressionsgewichte als Prädiktoren der Schwierigkeitsparameter

	Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
Adj. R^2 = 0.48	Regressionskoeffizient	Standardfehler	β	*T*	*p*
(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 3. Schwellenwerte, Kurzbeschreibungen der Kompetenzniveaus und Häufigkeitsverteilung der Testteilnehmenden

Kompetenz-niveau	Logit	Kurzbeschreibung	Häufigkeit	%
< I	< -0.463	Bearbeitungsprozesse von niedriger Komplexität können nicht mit hinreichender Wahrscheinlichkeit gelöst werden	3	3
I	-0.463	Bearbeitungsprozesse von niedriger Komplexität können mit hinreichender Wahrscheinlichkeit gelöst werden	10	9
II	0.491	Bearbeitungsprozesse von mittlerer Komplexität können mit hinreichender Wahrscheinlichkeit gelöst werden	95	84
III	2.363	Bearbeitungsprozesse von hoher Komplexität können mit hinreichender Wahrscheinlichkeit gelöst werden	5	4

Tabelle 4. Mittelwerte, Standardabweichungen und Standardfehler der WLE-Scores (Weighted Maximum Likelihood Estimation) zu den situationsspezifischen, videobasierten Tests TEDS-FU_{P_PID} (Teacher Education and Development Study – Follow Up; pädagogische Subskala „Pedagogy: Perception, Interpretation and Decision-Making") und CME (Classroom Management Expertise), nach Kompetenzniveau

Kompetenzniveau		TEDS-FU_{P_PID}			CME		
	N	*M*	*SD*	*SE*	*M*	*SD*	*SE*
< I	3	-0.18^{3}	0.29	0.17	-1.28^{III}	0.87	0.51
I	10	$-0.46^{II, III}$	0.76	0.24	$-0.69^{III, 2}$	0.92	0.29
II	95	0.50^{I}	0.58	0.06	0.40^{1}	0.94	0.10
III	5	$1.07^{I, < 1}$	0.74	0.33	$0.24^{< I, I}$	0.49	0.22

Anmerkungen: Die Indices $^{<I, I, II, III}$ 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 ($p < .10$).

Validierung der qualitativen Testwertinterpretation

Um die Validität der qualitativen Interpretation der Testwerte sicherzustellen, wird abschließend geprüft, inwiefern das Erreichen bestimmter Kompetenzniveaus auch situationsspezifische, pädagogische Fähigkeitswerte erklären kann. Es kann angenommen werden, dass sich die Fähigkeit im Bereich des pädagogischen Wissens kognitiv komplexere Prozesse durchzuführen auch in höheren situationsspezifischen Fähigkeiten niederschlägt.

Zur Prüfung dieser Annahme wurden zwei einfaktorielle Varianzanalysen durchgeführt, die das Kompetenzniveau als Faktor und die WLE-Scores der Testpersonen im TEDS-FU_{P_PID}- bzw. CME-Videotest als abhängige Variable berücksichtigen. Die Ergebnisse (s. Tabelle 4) zeigen für beide Tests, dass sich die Kompetenzniveaugruppen signifikant in den Mittelwerten der situationsspezifischen Tests unterscheiden und es sich nach Cohen (1988) um Effekte mittlerer bzw. geringer Stärke handelt (TEDS-FU_{P_PID}: $F_{3, 109} = 9.17$, $p < .001$, $\eta^2 = .20$; CME: $F_{3, 109} = 4.98$, $p = .003$, $\eta^2 = .12$).[1]

Post-hoc-Tests mit Bonferroni-Korrektur unterstreichen für beide Tests, dass Lehrpersonen, die in der Lage sind, Items mittlerer und hoher Komplexität zu lösen, über höhere situationsspezifische Fähigkeiten verfügen als Lehrpersonen, die nur niedrig komplexe Items lösen können. Die Mittelwerte der Personen auf Kompetenzniveau I fallen in beiden Tests signifikant niedriger aus als die der Personen auf Niveau II (TEDS-FU_{P_PID}: $p < .001$; CME: $p = .05$) bzw. der Personen auf Niveau III (TEDS-FU_{P_PID}: $p < .001$; CME: $p = .02$). Auch zwischen dem untersten (< I) und dem höchsten Niveau (III) zeigen sich signifikante Unterschiede bzw. statistische Trends zu Mittelwertunterschieden (TEDS-FU_{P_PID}: $p = .08$; CME: $p = .04$). Die Unterschiede zwischen den anderen Kompetenzniveaus fallen nicht signifikant, jedoch in ihrer Richtung erwartungskonform aus. Einzig beim CME-Test zeigt sich ein erwartungswidriges Ergebnis: Hier weisen Personen, die Niveau III zugeordnet wurden, niedrigere

[1] Aufgrund der niedrigen Gruppengrößen wurden zudem zwei Kruskal-Wallis-Tests durchgeführt, die die Ergebnisse der parametrischen Verfahren bestätigen (TEDS-FU_{P_PID}: $\chi^2(3) = 19.95$, $p < .001$; CME: $\chi^2(3) = 11.55$, $p = < .01$).

Testwerte auf als die, die Niveau II zugeordnet wurden. Insgesamt stützen die Ergebnisse der Varianzanalysen die Validität der qualitativen Interpretation der Fähigkeitswerte hinsichtlich der angenommenen kognitiven Prozesse.

Diskussion

Ziel dieses Beitrags war es, zu prüfen, ob sich der im Rahmen der TEDS-M-Studie entwickelte Test für pädagogisches Wissen eignet, um das pädagogische Wissen von berufstätigen Mathematiklehrkräften quantitativ und qualitativ im Hinblick auf zugrunde liegende kognitive Prozesse zu interpretieren. Es wurde zunächst der Frage nachgegangen, ob der Test das Wissen reliabel und differenziert erfasst. Ferner wurde untersucht, ob der Test kognitive Prozesse erfordert, die Mathematiklehrkräfte laut König (2009) und Klemenz und König (2019) für die erfolgreiche Bewältigung unterrichtsbezogener, pädagogischer Aufgaben bei der Bearbeitung pädagogischer, unterrichtsbezogener Testaufgaben anwenden – und somit eine entsprechende qualitative Testwertinterpretation erlaubt.

Zunächst ist festzuhalten, dass sich der Test auch in dieser spezifischen Gruppe von Lehrpersonen als geeignet erweist, pädagogisches Wissen zu erfassen: Die Ergebnisse der IRT-Skalierung weisen zufriedenstellende Reliabilitäts- und Item-Fit-Werte auf. Zudem konnten Validitätshinweise 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 Itemschwierigkeiten lässt sich bedeutsam durch die Komplexität der erforderlichen, kognitiven Bearbeitungsprozesse aufklären. Dies stützt die Annahme, dass während der Testbearbeitung Wissenselemente abgerufen werden, die durch die im Kodiermanual definierten Wertungskategorien so interdependent konzipiert sind, dass ihre Nennungen als Verknüpfungen interpretiert werden können. Die Ergebnisse können daher als erstes Indiz für die Konstruktrepräsentation interpretiert werden.

Die kognitive Leistung, einzelne Wissenselemente partiell zu verknüpfen, konnte in unserer Stichprobe für den Großteil der Lehrkräfte belegt werden. Die Leistung, Wissenselemente zu einem ganzheitlichen Konzept zusammenzuführen, ließ sich nur für einen kleinen Teil der Gruppe zeigen (vgl. Tabelle 3). Da wir angesichts eines relativ niedrigen Mittelwerts der zweiten Staatsexamensnoten (M = 1.97) von einer tendenziell positiv verzerrten Stichprobe ausgehen, werten wir das Fehlen starker Deckeneffekte als Hinweis, dass das aufgestellte Niveaumodell für berufstätige Lehrkräfte gut erreichbar, aber nicht zu einfach gehalten ist. Denkbar ist auch, dass das Lösen hoch komplexer Aufgaben besonderes Durchhaltevermögen der Testpersonen voraussetzt. Da die Testergebnisse keine Folgen für die Lehrkräfte hatten, kann eine Einschränkung bei etwa der Teilnahme- oder Durchführpersistenz nicht vollständig ausgeschlossen werden. Eine solche Interpretation gälte es jedoch in zukünftigen Studien zu prüfen, beispielsweise indem zusätzlich die Leistungsmotivation der Testteilnehmenden erhoben würde. Alternativ könnte auch eine fallanalytische Untersuchung der Lehrkräfte mit den höchsten (Niveau III) und den niedrigsten Leistungen (Niveau < I) lohnenswert sein, um Hinweise darauf zu generieren, was sie von der Mehrheit der getesteten Lehrkräfte unterscheidet.

Methodisch betrachtet konnte die Operationalisierung kognitiver Komplexität so weiterentwickelt werden, dass im Gegensatz zum bisherigen Ansatz bei König (2009) und Klemenz und König (2019) keine Dichotomisierung der komplexen Antwortitems mehr notwendig war. Somit konnten zwei wesentliche Nachteile der Dichotomisierung behoben werden: Zum einen wurde durch die Anwendung des Partial-Credit-Modells bei der Skalierung und Ermittlung der Itemschwierigkeiten ein Informationsverlust vermieden. Zum anderen ermöglichte es die anschließende Verwendung virtueller Fälle der Partial-Credit-Antwortitems, kognitive Komplexität in mehr als zwei Niveaus auf der Itemschwierigkeitsskala abzubilden.

Die IRT-Skalierung erlaubte es, von dem Fähigkeitswert einer Lehrperson auf die unterschiedlichen kognitiven Bearbeitungsprozesse zu schließen, die sie mit hinreichender Wahrscheinlichkeit durchführen konnte. Hinweise für die Validität dieser qualitativen Testwertinterpretation lieferten Varianzanalysen zwischen Kompetenzniveau und situationsspezifischen Fähigkeiten, die zeigten, dass situationsspezifische Fähigkeiten in pädagogischen Bereichen durch das Kompetenzniveau im pädagogischen Wissen erklärt werden konnten. Aus den Post-hoc-Tests wurde auch ersichtlich, dass Personen, die in der Lage waren Items niedriger Komplexität zu lösen, sich signifikant von solchen unterschieden, die Items mittlerer Komplexität lösen konnten. Es scheint also empfehlenswert, Prozesse mittlerer kognitiver Komplexität, wie das partielle Verknüpfen mehrerer Wissenselemente, nicht mit Prozessen niedriger Komplexität, wie dem Anwenden einzelner Wissenselemente, gleichzusetzen, wie es im Rahmen der Dichotomisierung in bisherigen Arbeiten umgesetzt wurde. Stattdessen legen die Ergebnisse nahe, beide Teilprozesse als eigenständige Niveaus 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 Stichprobengröße als eine Limitierung unserer Studie zu benennen. In zukünftigen Analysen sollte das weiterentwickelte Modell auf größere Gruppen angewendet werden, um weitere Hinweise für die Validität der qualitativen Testwertinterpretation zu erbringen. 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 Kompetenzniveaumodells 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-Tests bei berufstätigen Lehrkräften – auch anderer Unterrichtsfächer – einzusetzen, um auch diese anhand des hier dargestellten Vorgehens hinsichtlich der Konstruktrepräsentation zu prüfen.

Eine breitere Testpersonen- und Item-Basis böte zudem die Möglichkeit, weitere kognitive Teilprozesse 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 Wissenselemente auch die Anzahl der unterschiedlich gearteten, einander übergeordneten Operationen berücksichtigt würden, die zur Bearbeitung konsekutiv durchlaufen werden müssen. Schließlich wäre es interessant zu prüfen, inwiefern sich kognitive Komplexität auch eignet, um die Konstruktrepräsentation in anderen pädagogischen Kompetenztests zu stützen und möglicherweise als Basis für zukünftige Test- und Itementwicklungen zu dienen.

Elektronische Supplemente (ESM)

Die elektronischen Supplemente sind mit der Online-Version dieses Artikels verfügbar unter https://doi.org/10.1026/0012-1924/a000250

ESM 1. Design-Matrix zum Test des pädagogischen Wissens aus der Studie TEDS-M (Teacher Education and Development Study: Learning to Teach Mathematics; König & Blömeke, 2009), mit Angaben zur Anzahl von (Teil-)Antwortitems je Subdimension in der Test-Kurzversion

ESM 2. Offene Aufgabe mit beispielhaften Antworten zur Erfassung des pädagogischen Wissens zu den Anforderungen „Strukturierung von Unterricht“ und „Erinnern“ (a) bzw. „Verstehen / Analysieren“ (b) (Blömeke, Kaiser & Lehmann, 2010)

ESM 3. Beispielitems des Classroom Management Expertise-Tests zur Erfassung der Dimensionen „Genauigkeit der Wahrnehmung“ (a), „holistische Wahrnehmung“ (b) und „Interpretieren“ (c) (König, 2016)

ESM 4. Beispielitems aus dem Videoinstrument der Studie Teacher Education and Development Study – Follow Up, mit der Anforderung des Entscheidens (Blömeke et al., 2014; König, 2016)

ESM 5. Kodierungsverlauf bei offenen Items am Beispiel von Aufgabe PK39 (König & Blömeke, 2010)

ESM 6. Itemschwierigkeitsparameter der eindimensionalen Skalierung sowie Angaben zum Weighted Item Fit (Mean Square, Konfidenzintervall, T-Wert, Trennschärfe, erforderte Wissenselemente)

ESM 7. Kodierungsverlauf bei Aufgabe PK39 anhand beispielhafter Antworten (König & Blömeke, 2010)

ESM 8. Anzahl der Aufgaben, Wertungskategorien, Antwortitems und virtuellen (Antwortitem-)Fälle nach Komplexitätsausprägung in der Kurzfassung des Tests für pädagogisches Wissen aus der Studie TEDS-M (Teacher Education and Development Study: Learning to Teach Mathematics)

Literatur

Altrichter, H. & Posch, P. (2007). *Lehrerinnen und Lehrer erforschen ihren Unterricht. Unterrichtsentwicklung und Unterrichtsevaluation durch Aktionsforschung* (4. Aufl.). Bad Heilbrunn: Julius Klinkhardt.

Anderson, L., Krathwohl, D. R. Airasian, P. W., Cruikshank, K., A., Mayer, R. E. & Pintrich, P. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Baumert, J. & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften. *Zeitschrift für Erziehungswissenschaft*, *9*, 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.

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*, 287 – 308. https://doi.org/10.1007/s10763-015-9619-4

Blömeke, S., Kaiser, G. & Lehmann, R. (Hrsg.). (2010). *TEDS-M 2008. Professionelle Kompetenz und Lerngelegenheiten angehender Mathematiklehrkräfte für die Sekundarstufe I im internationalen Vergleich.* Münster: Waxmann.

Blömeke, S., König, J., Busse, A., Suhl, U., Benthien, J., Döhrmann, M. et al. (2014). Von der Lehrerausbildung in den Beruf – Fachbezogenes Wissen als Voraussetzung für Wahrnehmung, Interpretation und Handeln im Unterricht. *Zeitschrift für Erziehungswissenschaft*, *17*, 509 – 542. https://doi.org/10.1007/s11618-014-0564-8

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* (S. 105 – 134). Erste Ergebnisse zur Wirksamkeit der Lehrerausbildung. Münster: 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. & Heerden, J. van (2004). The concept of validity. *Psychological Review*, *111*, 1061 – 1071. https://doi.org/10.1037/0033-295X.111.4.1061

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Commons, M. L., Goodheart, E. A., Pekker, A., Dawson, T. L., Draney, K. & Adams, K. M. (2008). Using Rasch scaled stage scores to validate orders of hierarchical complexity of balance beam task sequences. *Journal of Applied Measurement*, *9*, 182 – 199.

Cronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281 – 302.

Daniel, R. C. & Embretson, S. (2010). Designing cognitive complexity in mathematical problem-solving items. *Applied Psychological Measurement*, *34*, 348 – 364. https://doi.org/10.1177/0146621609349801

Embretson, S. (1983). Construct validity. Contruct representation versus the nomothetic span. *Psychological Bulletin*, *93*, 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

Es, E. A. van & Sherin, M. G. (2002). Learning to notice: scaffolding new teachers' interpretations of classroom interactions. *Journal of Technology and Teacher*, *10*, 571 – 596.

Eye, A. von (1999). Kognitive Komplexität – Messung und Validität. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *20* (2), 81 – 96.

Gindele, V. & Voss, T. (2017). Pädagogisch-psychologisches Wissen. Zusammenhänge mit Indikatoren des beruflichen Erfolgs angehender Lehrkräfte. *Zeitschrift für Bildungsforschung*, *108*, 255 – 272. 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) (S. 83 – 99). Weinheim: Beltz.

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau*, *63*, 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. 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*, 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.

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), 355 – 377.

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*, 244 – 262. https://www.pedocs.de/volltexte/2018/14703/pdf/LbP_2010_2_Koenig_zur_Bildung_von_Kompetenzniveaus.pdf

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*, 999 – 1028. https://doi.org/10.1007/s10763-013-9420-1

König, J. (2014). *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.

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. (2016). Professionelle Kompetenz von Lehrkräften. Videobasierte Messung situationsspezifischer Fähigkeiten. In S. Blömeke, M. Caruso, S. Reh, U. Salaschek & J. Stiller (Hrsg.), *Traditionen und Zukünfte. Beiträge zum 24. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft* (Schriften der Deutschen Gesellschaft für Erziehungswissenschaft (DGfE), S. 215 – 228). Leverkusen-Opladen: Barbara Budrich.

König, J. & Blömeke, S. (2009). Pädagogisches Wissen von angehenden Lehrkräften. Erfassung und Struktur von Ergebnissen der fachübergreifenden Lehrerausbildung. *Zeitschrift für Erziehungswissenschaft*, *12*, 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*, 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*, 188 – 201. https://doi.org/10.1177/0022487110388664

König, J., Bremerich-Vos, A., Buchholtz, C., Fladung, I. & Glutsch, N. (2020). Pre-service teachers' generic and subject-specific lesson-planning skills: On learning adaptive teaching during initial teacher education. *European Journal of Teacher Education, 43* (2), 131 – 150. https://doi.org/10.1080/02619768.2019.1679115

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*, 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 Mathematics Education*, *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. (2012). *Lehramtsstudierende erwerben pädagogisches Professionswissen. Ergebnisse der Längsschnittstudie LEK zur Wirksamkeit der erziehungswissenschaftlichen Lehrerausbildung.* Münster: Waxmann.

Lauermann, 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.

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.

R Core Team. (2018). R. *A language and environment for statistical computing* (Version 3.5.0) [Computer software]. Wien: R Foundation for Statistical Computing. http://www.R-project.org

Rost, J. (2004). Psychometrische Modelle zur Überprüfung von Bildungsstandards anhand von Kompetenzmodellen. *Zeitschrift für Pädagogik*, *50*, 662 – 678.

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

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 Irribarra, D. & Freund, R. (2016). *Wright Map. IRT item-person map with ConQuest integration* (Version 1.2.1) [Computer software]. http://github.com/david-ti/wrightmap

Voss, T., Kunina-Habenicht, O., Hoehne, V. & Kunter, M. (2015). Stichwort Pädagogisches Wissen von Lehrkräften. Empirische Zugänge und Befunde. *Zeitschrift für Erziehungswissenschaft*, *18*, 187 – 223. https://doi.org/10.1007/s11618-015-0626-6

Wipperfürth, M. (2015). *Professional vision in Lehrernetzwerken. Berufssprache als ein Weg und ein Ziel von Lehrerprofessionalisierung* (Münchener Arbeiten zur Fremdsprachen-Forschung, Bd. 32). Münster: Waxmann.

Wu, M. L., Adams, R. & Wilson, M. (2007). *ACER ConQuest version 2.0: generalised item response modelling software.* Camberwell: Australian Council for Educational Research.

Wu, M. L., Tam, H. P. & Jen, T.-H. (2016). *Educational measurement for applied researchers.* Singapore: Springer. https://doi.org/10.1007/978-981-10-3302-5

Yinger, R. (1987). Learning the language of practice. *Curriculum Inquiry*, *17*, 293 – 318.

ORCID
Caroline Felske
https://orcid.org/0000-0003-2966-7514
Johannes König
https://orcid.org/0000-0003-3374-9408

Caroline Felske
Prof. Dr. Johannes König
Stefan Klemenz
Humanwissenschaftliche Fakultät
Department Erziehungs- und Sozialwissenschaften
Empirische Schulforschung, Quantitative Methoden
Universität zu Köln,
Gronewaldstraße 2a
50931 Köln
caroline.felske@posteo.de

Prof. Dr. Gabriele Kaiser
Natalie Ross
Fakultät für Erziehungswissenschaft
Didaktik der gesellschaftswissenschaftlichen und
mathematisch-naturwissenschaftlichen Fächer (EW 5)
Universität Hamburg
Von-Melle-Park 8
20146 Hamburg

Prof. Dr. Gabriele Kaiser
Institute for Learning Sciences & Teacher Educaion
Australian Catholic University
Level 4/229 Elizabeth Street
Brisbane Qld, 4000
Australien

Prof. Dr. Sigrid Blömeke
Centre for Educational Measurement
University of Oslo
Gaustadalléen 21
0373 Oslo
Norwegen