

Mit Kollaboration zum Standard: Empirisch basierte Bestimmung von Lehrkräfte-Expertise im Bereich Deutsch als Zweitsprache (DaZ)

Sarah-Larissa Hecker  · **Stephanie Falkenstern** · **Svenja Lemmrich** · **Timo Ehmke** · **Barbara Koch-Priewe** · **Anne Köker** · **Udo Ohm**

Eingegangen: 11. Juni 2021 / Überarbeitet: 12. August 2022 / Angenommen: 20. September 2022 /
Online publiziert: 19. Januar 2023
© Der/die Autor(en) 2023

✉ Dr. Sarah-Larissa Hecker

Fach Deutsch als Fremd- und Zweitsprache, Institut für Germanistik und Medienkulturen, Technische Universität Dresden, Wiener Straße 48, 01219 Dresden, Deutschland
E-Mail: sarah-larissa.hecker@tu-dresden.de

Stephanie Falkenstern · Prof. Dr. Udo Ohm

Deutsch als Fremd- und Zweitsprache, Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Deutschland

Stephanie Falkenstern

E-Mail: stephanie.falkenstern@uni-bielefeld.de

Prof. Dr. Udo Ohm

E-Mail: udo.ohm@uni-bielefeld.de

Dr. Svenja Lemmrich · Prof. Dr. Timo Ehmke

Erziehungswissenschaft/Empirische Bildungsforschung, Fakultät Bildung, Institut für Bildungswissenschaft, Leuphana Universität Lüneburg, Universitätsallee 1, 21335 Lüneburg, Deutschland

Dr. Svenja Lemmrich

E-Mail: lemmrich@leuphana.de

Prof. Dr. Timo Ehmke

E-Mail: tehmke@leuphana.de

Prof. Dr. Barbara Koch-Priewe · Dr. Anne Köker

AG 4 Schulentwicklung und Schulforschung, Fakultät für Erziehungswissenschaft, Universität Bielefeld, Universitätsstraße 25, 33615 Bielefeld, Deutschland

Prof. Dr. Barbara Koch-Priewe

E-Mail: bkoch-priewe@uni-bielefeld.de

Dr. Anne Köker

E-Mail: anne.koeker@uni-bielefeld.de

Zusammenfassung Der Beitrag befasst sich mit dem kollaborativen und zugleich empirisch basierten Prozess des Standardsettings in einem interdisziplinären Forschungsprojekt, welches Lehrkräfte-Expertise im Bereich Deutsch als Zweitsprache (DaZ) mithilfe von Aufgaben zu kurzen Unterrichtsvideos erfasst hat. Ausgehend von Annahmen aus der Expertiseforschung steht dabei die Frage im Fokus, wie sich anhand der Testergebnisse von $N=295$ angehenden und praktizierenden Lehrkräften sowie unter der aktiven Beteiligung von einschlägigen Fachexpert*innen mithilfe der Bookmark-Methode theorie- und datenbasiert Schwellen von Kompetenz ermitteln lassen. Eine inhaltliche Beschreibung der im Beitrag hergeleiteten Kompetenzniveaus entlang eines Beispiel-Testitems schließt die Darstellung des Prozesses ab. Interpretiert werden die in der Stichprobe erreichten Kompetenzniveaus mit Bezug auf die Lehrer*innenbildung im Bereich DaZ.

Schlüsselwörter DaZ · Expertise · Lehrkräftebildung · Standard-Setting · Testinstrument

Collaborating for the standard: empirically-based determination of teacher expertise in the field of German as a second language (GSL)

Abstract The article deals with the collaborative and at the same time empirically based standard setting procedure in an interdisciplinary research project. The project assessed teacher expertise in the field of German as a Second Language (GSL) using tasks on short instructional videos. Based on assumptions from expertise research, we determine thresholds of competence in a theory- and data-based manner using the bookmark method. The procedure uses $N=295$ prospective and practising teachers' test results while also depending on the active participation of relevant experts. We conclude the procedure by describing the competence levels derived in the article along an example test item. Finally, we interpret the competence levels achieved in the sample with reference to teacher training in the field of GSL.

Keywords Second Language · Expertise · Teacher Education · Standard Setting · Test Instrument

1 Einleitung

Da erst kürzlich der Blick hinter die Kulissen von Kompetenzmodellierungen und -messungen als ein besonders lohnender beschrieben wurde (vgl. Schreiner et al. 2020, S. 407), will der vorliegende Beitrag einen Einblick in das Standardsetting-verfahren des *DaZKom-Video*-Projekts gewähren. Er liefert dabei theoretische Hintergründe sowie Ergebnisse der empirisch basierten Bestimmung von Lehrkräfte-Expertise im Bereich Deutsch als Zweitsprache (DaZ) mittels eines kollaborativ-diskursiven Standardsettings nach der Bookmark-Methode (vgl. Lewis et al. 2012) unter Fachexpert*innen.

Die Frage, über welche Kompetenzen Lehrer*innen aller Fächer verfügen müssen, um einen sprachbewussten Fachunterricht zu gestalten, ist trotz der jahrelangen

Diskussion um die Notwendigkeit und Forderung einer „durchgängigen Sprachbildung“ (vgl. Gogolin et al. 2011) aller Schüler*innen vergleichsweise verzögert gestellt worden. Der umfassenden Herausforderung, diese professionellen Kompetenzen zunächst theoretisch zu modellieren und dann auch auf verschiedenen Kompetenzstufen empirisch messbar zu machen, stellte sich das Team der Projekte *DaZKom – Professionelle Kompetenzen angehender Lehrerinnen und Lehrer (Sek I) im Bereich Deutsch als Zweitsprache (DaZ)* und *DaZKom-Video – Performanznahe Messung von Deutsch-als-Zweitsprache-Kompetenz bei (angehenden) Lehrkräften*. Die Modellierung erhob zunächst den Anspruch, sowohl die Struktur als auch die Entwicklung von DaZ-Kompetenz abzubilden. Dabei orientierte sie sich einerseits am theoretischen Diskurs, akademischen Curricula und der Validierung durch Hochschul-Expert*innen, andererseits an dem *Five-Stage Model of Adult Skill Acquisition* von Dreyfus und Dreyfus (1986). Zudem wurde ein Testinstrument zur Messung von DaZ-Kompetenz bei angehenden Lehrkräften entwickelt, das sich auf die ersten drei von fünf der angenommenen Kompetenzstufen bezog.

Das *DaZKom-Video*-Projekt (2017–2020), welches in diesem Beitrag im Fokus steht, verfolgte daran anschließend das Ziel, DaZ-Kompetenz auf Expert*innenniveau bei (angehenden) Lehrkräften messbar zu machen, also die Stufen 4 und 5 des fünfstufigen Modells von Dreyfus und Dreyfus (1986). Wenngleich Lehrkräfte-Expertise in der Querschnittsdomäne DaZ noch wenig untersucht worden war (vgl. Hecker und Nimz 2020), lagen hierzu domänenübergreifend bereits umfassende und empirisch begründete Erkenntnisse vor. So dienten bei der Konzeption des neuartigen Testinstruments neben dem Modell von DaZ-Kompetenz einschlägige Ergebnisse der Expertiseforschung als Ausgangspunkt. Diesen zufolge nehmen Expert*innen Unterrichtssituationen beispielsweise holistisch wahr und zeichnen sich durch besonders intuitives Handeln aus (vgl. bspw. Bromme 2008). Diese wichtigen Expertisemerkmale wurden mithilfe eines neuartigen videobasierten Testinstruments operationalisiert, das authentische Unterrichtssituationen aus dem Bereich DaZ präsentiert und (angehende) Lehrkräfte dazu auffordert, ihre Wahrnehmung und Handlungsoptionen – sogenannte situationsspezifische Fähigkeiten (vgl. Blömeke et al. 2015) – spontan mündlich zu verbalisieren.

Der vorliegende Beitrag gibt zunächst in Kap. 2 einen kurzen Einblick in die Grundlagen des *DaZKom*-Modells und die Genese des videobasierten Testinstruments zur performanznahen Erfassung von DaZ-Kompetenz auf Expert*innen-Niveau, bevor in Kap. 3 die methodischen Schritte, d. h. die Normierungsstudie und das Standardsetting beschrieben werden. Im Rahmen des Standardsettings mit einschlägigen Fachexpert*innen ($N=9$) wurde bezogen auf die Daten von $N=295$ angehenden und praktizierenden Lehrkräften aller Fächer zunächst kollaborativ festgelegt, welche Antworten in den Aufgaben (*Items*) zum Videotest mindestens gegeben werden mussten, um eine der höchsten DaZ-Kompetenzstufen zu erreichen. So konnten die Testergebnisse der angehenden und praktizierenden Lehrkräfte in der Stichprobe einem Kompetenzniveau nach dem *DaZKom*-Modell (Ohm 2018) bzw. Dreyfus und Dreyfus (1986) zugeordnet werden. Im Zentrum des Beitrags stehen die Ergebnisse des Standardsettings in Kap. 4. Berichtet wird also, wie sich die von den $N=9$ Fachexpert*innen im Standardsetting datenbasiert und kollaborativ festgelegten Niveaus inhaltlich beschreiben lassen und wie sich die Stichprobe von $N=295$ angehenden

und praktizierenden Lehrkräften auf die Kompetenzniveaus verteilt. Der Beitrag schließt mit einer Diskussion der Ergebnisse (Kap. 5) in ihrer Bedeutung für die Lehrkräftebildung im Bereich DaZ.

2 Modellierung und performanznahe Messung von Kompetenz bei Lehrkräften im Bereich Deutsch als Zweitsprache

2.1 Modellierung: Das DaZKom-Entwicklungs- und Strukturmodell

DaZ-Kompetenz wird hier als generische, fächerübergreifende Kompetenz von Lehrer*innen verstanden, ihren Unterricht sprach- bzw. registersensibel zu gestalten. Diese bildet das *DaZKom*-Modell in drei Dimensionen ab: *Fachregister*, *Mehrsprachigkeit* und *Didaktik* (vgl. Köker et al. 2015; Ohm 2018). Alle drei Dimensionen sind wiederum unterteilt in Subdimensionen, welche ihrerseits durch inhaltliche Facetten näher beschrieben werden (vgl. Abb. 1).

Die Dimension *Fachregister* fokussiert auf die Rolle der deutschen Sprache zur Konstruktion von und Partizipation an Wissen im Fachunterricht (vgl. de Oliveira und Schleppegrell 2015; Schleppegrell 2007). Alle Schüler*innen müssen sich den für das jeweilige Fachregister spezifischen Gebrauch der deutschen Sprache erschließen können. Somit ist die deutsche Sprache nicht nur das zentrale Medium zur Vermittlung und Aneignung von Fachinhalten, sondern auch selbst ein Lernge-

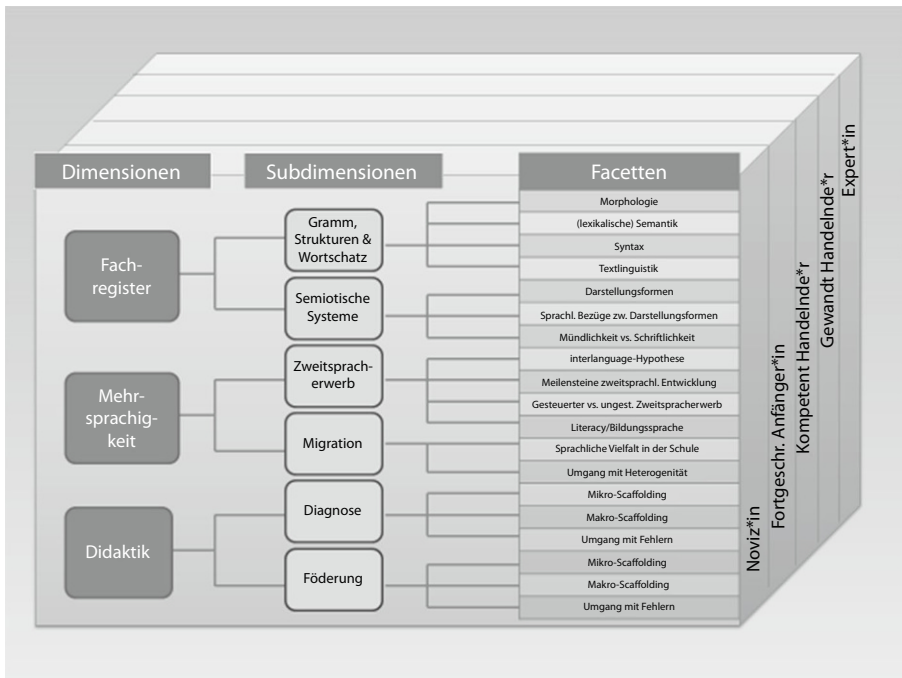


Abb. 1 Das DaZKom-Entwicklungs- und Strukturmodell

genstand (vgl. Ohm 2018, S. 75 f.). Die Dimension *Mehrsprachigkeit* nimmt den Lernprozess der Schüler*innen in den Blick. Um mehrsprachige Schüler*innen unter Berücksichtigung ihrer bereits vorhandenen (sprachlichen) Kompetenzen und Ressourcen optimal unterstützen zu können, brauchen Lehrkräfte unter anderem Kenntnisse im Bereich des Zweitspracherwerbs (vgl. Ohm 2018, S. 79). Gleichzeitig wird in dieser Dimension der Umgang mit Vielfalt an der Schule fokussiert (vgl. Falkenstern et al. 2021; Ohm 2018, S. 80). Mit der Dimension *Didaktik* fokussiert das Strukturmodell darüber hinaus auf den Lehrprozess, welcher sich jedoch lediglich analytisch vom Lernprozess der Schüler*innen und der Rolle des Fachregisters trennen lässt (vgl. Ohm 2018, S. 74). Schließlich gilt es im Sinne des didaktischen Prinzips des Scaffoldings, die Entwicklungspotenziale der Lerner*innen einschätzen zu können (*Diagnose*), um adäquate sprachliche und fachliche Anforderungen und Unterstützungsmaßnahmen auszuwählen (*Förderung*) (vgl. Gibbons 2002; Ohm 2018, S. 81). Die inhaltliche Beschreibung von DaZ-Kompetenz bei (angehenden) Lehrkräften liegt dem in Abschn. 3.1 beschriebenen Testinstrument ebenso zugrunde wie der Entwicklungsaspekt des Modells. Der Entwicklungsaspekt bezieht sich auf Dreyfus' und Dreyfus' (1986) fünfstufiges, dabei als Kontinuum zu verstehendes Modell vom Novizen hin zur Expertin, welches nachfolgend weiter ausgeführt wird.

2.2 Domänenübergreifende Merkmale von Expertise als höchster Kompetenzstufe

Ziel des verwendeten Testinstruments war es, die höchsten beiden Stufen von DaZ-Kompetenz, nach Dreyfus und Dreyfus (1986) *Gewandtheit* und *Expertise*, beschreiben und messen zu können. Hierfür wurden zunächst domänenübergreifende Merkmale von Lehrkräfte-Expertise definiert. Anschließend wurden diese Merkmale auf Anforderungen der inhaltlich durch das *DaZKom*-Modell bereits beschriebenen Domäne DaZ/Sprachbildung bezogen (Abschn. 2.4 und Kap. 4).

Die Kompetenzentwicklung verläuft laut empirischen Studien aus der Expertiseforschung beginnend mit der Stufe der Noviz*innen in bis zu fünf Stufen zur Expertise (vgl. Dreyfus und Dreyfus 1986). Die Beschreibung dieser Stufen kann bei der Messung von Kompetenz unterstützen, indem den modellierten Stufen empirisch erfassbare konkrete berufliche Anforderungen zugeordnet werden (vgl. ähnlich Piwowar 2013). So wissen wir beispielsweise über Expertenlehrer*innen, dass diese sich bei der Problemidentifizierung und -lösung im Klassenzimmer, anders als Noviz*innen, häufig nicht mehr auf ihr Regelwissen oder auf reflektierte Pläne verlassen (vgl. Bromme 2014; Dreyfus und Dreyfus 1986; Neuweg 2015). Stattdessen zeichnet sie eine präzise holistische Wahrnehmung relevanter Situationscharakteristika aus, durch die es ihnen auch in komplexen Unterrichtssituationen gelingt, situationsspezifisch intuitiv und angemessen zu handeln. Dies lässt sich auch mit der Entstehung von Expertise in Verbindung bringen. Es wird davon ausgegangen, dass Expertise erfahrungsbasiert ist und durch das wiederholt erfolgreiche Bewältigen von Situationen mit ähnlichen Anforderungen entsteht (vgl. Bromme 2008). Expertise ist somit an Situationen in der jeweiligen Domäne gebunden und damit domänenspezifisch.

Die Erfahrungsabhängigkeit von Expertise ist auch der Grund dafür, dass es bei der Erfassung der höchsten Kompetenzstufen nicht genügt, lediglich die Wissensbasis als Unterscheidungskriterium zwischen Expert*innen und Noviz*innen heranzuziehen und entsprechend Testaufgaben zu formulieren, die auf Wissen abzielen (vgl. Hecker und Nimz 2020; Hecker et al. 2020a, b; Lindmeier et al. 2013). Denn die Wissensbasis, dies legen Ergebnisse aus der Expertiseforschung nahe, wird mit zunehmender Erfahrung mit ähnlichen Situationen umstrukturiert, Verbindungen zwischen Wissens-elementen werden geschaffen, Schemata und Skripts werden nach häufiger erfolgreicher Bewältigung vergleichbarer Situationen angelegt (vgl. Berliner 1992; Bromme 2001; Neuweg 2015; Wahl 2002). Eine in diesem Sinne umstrukturierte Wissensbasis korrespondiert mit einer holistischen Wahrnehmung ebenso wie mit intuitivem Handeln, weil der Zugriff auf das Wissen besser und schneller gelingt. Expertise ist also nicht auf ein besonders ausgeprägtes, abrufbares Regelwissen reduzierbar. Vielmehr wird Expertise vermutlich erst in dem beruflichen Setting sichtbar, in dem sie auch entstanden ist: also in unserem Kontext im Handeln einer Lehrkraft in einer Unterrichtssituation. Wie das *DaZKom-Video-Test*-instrument diese Erkenntnisse aus der Expertiseforschung zur Erfassung von DaZ-Kompetenz auf Expertiseniveau nutzt, wird nachfolgend beschrieben.

2.3 Performanznahe Messung von DaZ-Expertise als höchster Kompetenzstufe

Eine performanznahe Messung von Kompetenz auf den Stufen 4 und 5 des Dreyfus'schen Modells (1986) kann ausgehend von den oben skizzierten typischen Expertisemerkmalen dann erfolgen, wenn die domänenspezifischen Anforderungen von Situationen in Verbindung mit gewohnten Handlungen abgebildet werden. Das *DaZKom-Video-Test*-instrument berücksichtigt dies, indem es Videovignetten von Fachunterrichtssituationen als Impulse nutzt, die laut Expert*innenbefragungen ($N=3$) DaZ-domänenspezifische Anforderungen authentisch abbilden (vgl. Hecker und Nimz 2020; s. a. Abschn. 3.1). Darüber hinaus enthält es Aufgaben, die es erlauben, zwei der typischen Expertisemerkmale zu simulieren: Proband*innen versetzen sich in die gesehene Situation hinein, um zuerst ihre (möglichst holistische) *Wahrnehmung* der Situation (*Was nehmen Sie wahr?*) zu verbalisieren, bevor sie zu einer spontanen Reaktion aufgefordert werden, die möglichst nah an einer tatsächlichen (*Unterrichts-*) *Handlungsreaktion* angesiedelt ist (1. *Sie sind die Lehrkraft in dieser Situation, wie reagieren Sie wörtlich?* oder 2. *Wie würden Sie anstelle der Lehrkraft in dieser Situation handeln?*). Die Intuitivität der Handlungsreaktionen als wichtiges Expertisemerkmal wird ermöglicht, indem im Test statt schriftlicher Antworten mündliche Antworten verlangt werden. So wird vermieden, dass durch die beim Verschriftlichen üblicherweise ablaufenden Planungs-, Korrektur- und Nachdenkprozesse die gewünschte Spontaneität der Reaktion verloren geht (vgl. dazu Hecker et al. 2020a). Die auf diese Weise in den Testitems elizitierten sogenannten situationsspezifischen Fähigkeiten¹ *Wahrnehmung* und *Handlungsreaktion* gelten in der Kompetenzforschung als besonders performanznahe Indikatoren für Expertise (vgl. bspw. Blömeke et al. 2015). Die Auswertung der Antworten der

¹ Bei Blömeke et al. (2015) ähnlich: *perception* und *decision-making*.

(angehenden) Lehrkräfte auf die Unterrichtssituationen soll somit die Zuordnung der Testpersonen zu den höchsten beiden Kompetenzstufen ermöglichen.

2.4 Vorstellungen von Kompetenzniveaus: Theorie und Praxis von DaZ-Expertise

Nun ist allerdings nicht davon auszugehen, dass die mit dem Videotestinstrument getesteten Personen sich nur den höchsten beiden Kompetenzstufen zuordnen lassen, sondern es ist im Gegenteil zu vermuten, dass viele angehende und praktizierende Lehrkräfte in amtlich deutschsprachigen Regionen sich eher auf niedrigeren Kompetenzstufen befinden. Denn übertragen wir die oben dargelegten Annahmen zur Genese von Expertise auf die Domäne DaZ im deutschsprachigen Raum, dürften Lehrkräfte in der Praxis bislang nur selten zu intuitiv handelnden, holistisch wahrnehmenden Expert*innen im Sinne von Dreyfus und Dreyfus geworden sein. Dies hängt mit der besonderen Entwicklung des Fachs DaZ und des Bereichs sprachsensibler Fachunterricht zusammen: Inhalte, die angehende Lehrkräfte auf einen durchgängig sprach- und registersensiblen Fachunterricht vorbereiten, sind an den Universitäten in Deutschland noch nicht in allen Bundesländern verpflichtender Teil der Lehrkräftebildung (vgl. Baumann 2017; Hecker et al. 2020b). Dies bedeutet zweierlei: Erstens, dass viele aktuell praktizierende Lehrkräfte kaum oder wenige Lehrveranstaltungen in Aus- und Weiterbildung besucht haben dürften, in denen sie Wissen in den DaZ-Kompetenz-Dimensionen Fachregister, Mehrsprachigkeit und Didaktik hätten erwerben können. Sollten Lehrkräfte also DaZ-Expert*innen sein, dürften sie diese Expertise vielfach hauptsächlich durch eigenes Interesse am Thema, selbst angestoßene Reflexionsprozesse und individuell angeeignetes Wissen erreicht haben.

Zweitens bedeutet dies aber auch, dass dank zahlreicher curricularer Änderungen in der jüngeren Vergangenheit angehende Lehrkräfte vermehrt Lehrveranstaltungen zu Themen im Bereich Sprachbildung/DaZ besucht haben (vgl. Becker-Mrotzek und Woerfel 2020). Diese Personen wiederum dürften vielfach noch an den Universitäten studieren oder erst vor kurzem in den Lehrkräfteberuf eingestiegen sein. Somit dürften diese Personen zwar noch keine oder wenig Berufserfahrung – eine wichtige Komponente bei der Entwicklung von Expertise (vgl. Abschn. 2.2) – haben, doch aber im Studium erworbenes Wissen, welches ebenfalls relevant bei der Expertiseentwicklung ist (Becker-Mrotzek und Woerfel 2020). Kurzgefasst: Expertise im Bereich DaZ bei Lehrkräften dürfte bisher noch in nur sehr wenigen Fällen auf die Art und Weise – also, stark vereinfacht, durch Wissen und Berufserfahrung – entstanden sein, wie es die Expertiseforschung nahelegt. Für den Versuch einer performanznahen Erfassung bedeutete dies, dass die Stichprobe, in der DaZ-Expertise erwartet werden kann, sich sehr heterogen zusammensetzen würde, nämlich aus Lehramtsstudierenden (mit potenziell viel Wissen im Bereich DaZ, aber ohne Berufserfahrung) und praktizierenden Lehrkräften (mit mehr oder weniger viel Berufserfahrung und ebenso viel oder wenig in Lehrveranstaltungen erworbenem DaZ-Wissen) (vgl. Hecker 2021).

Die Heterogenität der Stichprobe und die heterogenen Wege zur Genese von DaZ-Expertise im heutigen Deutschland könnten sich auch auf die festzustellen-

den Kompetenzniveaus auswirken, die mit dem performanznahen Testinstrument identifiziert werden können. So ist davon auszugehen, dass nicht nur die obersten beiden Kompetenzstufen erreicht werden, sondern auch mindestens eine zunächst nicht spezifizierbare weitere, untere Kompetenzstufe. Wie sich dies auf die Ergebnisse des Standardsettings des Testinstruments für Expertise auswirken kann, wird unten weiter ausgeführt (Kap. 4 und 5).

3 Methodisches Vorgehen

3.1 Die Normierungsstudie mit dem DaZKom-Video-Testinstrument

Der *DaZKom-Video-Test* operationalisiert die DaZ-Kompetenz auf den Stufen 4 und 5 des fünfstufigen Dreyfus'schen Modells. Erprobt und weiterentwickelt wurde der Test mithilfe der Ergebnisse zweier Pilotierungsstudien ($N=134$ bzw. $N=183$) (vgl. Lemmrich et al. 2019, 2020) und einer Normierungsstudie ($N=295$) (Lemmrich und Ehmke, eingereicht), an denen angehende und praktizierende Lehrkräfte aller Fächer teilnahmen. Das Standardsetting des Videotests, das in diesem Beitrag im Fokus steht, nutzte die Ergebnisse der Normierungsstichprobe, also die Antworten von $N=295$ angehenden und praktizierenden Lehrkräften aus 16 Städten aller Teile Deutschlands. Erhoben wurde z. B. an lehrkräftebildenden Universitätsstandorten oder in Lehrkräftefortbildungen (an Universitäten oder in schulinternen Fortbildungen). Die Stichprobe ($N=295$) umfasste 57 % angehende Lehrkräfte, 39 % Lehrkräfte und 4 % andere (z. B. Forscher*innen, Lehrkräfteausbilder*innen). In der Stichprobe waren knapp 81 % der Proband*innen weiblich, fast 90 % hatten Deutsch als Erstsprache und knapp 66 % hatten bereits Lehrerfahrung im Bereich Deutsch als Zweitsprache (Lemmrich und Ehmke, eingereicht). Das Fach Mathematik studierten knapp 28 %, Deutsch 60 % und Englisch 14 % der Proband*innen (Lemmrich und Ehmke, eingereicht). Von den Proband*innen unterrichteten etwa 13 % Mathematik, knapp 25 % Deutsch und fast 11 % Englisch. Zudem waren alle möglichen Bildungsgänge in der Stichprobe vertreten: Die meisten Testpersonen (32 %) studierten Grundschullehramt oder unterrichteten an dieser Schulform, jede*r Fünfte studierte ein Lehramt mit Bezug auf die Sekundarstufe I oder unterrichtete bereits an Haupt-, Real-, Ober- oder Stadtteilschulen. 15 % unterrichteten an Gymnasien oder planten dies, 13 % an (integrierten) Gesamtschulen. Berufsschulen und Förderschulen waren mit jeweils 4 % bzw. 5 % vertreten. Im Durchschnitt waren die Teilnehmer*innen knapp 31 Jahre alt.

Die Testpersonen verbalisierten zu jeweils zwölf DaZ-relevanten videographierten Unterrichtsszenen, was sie in der Videosequenz wahrgenommen hatten und wie sie anstelle der im Video gezeigten Lehrkraft handeln würden. Die DaZ-Relevanz, die Passung zu einer der drei Modelldimensionen sowie die Authentizität aller Szenen wurden in einem Expert*innen-Rating mit $N=3$ einschlägigen universitären Fachexpert*innen, die unabhängig voneinander befragt wurden, bestätigt (vgl. Lemmrich et al. 2019).

Der Ablauf der Testsituation verlief standardisiert. Den (angehenden) Lehrkräften wurden zunächst die zwischen 30 Sekunden und zwei Minuten langen Videosequen-

Tab. 1 Psychometrische Kennwerte des Testinstruments

Merkmale	Kennwerte
<i>EAP/PV Reliabilität</i>	0,76
<i>Trennschärfe</i>	MW = 0,42; SD = 0,07 MIN = 0,26; MAX = 0,53
<i>Itemschwierigkeit</i>	MW = 0,52; SD = 0,26 MIN = -0,08; MAX = 1,02
<i>Weighted Item fit</i>	MW = 1,01; SD = 0,05 MIN = 0,91; MAX = 1,14

zen auf Tablets gezeigt; jeweils anschließend an eine Sequenz sprachen die Testpersonen dann mittels Headsets ihre Wahrnehmung und Handlungsreaktion mündlich ein. Abschließend füllten die Testpersonen einen Fragebogen zu ihren soziodemographischen Merkmalen sowie ihren DaZ-relevanten Lerngelegenheiten in Studium und Beruf aus (vgl. dazu Ehmke und Lemmrich 2018; vgl. die Ergebnisse in Lemmrich et al. 2020 und Hecker et al. 2020b). Kodiert wurden die Antworten in den Pilotierungen sowie der Normierungsstudie mittels detaillierter Kodierleitfäden, die jeweils auf Basis von Fachexpert*innenantworten aus Universität und Schulpraxis ($N=6$) generiert und mithilfe der Ergebnisse der ersten Pilotierungsstudie weiterentwickelt worden waren (Kodierer*innenübereinstimmung $P\ddot{U}=88,84$, Cohens Kappa $\kappa=0,76$, vgl. Hecker et al. 2020a). Die in der Normierungsstudie verwendeten Kodierleitfäden erlauben eine Vergabe der Codes 0, 1, 2 und 3. Dabei wurde Code 0 allen nicht-situationsspezifischen, nicht DaZ-relevanten Antworten zugeordnet. Code 1 wurde für unspezifische, aber DaZ-relevante Antworten vergeben, Code 2 für Antworten, die situationsspezifisch und DaZ-relevant waren, dabei jedoch unkonkret bleiben. Code 3 wiederum erhielten situationsspezifische, DaZ-relevante, konkrete und über die Situation hinausgehende Antworten. Anhand der kodierten Antworten aus der Normierungsstichprobe mit dem Testinstrument wurden empirisch die Item-Schwierigkeiten der Aufgaben ermittelt, die die Basis des nachfolgend thematisierten Standardsettings bildeten.

Tab. 1 gibt einen Überblick über die psychometrischen Kennwerte des Testinstruments: die Reliabilität, die mittlere Trennschärfe, die mittlere Itemschwierigkeit und der mittlere Weighted Item Fit. Die Dimensionalität des Konstruktes wurde mithilfe von IRT-Modellen geprüft. Für das dreidimensionale Modell wurde angenommen, dass sich die drei Dimensionen entlang des DaZKom-Modells (Fachregister, Mehrsprachigkeit, Didaktik) abbilden lassen. Als Modellvergleichsmaß wurde unter anderem das Akaike Information Criterion (AIC) herangezogen. Die Fit-Indices des Modellvergleichs deuten auf ein eindimensionales Konstrukt hin (vgl. Lemmrich et al. 2020). Differenzierte Befunde zu den psychometrischen Kennwerten des Testinstruments finden sich auch in Lemmrich und Ehmke (eingereicht).

3.2 Das Standardsetting

3.2.1 Zielsetzung des Standardsettings

Das Standardsetting ist ein Verfahren, in dem auf Basis der fachlichen Anforderungen der Testaufgaben Fachexpert*innen Schwellen auf einer kontinuierlichen Kompetenzskala setzen, um diese in unterschiedliche Kompetenzstufen einzuteilen.

Bei der Darstellung von Testergebnissen in Studien lassen sich so Proband*innen entsprechend klassifizieren und die Verteilung der Personen auf die unterschiedlichen Kompetenzniveaus kann kriterial interpretiert werden (vgl. Abschn. 4) (vgl. Lewis et al. 2012). Das Standardsetting wurde in diesem Projekt in Anlehnung an das Prozedere im *DaZKom*-Projekt (vgl. Gültekin-Karakoç et al. 2016) mit der Bookmark-Methode durchgeführt (vgl. Lewis et al. 2012). Dabei werden die Items nach empirisch ermittelter Schwierigkeit aufsteigend (niedrigste–höchste Anforderung) sortiert und in einem Ordered-Item-Booklet (OIB) zusammengefasst. Ziel ist dann, Schwellen festzulegen (also *bookmarks* zwischen den Items zu setzen), die zwischen denjenigen Personen trennen, die dieses Item mit einer hohen Wahrscheinlichkeit korrekt beantworten können, und jenen, die es nicht können. Die Schwellen definieren dabei die Proband*innen mit der minimalen Kompetenz im entsprechenden Anforderungsniveau (vgl. Haschke et al. 2017; Lewis et al. 2012). In der internationalen Praxis werden hauptsächlich zwei Ansätze zur Festlegung von Bewertungsstandards verwendet: die Angoff-Methode (Cizek und Bunch 2007) und die Bookmark-Methode (Buckendahl et al. 2002). Im Gegensatz zur Angoff-Methode erfordert die Bookmark-Methode, dass der Test vor der Festlegung von Cut-Scores auf der Grundlage von Testdaten durchgeführt wird. Dies verleiht der Validität des Standardsettings zusätzliches Gewicht und gibt den beteiligten Expert*innen während des Prozesses Informationen über die empirisch ermittelten Schwierigkeiten der Aufgaben.

Am Standardsetting des *DaZKom-Video-Testinstruments* waren sechs einschlägige DaZ-Fachexpert*innen aus Schule und Hochschule beteiligt. Fünf der Expert*innen hatten eine Professur an einer Universität inne, eine Expertin hatte langjährige Erfahrungen im Bereich DaZ an Schulen. Auch unter den Professor*innen hatten mindestens drei Teilnehmende Erfahrung als Lehrkräfte an Schulen. Die Kombination von Forschungs- und Schulpraxisexpertise sollte die Genese von Expertise im Bereich DaZ berücksichtigen, die laut Expertiseforschung auch auf Basis von Berufserfahrung in entsprechenden Situationen, hier dem sprachsensiblen Unterricht, zustande kommt (vgl. Hecker und Nimz 2020). Zusätzlich nahmen drei Mitglieder der Projektgruppe² an dem Standardsetting teil.

Durch das Standardsetting soll die Kompetenzskala in mehrere Abschnitte unterteilt werden, angefangen bei *keine Kompetenz* (Niveau 0) bis hin zu *sehr hohe Kompetenz* (Niveau 3). Da nicht alle angehenden und praktizierenden Lehrkräfte in

² Udo Ohm (Professor für Deutsch als Zweit- und Fremdsprache an der Universität Bielefeld, Fakultät für Linguistik und Literaturwissenschaft), Anne Köker (wissenschaftliche Mitarbeiterin an der Universität Bielefeld, Fakultät für Erziehungswissenschaft) und Sarah-Larissa Hecker (zum Zeitpunkt des Standardsettings wissenschaftliche Mitarbeiterin an der Universität Bielefeld).

der Stichprobe eine hohe oder sehr hohe DaZ-Kompetenz auszeichnete, wurden nicht nur die Stufen 4 und 5 des Dreyfus’schen Modells – DaZ-Expertise – durch die Daten abgebildet, sondern auch niedrigere Kompetenzstufen. Dabei sind diese empirisch ermittelten Stufen nicht unbedingt gleichbedeutend mit den Kompetenzstufen nach Dreyfus und Dreyfus (1986) und werden daher von diesen abweichend zunächst als Kompetenzniveaus bezeichnet. Die Stufen nach Dreyfus und Dreyfus wurden aber bereits in der Test- und Itementwicklung herangezogen. Auch der Kodierleitfaden orientierte sich entsprechend an ihren Stufen. Die im Standardsetting festgelegten Kompetenzniveaus leiten sich dagegen aus den Ergebnissen der konkreten Datenerhebung mit dem *DaZKom-Video*-Testinstrument und der erfolgten Schwellensetzung durch Fachexpert*innen ab. Inwiefern die durch Schwellensetzung identifizierten Niveaus mit den Stufen von Dreyfus und Dreyfus korrespondieren dürften, zeigen die Beschreibungen der Ergebnisse in Abschnitt 4.2.

3.2.2 Durchführung des Standardsettings

Im Ordered-Item-Booklet (OIB) wurden alle möglichen Item-Antwortcodes (mit Ausnahme von 0) nach aufsteigenden empirischen Schwierigkeiten (von niedrig bis hoch) geordnet zusammengestellt. Die Antwortcodes ergaben sich aus der Anlage des Testinstruments wie folgt (vgl. Abb. 2): Das finale Testinstrument enthält zwölf Videos mit Unterrichtsszenen, zu denen die Testpersonen jeweils zwei Items bearbeiten mussten. Diese entsprachen immer denselben beiden o. g. Fragetypen zur *Wahrnehmung* und *Handlungsreaktion*. Zu jedem dieser beiden Items gab es drei mögliche Antwortcodes. *Falsche* Antworten (Antwortcode 0) wurden im Booklet nicht mitaufgeführt. So ergaben 12 Videos mal 2 Fragetypen mal 3 Antwortcodes (1 = unspezifische, aber DaZ-relevante Antworten; 2 = Antworten, die situationsspezifisch und DaZ-relevant waren, dabei jedoch unkonkret bleiben; 3 = situationsspezifische, DaZ-relevante, konkrete und über die Situation hinausgehende Antworten) insgesamt 72 geordnete Item-Antwortcodes, die das gesamte Schwierigkeitsspektrum abbildeten. Die jeweils ausformulierten Kodieranweisungen zu den 72 Item-Antwortcodes samt Ankerbeispielen wurden den Fachexpert*innen im OIB also gesammelt und nach Schwierigkeit geordnet zur Verfügung gestellt.

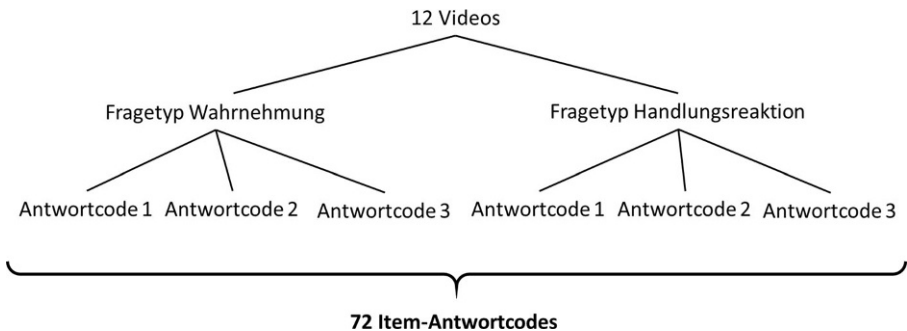
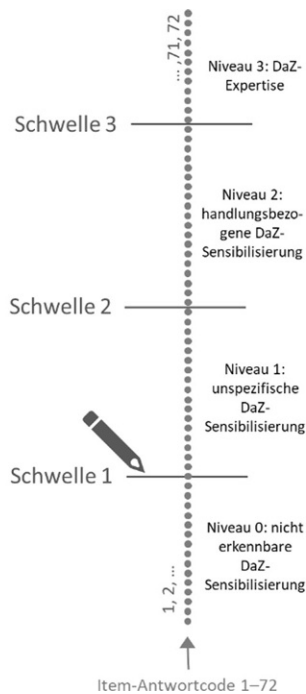


Abb. 2 Übersicht über die Item-Antwortcodes im OIB

Abb. 3 Markierung der Schwellen



Die $N=9$ Fachexpert*innen, die am Standardsetting teilnahmen, erhielten das OIB bereits vorab mit der Bitte, sich zur Vorbereitung des Standardsettings mit den Videostimuli, den Fragetypen und den Kodieranweisungen samt Ankerbeispielen zu den 72 Item-Antwortcodes vertraut zu machen. Außerdem markierten sie Schwellen an den entsprechenden Stellen im OIB, an denen sie die Übergänge zwischen den Anforderungsbereichen verorten würden (vgl. Abb. 3).

In der individuellen Vorbereitungsphase vor dem Treffen wurden den Fachexpert*innen dafür zunächst zwei Leitfragen zur Verfügung gestellt, die im Prozess des Standardsettings dann allerdings gemeinsam neu formuliert und erweitert wurden. In der Vorbereitung zielten die Leitfragen zur Setzung von zwei Schwellen (zwischen Niveau 1 und 2 sowie Niveau 2 und 3) darauf ab, sich über die subjektiven Expert*inneneinschätzungen zu minimalen und wünschenswerten Anforderungen sprachsensiblen Handelns im Fachunterricht den Kompetenzniveaus anzunähern. Dabei wurden bewusst keine Fragen gestellt, die sich auf Begriffe oder Annahmen aus der Expertiseforschung bezogen, denn im Fokus standen die subjektiven Einschätzungen der $N=9$ Expert*innen, die diese aufgrund ihrer eigenen Berufserfahrung in der relevanten Domäne entstandenen Expertise vornehmen mussten. Durch die Leitfragen sollte also zunächst ein Bezugsrahmen aufgespannt werden:

1. Zeigt eine Lehrkraft, die eine bzw. mehrere der abgebildeten Antworten gibt, dass sie die minimalen Anforderungen eines sprachsensiblen Fachunterrichts erfüllen kann? Setzen Sie die erste Schwelle bei dem Antwortcode, bei dem Sie diese Frage erstmals bejahen können.

2. *Zeigt eine Lehrkraft, die eine bzw. mehrere abgebildete Antworten gibt, dass sie die wünschenswerten Anforderungen eines sprachsensiblen Fachunterrichts erfüllen kann? Setzen Sie die zweite Schwelle bei dem Antwortcode, bei dem Sie diese Frage erstmals bejahen können.*

Während des Treffens wurden diese vorab in der Vorbereitungsphase gesetzten Schwellen ausgewertet und mit Bezug auf Expertise diskutiert, bevor dann erneut Schwellen gesetzt wurden. Nachdem im Laufe des Standardsetting-Prozesses allerdings deutlich geworden war, dass die ursprünglich angesetzten zwei Schwellen aufgrund der Heterogenität der Stichprobe (vgl. Abschn. 2.4) nicht ausreichen, um die in den Item-Antwortcodes vorhandenen Kompetenzniveaus abzubilden, wurden im Prozess gemeinsam mit den Expert*innen folgende drei neue Fragen für die Schwellensetzung festgelegt, in denen das Expertise-Niveau nun auch direkt angesprochen wurde:

1. *Welche Anforderung können wir als unspezifische DaZ-Sensibilisierung bezeichnen?*
2. *Welche Anforderung soll als ein DaZ-sensibilisiertes Niveau gelten?*
3. *Ab welcher Anforderung sprechen wir von einem DaZ-Expertise-Niveau?*

Nach drei Gesprächsrunden wurden schließlich final drei Schwellen festgesetzt, die vier Kompetenzniveaus auf Basis der dann transparent gewordenen qualitativen Anforderungen benannten (Abschn. 4.1) und inhaltlich beschrieben (Abschn. 4.2).

4 Ergebnisse

4.1 Festlegung der Schwellen

Die Abb. 4 zeigt das Ergebnis des Standardsettings, also die durch die Fachexpert*innen anhand der tatsächlichen Item-Anforderungen festgelegten Schwellen. Dargestellt ist eine sogenannte Wright Map, in der die Verteilung der Personenfähigkeiten und der Item-Schwierigkeiten auf einer Mess-Skala (Logit-Skala) abgebildet sind. Die Personenfähigkeiten variieren in der Normierungsstichprobe von etwa $-1,0$ Logits bis $+1,0$ Logits. Die Kreuze (X) in der Abbildung stehen für Proband*innen mit gleicher Personenfähigkeit. Die Aufgabenschwierigkeiten streuen von etwa $-1,0$ Logits bis $+2,0$ Logits. Für das Standardsetting wurden die Items nach den Logitwerten für die Schwellen mit 50% Antwortwahrscheinlichkeit verwendet. Die Testaufgaben sind nach dem Partial-Credit-Modell skaliert, das heißt, für jedes Item kann der Antwortscore zwischen 0 Score-Punkten (Antwort nicht situationsspezifisch, nicht DaZ-relevant) und 3 Score-Punkten (situationsspezifische, DaZ-relevante, konkrete und über die Situation hinausgehende Antworten) betragen. Entsprechend sind für jedes Item in der Wright Map drei Item-Schwierigkeiten dargestellt, die die Schwellen zwischen den vier möglichen Antwortscores pro Item aufzeigen. In der Abb. 4 sind diese Item-Parameterschwellen für beispielsweise das Item 1 mit 1,1, 1,2 und 1,3 bezeichnet. Die Item-Parameterschwelle 1,1 liegt bei

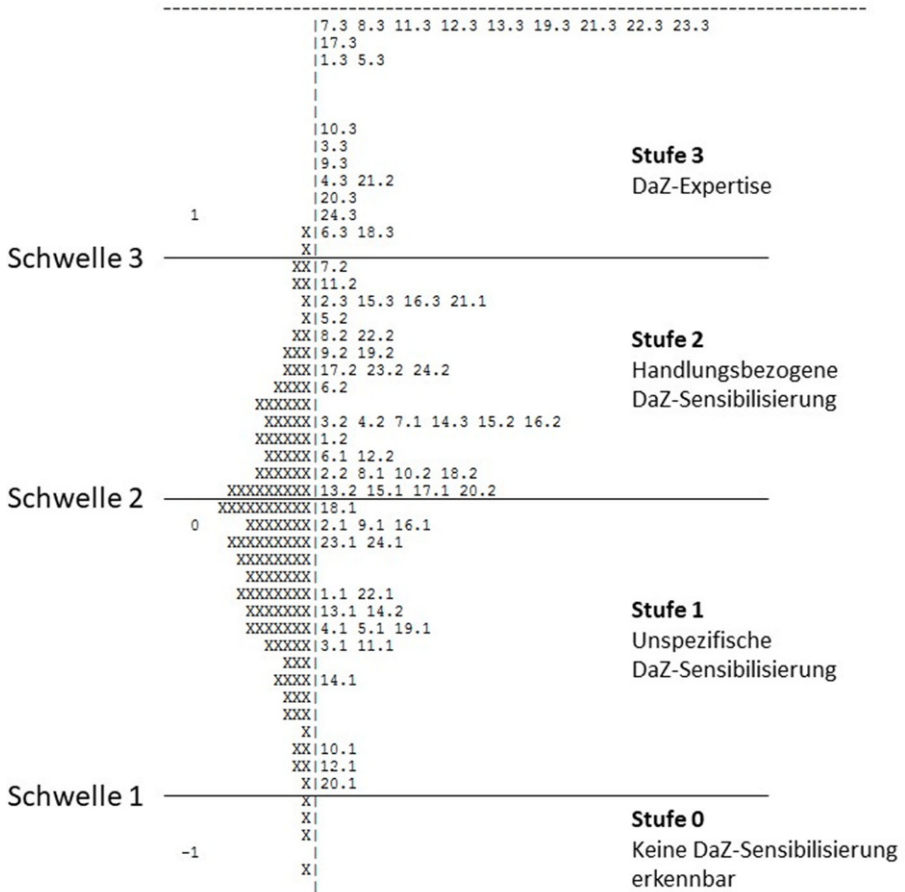


Abb. 4 Wright-Map mit Cut-Scores

-0,2 Logits und stellt die Grenze bzw. Schwelle zwischen dem Antwortscore=0 und dem Antwortscore=1 dar.

Die Abb. 4 zeigt, dass es eine akzeptable Passung zwischen der Verteilung der Personenfähigkeiten und der Aufgabenschwierigkeiten gibt. Allerdings können nur sehr wenige Proband*innen die Testaufgaben vollkommen lösen. Die WLE-Reliabilität beträgt $\alpha=0,76$. Der mittlere WLE-Standardfehler beträgt $MW=0,25$ ($SD=0,08$) und umfasst damit etwa ein Drittel der Breite der beiden mittleren Kompetenzstufen. Ausgegangen wird demnach von einer recht hohen Klassifikationsgenauigkeit. Auffällig ist, dass es zwei Niveaustufen gibt (die oberste und die unterste), auf denen jeweils nur sehr wenige Personen sind, so dass ca. 90% der Personen auf den beiden mittleren Niveaustufen verortet werden können. Die nach Abschluss des oben beschriebenen Standardsettingverfahrens festgelegten drei Schwellenwerte sind in der Abb. 4 durch schwarze horizontale Strecken veranschaulicht. Sie unterteilen die Messskala der DaZ-Kompetenz in vier Abschnitte. Diese werden von den

$N=9$ Fachexpert*innen als Niveau 0 (*keine DaZ-Sensibilisierung erkennbar*), Niveau 1 (*Unspezifische DaZ-Sensibilisierung*), Niveau 2 (*Handlungsbezogene DaZ-Sensibilisierung*) und Niveau 3 (*DaZ-Expertise*) bezeichnet.

4.2 Illustration und inhaltliche Beschreibung der Kompetenzniveaus

Ebenso wie die Cut-Scores, die von der Standardsetting-Expert*innengruppe „in einem iterativen Verfahren aus Einzelurteilen und Gruppendiskussionen“ (Pant et al. 2010, S. 176) bestimmt wurden, unterlag auch die Beschreibung der zwischen den Schwellen liegenden Stufen einem kollaborativen, diskursiven Prozess. Wengleich die leitenden Fragen der Instruktion zur Bestimmung der Cut-Scores im Vorfeld einen Bezugsrahmen für die Standardsetting-Expert*innen geliefert hatten, galt es die bei den Standardsetting-Expert*innen hierdurch evozierten Konzepte von Kompetenzniveaus und Expertise nachträglich im Zusammentreffen zu bündeln und dadurch inhaltlich auszuhandeln. Das Ergebnis dieses Prozesses sind vier Niveaubeschreibungen, die im Folgenden dargestellt und anhand eines Beispielimtems illustriert werden. Ein Beispiel für Repräsentationsformen aller vier Stufen bietet das Item *Obst*.

Beispielimtem *Obst*

Der Ausschnitt des im Test präsentierten Videos zeigt eine Szene aus dem Unterricht eines Kurses der Sekundarstufe I im Fach Hauswirtschaftslehre. Die an Gruppentischen sitzenden Schüler*innen werden von der Lehrerin aufgefordert, die auf dem vor ihnen liegenden Arbeitsblatt abgebildeten Obst- und Gemüsesorten der jeweiligen Kategorie zuzuordnen. Die Lehrerin geht im Raum umher, kommt auf eine Schülerfrage hin zu einem Gruppentisch und beantwortet mehrere Fragen der Schüler*innen. Am Ende der Szene fragt ein Schüler: „Wie kann Khaled sagen, dass eine Banane Gemüse ist?“

Bei der Frage nach der Wahrnehmung dieser Situation bietet das Item die Möglichkeit, einen Aspekt des Konzepts von Mehrsprachigkeit der Proband*innen zu evaluieren. Es fokussiert die Offenheit von Lehrpersonen gegenüber unterschiedlichen Kategorisierungen von Gegenständen, die in Sprachen angetroffen werden können; hier also, dass eine Banane in den unterschiedlichen Weltregionen und somit auch Sprachen unterschiedlich kategorisiert werden kann (Obst- und Gemüsebananen).

Als *0-Niveau* wurde von den Fachexpert*innen eine *nicht erkennbare DaZ-Sensibilisierung* definiert. Die Antworten der Testpersonen der Normierungsstichprobe, die dem Niveau 0 zugeordnet werden können, enthalten lediglich eine allgemeine Situationsbeschreibung, Paraphrasen der Aufgabenstellung oder Beschreibungen des Verhaltens der Lehrerin im Video. (Angehende) Lehrpersonen, die auf dieser Stufe zu verorten sind, bringen eine offensichtlich DaZ-relevante Unterrichtssituation nicht in einen solchen Zusammenhang. Es ist zu vermuten, dass sie bislang keinen wesentlichen Zugang zur Thematik sprachsensiblen Fachunterrichts und noch keine oder kaum (Lern-) Gelegenheiten hatten, ihre DaZ-Kompetenz auszubilden. Dieses Niveau dürfte entsprechend noch mit keiner Stufe nach Dreyfus und Dreyfus (1986)

korrespondieren, da sich bei den Testpersonen offenbar gar keine DaZ-Kompetenz ausbilden konnte.

Dagegen verfügen Lehrpersonen, die sich auf dem *Niveau 1: unspezifische DaZ-Sensibilisierung* befinden, bereits über die notwendige DaZ-Kompetenz, um in Ansätzen sensibilisiert auf eine sprachliche Benachteiligung von mehrsprachigen Schüler*innen im Fachunterricht reagieren zu können. Die Testpersonen der Normierungsstichprobe, deren Antworten zum Beispielitem dem Niveau 1 zugeordnet werden können, äußern sich unspezifisch zur mangelnden Unterstützung durch die Lehrerin und/oder spekulieren über den Sprachstand der Schüler*innen. Sie erkennen somit die DaZ-Relevanz einer wahrgenommenen Unterrichtssituation, antworten aber diesbezüglich unspezifisch, ungenau oder gar falsch. Ihre begrifflich-theoretische Einordnung der Situation ist unzulänglich oder diffus. Nach Dreyfus und Dreyfus dürfte dies etwa den Kompetenzstufen 1 bis 3 (von 5) entsprechen, eine genaue Zuordnung ist aber mit dem auf höhere Kompetenzstufen ausgerichteten Testinstrument erst einmal nicht möglich.

Das *Niveau 2* repräsentiert Lehr-/Testpersonen, die eine *handlungsbezogene DaZ-Sensibilisierung* aufweisen. Die Testpersonen, deren Antworten zur Wahrnehmung der Videoszene sich auf dem Niveau 2 finden, beziehen sich einerseits auf fehlende vorentlastende Hilfsmittel oder eine mangelnde sprachliche Unterstützung des Unterrichtsmaterials. Andererseits registrieren Testpersonen den despektierlichen und defizitorientierten Umgang der Schüler*innen untereinander. Sie beziehen sich nicht nur auf die spezifische Situation, sondern sind auch in der Lage, sie differenziert zu erfassen, zu analysieren und zu interpretieren, und letztendlich hieraus konkrete Handlungsoptionen abzuleiten. Das Niveau 2 im Standardsetting korrespondiert damit mit der Beschreibung der Stufe 4 nach Dreyfus und Dreyfus (1986).

DaZ-Expertise wird den Antworten von Testpersonen attestiert, die das *Niveau 3* erreichen. Wenige Testpersonen nehmen wahr, dass unter Umständen das Konzept des Schülers Khaled nicht dem in der Unterrichtssprache Deutsch dominanten und von seinem Mitschüler einzig bekannten Konzept von „Banane“ entspricht, aber ebenso wie dieses im Fachunterricht Gültigkeit beansprucht und thematisiert werden müsste. Diese Lehrpersonen können aufgrund eines sehr hohen Grades an DaZ-Kompetenz Unterrichtsinteraktionen und Lernsituationen im Fachunterricht adäquat holistisch wahrnehmen, diese fachlich-kategorial einordnen und begrifflich-theoretisch diskutieren. Sie zeichnen sich durch Selbstreflexivität und einen breiten Blick aus, der mehrere Optionen lernförderlicher Handlungsentscheidungen für DaZ-Schüler*innen zulässt. Personen auf diesem Niveau charakterisiert also Expertise, Kompetenzstufe 5, im Sinne von Dreyfus und Dreyfus (1986).

4.3 Zuweisung von Kompetenzniveaus

Tab. 2 zeigt, dass in der Normierungsstichprobe knapp fünf Prozent der getesteten Personen keine DaZ-Sensibilisierung aufweisen, also noch überhaupt keine Kompetenz im Bereich DaZ entwickeln konnten. Die größte Gruppe von knapp 52% der Stichprobe zeigt eine DaZ-Kompetenz auf dem ersten Niveau, das eine unspezifische DaZ-Sensibilisierung repräsentiert und vermutlich mit den Kompetenzstufen 1–3 nach Dreyfus und Dreyfus korrespondiert. Genauer ließe sich dies mit dem

Tab. 2 Niveaus der DaZ-Kompetenz

Niveau	Beschreibung	Cut-Scores	Häufigkeit			Prozent		
			Gesamt	Angehnd. Lehrkraft	Praktiz. Lehrkraft	Gesamt	Angehnd. Lehrkraft	Praktiz. Lehrkraft
3	DaZ-Expertise	>0,84	10	3	7	3,7	1,8	6,5
2	Handlungsbezogene DaZ-Sensibilisierung	[0,07; 0,84]	109	68	41	40,0	41,2	38,3
1	Unspezifische DaZ-Sensibilisierung	[-0,80; 0,07]	140	91	49	51,5	55,2	45,8
0	Keine DaZ-Sensibilisierung erkennbar	<-0,80	13	3	10	4,8	1,8	9,4
-	Gesamt	-	272 ^a	165	107	100,0	100,0	100,0

^aIn dieser Tabelle werden nur die angehenden und praktizierenden Lehrkräfte mit vollständigen Werten betrachtet, daher die leicht verringerte Gesamtanzahl

Testinstrument des Vorgängerprojektes (vgl. Ehmke et al. 2018) feststellen, welches die Kompetenzstufen 1–3 erfasst. Etwa 40 % der getesteten Personen erreichen hingegen eine handlungsbezogene DaZ-Sensibilisierung, befinden sich also auf einem Niveau, das mit der Kompetenzstufe 4 nach Dreyfus und Dreyfus korrespondiert. Erwartungsgemäß erreicht aber nur ein sehr geringer Teil der Stichprobe, etwas unter vier Prozent, das Niveau der DaZ-Expertise.

Wird in den Ergebnissen zwischen angehenden und praktizierenden Lehrkräften differenziert, unterscheiden sich die beiden Gruppen in ihrer Verteilung auf die Niveaus nicht sehr stark voneinander. Allerdings erreicht unter den praktizierenden Lehrkräften ein prozentual höherer Anteil das Niveau *DaZ-Expertise*. Gleichzeitig aber erreichen auch 9,4 % der praktizierenden Lehrkräfte keine erkennbare DaZ-Sensibilisierung. Bei den angehenden Lehrkräften befinden sich nur 4,8 % auf diesem niedrigsten Niveau. Mehr als jede zweite angehende Lehrkraft erreicht dagegen das Niveau 1, 41,2 % gelten sogar als handlungsbezogen DaZ-sensibilisiert. Damit erreichen 96,4 % der Studierenden und knapp 85 % der Lehrkräfte die Niveaus 1 oder 2.

5 Diskussion

Ziel des Beitrags war es, einen inhaltlich definierten Standard von DaZ-Kompetenz auf Expertiseniveau empirisch, also datenbasiert sowie kollaborativ-diskursiv zu finden und dabei zugleich den dorthin führenden, teils kollaborativen Prozess transparent zu machen. Hierfür wurde zunächst die Konzeption des videobasierten Testinstruments auf Basis einschlägiger Annahmen aus der Expertiseforschung und einem Modell von DaZ-Kompetenz dargelegt, mit dessen Hilfe sich bekannte Expertisemerkmale – die holistische Wahrnehmung sowie intuitives Handeln – bei angehenden und praktizierenden Lehrkräften im Bereich DaZ performanznah

erfassen lassen. Die Ergebnisse einer Normierungsstudie des Testinstruments mit $N=295$ (angehenden) Lehrkräften und zwölf Unterrichtsvideoausschnitten stellten die Grundlage des anschließenden und hier beschriebenen Standardsettingverfahrens dar. Das Standardsetting beruhte auf den Antworten dieser 295 angehenden und praktizierenden Lehrkräften aller Fächer, denen nach dem Prozess ein Kompetenzniveau zugewiesen werden konnte. Im Verfahren wurden entlang der einzelnen Item-Schwierigkeiten in einem detailliert beschriebenen iterativen und kollaborativen Prozess in einer DaZ-Fachexpert*innen-Runde erst zwei, dann drei Schwellen gesetzt, anhand derer schließlich vier Kompetenzniveaus definiert und inhaltlich ausdifferenziert werden konnten. Dabei gehen wir davon aus, dass das höchste im Standardsetting definierte Niveau mit der Stufe 5, Expertise, nach Dreyfus und Dreyfus (1986) korrespondiert. Das unterste Niveau hingegen entspricht noch keiner DaZ-Kompetenz, während das erste Niveau die Kompetenzstufen 1–3 nach Dreyfus und Dreyfus unspezifisch abbildet. Mithilfe des Testinstruments aus dem Vorgängerprojekt würden sich die Personen auf dem Niveau 1 vermutlich differenzierter einer Kompetenzstufe zuordnen lassen; dies allerdings müssen Nachfolgestudien zeigen. Das Niveau 2 dürfte mit der Kompetenzstufe 4 nach Dreyfus und Dreyfus korrespondieren. Zwischen diesen vier Niveaus also differenziert der *DaZKom-Video-Test* nach dem Standardsetting.

In den Daten zeigte sich, dass sich der Großteil der Testpersonen in der deutschlandweit erhobenen Normierungsstichprobe auf dem ersten oder zweiten Kompetenzniveau befand, während ein nicht erkennbares DaZ-Niveau beinahe so selten vorkam wie das höchste Niveau der DaZ-Expertise. Die Ergebnisse lassen sich dabei auf die spezifische Situation der fachlichen DaZ-Inhalte im Lehramtsstudium beziehen. Erst seit kurzer Zeit und auch nur in wenigen Bundesländern sind DaZ-Inhalte verpflichtender Anteil des Lehramtsstudiums. In dieser kurzen Zeit aber wurde die Lehrkräfteausbildung im Bereich DaZ an vielen Universitäten umfassend erweitert und ausgebaut (vgl. Becker-Mrotzek und Woerfel 2020). Zum anderen rückt die Theorie-Praxis-Verzahnung immer mehr in den Fokus der Lehrkräftebildung, um professionelle Kompetenzen bei angehenden Lehrkräften ausbilden zu können (Rothland 2020). Das bedeutet, dass die realen Praxisanteile (betreute Praktika, Hospitationen, Unterrichtsbeobachtungen, Praxisprojekte u. Ä.) im Lehramtsstudium immer mehr zunehmen und gleichzeitig vermehrt Methoden gewählt werden, die praxisnahes Lernen ermöglichen, wie z. B. Videovignetten oder Simulationen (vgl. Berkel-Otto und Stander-Dulisch 2021; Peuschel et al. 2021). Viele seit langem praktizierende Lehrkräfte – auf Grund der Erfahrungsabhängigkeit von Expertise eigentlich Kernzielgruppe des Tests (vgl. Abschn. 2) – dürften also noch gar keine oder nur wenige DaZ-Lehrveranstaltungen absolviert haben, während viele Lehramtsstudierende in jüngster Zeit vermehrt entsprechende Angebote in Anspruch genommen haben dürften (vgl. Hecker 2021). Daher ist es nicht verwunderlich, dass sich nur sehr wenige Testpersonen in der Normierungsstichprobe auf einem Expertiseniveau befinden (vgl. Tab. 2), davon ein höherer Anteil praktizierender Lehrkräfte. Aus demselben Grund ist es plausibel, dass einige Testpersonen noch über keine messbare DaZ-Kompetenz verfügen (Niveau 0); davon ebenfalls ein höherer Anteil praktizierender Lehrkräfte. Es ist möglich, dass diese in ihrer Lehramtsausbildung

(noch) nicht mit der Thematik konfrontiert wurden und deshalb auch mit praktischer Erfahrung (noch) keine DaZ-Expertise aufbauen konnten.

Positiv überraschend ist dagegen die hohe Anzahl von Testpersonen in beiden Gruppen (angehende und praktizierende Lehrkräfte), die der Niveaustufe 2 bzw. der damit korrespondierenden Stufe 4 nach Dreyfus und Dreyfus zugeordnet werden konnten. Möglicherweise lässt sich die vergleichsweise hohe Anzahl an Lehrkräften auf diesem Niveau damit begründen, dass viele der Testpersonen an freiwillig besuchten Weiterbildungsveranstaltungen im Bereich DaZ teilnahmen. In diesem DaZ-Weiterbildungsrahmen fanden mehrfach Erhebungen mit praktizierenden Lehrkräften statt. Dies lässt bei den Teilnehmenden ein großes Interesse am Themengebiet vermuten, welches auch bereits in der jeweiligen Berufsvorgängerheit zu Reflexionsprozessen geführt haben könnte. Denkbar ist auch, dass es Anknüpfungspunkte zu anderen Fächern und Fachbereichen gibt, in denen die Testpersonen versiert sind, und deren Kenntnisse sie bei der Wahrnehmung und Bewältigung der videographierten Unterrichtssituationen unterstützt haben (vgl. Hecker et al. 2020b). Konform mit den oben beschriebenen Annahmen zur Ausbildungssituation ist es, dass sich auf den Niveaustufen 1 und 2 prozentual jeweils ein höherer Anteil Lehramtsstudierender als praktizierender Lehrkräfte befindet.

Weitergehende Analysen müssen zeigen, welche Personen mit welchem Hintergrund die im Standardsetting ermittelten Kompetenzniveaus erreicht haben. Auch inwiefern die universitären Lerngelegenheiten im Bereich DaZ mit dem Abschneiden im Test tatsächlich zusammenhängen – ob sie also etwa zum Erreichen eines der höheren Kompetenzniveaus beitragen konnten –, müssen weitergehende Auswertungen offenlegen. Da im Test auch ein Fragebogen zu Anzahl und Inhalt der DaZ-bezogenen universitären und außeruniversitären Lerngelegenheiten der Proband*innen integriert war (vgl. Ehmke und Lemmrich 2018), können und werden statistische Zusammenhänge noch berechnet werden. Ergebnisse der vorangegangenen Pilotierungsstudien zeigten hier zwar bislang ein recht uneinheitliches Bild und boten noch viele Interpretationsmöglichkeiten (vgl. Hecker et al. 2020b; Lemmrich et al. 2019, 2020). In künftigen Auswertungen aber werden dank des in diesem Beitrag beschriebenen gelungenen Standardsettings nun auch genauere Aussagen dazu möglich, welche Rolle den absolvierten DaZ-Lerngelegenheiten in Bezug auf das Erreichen ganz bestimmter DaZ-Kompetenzniveaus zugeschrieben werden kann. Dies erlaubt perspektivisch gezieltere Empfehlungen zum Mindestumfang und Inhalt universitärer DaZ-Lerngelegenheiten im Rahmen einer kompetenzorientierten Lehrkräftebildung, womit ein wichtiger Beitrag für eine empirisch fundierte Verbesserung geleistet werden kann. Diese entspricht gleichzeitig dem Ziel der Projekte *DaZKom* und *DaZKom-Video*, wonach die Testinstrumente ausschließlich eingesetzt werden sollen und können, um die Relevanz und die Qualität der Lerngelegenheiten in der Lehrer*innenausbildung zu dokumentieren bzw. zu evaluieren. Individuelle Kompetenzdiagnosen dagegen sind nicht geplant.

Auffällig ist, dass die resultierenden Schwellen (vgl. Abb. 4) eine hohe Entsprechung der im Prozess des Standardsettings ausgehandelten Stufen mit den Bewertungen der Aufgaben, wie sie im Kodierleitfaden angelegt sind, zeigen. Die empirisch ermittelte Schwierigkeit der Items bildete also die in den Kodierleitfäden angelegte Differenzierung ab; es war bei fast jedem Item am schwierigsten – benötigte also

die höchste Kompetenz –, einen Code 3 zu erhalten. So liegen fast alle Schwellen für Code 1 in Stufe 1, die Schwellen für Code 2 in Stufe 2, und so weiter. Diese Korrespondenz dürfte teilweise dadurch begründet sein, dass die Methodik des Standardsettings zu einem gewissen Grad von den bereits im Vorfeld des Settings festgelegten Differenzierungen, hier in Bezug auf die Bewertung der Aufgaben, abhängt. Denn die Differenzierung nach Schwierigkeiten wurde implizit durch die Bewertungsmaßstäbe in den Kodierleitfäden, an denen ebenfalls (andere) Fachexpert*innen beteiligt waren, festgelegt. Die Schwellensetzung der Fachexpert*innen im Standardsetting musste sich dadurch der Differenzierung in den Kodierleitfäden anschließen, da die Reihenfolge, in der die Item-Antwortcodes im OIB präsentiert wurden, sich nach der empirisch ermittelten Schwierigkeit richtete, die wiederum auf den Kodierleitfäden beruhte. Wo genau die Fachexpert*innen allerdings die Schwellen setzten, wurde im Setting auf Basis der inhaltlichen Anforderungen der Items ausgehandelt. Theoretisch wäre es möglich gewesen, dass die Expert*innen die erste Schwelle erst bei Antwortcodes gesetzt hätten, die im Kodierleitfaden den Code 2 erhalten hatten. Oder, dass sie die zweite Schwelle so gesetzt hätten, dass Antworten, die mit Code 2 bewertet wurden, dem höchsten Niveau zugeordnet worden wären. So hat das Standardsetzungsverfahren also eine zusätzliche, nochmals ausdifferenzierende Funktion. Die Stärke des Standardsettings im Gegensatz zu einer vereinfachten Festsetzung von Bewertungsstufen liegt damit darin, dass die Fachexpert*innengruppe aufgrund der inhaltlichen Anforderungen der Aufgaben sowie der kollaborativ-diskursiven Aushandlung zu der Stufenbildung gekommen ist.

Eine mögliche Einschränkung der hier berichteten Ergebnisse ist die bereits oben zur Erklärung herangezogene zu vermutende Positiv-Selektion der Stichprobe. Denn da die Teilnahme am Test stets freiwillig war, ist zu erwarten, dass insbesondere solche (angehenden) Lehrkräfte an der Studie teilnahmen, die ein Interesse am Thema DaZ auszeichnete. Trotz der deutschlandweit je Bundesland in Anzahl und Verpflichtung unterschiedlichen und teilweise kaum vorhandenen DaZ-Lerngelegenheiten überrascht es daher nicht, dass das Niveau 0 nur von vergleichsweise wenigen (angehenden) Lehrkräften erreicht wird. Auch die große Anzahl derjenigen auf den Niveaus 1 und 2 könnte mit der zu vermutenden Positiv-Selektion zusammenhängen. Künftige Auswertungen versprechen hier Klarheit.

Die Setzung der Schwellen kann ebenfalls diskutiert werden. Die Schwellen sind so gesetzt, dass eine Person auf Stufe 1 verortet wird, die im leichtesten Item mit einer Wahrscheinlichkeit von 50 % eine Bewertung von 1 erreicht, in allen anderen Items jedoch mit einer Wahrscheinlichkeit von mehr als 50 % noch null Punkte. Durch die Orientierung an den leichtesten Items kann zwar der Eindruck einer großzügigen Stufendefinition entstehen. Wir gehen jedoch davon aus, dass eine andere Setzung der Schwellen (z. B. mit Lösungswahrscheinlichkeiten von >60 % wie in PISA, vgl. OECD 2017) vermutlich die gleiche Stufung ergeben hätte, da sich alle Itemparameter in Richtung höhere Schwierigkeit verschoben hätten.

Eine Limitation der gewählten Methodik des Standardsettings ist es, dass der normative Prozess des Standardsettings und damit die gesetzten Standards abhängig von den teilnehmenden Fachexpert*innen sind. Um ein möglichst objektives Ergebnis zu erzielen, wurden deshalb für dieses Standardsetting unterschiedliche Expert*innen aus Wissenschaft und Praxis eingeladen. Auch wäre es rückblickend

denkbar gewesen, den am Standardsetting beteiligten Expert*innen die theoretischen Annahmen hinter dem zu standardisierenden Testinstrument so weit offenzulegen, dass eine korrespondierende Zuordnung von theoretisch angenommenen Stufen zu den diskursiv-kollaborativ bestimmten Niveaus erleichtert wird. Allerdings würde dann die eigene Expertise der beteiligten Fachexpert*innen nicht genutzt, was das Prozedere an sich wiederum infrage stellen würde.

Für weitere Forschungen lassen sich zwei Empfehlungen ableiten. Zunächst bleibt zu diesem Zeitpunkt noch unklar, inwieweit die im Rahmen des Standardsettings definierten Kompetenzniveaus tatsächlich einerseits, wie vermutet, mit den Stufen von Dreyfus und Dreyfus (1986), andererseits mit denen im *DaZKom*-Modell und dem *DaZKom*-Test des *DaZKom*-Projekts (vgl. Ehmke et al. 2018) korrespondieren. Auch ob das höchste im Rahmen dieser Studie definierte Kompetenzniveau, die DaZ-Expertise, qualitativ vergleichbar ist mit der Stufe der Expertise nach Dreyfus und Dreyfus, verlangt weitere Analysen. Ein Abgleich bekannter Expertisemerkmale mit den Ergebnissen künftiger qualitativer Analysen der Antworten derjenigen Proband*innen, die in der Normierungsstudie die Niveaus 2 und 3 erreicht haben, könnte hier zukünftig zur Aufklärung beitragen.

Die Pilotierungsstudien ließen den Schluss zu, dass die explizite Verbalisierung der Wahrnehmung im Test möglicherweise nicht identisch ist mit dem tatsächlichen impliziten Wahrnehmungsprozess (vgl. Hecker et al. 2020a; Hecker 2021); das könnte es erschwert haben, Expertise valide zu ermitteln. Dies ebenfalls im Rahmen von performanznaher Kompetenzerfassung weitergehend zu erforschen, stellt auch zukünftig eine spannende Herausforderung dar.

Es ist zu vermuten, dass sich das hier beschriebene standardisierte Testinstrument in der Zukunft noch weiter bewähren wird. Denn die Zielgruppe der Lehrkräfte mit DaZ-Expertise dürfte sich dank immer mehr Lerngelegenheiten an Universitäten, im Vorbereitungsdienst und in Weiterbildungen momentan in amtlich deutschsprachigen Regionen erst noch zahlreicher herausbilden.

Danksagung Wir bedanken uns herzlich bei den beteiligten Hilfskräften für die Unterstützung bei der Datenerhebung und bei den Lehrer*innen und Studierenden für ihre Teilnahme an unserer Studie.

Förderung Die vorliegende Studie entstand im Rahmen des Projekts *DaZKom-Video – Performanznahe Messung von Deutsch-als-Zweitsprache-Kompetenz bei (angehenden) Lehrkräften*. Dieses wurde mit Mitteln des Bundesministerium für Bildung und Forschung (BMBF) im Rahmen der Förderlinie *Kompetenzmodellierung und Kompetenzerfassung* (2017–2020) gefördert (01PK16001A und 01PK16001B).

Funding Open Access funding enabled and organized by Projekt DEAL.

Interessenkonflikt S.-L. Hecker, S. Falkenstern, S. Lemmrich, T. Ehmke, B. Koch-Priewe, A. Köker und U. Ohm geben an, dass kein Interessenkonflikt besteht.

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

Die in diesem Artikel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung

nicht nach gesetzlichen Vorschriften erlaubt ist, ist für die oben aufgeführten Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.

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

Literatur

- Baumann, B. (2017). Sprachförderung und Deutsch als Zweitsprache in der Lehrerbildung – ein deutschlandweiter Überblick. In M. Becker-Mrotzek, P. Rosenberg, C. Schroeder & A. Witte (Hrsg.), *Deutsch als Zweitsprache in der Lehrerbildung* (S. 9–26). Münster: Waxmann.
- Becker-Mrotzek, M., & Woerfel, T. (2020). Sprachsensibler Unterricht und Deutsch als Zweitsprache als Gegenstand der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König & M. Rothland (Hrsg.), *Handbuch Lehrerinnen- und Lehrerbildung* (S. 98–104). Bad Heilbrunn: Julius Klinkhardt.
- Berkel-Otto, L., & Stander-Dulisch, J. (2021). Szenarioorientierung als Befähigung zum unterrichtspraktischen Handeln im lehrkräftebildenden Modul „Sprachförderung und Transkulturelle Sensibilität“. In L. Berkel-Otto, K. Peuschel & S. Steinmetz (Hrsg.), *Theorie-Praxis-Verzahnung in der Lehrkräftebildung: Ergebnisse aus dem Netzwerk „Stark durch Diversität“* (S. 37–54). Münster: Waxmann.
- Berliner, D. C. (1992). The nature of expertise in teaching. In F. K. Oser, A. Dick & J.-L. Patry (Hrsg.), *Effective and responsible teaching: the new synthesis* (S. 277–248). San Francisco: Jossey-Bass.
- Blömeke, S., Gustafsson, J.-E., & Shavelson, R. (2015). Beyond dichotomies: competence viewed as a continuum. *Zeitschrift für Psychologie*, 223(1), 3–13. <https://doi.org/10.1027/2151-2604/a000194>.
- Bromme, R. (2001). Teacher expertise. In N. J. Smelser, P. B. Baltes & F. E. Weinert (Hrsg.), *International encyclopedia of the behavioral sciences: education* (S. 15459–15465). London: Pergamon.
- Bromme, R. (2008). Lehrerexpertise. In W. Schneider, M. Hasselhorn & J. Bengel (Hrsg.), *Handbuch der pädagogischen Psychologie* (Bd. 10, S. 159–167). Göttingen: Hogrefe.
- Bromme, R. (2014). *Der Lehrer als Experte. Zur Psychologie des professionellen Wissens* (Standardwerke aus Psychologie und Pädagogik, Bd. 7). Münster: Waxmann.
- Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, 39(3), 253–263. <https://doi.org/10.1111/j.1745-3984.2002.tb01117.x>.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting. A guide to establishing and evaluating performance standards on tests*. SAGE.
- Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine. The power of human intuition and expertise in the era of the computer*. Oxford: Blackwell.
- Ehmke, T., & Lemmrich, S. (2018). Bedeutung von Lerngelegenheiten für den Erwerb von DaZ-Kompetenz. In T. Ehmke, S. Hammer, A. Köker, U. Ohm & B. Koch-Priewe (Hrsg.), *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache* (S. 201–219). Münster: Waxmann.
- Ehmke, T., Hammer, S., Köker, A., Ohm, U., & Koch-Priewe, B. (Hrsg.) (2018). *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache*. Münster: Waxmann.
- Falkenstern, S., Hecker, S.-L., & Lemmrich, S. (2021). Lehrkräfteprofessionalisierung im Kontext von Heterogenität: DaZ-kompetenter Umgang mit Differenzordnungen im Unterricht. In I. Dirim & A. Wegner (Hrsg.), *Deutsch als Zweitsprache: Inter- und transdisziplinäre Zugänge* (S. 189–202). Opladen: Budrich. <https://doi.org/10.2307/j.ctv1hm8gfs.13>.
- Gibbons, P. (2002). *Scaffolding language, scaffolding learning. Teaching second language learners in the mainstream classroom*. Portsmouth: Heinemann.
- Gogolin, I., Dirim, I., Klinger, T., Lange, I., Lengyel, D., Michel, U., Neumann, U., Reich, H. H., Roth, H.-J., & Schwippert, K. (2011). *Förderung von Kindern und Jugendlichen mit Migrationshintergrund. Bilanz und Perspektiven eines Modellprogramms (FörMig Edition)*. Münster: Waxmann.
- Gültekin-Karakoç, N., Köker, A., Hirsch, D., Ehmke, T., Hammer, S., Koch-Priewe, B., & Ohm, U. (2016). Bestimmung von Standards und Stufen der Kompetenz angehender Lehrerinnen und Lehrer aller Fächer im Bereich Deutsch als Zweitsprache (DaZ). In B. Koch-Priewe & M. Krüger-Potratz (Hrsg.), *Qualifizierung für sprachliche Bildung. Programme und Projekte zur Professionalisierung von Lehrkräften und pädagogischen Fachkräften* (DDS-Die Deutsche Schule: Beiheft 13, S. 130–146). Münster: Waxmann.
- Haschke, L., Kampa, N., Hahn, I., & Köller, O. (2017). Setting standards to a scientific literacy test for adults using the item-Descriptor (ID) matching method. In S. Blömeke & J.-E. Gustafsson

- (Hrsg.), *Standard setting in education, methodology of educational measurement and assessment* (S. 319–339). Cham: Springer. https://doi.org/10.1007/978-3-319-50856-6_18.
- Hecker, S.-L. (2021). *Situationsspezifische Fähigkeiten im Bereich Deutsch als Zweitsprache bei (angehenden) Lehrkräften performanznah messen und fördern. Konzeptionelle Hürden sowie Ergebnisse eines videobasierten Testinstruments und Implikationen für die Lehrkräfteprofessionalisierung*. Bielefeld: Universität Bielefeld. <https://doi.org/10.4119/unibi/2951405>.
- Hecker, S.-L., & Nimz, K. (2020). Expertinnenratings zur Deutsch-als-Zweitsprache-Kompetenz von Lehrkräften. Eine Vorstudie zur Konzipierung eines videobasierten Testinstruments. In J. Stiller, C. Laschke, T. Nesyba & U. Salaschek (Hrsg.), *Berlin-Brandenburger Beiträge zur Bildungsforschung* (S. 39–64). Berlin: Peter Lang.
- Hecker, S.-L., Falkenstern, S., Lemmrich, S., & Ehmke, T. (2020a). Zum Verbalisierungsdilemma bei der Erfassung der situationsspezifischen Fähigkeiten von Lehrkräften. Am Beispiel der Domäne Deutsch als Zweitsprache. *Zeitschrift für Bildungsforschung*. <https://doi.org/10.1007/s35834-020-00268-1>.
- Hecker, S.-L., Falkenstern, S., & Lemmrich, S. (2020b). Zwischen DaZ-Kompetenz und Performanz. Zur Relevanz situationsspezifischer Fähigkeiten und zum Stand ihrer Ausbildung bei Lehrkräften aller Fächer in der Domäne DaZ. *HLZ*, 3(1), 565–584. <https://doi.org/10.4119/hlz-3374>.
- Köker, A., Rosenbrock-Agyei, S., Ohm, U., Carlson, S. A., Ehmke, T., Hammer, S., Koch-Priewe, B., & Schulze, N. (2015). DaZKom – Ein Modell von Lehrerkompetenz im Bereich Deutsch als Zweitsprache. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* (S. 177–205). Bad Heilbrunn: Julius Klinkhardt.
- Lemmrich, S., Hecker, S.-L., Klein, S., Ehmke, T., Köker, A., Koch-Priewe, B., & Ohm, U. (2019). Performanznahe und videobasierte Messung von DaZ-Kompetenz bei Lehrkräften: Skalierung und dimensionale Struktur des Testinstruments. In T. Ehmke, P. Kuhl & M. Pietsch (Hrsg.), *Lehrer. Bildung. Gestalten. Beiträge zur empirischen Forschung in der Lehrerbildung* (S. 190–204). Weinheim: Beltz Juventa.
- Lemmrich, S., Hecker, S.-L., Klein, S., Ehmke, T., Koch-Priewe, B., Köker, A., & Ohm, U. (2020). Linguistically responsive teaching in multilingual classrooms. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper & C. Lautenbach (Hrsg.), *Student learning in German higher education* (S. 125–140). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-27886-1_7.
- Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Hrsg.), *Setting performance standards. Foundations, methods, and innovations* (Bd. 2, S. 225–254). New York: Routledge.
- Lindmeier, A. M., Heinze, A., & Reiss, K. (2013). Eine Machbarkeitsstudie zur Operationalisierung aktionsbezogener Kompetenz von Mathematiklehrkräften mit videobasierten Maßen. *Journal für Mathematik-Didaktik*, 34(1), 99–119. <https://doi.org/10.1007/s13138-012-0046-6>.
- Neuweg, G. H. (2015). Kontextualisierte Kompetenzmessung. Eine Bilanz zu aktuellen Konzeptionen und forschungsmethodischen Zugängen. *Zeitschrift für Pädagogik*, 61(3), 377–383.
- OECD (2017). *PISA 2015 technical report*. Paris: OECD.
- Ohm, U. (2018). Das Modell von DaZ-Kompetenz bei angehenden Lehrkräften. In T. Ehmke, S. Hammer, A. Köker, U. Ohm & B. Koch-Priewe (Hrsg.), *Professionelle Kompetenzen angehender Lehrkräfte im Bereich Deutsch als Zweitsprache* (S. 73–91). Münster: Waxmann.
- de Oliveira, L. C., & Schleppegrell, M. J. (2015). *Focus on grammar and meaning*. Oxford: University Press.
- Pant, H. A., Tiffin-Richards, S. P., & Köller, O. (2010). Projekt Standard-Setting. Standard-Setting für Kompetenztests im Large-Scale-Assessment. In E. Klieme, D. Leutner & M. Kenk (Hrsg.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes* (Zeitschrift für Pädagogik: Beiheft 56, S. 175–188). Weinheim: Beltz.
- Peuschel, K., Petersen, I., Reble, R., Haberzettl, S., Otto, L., Niederhaus, C., Steinmetz, S., Rojahn, P., & Sondershaus, E. (2021). Der besondere Moment – Ergebnisse aus der Evaluation der Projekte „Sprachförderung PLUS“ im Netzwerk „Stark durch Diversität“. In L. Berkel-Otto, K. Peuschel & S. Steinmetz (Hrsg.), *Theorie-Praxis-Verzahnung in der Lehrkräftebildung: Ergebnisse aus dem Netzwerk „Stark durch Diversität“* (S. 13–36). Münster: Waxmann.
- Piowar, V. (2013). *Konzeptualisierung, Erfassung und Entwicklung von Kompetenzen im Klassenmanagement*. Freie Universität, Berlin. https://refubium.fu-berlin.de/bitstream/handle/fub188/13331/Piowar_Valentina.diss.pdf?sequence=1&isAllowed=y. Zugegriffen: Sep. 2022.
- Rothland, M. (2020). Theorie-Praxis-Verhältnis in der Lehrerinnen- und Theorie-Praxis-Verhältnis in der Lehrerinnen- und Lehrerbildung. In C. Cramer, J. König & M. Rothland (Hrsg.), *Handbuch Lehrerinnen- und Lehrerbildung* (S. 133–140). Bad Heilbrunn: Julius Klinkhardt.

- Schleppegrell, M.J. (2007). The linguistic challenges of mathematics teaching and learning: a research review. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, 23(2), 139–159.
- Schreiner, C., Harych, P., & Wiesner, C. (2020). Kompetenzstufen in Studien zur Kompetenzmessung im Vergleich: Konzepte, Entwicklung und Interpretation. In I. U. Greiner, F. Hofmann, C. Schreiner & C. Wiesner (Hrsg.), *Bildungsstandards. Kompetenzorientierung, Aufgabenkultur und Qualitätsentwicklung im Schulsystem* (S. 388–409). Münster: Waxmann.
- Wahl, D. (2002). Mit Training vom trägen Wissen zum kompetenten Handeln? *Zeitschrift für Pädagogik*, 48(2), 227–241.