UNIVERSITÄT LEIPZIG
Erziehungswissenschaftliche Fakultät

Bericht

Entwicklung eines Instruments zur Erfassung der kognitiven Komponente von Einstellungen zu Diagnostik im Schulkontext (EiDiLe)

Katrin Gottlebe, Jan Dörer, Isabel Berger & Brigitte Latzko

Universität Leipzig

Erziehungswissenschaftliche Fakultät

Marschnerstraße 31

04109 Leipzig

katrin.gottlebe@uni-leipzig.de

Abstract

Professional teaching competence includes knowledge, motivation, beliefs, and attitudes (Baumert & Kunter, 2013; Weinert, 2001). While teachers' diagnostic competence has gained increasing attention, attitudinal components have received relatively little consideration, despite their influence on professional action and knowledge acquisition (Pajares, 1992; Ajzen, 1991). This study addresses cognitive attitudes toward educational assessment as a component of diagnostic competence, understood in Weinert's (2001) sense of "wanting" alongside knowing and being able to act.

We present a newly developed questionnaire consisting of 22 items across four scales: *Functional beliefs* (usefulness for school-level decision-making), *Quality and consequences* (trust, accuracy, fairness), *Impact on teaching and learning* (support for instructional differentiation), and *Skepticism* (critical views of the practical relevance of assessment). The scale structure integrates both German-language frameworks and the internationally established TCoA model by Brown (2006).

The sample comprised N = 148 pre-service teachers (predominantly female, age 19–55, median: 23) from various school tracks. Confirmatory factor analysis indicated acceptable model fit (CFI = 0,85; TLI = 0,83; RMSEA = 0,057; SRMR = 0,083). All factor loadings were significant, with most items showing moderate to high values. Ten items with lower loadings were retained due to their theoretical relevance.

Reliability analyses showed acceptable to good internal consistency (Cronbach's α = 0,73 – 0,81), except for the Skepticism scale (α = 0,52), which was retained due to its conceptual importance. Initial evidence of construct validity emerged through expected correlations with Need for Cognition and Need to Evaluate.

The instrument enables a nuanced assessment of diagnostic beliefs and can be used to evaluate digital learning programs, extending beyond the measurement of knowledge gain. To support professional development, teacher education should explicitly address and reflect on diagnostic attitudes.

Zusammenfassung

Professionelle Unterrichtskompetenz umfasst Wissen, Motivation, Einstellungen und Überzeugungen (Baumert & Kunter, 2013; Weinert, 2001). Während die diagnostische Kompetenz von Lehrkräften zunehmend in den Fokus rückt, wurden Einstellungsaspekte bislang nur wenig berücksichtigt, obwohl sie für das professionelle Handeln und die Aneignung neuen Wissens zentral sind (Pajares, 1992; Ajzen, 1991). Die vorliegende Studie widmet sich daher der Erfassung kognitiver Einstellungen zur pädagogisch-psychologischen Diagnostik als Teil diagnostischer Kompetenz im Sinne des „Wollens" (Weinert, 2001).

Vorgestellt wird der Prozess der Neuentwicklung eines Fragebogens mit 22 Items, die auf vier Skalen verteilt sind: Funktionale Überzeugungen (Nutzen für Steuerung und Evaluation), Qualität & Fairness (Vertrauen, Genauigkeit, Fairness), Nutzen für Lehren und Lernen (diagnostisch unterstützte Unterrichtsgestaltung) und Alltagsskepsis (kritische Haltungen zur Relevanz von Diagnostik im Schulalltag). Die Skalen basieren auf deutschsprachigen Konzepten sowie dem international etablierten TCoA-Ansatz von Brown (2006).

Die Stichprobe umfasste N = 148 Lehramtsstudierende (überwiegend weiblich, Alter 19–55, Median: 23 Jahre) aus verschiedenen Schulformen. Eine konfirmatorische Faktorenanalyse ergab eine akzeptable Modellpassung (CFI = 0,85, TLI = 0,83, RMSEA = 0,057, SRMR = 0,083). Alle Faktorladungen waren signifikant, der Großteil der Items zeigte moderate bis hohe Ladungen. Zehn Items mit geringeren Ladungen wurden aufgrund ihrer inhaltlichen Relevanz im Modell belassen.

Die Reliabilitätsanalyse ergab akzeptable bis gute interne Konsistenzen (Cronbach's α = 0,73–0,81), mit Ausnahme der Skala Alltagsskepsis (α = 0,52), die jedoch beibehalten wurde. Erste Hinweise auf die Validität zeigten sich in erwartungskonformen Zusammenhängen mit Need for Cognition und Need to Evaluate.

Das Instrument erlaubt eine differenzierte Erfassung diagnostischer Überzeugungen und kann auch zur Evaluation von Lernangeboten eingesetzt werden, über den reinen Wissenserwerb hinaus. Zur Förderung professioneller Entwicklung sollten Einstellungen zur Diagnostik frühzeitig thematisiert und reflektiert werden.

Inhalt

Tabellenverzeichnis 5
1 Einleitung 6
2 Theoretischer Hintergrund 7
2.1 Skalenkonzept 7
2.2 Pilotierung 8
3 Methodisches Vorgehen 10
3.1 Durchführung 10
3.2 Stichprobe 10
3.4 Weitere Instrumente 11
3.4.1 Soziodemografische Daten 11
3.4.2 Konstruktvalidität 11
3.5 Analyse 11
4 Ergebnisse der Skalenkonstruktion 12
4.1 Pilotversion II 12
4.1.1 Itemkennwerte 12
4.1.2 Messtheoretische Prüfung 12
4.1.3 Itemauswahl für die finale Version EiDiLe 13
4.2 Finales Instrument EiDiLe 14
4.2.1 Item- und Skalenkennwerte 14
4.2.2 Messtheoretische Prüfung 15
4.2.3 Reliabilität 17
4.2.4 Validität 18
4.2.5 Auswertungshinweise 19
5 Kurzfassung 20
Literaturverzeichnis 21
Anhang 24

Tabellenverzeichnis

Tab. 1: Pilotversion II 9
Tab. 2: Pilot II - Itemkennwerte der Unterskalen 12
Tab. 3: Pilot II - Skalenkennwerte 13
Tab. 4: EiDiLe - Itemkennwerte 15
Tab. 5: EiDiLe - Faktorladungen der CFA 16
Tab. 6: EiDiLe – Mittelwerte der Unterskalen nach Geschlecht 17
Tab. 7: EiDiLe - Reliabilitätskennwerte 17
Tab. 8: EiDiLe - Korrelationen mit Need for Cognition (NfC), Need to Evaluate (NtE), und Selbstwirksamkeitserwartung (ASKU) 18

1 Einleitung

Professionelle Unterrichtskompetenz wird als Zusammenspiel von professionellem Wissen, motivationalen Orientierungen, Einstellungen und Überzeugungen verstanden (z. B. Baumert & Kunter, 2006). Nach dem sogenannten PISA-Schock in Deutschland Anfang der 2000er-Jahre kam es zu einem deutlichen Anstieg an Forschung zur diagnostischen Kompetenz und Bewertungskompetenz von Lehrkräften. Diese Forschung beruhte auf der Annahme, dass Defizite in diesen Bereichen zu Wissenslücken bei Schüler:innen beitragen (Abs, 2007). In der Folge wurden verschiedene Modelle zur Beschreibung diagnostischer Kompetenzen entwickelt (z. B. Kaiser et al., 2013; Klug et al., 2013) sowie Studien zur diagnostischen Kompetenz von Lehrkräften durchgeführt (z.B. Urhahne & Wijnia, 2021). Der Forschungsschwerpunkt lag bislang jedoch vor allem auf der Genauigkeit von Lehrerurteilen und dem Wissensaspekt von Kompetenz (vgl. Weinert, 2001). Es wurde untersucht, was Lehrkräfte über Leistungsbewertung wissen, was Lehramtsstudierende im Studium dazu lernen und wie dieses Wissen in der Aus- und Fortbildung gefördert werden kann (z. B. DeLuca & Klinger, 2010).

In jüngerer Zeit richtet sich die Aufmerksamkeit verstärkt auf professionelle Kompetenzen, die für die Umsetzung von Digitalisierung und Inklusion in Schulen erforderlich sind (Pit-ten Cate et al., 2018). Wir argumentieren, dass diagnostische Kompetenzen eine zentrale Komponente professioneller Kompetenz in inklusiven Bildungskontexten darstellen und daher weiterhin intensiv beforscht werden sollten (Latzko & Gottlebe, 2022). Trotz der bisherigen Forschung lag der empirische Fokus stark auf dem Wissen über Diagnostik, während Einstellungen bisher wenig berücksichtigt wurden.

Allgemein wird angenommen, dass über die Quantität und Qualität von Lerngelegenheiten hinaus auch deren Nutzung, beeinflusst durch individuelle Faktoren wie Vorwissen und Einstellungen, entscheidend für den Kompetenzerwerb in der Lehramtsausbildung ist (Baier et al., 2019). Wissen über schulische Leistungsbewertung steht in einem moderaten Zusammenhang mit aktuellen Einstellungen zur Leistungsbewertung (Quilter & Gallini, 2000). Studien deuten darauf hin, dass einstellungs- und motivationsbezogene Faktoren auch den Einsatz digitaler Technologien im Unterricht beeinflussen (Bürger et al., 2021).

Forschung zu den Einstellungen von Lehrkräften zur Leistungsbewertung befasst sich mit Überzeugungen zur Rolle von Tests, zu Benotungspraktiken und zur Wirkung von Bewertungen auf das Lernen und die Motivation von Schüler:innen (z. B. Brown, 2006). Bislang ist jedoch kaum erforscht, inwieweit spezifische Einstellungen und Überzeugungen zu Leistungsbewertung und Unterstützung im Studium und in Praktika beeinflusst werden können und die bisherigen Befunde hierzu sind uneinheitlich (Brown, 2008; Xu & Brown, 2016). In Deutschland wurde bislang nur wenig zur Einstellung von Lehrkräften gegenüber Leistungsbewertung geforscht, was sich auch in der fehlenden Verfügbarkeit etablierter Erhebungsinstrumente widerspiegelt. Auf Grundlage dieses Forschungsdesiderats wurde im vorliegenden Projekt ein Fragebogen entwickelt, der die Einstellungen zu pädagogisch-psychologischer Diagnostik von Personen im Lehramt systematisch erfasst.

2 Theoretischer Hintergrund

2.1 Skalenkonzept

Einstellungen werden als mehrdimensionales Konstrukt betrachtet, das aus kognitiven, affektiven und verhaltensbezogenen Komponenten bestehen (Rosenberg & Hovland, 1960; Zanna & Rempel, 1988). Sie beschreiben relativ stabile Bewertungen eines Objekts und umfassen Überzeugungen, Gefühle und Handlungstendenzen gegenüber diesem Objekt (Eagly & Chaiken, 1993). In diesem Fragebogen des Projekts wird vor allem die kognitive Komponente von Einstellungen erfasst, also das Wissen und die Überzeugungen von Lehramtsstudierenden über pädagogisch-psychologische Diagnostik.

Einstellungen gelten als zentrale Einflussfaktoren für die Intentionsbildung und das Verhalten von Lehrkräften (Ajzen, 1991). Sie beeinflussen nicht nur diagnostisches Handeln, sondern auch die Organisation und Aneignung neuen Wissens (Pajares, 1992). Damit tragen sie wesentlich zum professionellen Lernen und zur Entwicklung diagnostischer Kom-petenzen bei.

Vor diesem Hintergrund wird der Einstellungsaspekt als Bestandteil diagnostischer Kompetenz im Sinne des Kompetenzmodells von Weinert (2001) verstanden, das Wissen, Können und Wollen als zentrale Komponenten professionellen Handelns umfasst. In Anlehnung daran lassen sich Einstellungen zu Diagnostik als Ausdruck des „Wollens“ begreifen und stellen somit eine eigenständige, theoretisch fundierte Teilkompetenz im Bereich pädagogisch-psychologischer Diagnostik dar (vgl. Latzko & Gottlebe, 2022).

Inhaltlich bezieht sich die Einstellung auf den Gegenstand der schulischen Diagnostik, der im deutschsprachigen Raum und auch internationale sowohl pädagogische als auch gesellschaftliche Zielsetzungen umfasst (Brookhart, 2011; G. T. L. Brown, 2004; Hesse & Latzko, 2017; Ingenkamp & Lissmann, 2008; Schrader, 2013). Während diagnostische Maßnahmen einerseits der Optimierung von Lernprozessen dienen (improvement), erfüllen sie andererseits Selektions- und Qualifikationsfunktionen (student accountability). Darüber hinaus sind auch Funktionen der Schulevaluation (school accountability) zu berücksichtigen, etwa in Bezug auf die Bewertung von Unterrichtsqualität.

Gerade weil Diagnostik in Ausbildung und Berufspraxis eine zentrale, aber auch potenziell konfliktträchtige Rolle einnimmt, ist davon auszugehen, dass Lehramtsstudierende bereits mit bestimmten Überzeugungen und Bewertungen in das Studium eintreten.

Die inhaltliche Strukturierung der Items erfolgte theoriegeleitet und basiert auf einer Kombination deutschsprachiger und internationaler konzeptioneller Zugänge zu pädagogisch-psychologischer Diagnostik. Einerseits wurde einschlägige deutschsprachige Fachliteratur zu Zielen und Aufgaben schulischer Diagnostik herangezogen (Hesse & Latzko, 2017; Ingenkamp & Lissmann, 2008). Andererseits wurde der international etablierte Teachers’ Conceptions of Assessment (TCoA)-Ansatz von Brown (2006) berücksichtigt, der diagnostische Einstellungen im Bildungskontext als mehrdimensionale Konstrukte modelliert. Diese Perspektive ist besonders anschlussfähig an aktuelle Forschungsarbeiten im englischsprachigen Raum (vgl. Yan & Cheng, 2015) und erlaubt eine differenzierte Erfassung verschiedener Überzeugungsmuster.

Aus dem TCoA-Ansatz wurden zentrale Dimensionen übernommen und in den deutschen schulischen Kontext übersetzt und angepasst. Dazu zählen unter anderem:

- Verantwortung von Schule und Lehrkräften: die Überzeugung, dass Diagnostik der institutionellen Rechenschaftspflicht dient
- Verantwortung der Lernenden: die Vorstellung, dass Diagnostik vor allem der Leistungsmessung der Schüler:innen dient
- Förderung: die Annahme, dass Diagnostik einen positiven Einfluss auf Unterricht und Lernprozesse hat
- Irrelevanz: die Sichtweise, dass Diagnostik im schulischen Alltag wenig hilfreich oder sogar hinderlich ist

Diese Dimensionen wurden für den Fragebogen herangezogen, adaptiert und im Rahmen einer Pilotierung mit deutschsprachigen Lehramtsstudierenden auf ihre Verständlichkeit und Passung hin überprüft.

2.2 Pilotierung

Die erste Pilotierung des Instruments fand im Sommer 2022 mit 145 Lehramtsstudierenden statt. Die Ergebnisse zeigten, dass sich das Modell von Brown nicht auf die Stichprobe übertragen ließ. Zudem wiesen einige Items ungünstige Itemkennwerte auf (Berger & Latzko, 2022; Berger et al., 2023).

Nachdem die erste Pilotierung weiteren Anpassungsbedarf ergab, wurden Items entfernt, modifiziert oder umformuliert, um besser in den deutschsprachigen Kontext zu passen. Der Itempool wurde so ergänzt, dass die postulierten vier Dimensionen der kognitiven Überzeugungen von Lehrkräften gegenüber Diagnostik: (1) Bilanzierung von Leistungen, (2) Qualität und Konsequenz, (3) Einfluss auf Lehren und Lernen, (4) Diagnostik wird ignoriert, inhaltlich passend beschrieben werden. So sollte sichergestellt werden, dass die Skalen über ausreichend Items für den zweiten Pilot verfügten.

Die hier im Bericht vorgelegte Version wird dementsprechend als Pilot II (siehe Tab. 1) bezeichnet, aus der die finale Version des Instruments EiDiLe (siehe Abschnitt 4.2) erstellt wird.

Tab. 1: Pilotversion II

Unterskala		Inhalt
1 Bilanzierung von Leistung	1	Durch Diagnostik kann die Qualität des Unterrichts überprüft werden.
	2	Diagnostik ist eine gute Möglichkeit, um die Qualität einer Schule zu evaluieren.
	3	Durch Diagnostik wird festgestellt, ob die Schüler:innen Bildungsstandards erreicht haben.
	4	Durch Diagnostik wird festgestellt, wie viel die Schüler:innen im Unterricht gelernt haben.
	5	Durch Diagnostik werden die Leistungen von Schüler:innen bewertet.
	6	Durch Diagnostik können Lernfortschritte sichtbar gemacht werden.
	7	Durch Diagnostik wird die Leistung der Lehrkräfte beurteilt.
2 Qualität und Konsequenz	*8*	*Durch Diagnostik werden Schüler:innen in Kategorien eingeteilt. (*)*
	9	Diagnostische Ergebnisse sind vertrauenswürdig.
	10	Diagnostik ist unfair gegenüber Schüler:innen. (*)
	11	Diagnostik hindert Lehrkräfte daran, andere wichtige Dinge zu tun. (*)
	12	Diagnostik ist meist ungenau. (*)
	13	Auf diagnostische Ergebnisse kann man sich verlassen.
	14	*Diagnostik sollte vor jeder pädagogischen Entscheidung durchgeführt werden.*
3 Einfluss auf Lehren und Lernen	15	Diagnostik ist ein wichtiger Bestandteil des Unterrichts.
	16	Durch Diagnostik wird der Unterricht effektiver.
	17	Durch Diagnostik wird der Unterricht adaptiver gestaltet.
	18	Diagnostik ermöglicht Differenzierung im Unterricht.
	19	Diagnostik hat kaum Auswirkungen auf den Unterricht. (*)
	20	Diagnostik ist für Lernprozesse nicht so wichtig. (*)
	21	Ohne Diagnostik würden Schüler:innen besser lernen. (*)
	22	Diagnostik ist für guten Unterricht nicht so wichtig. (*)
4 Diagnostik wird ignoriert	*23*	*Lehrkräfte führen zwar Diagnostik durch, nutzen die Ergebnisse aber kaum.* (*)
	24	Diagnostische Ergebnisse werden abgeheftet und ignoriert. (*)
	25	Diagnostik wird oft zu wichtig genommen. (*)

Anmerkung: Die Items werden jeweils mittels einer fünfstufigen Antwortskala von "stimme nicht zu“ (1), „stimme eher nicht zu“ (2), „weder noch“ (3), „stimme eher zu“ (4), „stimme zu“ (5) beantwortet; mit (*)-markierten Items wurden invertiert

3 Methodisches Vorgehen

3.1 Durchführung

Um eine möglich repräsentative Stichprobe zu gewährleisten, wurden Teilnehmer aus verschiedenen Studiengängen und Fachrichtungen gewählt. Hierbei sind sowohl Grundschullehramt, Lehramt am Gymnasium, Oberschule und der Sonderpädagogik vertreten. Die Befragten wurden über verschiedene Kanäle zur Teilnahme an der Umfrage eingeladen. Hierzu gehörten verschiedene Moodle-Kurse und die sozialen Medien. Die Teilnahme ist freiwillig und anonym, um eine möglich ehrliche und unverfälschte Beantwortung der Fragen zu gewährleisten. Der Erhebungszeitraum fand vom 10.07.2023 bis zum 14.12.2023 statt.

Die Befragung erfolgte durch die Online-Umfrage-Applikation LimeSurvey. Durchschnittliche Bearbeitungsdauer des Fragebogens beträgt 15 Minuten.

Der verwendete Fragebogen folgt einem kontingenten Ablauf, bei der man nur zur nächsten Fragestellung kommt, wenn man die vorherige Frage beantwortet hat. Aufgrund dessen entstehen auch viele Missings, deshalb wurden in der Analyse auch nur Befragte betrachtet, die mindestens den ersten Fragekomplex bezüglich der kognitiven Komponente vollständig beantwortet haben.

3.2 Stichprobe

Insgesamt wurden N = 228 Lehramtsstudierende an der Universität Leipzig befragt. In die Analysen einbezogen wurden nur Teilnehmende, die den ersten Fragekomplex bezüglich der kognitiven Komponente vollständig beantwortet haben. Nach der Bereinigung fehlender Werte beinhaltet die Stichprobe n = 148 Personen.

Die Teilnehmenden waren zwischen 19 und 55 Jahre alt. Der Median lag bei 23 Jahren (n = 33), die häufigste Altersgruppe bildeten die 22-Jährigen (n = 40). Die jüngste befragte Person war 19, die älteste 55 Jahre alt.

Von den Befragten (n = 148) gaben 118 an, weiblich zu sein; 26 identifizierten sich als männlich. Vier Personen wählten die Angabe "divers".

Hinsichtlich des Bildungshintergrundes ergibt sich folgendes Bild: Die Mehrheit der Befragten studierte Grundschullehramt (n = 57) oder Gymnasiallehramt (n = 57). Weitere Teilnehmende gehörten zu den Studienrichtungen Oberschule (n = 15) und Sonderpäda-gogik (n = 19).

Die meisten Teilnehmenden befanden sich im 8. Fachsemester (n = 45), gefolgt vom 6. Semester (n = 39). Besonders häufig wurden Sprachfächer gewählt (n = 71). Die häufigsten Fächer waren: Sprache (n = 84), Musik/Kunst/Werken/Sport (n = 48), Naturwissenschaft (n = 44), Geschichte/Gemeinschaftskunde/Recht/Wirtschaft (n = 31), Religion/Ethik (n = 11) und sonstige Fächer (n = 28).

Von den insgesamt 148 Befragten gaben 146 an, das Modul „BWI 06 – Diagnostik, Förderung, Beratung“ bereits absolviert zu haben; zwei Personen hatten es noch nicht belegt.

3.4 Weitere Instrumente

3.4.1 Soziodemografische Daten

Folgende Variablen wurden erfasst: Alter, Geschlecht, studiertes Lehramt, aktuelles Fachsemester, studierte Fächer, ob das bildungswissenschaftliche Modul Diagnostik bereits belegt wurde und in welchem Semester sich die Befragten befinden.

3.4.2 Konstruktvalidität

Zur Überprüfung der Konstruktvalidität wurden Zusammenhänge der entwickelten Skalen mit etablierten psychologischen Konstrukten untersucht. Neben fachlich verwandten Maßen kamen auch theoretisch abgrenzbare Konstrukte zum Einsatz, um Hinweise auf konvergente und diskriminante Validität zu erhalten.

Für die konvergente Validierung wurden Skalen zur Allgemeinen Selbstwirksamkeitserwartung (Beierlein et al., 2014), zum Need for Cognition (Beißert et al., 2014) sowie zum Need to Evaluate (Collani, 2003) herangezogen. Es wurde angenommen, dass insbesondere positive diagnostische Überzeugungen (z. B. funktionale Nutzenüberzeugungen, wahrgenommene Qualität) moderat positiv mit Selbstwirksamkeit und kognitiver Aktivierungsbereitschaft korrelieren sollten. Zugleich wurde erwartet, dass skeptischere Haltungen (z. B. Alltagsskepsis) geringe oder negative Zusammenhänge mit diesen Konstrukten aufweisen. Angaben zu den Itemcharakteristika und Model Fit sind im Anhang zu finden (Tab. A1).

3.5 Analyse

Alle statistischen Analysen wurden in R (Version 4.3.0) durchgeführt. Die Reliabilität der Skalen wurden mithilfe von Cronbachs Alpha bestimmt. Zusätzlich wurden Itemkennwerte berechnet, um die psychometrischen Eigenschaften der Skalen zu evaluieren.

Für die Datenanalyse wurde zunächst die deskriptive Statistik auf Skalen- und Itembasis durchgeführt, um zentrale Tendenzen (Mittelwert und Median), Streuungen (SD) sowie Quartile zu bestimmen. Zur Überprüfung der internen Struktur des Instruments wurde eine konfirmatorische Faktorenanalyse (CFA) mit dem R-Paket lavaan durchgeführt (Rosseel, 2012). Anpassungen an der Modellstruktur wurden dann unternommen, wenn Items eine niedrige Faktorenladung aufwiesen (<0.30) (Boateng et al., 2018; T. A. Brown, 2015). Diesbezüglich wurden Items entfernt und mit Hilfe von Modifikationsindizes Modellverbesserungen vorgenommen, indem Items in die jeweiligen passenden Subskalen hinzugenommen wurden (Boateng et al. 2018). Die Bewertung der Modellgüte erfolgte anhand folgender Fit-Indizes und empfohlener Grenzwerte CFI ≥ 0,95 (Hu & Bentler, 1999), TLI ≥ 0,90 (Bentler, 1992), RMSEA ≤ 0,08 (Browne & Cudeck, 1992) und SRMR ≤ 0,8 (Hu & Bentler, 1999).

4 Ergebnisse der Skalenkonstruktion

4.1 Pilotversion II

4.1.1 Itemkennwerte

Die Schwierigkeitsindizes liegen zwischen p_i = 2,4 und 4,6. Die Itemvarianzen liegen zwischen var_i = 0,31 und 1,52 (vgl. Tab. A2). Bezogen auf die 5-stufige Antwortskala gibt insbesondere Items in der Skala 3, die eher leicht sind und damit eher viel Zustimmung erhalten.

Die Trennschärfekoeffizienten liegen zwischen r_i = -0,04 und 0,62. Ungefähr die Hälfte der Items liegt damit oberhalb der angestrebten Schwelle von 0,4. Die betreffenden Items sind hauptsächlich in den Unterskalen 2 und 3 zu finden. In Unterskala 4 finden sich ausschließlich Items unterhalb der Schwelle (vgl. Tab. 2). Diese Items erscheinen ungeeignet, um zwischen Personen mit unterschiedlichen Ausprägungen des gemessenen Merkmals zu unterscheiden.

Tab. 2: Pilot II - Itemkennwerte der Unterskalen

	Mittelwert (*SD*)	Itemschwierigkeiten
Leistungsbilanzierung	3,16 (0,67)	2,4 - 3,7
Qualität und Konsequenz	3,47 (0,60)	3,3 - 4,2
Einfluss auf Lehren und Lernen	4,13 (0,46)	3,8 - 4,6
Unwichtig	3,87 (0,53)	2,8 - 4,0

Anmerkung: ausführlich in Tab. A2 im Anhang

4.1.2 Messtheoretische Prüfung

Die konfirmatorische Faktorenanalyse (CFA) wurde mit einer Stichprobe von N = 148 Beobachtungen und 56 Modellparametern durchgeführt. Das User-Model zeigte mit einem Chi-Quadrat-Wert von $\chi^2(269) = 411{,}32$, $p < 0{,}05$, eine signifikante Abweichung von perfekter Modellanpassung.

Die Fit-Indizes deuten auf eine akzeptable, jedoch verbesserungswürdige Modellpassung hin: Der Comparative Fit Index (CFI) lag bei 0,82, der Tucker-Lewis Index (TLI) bei 0,79, somit beide unterhalb des empfohlenen Schwellenwertes von 0,90. Der Root Mean Square Error of Approximation (RMSEA) betrug 0,060 (90-%-KI = [0,048; 0,071]), der zugehörige p-Wert für H_0: RMSEA ≤ 0,050 lag bei 0,084. Der Standardized Root Mean Square Residual (SRMR) betrug 0,088 und lag damit knapp über der empfohlenen Obergrenze von 0,08.

Alle Faktorladungen waren signifikant ($p < 0{,}05$). Acht Items zeigten geringe standardisierte Ladungen ($\lambda < 0{,}50$), darunter Items 14 ($\lambda = 0{,}23$), 23 ($\lambda = 0{,}31$) und 8 ($\lambda = 0{,}38$) (siehe Tab. A3 im Anhang).

Die deskriptiven Auswertungen der Gesamt- und Unterskalen zeigen linksschiefe Verteilungen (Tab. 3).

Tab. 3: Pilot II - Skalenkennwerte

	Mittelwert (*SD*)	Min-Max	Q1; Q3
EiDiLe	3,66 (0,41)	2,75 – 4,54	3,42; 3,96
Leistungsbilanzierung	3,16 (0,67)	1,43 - 4,71	2,71; 3,71
Qualität und Konsequenz	3,47 (0,60)	1,50 - 4,83	3,17; 3,83
Einfluss auf Lehren und Lernen	4,13 (0,46)	2,25 - 5,00	3,88; 4,50
Unwichtig	3,87 (0,53)	2,00 - 5,00	3,33; 4,33

4.1.3 Itemauswahl für die finale Version EiDiLe

Die Auswahl der Items für die finale Version erfolgte vor allem aufgrund von theoretischen Überlegungen und von Gruppendiskussionen zur Inhaltsvalidität, nur sekundär aufgrund der empirischen Itemkennwerte. Es war nicht das Ziel, eine stark homogene Skala zu entwickeln und die interne Konsistenz zu maximieren. Vielmehr erschien es sinnvoll, die Optimierung der Validität zu betonen, indem relativ heterogene Handlungsbereiche im Berufsleben von Lehrern angesprochen werden.

Das Item „Diagnostik hindert Lehrkräfte daran, andere wichtige Dinge zu tun.“ (Item 11) verbleibt trotz der vergleichsweisen geringen Ladung im Modell, wird aber in Skala 4 aufgenommen. Ausschlaggebend hierfür waren inhaltliche Überlegungen sowie der Umstand, dass diese Skala ansonsten lediglich aus zwei Items bestanden hätte. Gleichwohl sollte dieses Item, ebenso wie andere Items mit grenzwertiger Ladung, im Rahmen zukünf-tiger Modellanpassungen erneut kritisch geprüft werden.

Die genannten Items verbleiben vorläufig im Modell, bedürfen jedoch einer sorgfältigen Prüfung im Rahmen weiterführender Analysen. Items 14 („Diagnostik sollte vor jeder pädagogischen Entscheidung durchgeführt werden.“) und 23 („Lehrkräfte führen zwar Diagnostik durch, nutzen die Ergebnisse aber kaum.“) wiesen besonders niedrige Faktorladungen auf und konnten auch im Rahmen alternativer Modellrespezifikationen keinem anderen latenten Faktor sinnvoll zugeordnet werden. Bei Item 23 kommt erschwerend hinzu, dass es zwei inhaltlich unterschiedliche Aspekte, die Durchführung diagnostischer Maßnahmen und die Nutzung der Ergebnisse, in einer Frage kombiniert, was die inhaltliche Klarheit und Messgenauigkeit beeinträchtigt. Das Item „Durch Diagnostik werden Schüler:innen in Kategorien eingeteilt“ (Item 8) wurde aus dem Modell ausgeschlossen, da es eine auffällig hohe negative Residualkorrelation mit anderen Items derselben Skala aufwies. Dies deutet auf eine inhaltliche Inkonsistenz bzw. eine abweichende Interpretation im Vergleich zu den übrigen Items hin. Aufgrund dieser empirischen und inhaltlichen Erwägungen wurden die drei Items aus dem Modell ausgeschlossen.

Nach der Überarbeitung der Skalenstruktur, einschließlich Itemausschluss und Neuzuordnungen, wurden die Skalenbezeichnungen im Vergleich zur Pilotversion klarer und treffender formuliert (siehe Abschnitt 4.2).

4.2 Finales Instrument EiDiLe

Das finale Instrument EiDiLe besteht aus 22 Items, die sich vier inhaltlich unterschiedlichen Skalen zuordnen lassen.

Skala 1 „Funktionale Überzeugungen“ (7 Items) legt den Fokus auf die Funktionen diagnostischer Verfahren im Schulsystem. Diese Skala erfasst Überzeugungen darüber, welche Funktionen diagnostische Verfahren im schulischen Kontext erfüllen können. Im Mittelpunkt stehen Vorstellungen zur Nutzung von Diagnostik zur Bewertung individueller Leistungen von Schüler:innen, zur Überprüfung von Lernzuwachs und Bildungsstandards sowie zur Evaluation von Unterrichts- und Schulqualität. Die Items spiegeln eine systemische Perspektive auf Diagnostik wider, die über die rein individuelle Leistungsrückmeldung hinausgeht.

Skala 2 „Qualität & Fairness“ (4 Items) setzt den Fokus auf die Überzeugungen bezüglich Qualität, Genauigkeit und Fairness diagnostischer Verfahren. Diese Skala erfasst die subjektive Einschätzung der Genauigkeit, Fairness und Vertrauenswürdigkeit von Diagnostik im schulischen Kontext. Die Items spiegeln Bewertungen diagnostischer Verfahren wider, sowohl im Hinblick auf ihre technische Güte als auch auf ihre Gerechtigkeit gegenüber Schüler:innen. Im Gegensatz zur ersten Skala, die sich auf die Funktionen diagnostischer Maßnahmen konzentriert, thematisiert diese Skala Zweifel oder Vertrauen in die Qualität der diagnostischen Ergebnisse selbst.

Skala 3 „Nutzen für Lehren und Lernen“ (8 Items) fokussiert auf den Nutzen diagnostischer Verfahren für die Gestaltung von Lehr-Lern-Prozessen. Diese Skala erfasst Überzeugungen zum konkreten Beitrag von Diagnostik zur Unterrichtsgestaltung und zur Förderung von Lernprozessen. Sie misst, inwieweit Diagnostik als nützlich für Differenzierung, Individualisierung und Effektivitätssteigerung im Unterricht wahrgenommen wird. In die Skala fließen auch kritische Einschätzungen ein, die den Nutzen von Diagnostik für das Lernen infrage stellen (invertierte Items). Im Unterschied zur ersten Skala, die abstraktere Funktionen von Diagnostik beleuchtet, liegt hier der Fokus auf der unmittelbaren Relevanz im Klassenzimmer.

Skala 4 „Alltagsskepsis“ (3 Items) greift praktische Einwände gegen den Einsatz von Diagnostik auf. Diese Skala erfasst kritische Haltungen gegenüber dem Einsatz von Diagnostik im schulischen Alltag. Die Items spiegeln Überzeugungen wider, dass Diagnostik im pädagogischen Alltag eher hinderlich sei, überbewertet werde oder letztlich folgenlos bleibe. Damit adressiert die Skala typische Einwände, die Lehrkräfte im Zusammenhang mit der Umsetzung diagnostischer Maßnahmen äußern. Im Unterschied zu anderen Skalen, die auf theoretische Funktionen oder qualitative Merkmale diagnostischer Verfahren fokussieren, stehen hier die wahrgenommene praktische Relevanz und Umsetzbarkeit im Vordergrund.

4.2.1 Item- und Skalenkennwerte

Die Schwierigkeitsindizes für die finale Skala EiDiLe liegen zwischen p_i = 2,4 und 4,6. Bezogen auf die 5-stufige Antwortskala gibt es nach wie vor Items, die eher leicht sind und damit eher viel Zustimmung erhalten.

Die Trennschärfekoeffizienten liegen zwischen r_i = 0,07 und 0,68. Ungefähr die Hälfte der Items liegt damit oberhalb der angestrebten Schwelle von 0,4 (Hu & Bentler 1999). Die

betreffenden Items sind hauptsächlich in den Unterskalen 2 und 3 zu finden. In Unterskala 4 finden sich ausschließlich Items unterhalb der Schwelle (vgl. Tab. 4). Diese Items erscheinen ungeeignet, um zwischen Personen mit unterschiedlichen Ausprägungen des gemessenen Merkmals zu unterscheiden.

Tab. 4: EiDiLe - Itemkennwerte

	Itemschwierigkeiten	Trennschärfen
EiDiLe-Gesamtskala	2,4 - 4,6	0,07 – 0,68
Funktionale Überzeugungen	2,4 - 4,3	0,07 – 0,45
Qualität & Fairness	3,0 - 4,2	0,24 – 0,48
Nutzen für Lehren und Lernen	3,8 - 4,6	0,42 – 0,68
Alltagsskepsis	3,9 - 4,0	0,37 – 0,46

Anmerkung: detaillierte Itemschwierigkeiten in Tab. A2 im Anhang

4.2.2 Messtheoretische Prüfung

Die überarbeitete CFA mit N = 148 und 52 Modellparametern ergab eine verbesserte Modellpassung: $\chi^2(224) = 330{,}45$, $p < 0{,}05$ (vgl. Tab. A4 im Anhang). Die Fit-Indizes verbesserten sich: CFI =0,85 (zuvor 0,82), TLI = 0,83 (zuvor 0,79). Der RMSEA sank leicht auf 0,057 (90-%-KI = [0,043; 0,069]), mit einem p-Wert von 0,197 für H_0: RMSEA ≤ 0,050. Der SRMR lag bei 0,083 und näherte sich damit der empfohlenen Grenze an (Hu & Bentler 1999).

Alle Faktorladungen blieben signifikant ($p < 0{,}05$), die meisten Items zeigten moderate bis hohe Ladungen. Zehn Items wiesen jedoch weiterhin geringe Ladungen zwischen 0,45 und 0,5 auf. Aufgrund der inhaltlichen Bedeutung der Items wurde entschieden, diese im Fragebogen zu belassen. Diese Items sollten im Rahmen zukünftiger Modellrespezifikationen erneut kritisch geprüft werden.

Tab. 5: EiDiLe - Faktorladungen der CFA

Skala		Itemwortlaut	Faktor-ladung	std. λ
Funktionale Überzeugungen	1	Durch Diagnostik kann die Qualität des Unterrichts überprüft werden.	0,49	0,48
	2	Diagnostik ist eine gute Möglichkeit, um die Qualität einer Schule zu evaluieren.	0,61	0,52
	3	Durch Diagnostik wird festgestellt, ob die Schüler:innen Bildungsstandards erreicht haben.	0,70	0,64
	4	Durch Diagnostik wird festgestellt, wie viel die Schüler:innen im Unterricht gelernt haben.	0,74	0,63
	5	Durch Diagnostik werden die Leistungen von Schüler:innen bewertet.	0,61	0,49
	6	Durch Diagnostik können Lernfortschritte sichtbar gemacht werden.	0,42	0,49
	7	Durch Diagnostik wird die Leistung der Lehrkräfte beurteilt.	0,49	0,45
Qualität & Fairness	9	Diagnostische Ergebnisse sind vertrauenswürdig.	0,38	0,49
	10	Diagnostik ist unfair gegenüber Schüler:innen. (*)	0,51	0,52
	12	Diagnostik ist meist ungenau. (*)	0,45	0,49
	13	Auf diagnostische Ergebnisse kann man sich verlassen.	0,39	0,51
Nutzen für Lehren und Lernen	15	Diagnostik ist ein wichtiger Bestandteil des Unterrichts.	0,52	0,61
	16	Durch Diagnostik wird der Unterricht effektiver.	0,60	0,68
	17	Durch Diagnostik wird der Unterricht adaptiver gestaltet.	0,39	0,52
	18	Diagnostik ermöglicht Differenzierung im Unterricht.	0,68	0,68
	19	Diagnostik hat kaum Auswirkungen auf den Unterricht. (*)	0,40	0,47
	20	Diagnostik ist für Lernprozesse nicht so wichtig. (*)	0,45	0,61
	21	Ohne Diagnostik würden Schüler:innen besser lernen. (*)	0,28	0,5
	22	Diagnostik ist für guten Unterricht nicht so wichtig. (*)	0,53	0,67
Alltagsskepsis	11	Diagnostik hindert Lehrkräfte daran, andere wichtige Dinge zu tun. (*)	0,45	0,49
	24	Diagnostische Ergebnisse werden abgeheftet und ignoriert. (*)	0,46	0,47
	25	Diagnostik wird oft zu wichtig genommen. (*)	0,56	0,59

Anmerkung: (*)- markierte Items sind invertiert; Items 8, 14 und 23 sind nicht Teil der finalen Version

Die deskriptiven Auswertungen der Gesamt- und Unterskalen zeigen linksschiefe Verteilungen (Tab. 6). Frauen (M = 3,70) erzielten auf der EiDiLe Gesamtskala signifikant höhere Werte als Männer (M = 3,50), $t(39{,}60) = -2{,}51$, $p = 0{,}0016$, 95%-KI [-0,36, -0,04]. Höhere Werte für Frauen zeigten sich auch auf den Subskalen „Qualität & Fairness“ ($t(39{,}60) = -2{,}81$, $p = 0{,}008$) und „Lehren und Lernen“ ($t(36{,}03) = -2{,}58$, $p = 0{,}014$). Für „Funktionale Überzeugungen“ ($t(38{,}43) = -0{,}22$, $p = 0{,}829$) und „Alltagsskepsis“ ($t(34{,}91) = -0{,}86$, $p = 0{,}395$) bestanden keine signifikanten Unterschiede. Zur Berechnung des t-Tests wurden nur männliche und weibliche Teilnehmende erfasst.

Tab. 6: EiDiLe – Mittelwerte der Unterskalen nach Geschlecht

	Mittelwert (*SD*)	Weiblich	Männlich	T-Test *t*(df), *p*
EiDiLe-Gesamtskala	3,65 (0,40)	3,70 (0,40)	3,50 (0,35)	-2,51(40,35), *p* = 0,016
Funktionale Überzeugungen	3,16 (0,68)	3,18 (0,69)	3,15 (0,65)	-0,22(38,43), *p* = 0,829
Qualität & Fairness	3,47 (0,54)	3,53 (0,54)	3,22 (0,49)	-2,81(39,60), *p* = 0,008
Nutzen für Lehren und Lernen	4,13 (0,51)	4,20 (0,50)	3,91 (0,52)	-2,58(36,03), *p* = 0,014
Alltagsskepsis	3,87 (0,67)	3,90 (0,66)	3,77 (0,72)	-0,86(34,91), *p* = 0,395

4.2.3 Reliabilität

Zur Überprüfung der internen Konsistenz wurde Cronbachs Alpha für die Gesamtskala sowie die vier Subskalen berechnet. Die Ergebnisse sind in Tabelle 6 dargestellt.

Tab. 7: EiDiLe - Reliabilitätskennwerte

	Cronbachs α	Item-Item-Korrelation	Itemanzahl
EiDiLe-Gesamtskala	0,80	0,17	22
Funktionale Überzeugungen	0,73	0,28	7
Qualität & Fairness	0,63	0,30	4
Nutzen für Lehren und Lernen	0,81	0,35	8
Alltagsskepsis	0,52	0,27	3

Die EiDiLe-Gesamtskala weist mit einem Cronbachs Alpha von $\alpha = 0{,}80$ eine gute interne Konsistenz auf. Die durchschnittliche Inter-Item-Korrelation liegt bei $r = 0{,}17$.

Die Subskala Funktionale Überzeugungen zeigt mit $\alpha = 0{,}73$ eine akzeptable interne Konsistenz. Die durchschnittliche Inter-Item-Korrelation beträgt $r = 0{,}28$. Eine Itemanalyse ergab, dass der Ausschluss einzelner Items die Reliabilität nicht wesentlich verbessern würde.

Für die Subskala Qualität & Fairness ergibt sich ein Alpha-Wert von $\alpha = 0{,}63$, was eine eher mäßige interne Konsistenz darstellt. Die durchschnittliche Korrelation zwischen den Items liegt bei $r = 0{,}30$. Auch hier würde das Entfernen einzelner Items die Reliabilität nicht nennenswert steigern.

Die Subskala Nutzen für Lehren und Lernen weist mit $\alpha = 0{,}81$ eine gute interne Konsistenz auf. Die durchschnittliche Inter-Item-Korrelation liegt bei $r = 0{,}35$. Keines der Items beeinträchtigt die Reliabilität negativ.

Die Subskala Alltagsskepsis zeigt mit $\alpha = 0{,}52$ eine geringe interne Konsistenz. Die durchschnittliche Inter-Item-Korrelation liegt bei $r = 0{,}27$. Eine Analyse der Itemauswirkungen ergab, dass insbesondere das Item 25 die interne Konsistenz schwächt (α bei Ausschluss: 0,36). Auch das Item 11 senkt die Reliabilität deutlich (α bei Ausschluss: 0,38). Aufgrund des theoretischen Gehalts und der geringen Itemanzahl wurde jedoch auf eine Itemreduktion verzichtet. Die Werte sind daher mit Vorsicht zu interpretieren.

4.2.4 Validität

Zur Überprüfung der Einsetzbarkeit der Validierungsinstrumente wurde zunächst geprüft, ob die jeweiligen Modellannahmen erfüllt sind (z. B. Skalenstruktur, Reliabilität). Die vorliegenden Ergebnisse deuten darauf hin, dass die eingesetzten Skalen für die angestrebte Validierung geeignet sind und keine methodischen Bedenken hinsichtlich der Anwendbarkeit der Forschungsinstrumente im gegebenen Stichprobenkontext bestehen (siehe Tabelle A1 im Anhang).

Tab. 8: EiDiLe - Korrelationen mit Need for Cognition (NfC), Need to Evaluate (NtE), und Selbstwirksamkeitserwartung (ASKU)

	ASKU	NfC	NtE
EiDiLe-Gesamtskala	$r = 0{,}16$; $p = 0{,}10$	$r = 0{,}22$; $p = 0{,}02$ *	$r = 0{,}27$; $p = 0{,}01$ **
Funktionale Überzeugungen	$r = 0{,}02$; $p = 0{,}84$	$r = 0{,}18$; $p = 0{,}06$	$r = 0{,}32$; $p = 0{,}001$ ***
Qualität & Fairness	$r = 0{,}15$; $p = 0{,}13$	$r = 0{,}22$; $p = 0{,}03$ *	$r = 0{,}07$; $p = 0{,}48$
Nutzen für Lehren und Lernen	$r = 0{,}12$; $p = 0{,}24$	$r = 0{,}05$; $p = 0{,}60$	$r = 0{,}11$; $p = 0{,}26$
Alltagsskepsis	$r = 0{,}24$; $p = 0{,}01$**	$r = 0{,}17$; $p = 0{,}08$	$r = 0{,}15$; $p = 0{,}13$

Anmerkung: N = 106

Die beobachteten Korrelationen stützen die Annahmen zur konvergenten und diskriminanten Validität teilweise (vgl. Tab. 7). Die Gesamtskala EiDiLe zeigt signifikante positive Zusammenhänge mit dem Need for Cognition (NfC) ($r = 0{,}22$, $p = 0{,}02$) und dem Need to Evaluate (NtE) ($r = 0{,}27$, $p = 0{,}01$), nicht jedoch mit der Allgemeinen Selbstwirksam-

keitserwartung (r = 0,16, p = 0,09). Auf Subskalenebene korreliert Funktionale Überzeu-gungen erwartungsgemäß moderat positiv mit NtE (r = 0,32, p = 0,02), während Qualität & Fairness insbesondere mit NfC (r = 0,22, p = 0,02) zusammenhängt. Für die Subskalen Nutzen für Lehren und Lernen sowie Alltagsskepsis zeigen sich keine konsistenten oder signifikanten Zusammenhänge mit den Validierungsskalen.

Insgesamt ergeben sich somit Hinweise auf eine partielle konvergente Validität, insbesondere in Bezug auf kognitive Bewertungs- und Verarbeitungsstile.

4.2.5 Auswertungshinweise

Zur Berechnung der individuellen Ausprägung der EiDiLe wird aus den 22 Rohwertpunkten der Antworten ein Mittelwert (Skalenwert) gebildet. Bei der Vergabe der Rohwertpunkte ist darauf zu achten, dass die Items 10, 11, 12, 19, 20, 21, 22, 24, 25 invers kodiert werden. Für die Berechnung des Mittelwerts werden anschließend die 22 Antworten zu einem Gesamtrohwert aufsummiert und durch die Anzahl der bearbeiteten Items geteilt. Der Wertebereich für die Mittelwerte liegt zwischen 1 und 5.

5 Kurzfassung

Diagnostische Zielsetzung

Der Fragebogen erfasst die kognitiven Einstellungen von Lehramtsstudierenden zur pädagogisch-psychologischen Diagnostik als Teil diagnostischer Kompetenz im Sinne des „Wollens“ (Weinert, 2001). Das Instrument kann auch zur Evaluation von Lernangeboten im Lehramtsstudium und Fortbildungen eingesetzt werden.

Aufbau

Das Instrument EiDiLe misste die Ausprägung von Einstellung zu Diagnostik im Schulkontext anhand von 22 Items auf vier Skalen, die unterschiedliche Aspekte diagnostischer Überzeugungen bei Lehrkräften abbilden: Funktionale Überzeugungen, Qualität & Fairness, Nutzen für Lehren und Lernen sowie Alltagsskepsis. Die Items werden jeweils mittels einer fünfstufigen Antwortskala von "stimme nicht zu“ (1), „stimme eher nicht zu“ (2), „weder noch“ (3), „stimme eher zu“ (4), „stimme zu“ (5) beantwortet Zur Berechnung der individuellen Ausprägung Der EiDiLe wird aus den Rohwertpunkten der Antworten auf den 22 Items ein Mittelwert gebildet. Bei der Vergabe der Rohwertpunkte ist darauf zu achten, dass die Items 10, 11, 12, 19, 20, 21, 22, 24, 25 invers kodiert wurden und für die Bildung des Skalenwerts umkodiert werden müssen.

Grundlagen und Konstruktion

Die Konstruktion der EiDiLe basiert auf internationalen (Brown, 2006) und deutschsprachigen Konzepten schulischer Diagnostik und differenzieren zwischen funktio-nalem Nutzen, Qualität, Beitrag zum Unterricht und alltagsskeptischen Haltungen. Die EiDiLe wurde anhand einer Stichprobe von Lehramtsstudierenden evaluiert und wird in deutschsprachiger Form vorgelegt.

Gütekriterien

Reliabilität: Drei der vier Subskalen sowie die Gesamtskala weisen mit Cronbachs Alpha-Werten zwischen 0,73 und 0,81 eine akzeptable bis gute interne Konsistenz auf. Die durchschnittlichen Inter-Item-Korrelationen liegen im erwartbaren Bereich und sprechen für eine inhaltlich stimmige Skalenstruktur. Die Subskala Alltagsskepsis zeigt mit $\alpha = 0{,}52$ eine geringe Reliabilität, was auf die geringe Itemanzahl und inhaltliche Heterogenität zurückzuführen ist; sie wurde dennoch beibehalten, da sie zentrale Einwände von Lehrkräften adressiert.

Validität: Die Korrelationen mit externen Konstrukten stützen die Annahmen zur konvergenten und diskriminanten Validität nur teilweise. Die Gesamtskala zeigt signifikante positive Zusammenhänge mit dem Need for Cognition ($r = 0{,}22$, $p = 0{,}022$) und dem Need to Evaluate ($r = 0{,}27$, $p = 0.006$), nicht jedoch mit der Allgemeinen Selbstwirksamkeitserwartung ($r = 0{,}16$, $p = 0{,}095$). Auf Subskalenebene besteht ein signifikanter Zusammenhang zwischen „Qualität & Fairness“ und dem Need for Cognition sowie zwischen „Funktionale Überzeugungen“ und dem Need to Evaluate. Für die weiteren Subskalen ergeben sich keine konsistenten Zusammenhänge. Insgesamt weisen die Ergebnisse auf eine partielle konvergente Validität hin, insbesondere in Bezug auf kognitive Bewertungs- und Verarbeitungsstile.

Literaturverzeichnis

Abs, H. J. (2007). Überlegungen zur Modellierung diagnostischer Kompetenz bei Lehrerinnen und Lehrern. In M. Lüders & J. Wissinger (Hrsg.), *Forschung zur Lehrerbildung: Kompetenzentwicklung und Programmevaluation /* (S. 63–84). Waxmann.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Baier, F., Decker, A.-T., Voss, T., Kleickmann, T., Klusmann, U. & Kunter, M. (2019). What makes a good teacher? The relative importance of mathematics teachers' cognitive ability, personality, knowledge, beliefs, and motivation for instructional quality. *British Journal of Educational Psychology*, *89*(4), 767–786. https://doi.org/10.1111/bjep.12256

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

Beierlein, C., Kovaleva, A., Kemper, C. J. & Rammstedt, B. (2014). *Allgemeine Selbstwirksamkeit Kurzskala (ASKU).* https://doi.org/10.6102/zis35

Beißert, H., Köhler, M., Rempel, M. & Beierlein, C. (2014). *Eine deutschsprachige Kurzskala zur Messung des Konstrukts Need for Cognition: Die Need for Cognition Kurzskala (NFC-K)*, *2014/32*.

Bentler, P. M [P. M.] (1992). On the fit of models to covariances and methodology to the Bulletin. *Psychological bulletin*, *112*(3), 400–404. https://doi.org/10.1037/0033-2909.112.3.400

Berger, I. & Latzko, B. (2022, 21. Juni). *Measuring pre-service teachers attitudes.* Carl von Ossietzky University. European Association for Research on Learning and Instruction (EARLI) SIG 11, Oldenburg.

Berger, I., Latzko, B. & Gottlebe, K. (2023). Teachers attitudes towards classroom assessment - Constructing and testing an instrument. In European Association for Research on Learning and Instruction (Hrsg.), *JURE 2023 Book of Abstracts* (S. 16).

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R. & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, *Volume 6 - 2018*. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2018.00149

Brookhart, S. M. (2011). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3–12. https://doi.org/10.1111/j.1745-3992.2010.00195.x

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 301–318. https://doi.org/10.1080/0969594042000304609

Brown, G. T. L. (2006). Teachers' Conceptions of Assessment: Validation of an Abridged Version. *Psychological Reports*, *99*(1), 166–170. https://doi.org/10.2466/pr0.99.1.166-170

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2. ed.). *Methodology In The Social Sciences*. Guilford Press.

Browne, M. W. & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bürger, N., Haselmann, S., Baumgart, J., Prinz, G., Girnat, B., Meisert, A., Menthe, J., Schmidt-Thieme, B. & Wecker, C. (2021). Jenseits von Professionswissen: Eine systematische Überblicksarbeit zu einstellungs- und motivationsbezogenen Einflussfaktoren auf die Nutzung digitaler Technologien im Unterricht. *Zeitschrift für Erziehungswissenschaft*, *24*(5), 1087–1112. https://doi.org/10.1007/s11618-021-01050-3

Collani, G. von. (2003). *Bedürfnis nach Bewertung (Need to Evaluate).* https://doi.org/10.6102/zis59

DeLuca, C. & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 419–438. https://doi.org/10.1080/0969594X.2010.516643

Eagly, A. H. & Chaiken, S. (1993). *The psychology of attitudes*. *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Hesse, I. & Latzko, B. (2017). *Diagnostik für Lehrkräfte* (3., vollst. überarb. u. erw. Aufl.). Verlag Barbara Budrich (UTB).

Hu, L. & Bentler, P. M [Peter M.] (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ingenkamp, K. & Lissmann, U. (2008). *Lehrbuch der pädagogischen Diagnostik* (6. Aufl.). *Beltz Pädagogik*. Beltz. https://content-select.com/media/moz_viewer/519cc17d-a7bc-4831-b7a1-253d5dbbeaba

Kaiser, J., Retelsdorf, J., Südkamp, A. & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgments. *Learning and Instruction*, *28*, 73–84. https://doi.org/10.1016/j.learninstruc.2013.06.001

Klug, J., Bruder, S., Kelava, A., Spiel, C. & Schmitz, B. (2013). Diagnostic competence of teachers: A process model that accounts for diagnosing learning behavior tested by means of a case scenario. *Teaching and Teacher Education*, *30*, 38–46. https://doi.org/10.1016/j.tate.2012.10.004

Latzko, B. & Gottlebe, K. (2022). Der Nutzen der pädagogisch-psychologischen Diagnostik für eine diversitätssensitive und inklusive, pädagogische Praxis. In K. Kramer & B. Hoyer (Hrsg.), *FAU Lehren und Lernen - 8. „INDIVIDUELL FÖRDERN" - wissenschaftlicher Hintergrund sowie Ansatzpunkte aus und für die Praxis: Fostering INDIVIDUAL LEARNING - scientific background and starting points from and for professional teaching practice* (S. 43–67). FAU University Press.

Pajares, M. F. (1992). Teachers' Beliefs and Educational Research: Cleaning Up a Messy Construct. *Review of Educational Research*, *62*(3), 307–332. https://doi.org/10.3102/00346543062003307

Pit-ten Cate, I. M., Markova, M., Krischler, M. & Krolak-Schwerdt, S. (2018). Promoting Inclusive Education: The Role of Teachers' Competence and Attitudes. *Insights on Learning Disabilities: From Prevailing Theories to Validated Practices*, *15*, 49+. https://link.gale.com/apps/doc/A547630810/AONE?u=anon~df78d0d3&sid=googleScholar&xid=a79515a9

Quilter, S. M. & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, *36*(2), 115–131. https://doi.org/10.1080/08878730009555257

Rosenberg, M. J. & Hovland, C. I. (1960). Cognitive, affective and behavioral components of attitudes. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson & J. W. Brehm

(Hrsg.), *Attitude organization and change: An analysis of consistency among attitude components. (Yales studies in attitude and communication.), Vol. III. Attitude organization and change: An analysis of consistency among attitude components. (Yales studies in attitude and communication.), Vol. III* (S. 1–14). Yale Univer. Press.

Rosseel, Y. (2012). lavaan : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Schrader, F.-W. (2013). Diagnostische Kompetenz von Lehrpersonen. *Beiträge zur Lehrerinnen- und Lehrerbildung*, *31*(2), 154–165.

Urhahne, D. & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, *32*, 100374. https://doi.org/10.1016/j.edurev.2020.100374

Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Hrsg.), *Defining and selecting key competencies* (S. 45–65). Hogrefe & Huber.

Yan, Z. & Cheng, E. C. K. (2015). Primary teachers' attitudes, intentions and practices regarding formative assessment. *Teaching and Teacher Education*, *45*, 128–136. https://doi.org/10.1016/j.tate.2014.10.002

Zanna, M. P. & Rempel, J. K. (1988). Attitudes: A new look at an old concept. In D. Bar-Tal & A. W. Kruglanski (Hrsg.), *The social psychology of knowledge* (S. 315–354). Cambridge University Press.

Anhang

Tab. A1: Itemcharakteristika der Instrumente Need for Cognition (NFC), Need to Evaluate (NTE), und Selbstwirksamkeitserwartung (ASKU) 25
Tab. A2: Pilot II Itemkennwerte 25
Tab. A3: CFA für 4-faktorielles Modell nach der Pilotierung (Pilotversion II) 27
Tab. A4: Pilot II vs. finales Modell: Ergebnisse der CFA im Vergleich 28

Tab. A1: Itemcharakteristika der Instrumente Need for Cognition (NFC), Need to Evaluate (NTE), und Selbstwirksamkeitserwartung (ASKU)

	N	Mittelwert (SD)	Item-schwierigkeiten	Cronbachs Alpha	Inter-Item-Korrelation	Korrelation mit EiDiLe
NfC	72	4,73 (0,87)	3,9 - 6,0	0,60	0,27	r = 0,252 p = 0,03**
NtE	71	3,1 (0,46)	2,5 - 4,1	0,73	0,15	r = 0,28 p = 0,02**
ASKU	72	3,9 (0,54)	3,8 - 4,0	0,78	0,55	r = 0,148 p = 0,214

Anmerkung: *es finden sich keine signifikanten Geschlechtsunterschiede

Tab. A2: Pilot II Itemkennwerte

Unter-skala		Inhalt	Schwierig-keitsindex	Item-varianz	Itemtrenn-schärfe
1	1	Durch Diagnostik kann die Qualität des Unterrichts überprüft werden.	3,86	1,02	0,55
	2	Diagnostik ist eine gute Möglichkeit, um die Qualität einer Schule zu evaluieren.	2,89	1,42	0,57
	3	Durch Diagnostik wird festgestellt, ob die Schüler:innen Bildungsstandards erreicht haben.	3,09	1,22	0,57
	4	Durch Diagnostik wird festgestellt, wie viel die Schüler:innen im Unterricht gelernt haben.	3,11	1,39	0,62
	5	Durch Diagnostik werden die Leistungen von Schüler:innen bewertet.	2,64	1,52	0,45
	6	Durch Diagnostik können Lernfortschritte sichtbar gemacht werden.	4,15	0,74	0,54
	7	Durch Diagnostik wird die Leistung der Lehrkräfte beurteilt.	2,36	1,20	0,44
2	*8*	*Durch Diagnostik werden Schüler:innen in Kategorien eingeteilt. (*)*	*3,00*	*1,13*	*-0,04*
	9	Diagnostische Ergebnisse sind vertrauenswürdig.	3,70	0,62	0,3
	10	Diagnostik ist unfair gegenüber Schüler:innen. (*)	4,01	0,99	0,22
	11	Diagnostik hindert Lehrkräfte daran, andere wichtige Dinge zu tun. (*)	3,8	0,85	0,26
	12	Diagnostik ist meist ungenau. (*)	3,53	0,85	0,18
	13	Auf diagnostische Ergebnisse kann man sich verlassen.	3,28	0,61	0,45
	14	*Diagnostik sollte vor jeder pädagogischen Entscheidung durchgeführt werden.*	*3,28*	*1,37*	*0,40*

Unter-skala		Inhalt	Schwierig-keitsindex	Item-varianz	Itemtrenn-schärfe
3	15	Diagnostik ist ein wichtiger Bestandteil des Unterrichts.	3,97	0,74	0,42
	16	Durch Diagnostik wird der Unterricht effektiver.	3,91	0,77	0,57
	17	Durch Diagnostik wird der Unterricht adaptiver gestaltet.	3,8	0,55	0,42
	18	Diagnostik ermöglicht Differenzierung im Unterricht.	4,45	0,62	0,43
	19	Diagnostik hat kaum Auswirkungen auf den Unterricht. (*)	3,80	0,75	0,18
	20	Diagnostik ist für Lernprozesse nicht so wichtig. (*)	4,34	0,55	0,31
	21	Ohne Diagnostik würden Schüler:innen besser lernen. (*)	4,56	0,31	0,35
	22	Diagnostik ist für guten Unterricht nicht so wichtig. (*)	4,26	0,64	0,39
4	*23*	*Lehrkräfte führen zwar Diagnostik durch, nutzen die Ergebnisse aber kaum.* (*)	*2,78*	*0,88*	*0,06*
	24	Diagnostische Ergebnisse werden abgeheftet und ignoriert. (*)	3,83	0,85	0,10
	25	Diagnostik wird oft zu wichtig genommen. (*)	3,98	0,93	0,18

Anmerkung: mit (*)- markierte Items sind invertiert; kursiv gedruckte Items sind nicht Teil der finalen Version

Tab. A3: CFA für 4-faktorielles Modell nach der Pilotierung (Pilotversion II)

Skala		Item	Faktorladung	Standardisierte Faktorladung
Leistungs-bilanzierung	1	dia_quali	0,45	0,50
	2	dia_schulquali	0,70	0,60
	3	dia_bildstand	0,70	0,67
	4	dia_lern	0,84	0,72
	5	dia_leistung	0,49	0,41
	6	dia_fortschr	0,32	0,48
	7	dia_beurt	0,51	0,52
Qualität und Konsequenzen	8	dia_kat*	0,41	0,38
	9	dia_vertrau	0,43	0,59
	10	dia_unfair*	0,43	0,47
	11	dia_hindert*	0,38	0,43
	12	dia_ungenau*	0,50	0,57
	13	dia_verlass	0,45	0,60
	14	dia_entscheidung	0,27	0,23
Einfluss auf Lehren und Lernen	15	dia_bestand	0,55	0,67
	16	dia_effekt	0,55	0,64
	17	dia_adapt	0,43	0,63
	18	dia_diff	0,49	0,72
	19	dia_ausw*	0,48	0,53
	20	dia_lprozess*	0,48	0,65
	21	dia_besser*	0,31	0,56
	22	dia_nichtwichtig*	0,55	0,68
Diagnostik wird ignoriert	23	dia_ergeb*	0,28	0,31
	24	dia_igno*	0,40	0,44
	25	dia_wichtig*	0,57	0,60

Anmerkung: N = 96; mit (*) -markierte Items sind invertiert

Tab. A4: Pilot II vs. finales Modell: Ergebnisse der CFA im Vergleich

Kennwerte der CFA	Pilot II	Finales Modell
χ^2	411,32 df = 269, $p < 0.001$	330,45 df = 224, $p < 0.001$
$\Delta\chi^2$, df = 1	-	80,87
CFI	0,82	0,85
TLI	0,79	0,83
RMSEA	0,060	0,057
SRMR	0,088	0,083

Anmerkung: n = 148