Behrendt, Stefan; Köllner, Jan; Kögler, Kristina; Sälzer, Christine; Just, Andreas

Konstruktion und psychometrische Prüfung eines Tests zur Diagnostik mathematischer Studieneingangsleistungen

Zeitschrift für empirische Hochschulforschung : ZeHf 7 (2023) 1, S. 74-95

Quellenangabe/ Reference:
Behrendt, Stefan; Köllner, Jan; Kögler, Kristina; Sälzer, Christine; Just, Andreas: Konstruktion und psychometrische Prüfung eines Tests zur Diagnostik mathematischer Studieneingangsleistungen - In: Zeitschrift für empirische Hochschulforschung : ZeHf 7 (2023) 1, S. 74-95 - URN: urn:nbn:de:0111-pedocs-342208 - DOI: 10.25656/01:34220; 10.3224/zehf.v7i1.06

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-342208
https://doi.org/10.25656/01:34220

in Kooperation mit / in cooperation with:

https://www.budrich.de

Konstruktion und psychometrische Prüfung eines Tests zur Diagnostik mathematischer Studieneingangsleistungen

Stefan Behrendt, Jan Köllner, Kristina Kögler, Christine Sälzer, Andreas Just

Zusammenfassung: Eingangsvoraussetzungen im Bereich Mathematik sind für den Erfolg in der Studieneingangsphase von MINT-Studiengängen von übergeordneter Bedeutung. Dennoch werden mathematische Basisfähigkeiten aus der Sekundarstufe I in Wiederholungs- und Unterstützungsmaßnahmen häufig vernachlässigt. Gleichermaßen fehlen geeignete diagnostische, qualitätsgeprüfte Instrumente für diesen Zweck. Der Beitrag stellt ein reliables und sowohl inhaltlich als auch differenziell sowie prognostisch valides computerbasiertes Instrument zur Diagnose dieser Fähigkeiten vor. Eine Papier- und eine Online-Version messen dasselbe Konstrukt. Die Online-Version enthält ein Instant-Feedback, welches sowohl Leistungs- als auch Verbesserungsrückmeldungen integriert. Im Ausblick werden noch ausstehende, zentrale Entwicklungs- und Prüfungsschritte thematisiert.

Schlüsselwörter: Mathematik der Sekundarstufe I, Online-Self-Assessment, Instant-Feedback, Studieneingangstest, MINT-Studiengänge, Studieneingangsphase

Development and psychometric verification of a diagnostic test for mathematical study entry performance

Summary: Entry-level prerequisites in mathematics are of paramount importance for success in STEM undergraduate programs. Nevertheless, basic mathematical skills from lower secondary school are often neglected in repetition and support measures. Similarly, there is a lack of appropriate diagnostic, quality-controlled instruments for this purpose. This paper presents a reliable and both content as well as differentially and prognostically valid computer-based instrument for diagnosing these skills. A paper and an online version measure the same construct. The online version includes instant feedback that integrates both performance and improvement feedback. In the outlook, outstanding key development and testing steps are addressed.

Keywords: lower secondary school mathematics, online self-assessment, instant feedback, university entrance test, STEM courses, study entry phase

1 Ausgangslage und Zielsetzung

1.1 Relevanz mathematischer Basisfähigkeiten in der Studieneingangsphase

Ein vor dem Abschluss abgebrochenes Studium ist sowohl auf individueller Ebene als auch aus volkswirtschaftlicher Perspektive nicht wünschenswert: Es ist kostenintensiv und transportiert das Etikett des Scheiterns (in einer Übersicht von Neugebauer et al., 2019). Hochschulen investieren Ressourcen in Form von Lehrveranstaltungen und Infrastruktur in abge-

brochene Studien, die sich letztlich nicht in einer entsprechenden Anzahl von Absolvierenden auszahlen. Tatsächlich steigen die Studienabbruchquoten jedoch in den letzten Jahren in den meisten Studiengängen kontinuierlich bis auf jüngst 43% in Mathematik und Naturwissenschaften – was der höchsten Quote aller Fächer entspricht – und 35% in den Ingenieurwissenschaften (Heublein et al., 2020) – eine pädagogische wie hochschuldidaktische Herausforderung, der mit geeigneten Maßnahmen zu begegnen ist.

Die Ursachen für einen Studienabbruch können dabei vielfältig sein und variieren teilweise je nach Studiengang. Dabei ist nach Heublein et al. (2017) mangelndes fachliches Vorwissen für drei Viertel aller Studienabbrüche mitverantwortlich. Auch Petri (2020) sieht die Abiturnote – und damit mitunter das mangelnde Vorwissen – neben motivationalen und emotionalen Merkmalen als hauptursächlich. Insbesondere die Mathematikleistung stellt sich als direkt abhängig vom Vorwissen dar (Krawitz, 2020). Dabei sind mathematische Basisfähigkeiten gerade in den MINT-Studienfächern (Fächer der Gebiete Mathematik, Informatik, Naturwissenschaften und Technik) unerlässliches Grundhandwerk (Neumann et al., 2017; CoSH-Gruppe, 2021), welches entgegen dem im Abiturzeugnis enthaltenen Kompetenzversprechen jedoch von den Studienanfangenden nicht immer ausreichend sicher beherrscht wird. Die Hochschulen reagieren auf diese Passungsprobleme mit deutlichen curricularen Anpassungen (Bausch et al., 2014a). Dabei werden die Inhalte und Fähigkeiten der Sekundarstufe II mehr oder weniger intensiv in den entsprechenden Lehrveranstaltungen wiederholt. Die curricular vorgesehenen Inhalte in den Lehrveranstaltungen der ersten Semester der MINT-Studiengänge begrenzen jedoch die Möglichkeiten zum Wieder- und Nachholen, sodass eigentlich als bekannt vorauszusetzende Inhalte der Sekundarstufe I meist nur implizit integriert werden. Gleichzeitig werden Orientierungs- und Lernangebote auch schon in die Zeit vor Studienbeginn verlegt (Bausch et al., 2014a), welche sich aber ebenfalls auf die Sekundarstufe II fokussieren und die Sekundarstufe I nur implizieren. Dies wiederum erfordert eine zuverlässige individuelle Diagnostik, welche unmittelbar von den Studierenden selbst interpretiert und genutzt werden kann, damit individuell festgestellte Lücken auch ohne Anleitung durch Dozierende gezielt geschlossen werden können (Karst et al., 2017).

1.2 Potenziale technologiebasierter IRT-skalierter Diagnoseinstrumente

Im Kontext der Diagnostik studienfeldspezifischer Eingangsfähigkeiten spielen Online-Self-Assessments als Instrumente der Studienorientierung eine bedeutende Rolle, die von vielen Hochschulen auch im Bereich der Mathematik in der Phase der Studienwahl angeboten werden, aber in der Regel keinerlei Verbindlichkeit mit sich bringen. Sie dienen vielmehr lediglich dem Abgleich individueller Neigung und Eignung mit dem jeweiligen Studiengangsprofil. Damit sollen sie „Auskunft über die Passung der Erwartungen zum angestrebten Studienfach geben und Selbstselektionsprozesse anstoßen, indem die studienfachbezogenen Einstellungen verändert werden“ (Karst et al., 2017, S. 205). Die Ergebnisrückmeldungen sind häufig sehr einfach gehalten, etwa in Gestalt numerischer Summenscores und kurzer Erläuterungen, die zumeist nicht auf komplexen Skalierungsverfahren beruhen, sondern ähnlich wie bei schulischen Klausuren den Anteil korrekter Antworten rückmelden (Brunner, 2017). Individuelle Defizite, spezifischer Nachholbedarf oder festzustellende Kompetenzprofile werden damit selten erkannt. Entsprechend werden konkrete Maßnahmen mit Relevanz für die Studieneingangsphase im Sinne einer frühzeitigen Diagnostik individueller Defizite und Ablei-

tung geeigneter Fördermaßnahmen damit in aller Regel nicht verknüpft – es zeichnet sich somit mit Blick auf die sich verschärfende Problematik der frühen Studienabbrüche Handlungsbedarf ab.

Einen oftmals relevanten Baustein zum erfolgreichen Übergang in die Hochschule bilden Onlinebrückenkurse, die häufig von entsprechenden Einstufungstests flankiert werden, um die Studierenden entsprechend ihren Fähigkeiten bestmöglich zu fördern (Hanft et al., 2015). Die Qualität der Instrumente und des Feedbacks an Testteilnehmende variiert dabei deutlich. Zu kritisieren ist dabei insbesondere auch die unzureichend verfügbare Dokumentation der psychometrischen Qualität der Instrumente (Brunner, 2017), was darauf hinweisen kann, dass diese auch nur bedingt geprüft wurde. Skalierungsverfahren auf Basis der Item-Response-Theorie (IRT) bieten nicht nur die Möglichkeit, Personenfähigkeit und Itemschwierigkeit simultan auf einer Skala zu schätzen (Wilson, 2005), sondern eignen sich insbesondere auch für technologiebasierte Diagnoseinstrumente mit umfangreichen Feedbackformaten, da sie hohe Automatisierungspotenziale mit sich bringen und auch im Large-Scale-Bereich bei großen Stichproben effizient einsetzbar sind. Die technologiebasierte Diagnostik von Studieneingangsleistungen auf Basis komplexer Skalierungsverfahren ist ungeachtet dessen im Bereich der Mathematik noch nicht etabliert – ein Umstand, der mit dem hier vorgestellten Instrument adressiert werden sollte.

Hier setzt dieser Beitrag an und stellt ein selbst entwickeltes Instrument zur Diagnostik mathematischer Eingangsvoraussetzungen zu Studienbeginn vor, das auch Feedbackelemente enthält. Gerade die Feedbackkonzeption ist dabei bei Online-Self-Assessments besonders herausfordernd, da die Interpretation nicht von psychometrischen Fachleuten begleitet, sondern lediglich von den Testteilnehmenden gelesen und interpretiert wird. Der Gesamterfolg eines Assessments darf somit nicht lediglich auf die Instrumentenqualität zurückgeführt werden, sondern muss auch die Qualität des Feedbacks berücksichtigen (Kubinger et al., 2012). Dabei identifizieren Brunner et al. (2015) in einem Vergleich von über 40 verschiedenen Mathematik-Self-Assessment-Tools aus dem deutschsprachigen Raum „noch viel Potential für die Weiterentwicklung des Feedbacks“ (Brunner et al., 2015, S. 159). Wie genau sie dies quantifizieren und wo genau sie dies sehen, bleibt allerdings offen.

Aus Umfangsgründen sind die Ausgestaltung und erste Erkenntnisse zur Rezeption des Feedbacks nicht Teil dieses Beitrags. Ein zentrales Element des Feedbacks, welches hier aufgegriffen wird, ist eine einfach zu interpretierende Leistungsskala auf Basis eines Niveaumodells (Beaton & Allen, 1992). Dieses bietet die Möglichkeit, einerseits den dritten Feedback-Schritt nach Hattie und Timperley (2007), das Feed Forward, zu integrieren und andererseits auch neben der Aufgabenebene die Konstruktebene (Belcadhi, 2016) einzubeziehen.

1.3 Entwicklungsbedarf psychometrisch fundierter Testinstrumente

Zur Testentwicklung wurden im Vorfeld relevante Rahmenbedingungen näher in den Blick genommen. Dazu wurden bereits bestehende Testinstrumente und Lernmaterialien, insbesondere für den Übergang zwischen Schule und Studium, begutachtet, beispielsweise Schäfer et al. (1997), Knorrenschild (2004) und Dürrschnabel et al. (2019). Die daraus resultierenden Inhalte wurden unter Berücksichtigung des Mindestanforderungskatalogs an ein WiMINT-Studium (MINT-Fächer und Fächer der Wirtschaftswissenschaften) der CoSH-Gruppe kritisch gesichtet, und entsprechend wurde eine Auswahl der Themen vorgenommen (CoSH-

Gruppe, 2021). Inhaltlich fokussiert der Test Themen der Sekundarstufe I, welche explizit nicht Gegenstand universitärer Lehre sind und daher bereits vor Studienbeginn von Anfangenden beherrscht werden müssen.

Die Berücksichtigung bestehender Testinstrumente wie die Vergleichsstudien TIMSS (Schwippert et al., 2020) oder PISA (Reiss et al., 2019) erwies sich aufgrund der Konzeption dieser Instrumente als nur eingeschränkt sinnvoll. Die Recherche bestehender Tests zeigte den Bedarf zur Erstellung eigener Tests aus zwei zentralen Gründen auf: (1) Gängige Aufgabensätze fokussieren hierbei eher offene Antwortformate mit Rechenwegen (beispielsweise Neubrand et al., 2004). Dies macht allerdings eine automatische Auswertung zur Generierung eines Instant-Feedbacks schwierig und fehleranfällig. (2) In kontextbehafteten Aufgaben findet schnell eine Verschiebung des Fokus weg von mathematischen Schwierigkeiten hin zu einer nicht immer eindeutigen Umsetzung von Texten in mathematische Fragestellungen statt (Jahnke et al., 2014). Um diese Form mathematischer Modellierung zu umgehen, liegt der Fokus vorwiegend auf handwerklichen Fähigkeiten und innermathematischen Problemen. Die Aufgaben sollten kleinschrittige Probleme widerspiegeln, die in kurzer Zeit ohne Hilfsmittel bearbeitbar sind. Dies folgt auch dem Ansatz etablierter Brückenkursprojekte wie Ve&MINT (Bausch et al., 2014b) bzw. OMB+ (Krunke et al., 2012), welche von der Allianz führender Technischer Universitäten in Deutschland (TU9) auch ausdrücklich zur Studienvorbereitung empfohlen werden (TU9, 2020).

Aus diesen Gründen entwickeln viele Hochschulen entsprechende eigene Eingangs- und Eignungstests, welche aber häufig nicht umfangreich auf ihre psychometrische Güte geprüft sind (Brunner, 2017). Das hier vorgestellte Instrument bezieht bestehende Aufgabenstellungen aus den Eingangstests der Mathematik-Vorkurse und der Prüfungsaufgaben an der Universität Stuttgart ein. Letztere dienen dazu, Besonderheiten des späteren Prüfungsformats als erste Lernhilfe aufzuzeigen. Die Aufgaben legen hierbei besonderen Wert auf eine in den Grundlagenvorlesungen üblichen Fachsprache und Notation.

1.4 Zielsetzung und Fragestellung des Beitrags

Vor dem Hintergrund der gewürdigten Argumente und Desiderate wird deutlich, dass ein feedbackgenerierender computerbasierter Eingangstest bezüglich der mathematischen Anforderungen der MINT-Studiengänge in einer expliziten Ausrichtung auf die Sekundarstufe I dazu beitragen könnte, mittelfristig Studienabbrüche effektiv zu reduzieren. Ungeachtet einer breiten Verfügbarkeit mathematischer Online-Self-Assessments konnte für den betreffenden Inhaltsbereich bislang kein Instrument identifiziert werden, das die skizzierten Anforderungen an individualisierte und psychometrisch basierte Ergebnis- und Förderrückmeldungen einzulösen vermochte, wodurch eine Eigenentwicklung angestrebt wurde. Zur Sicherstellung der diagnostischen Eignung wurden im Zuge dessen die folgenden drei Ziele definiert, die mit dem Instrument erreicht werden sollten:

(1) Selbstreflektionsanlass für Studierende bezüglich ihrer individuellen Eignung,
(2) Identifikation und adressatengerechte Rückmeldung individueller Nachholbedarfe,
(3) Monitoring der Zielgruppe und von Kohortenveränderungen im Zeitverlauf.

Dabei wird bewusst ein inhaltlicher Schwerpunkt auf die relevanten Inhalte und Fähigkeiten der Mathematik der Sekundarstufe I gelegt, da diese häufig in der Studieneingangsphase

nicht in der unter den oben thematisierten Umständen gebotenen Ausführlichkeit wiederholt werden. Demgegenüber werden Inhalte und Fähigkeiten aus der Sekundarstufe II (zum Beispiel die Differential- und die Integralrechnung sowie die Vektor- und Matrizenalgebra) erneut grundlegend aufgegriffen und teilweise auch axiomatisch neu eingeführt (Meyberg & Vachenauer, 2001). Der Beitrag fokussiert dabei nachstehende Fragestellungen:

(1) Welche Struktur und welche Parametrisierung des Item-Response-Modells ist für eine reliable und valide Interpretation geeignet?
(2) Ist die Skala fair in Bezug auf das Geschlecht sowie invariant bezüglich des Darreichungsmodus und der Version?
(3) Unterscheiden sich die Testleistungen der Teilnehmenden in Abhängigkeit ihres gewählten Studiengangs?
(4) Kann aus der modellierten Skala ein inhaltlich valide interpretierbares Niveaumodell generiert werden?
(5) Haben die gemessenen Leistungen – über die Hochschulzugangsberechtigungsnote hinaus – eine Vorhersagekraft für den Erfolg im ersten Studiensemester?

2 Methode

2.1 Aufbau des Testinstruments

Das Testinstrument gliedert sich in vier Teile mit insgesamt 17 Aufgaben zu folgenden Teilgebieten. Beispielaufgaben sind in den Abbildungen 2, 3 und 4 (s. unten Abschnitt 3.2) dargestellt.

(1) Bruch- und Potenzrechnung (3 Aufgaben; 6 Minuten): Einfache Terme mit Zahlen; Binomische Formeln; Mustererkennung in Folgen.
(2) Termumformung (5 Aufgaben; 10 Minuten): Brüche mit Variablen; Binomische Formeln; Techniken zur Lösung gemischt quadratischer Gleichungen; quadratisches Ergänzen.
(3) Aspekte der Geometrie (4 Aufgaben; 10 Minuten): Räumliche und ebene Abstandprobleme; räumliche Vorstellung; Lineare Gleichungssysteme.
(4) Elementare Funktionen (5 Aufgaben; 10 Minuten): Definitionen und Rechenregeln der Exponential-, Logarithmus- und trigonometrischen Funktionen.

Dabei wird für jeden Testteil in Abhängigkeit der angenommenen Itemschwierigkeit und der studienfeldspezifischen Anforderungen eine Höchstbearbeitungszeit festgelegt, da eine ausreichend zügige Bearbeitung dieser Basisanforderungen relevant ist, um entsprechende Routinen dieser Basisfähigkeiten annehmen zu können.

Es wird die in den (mathematischen) Eingangsveranstaltungen verwendete Fachsprache, auch mit nichtverbalisierten Formelzeichen, verwendet (zum Beispiel „Für $t \in \mathbb{R}$ besitzt das Lineare Gleichungssystem […] eine eindeutige Lösung (x_1, x_2). Wie lautet diese?“), anstatt die teilweise vereinfachenden Ausdrücke der Schule (zum Beispiel „Bestimmen Sie die Lösung des Linearen Gleichungssystems […] in Abhängigkeit des Parameters t.“), um die implizite Orientierungsleistung des Tests zu erhöhen (Karst et al., 2017). Gleichzeitig wird im Sinne der *Opportunity to Learn* (AERA et al., 2014, 56f.) darauf geachtet, dass die Fachspra-

che auch mit dem curricular abgebildeten Vorwissen verstanden werden kann. Auf eine Kontextualisierung der Items wird auf Basis der Argumentation von Bach (2016) verzichtet, der prozessbezogene Aufgabenstellungen lediglich als eine Meta-Ebene der inhaltsbezogenen Aufgabenstellungen herausstellt.

Der Test existiert in drei parallelisierten Formen, um primär ein Abschreiben in Präsenzadministration zu vermeiden und sekundär auch Versionen für wiederholende Testungen verfügbar zu machen. Diese Formen unterscheiden sich meist lediglich in den verwendeten Zahlenwerten (die Beispielaufgabe in Abbildung 4 wird mit den Kombinationen $x = \log 2$ und $\log 8$; $x = \log 3$ und $\log 9$; $x = \log 2$ und $\log 4$ gestellt) oder in der Verwendung verwandter Definitionen (z.B. ist zu entscheiden, ob die Definition einer trigonometrischen Funktion korrekt ist, dabei werden $c \cdot \sin\alpha = b$; $c \cdot \cos\alpha = a$; $a \cdot \cos\alpha = c$ verwendet). Die stärksten Unterschiede zwischen den drei Versionen sind in den Abbildungen 2 und 3 (s. unten) dargestellt. Weiterhin existieren sowohl ein Darreichungsmodus als automatisch einlesbarer Papier-Test als auch als Online-Test in einer eigenentwickelten Plattform. Aufgabenstellungen und -zusammensetzungen sind für die Modi identisch.

Die Aufgabenformate umfassen Single-Choice-Aufgaben mit im Allgemeinen vier Antwortalternativen, komplexe Multiple-Choice-Aufgaben mit vier bis fünf zu entscheidenden Aussagen sowie ganzzahlig numerische Antwortformate.

2.2 Testentwicklung und analytisches Vorgehen

2.2.1 Testentwicklung

Der Test wurde von Hochschullehrenden der Mathematik unter Rückgriff auf bestehende Aufgaben und Instrumente erfahrungsgeleitet entwickelt. Psychometrische und testtheoretische Expertise wurde erst nach der ersten Datenerhebung einbezogen. Dies führt zu der im Folgenden sehr exploratorisch ausgelegten Methodenauswahl.

Die fachwissenschaftliche Strukturierung des Tests erfolgte vor demselben erfahrungsgeleiteten Hintergrund. Die Auswahl der Themen und Aufgaben adressiert bekannte Probleme in der Studieneingangsphase und ist durch die Aufteilung in kurze, zeitbegrenzte Blöcke so ausgelegt, dass diese perspektivisch auch erweitert werden kann. Die aufgegriffenen Themen decken sich mit wichtigen Themenfeldern, die in unterschiedlichen Betrachtungen der Studieneingangsphase identifiziert werden (vgl. die MaLeMINT-Studie; Neumann et al., 2017; und den Mindestanforderungskatalog; CoSH-Gruppe, 2021). Dabei wird, wie in Abschnitt 1.4 begründet, der Schwerpunkt auf die Mathematik der Sekundarstufe I gelegt und hierin aus Testumfangsgründen weiter selektiert.

Die Pilotierung mit $n = 1004$ Studienanfangenden wurde mit vergleichbaren Methoden wie im Folgenden skizziert analysiert. Es wurden insgesamt 6 von 16 Aufgaben identifiziert, die die psychometrischen Anforderungen nicht erfüllten – meist aufgrund einer deutlich zu hohen Schwierigkeit für die Zielgruppe. Diese wurden im folgenden Durchgang durch geeignetere Aufgaben ersetzt. Der grundlegende Aufbau und die Struktur des Tests konnten bestätigt werden.

2.2.2 Itemkodierung

Die Basiselemente aller Aufgaben (die Auswahl bei Single-Choice, der Eintrag in einem numerischen Freitextfeld sowie die Bewertung einer einzelnen Aussage bei komplexen Multiple-Choice) werden dichotom kodiert: 1 für die richtige Antwort und 0 für falsche Antworten. Fehlende Antworten werden dabei durchgängig als falsche Antworten gewertet, da diese im Sinne der Testkonstruktion als fähigkeitsmindernd zu interpretieren sind (Lüdtke et al., 2007). Diese Einzelbetrachtung wird gewählt, um zu vermeiden, dass unterscheidbare Fähigkeiten in einer Aufgabe zusammengefasst wurden und damit nicht sauber differenziert werden können (Tripp & Tollefson, 1985).

Zur Ermittlung sinnvoll skalierbarer Items werden diese Einzelkodierungen schrittweise zusammengefasst. Dabei werden unter Verwendung des 1PL-IRT-Modells sowohl die Itemfitstatistiken (Infit, Outfit sowie Trennschärfe; Wilson, 2005) als auch die paarweisen $Q_{3,*}$-Statistiken (Chen & Thissen, 1997) herangezogen und diejenigen Items zusammengefasst, welche die höchsten $Q_{3,*}$-Werte aufweisen, um die Grenzwertproblematik dieser Statistik (Chen & Thissen, 1997) zu umgehen. Dabei wird eine dichotome Kodierung beibehalten, wobei 1 für vollständig richtige Antworten steht.

Die mittlere Bearbeitungsdauer liegt bei 69% der verfügbaren Bearbeitungszeit (Rechnen 66%, Terme 71%, Geometrie 74% und Funktionen 63%). Weniger als 5% der Teilnehmenden hat die jeweils verfügbare Bearbeitungszeit ausgenutzt. Da außerdem nicht davon ausgegangen werden kann, dass die Teilnehmenden die Aufgaben in der gezeigten Reihenfolge bearbeiten, wurde auf eine spezielle Kodierung für fehlende Antworten aus Zeitgründen verzichtet.

2.2.3 Skalierung

Zur Skalierung werden Methoden der Item-Response-Theorie zu Grunde gelegt (Wilson, 2005). Der resultierende Itempool wird mittels Likelihood-Ratio-Tests bezüglich der notwendigen Parametrisierung (1PL, 2PL oder 3PL) sowie Between-Item-Multidimensionalität bezüglich der vier Testlets, also die abgegrenzt administrierten theoretisch konstruierten Themenbereiche (Rechnen, Terme, Geometrie und Funktionen) geprüft (Reise, 2015). Weiterhin wird das Informationskriterium nach Akaike unter Anpassung auf kleine Stichprobengrößen (da $N/p = 33.5 < 60$; Burnham & Anderson, 2002) verwendet. Alle hier angegebenen Parameter beziehen sich auf die *Difficulty*-Parametrisierung, also $\alpha_i \cdot (\theta - \beta_i)$. Ergänzend wird noch auf Itemebene die Notwendigkeit des Rateparameters der 3PL-Parametrisierung geprüft. Diese Prüfung scheint angebracht, da einzelne Items rein theoretisch eine hohe Ratewahrscheinlichkeit besitzen können. Die Passung des Modells auf die Daten wird mittels gängiger Kenn- und Grenzwerte überprüft: Itemfit-Statistiken ($\in [0.7,1.3]$; Wilson, 2005), EAP/PV-Reliabilität ($> .7$; Bond & Fox, 2015) und SRMSR ($< .08$; Maydeu-Olivares, 2013).

2.2.4 Prüfung der Parallelität und der Fairness

Die Fairness des Tests, d.h. das Fehlen einer systematischen Benachteiligung einzelner Gruppen, sowie die Parallelität bezüglich der drei Versionen und der beiden Modi wird auf Basis von Methoden des *Differential Item Functioning* (DIF) geprüft. Da bezüglich der Fairness der drei Testversionen mehr als zwei Gruppen miteinander verglichen werden, werden statt

Hypothesentests zwei Kennwerte je Gruppe betrachtet: (1) Die Kommunalität h^2 einer einfaktoriellen Hauptkomponentenanalyse (Jolliffe, 2002) der Itemschwierigkeitsparameter mit Grenzwert $> .90$ sowie (2) die *Mean Absolute Deviation* (MAD) gegenüber der Itemschwierigkeitsparameter des Gesamtmodells mit Grenzwert < 0.25 (Grisay & Monseur, 2007). Die Itemdiskriminisationsparameter werden dabei gruppenunabhängig geschätzt und somit nicht berücksichtigt. Zusätzlich wird das zweite Kriterium erweitert: Da ein extremer Ausreißer durch viele gute Items zu akzeptablen Kennwerten führen kann, wird die absolute itembezogene Abweichung einbezogen und lediglich bis 0.4 akzeptiert.

Da im Fall der Versionen und Modi DIF identifiziert wird, werden die betroffenen Items unter Verwendung unterschiedlicher Itemschwierigkeitsparameter im Modell integriert. Auf dieser Basis erfolgt ein anschließender Vergleich der Schwierigkeiten der einzelnen Testzusammensetzungen mittels zweifaktorieller Varianzanalyse unter Verwendung des Zusammenhangsmaßes partielles η^2.

2.2.5 Niveaumodellierung

Das absolute Leistungsfeedback erfordert eine individuell interpretierbare Skala. Aus diesem Grund wird die Personenfähigkeitsskala mittels Methoden von Beaton und Allen (1992) in ein Niveaumodell überführt. Die Niveaugrenzen werden dabei auf Häufungspunkte der Itemschwierigkeiten gesetzt. Die Beschreibung der Fähigkeiten erfolgte mittels qualitativer Anforderungsanalysen durch fünf Fachleute aus den Bereichen Mathematik, Mathematiklehramt sowie Erziehungswissenschaft. Dabei werden jeweils die Items herangezogen, welche von allen Studierenden des Niveaus *n* ausreichend sicher korrekt bearbeitet werden können ($P\left(X_i = 1 \middle| \theta_{n,\min}\right) > .50$), aber nicht von den Studierenden des Niveaus unterhalb ($P\left(X_i = 1 \middle| \theta_{n-1,\max}\right) < .65$). Die Verständlichkeit und Nutzbarkeit der Beschreibungen wurden durch Vorstellungen gegenüber unterschiedlichen Zielgruppen aus der Schulbildung, der universitären Bildung sowie der Fachwissenschaften überprüft und anhand der Rückmeldungen verbessert.

2.2.6 Gruppierung und Vergleich der Studiengänge

Das Gruppieren vergleichbarer Studiengänge erfolgt manuell, indem auf Basis der mittleren Fähigkeiten der einzelnen Studiengänge inhaltlich zusammengehörige Gruppen (z.B. Maschinenbau sowie Fahrzeug- und Motorentechnik) gebildet werden, innerhalb derer mittels einfaktorieller Varianzanalyse die Zusammengehörigkeit identifiziert und unter Verwendung des Zusammenhangsmaßes η^2 quantifiziert wird.

2.2.7 Prognostische Validität

Zur Prüfung der Vorhersageeffekte werden Modelle der logistischen linearen Regression bestimmt. Dabei wird als Kriterium der Erwerb des Scheins der Lehrveranstaltung Höhere Mathematik 1 herangezogen, welcher angesichts der üblichen Bestehensquoten von lediglich um die 70% eine kritische Hürde im ersten Studiensemester darstellt. Als standardisierte Effektgröße dient zur unmittelbaren Vergleichbarkeit der Varianzaufklärungskoeffizient nach Nagelkerke (1991) beziehungsweise dessen Differenz bei Modellvergleichen. Das maximale Erklärungsmodell wird über schrittweises Hinzufügen nach größter Erhöhung des Vari-

anzaufklärungskoeffizienten nach Nagelkerke ermittelt, bis keine statistisch signifikante Modellverbesserung im Likelihood-Ratio-Test mehr stattfindet (Madsen & Thyregod, 2010).

2.3 Stichprobe

2.3.1 Gesamtstichprobe

Der Test adressiert Studienanfangende in MINT-Studiengängen dreier baden-württembergischer Universitäten aus dem zugrundeliegenden Projektkontext. Er soll perspektivisch aber auch Studienanfangende aller WiMINT-Studiengänge in Deutschland zwischen Erwerb der Hochschulzugangsberechtigung und Beginn des Studiums adressieren können. Durch das Vorhandensein spezifischer Förderangebote schon vor Studienbeginn sowie die bedingte Erreichbarkeit bereits vor Hochschuleintritt wird eine Gütebestimmung am ersten Vorkurstag durchgeführt. Dabei werden drei baden-württembergische Universitäten sowie in der zweiten Kohorte die Duale Hochschule Baden-Württemberg einbezogen. Dabei wird auf die Nennung und Darstellung hochschulspezifischer Ergebnisse verzichtet, um nicht den Eindruck eines (nicht belastbaren) Rankings entstehen zu lassen.

Die Erhebungen starteten zum Wintersemester 2018/19 mit einem Pilotinstrument. Seit dem Wintersemester 2019/20 ist dieses einheitlich gestaltet. Zum Wintersemester 2019/20 wurde rein in Papier-Form gemessen. Zum Wintersemester 2020/21 aufgrund der coronabedingten Umstellungen erfolgte die Erhebung ausschließlich online. Da zu diesem Zeitpunkt aber noch von relativ geringen Auswirkungen auf die Fähigkeiten bezüglich der Mathematik der Sekundarstufe I auszugehen ist, verwenden wir diese Besonderheit zur Prüfung der Invarianz der Modi. Die drei Versionen wurden jeweils gleichverteilt zufällig zugewiesen.

Das Verhältnis der Anzahl der Vorkursteilnehmenden zu Studienanfangenden unterliegt einer Unschärfe, da die Zahl der Angemeldeten und die der tatsächlich Teilnehmenden in den Vorkursen teils deutliche Diskrepanzen aufweist. Im Jahr 2019 nahmen alle Anwesenden des ersten Vorkurstags an der Erhebung teil. Im Jahr 2020 zeigt sich hier ein deutlicher Einbruch an einer Hochschule, wohingegen die beiden anderen Hochschulen hinzugewannen. Inwieweit die Stichproben repräsentativ für die Vorkursteilnehmenden stehen, kann mangels Vergleichsdaten nicht bestimmt werden. Es sind insgesamt 84 verschiedene Bachelor-Studiengänge berücksichtigt.

In Summe können für die Skalierung sowie die Modell- und Güteprüfungen $N = 3{,}819$ Studienanfangende berücksichtigt werden. Dabei sind $n = 1730$ Studienanfangende (45%) im Jahr 2019 in Papier-Form und $n = 2{,}089$ Studienanfangende im Jahr 2020 in Online-Form integriert. $n = 2{,}929$ Teilnehmende gaben ihr Geschlecht an, wovon $n = 976$ weiblich sind (33%), was einen durchaus realistischen Anteil in MINT-Studiengängen darstellt. $n = 2{,}885$ Teilnehmende gaben ihr Geburtsjahr an, das mittlere Alter zum Testzeitpunkt beträgt 19.5 Jahre (SD = 1.7 Jahre). Lediglich 55% der Teilnehmenden begannen das betroffene Studium direkt nach dem Abitur, was das vergleichsweise hohe mittlere Alter erklärt. Ob bei den anderen Teilnehmenden zwischen Abitur und Studienbeginn ein abgebrochenes Studium, eine Berufsausbildung oder andere Tätigkeiten lagen, wurde nicht abgefragt.

2.3.2 Stichprobe zur Bestimmung der prognostischen Validität

Zur Bestimmung der prognostischen Validität werden zwei Stichproben zusammengezogen. Mittels ausreichender itembasierter Verankerung des Pilotinstruments mit dem finalen Instrument ($\approx 65\%$) können die Pilotdaten auf dieselbe Skala gebracht (Reise, 2015) und somit die Leistungsdaten zweier Gruppen vereinigt werden: Für eine der Hochschulen konnten – auf freiwilliger Basis – Leistungsdaten für $n = 363$ Studierende bezüglich des ersten Studiensemesters in den Kohorten 2018 und 2019 einbezogen werden. Danach war aufgrund der fehlenden Passung des Datenschutzkonzepts auf die Corona-Situation keine Erhebung von Leistungsdaten mehr möglich.

Tabelle 1 nennt die Elemente der Hochschulzugangsberechtigung (HZB), die neben der mittels EAP-Schätzern (*Expected A-Posteriori*) modellierten Fähigkeit einbezogen wurden, und greift grundlegende deskriptive Maße auf.

Tabelle 1: Elemente der Hochschulzugangsberechtigung (HZB)

Merkmal	Codes	N	M (SD)
Gesamtnote	1.0 … 3.9	353	2.17 (0.66)
Gesamtpunktzahl in Mathematik	≤ 4 (kodiert als 4) 5 … 15	158	10.5 (3.2)
Schulart, an der die HZB erworben wurde	1: Allgemeinbildendes Gymnasium (AG) 0: Berufliche Gymnasien	250 100	.71
Bundesland, in dem die HZB erworben wurde	1: Baden-Württemberg (BW) 0: Andere Bundesländer	291 56	.84
Zeitlicher Abstand zwischen Erwerb der HZB und Beginn des Studiums	1: direkter Einstieg 0: mindestens ein Jahr Abstand	171 187	.48
Vertiefungskurs, wenn die HZB am AG in BW erworben wurde	1: teilgenommen 0: nicht teilgenommen	82 190	.30

3 Ergebnisse und Interpretation

3.1 Struktur- und Parameterprüfung

Zur Struktur- und Parameterprüfung wird das Informationskriterium AIC_c herangezogen. Dabei ergeben sich für die vier relevanten Struktur-Parameter-Kombinationen die folgenden Werte. 1-faktoriell 1PL: $AIC_c = 99013$; 1-faktoriell 2PL: $AIC_c = 98416$; 4-faktoriell 1PL: $AIC_c = 98719$; 4-faktoriell 2PL: $AIC_c = 98188$. Somit weist das 4-faktorielle 2PL-Modell den niedrigsten Wert auf, wobei die Differenz durchgängig bedeutsam ist (> 10; Burnham & Anderson, 2002). Die paarweisen Likelihood-Ratio-Tests dieses Modells mit dem 4-faktoriellen 1PL-Modell ($\chi^2(30) = 593.2; p < .001$) sowie mit dem 1-faktoriellen 2PL-Modell ($\chi^2(6) = 239.9; p < .001$) bestätigen diese Präferenz. Ein Vergleich mit der 3PL-Parametrisierung erfolgt lediglich für die 1-faktorielle Modellierung, da ansonsten zu ungenaue Parameterschätzungen zu erwarten sind: $\chi^2(34) = 8.9; p = 1.000$. Wird jeweils ein Ratepara-

meter frei geschätzt und die anderen auf 0 fixiert, ergibt sich bei einem Item ein Parameter $\gamma_i = .031$, für alle anderen Items gilt $\gamma_i \leq .007$.

Zur einfacheren Kommunizierbarkeit der Ergebnisse sowie zur Reduktion der Ungenauigkeit aufgrund sehr geringer Testletlängen soll zusätzlich eine Gesamtfähigkeit modelliert werden. Da deren Existenz nicht theoretisch begründet werden kann, wird ein hierarchisches Konstrukt abgelehnt, sodass ein *Bifactor*-Modell (Reise, 2012) in der 2PL-Parametrisierung gewählt wird. Hierfür werden Personenfähigkeiten mittels EAP-Schätzern gebildet, da WLE-Schätzer (*Weighted Likelihood Estimates*) keine Within-Item-Multidimensionalität unterstützen (Robitzsch & Steinfeld, 2018). Die Gütekennwerte dieses finalen Modells liegen innerhalb der oben genannten Grenzen: $Q_{3,*} \in [-0.37, 0.18]$; Item-Outfit $\in [0.80, 1.14]$; Item-Infit $\in [0.97,1.01]$; SRMSR $= .036$; EAP/PV-Reliabilität des gemeinsamen Faktors = .77. Die Itemparameter der Difficulty-Parametrisierung liegen in den Bereichen $\alpha_{g,i} \in [0.31, 3.01]$ sowie $\beta_i \in [-4.6, 2.8]$. Abbildung 1 zeigt beispielhaft die zu Version A in der Papierform gehörige WrightMap unter Kennzeichnung der Zugehörigkeit zu den vier Residualfaktoren bezüglich der vier Inhaltsbereiche.

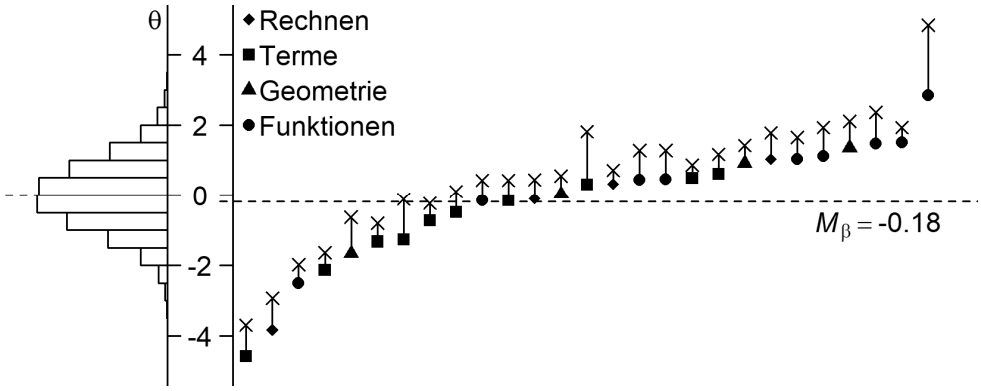


Abbildung 1: WrightMap des finalen Bifactor-Modells für Version *A Papier* unter Kennzeichnung der Zugehörigkeit zu den vier Residualfaktoren.

Anmerkung: Die beiden Itemparameter werden mit Hilfe des Bereichs der Lösungswahrscheinlichkeit von 50% bis 65% visualisiert.

Betrachtet man die vierfaktorielle Modellierung ohne den gemeinsamen Faktor, ergeben sich die in Tabelle 2 dargestellten Reliabilitäten sowie latenten Zusammenhänge. Gerade für individualdiagnostische Zwecke scheint diese Genauigkeit – insbesondere auf Basis der vier identifizierten Faktoren – noch deutlich optimierbar. Da die verfügbare Testzeit aus hochschulpolitischen Gründen kaum ausbaufähig erscheint, verbleiben drei Möglichkeiten: (1) Verbesserung der Genauigkeit des Itempools – unter Gefahr der Reduktion inhaltlicher Validität. (2) Entwicklung eines Feedbacksystems, welches die Messgenauigkeit kompensieren kann. Diese Optimierung wurde für das Wintersemester 2021/22 umgesetzt, ist aber nicht Teil dieses Beitrags. (3) Einsatz von Methoden des *Computerized Adaptive Testing* (van der Linden & Ren, 2019). Diese Optimierung wurde ebenfalls für das Wintersemester 2021/22 für den Teil zu Bruch- und Potenzrechnung umgesetzt.

Tabelle 2: EAP/PV-Reliabilität und latente Korrelationen im vierfaktoriellen Modell

Faktor	EAP/PV-Reliabilität	Latente Korrelation		
Rechnen	.66	Rechnen	Terme	Geometrie
Terme	.75	.76		
Geometrie	.66	.64	.81	
Funktionen	.71	.74	.86	.76

3.2 Invarianz bezüglich Modi und Versionen sowie Fairness bezüglich Geschlecht

Mangels inhaltlicher Vergleichbarkeit wurden zwei Items in den verschiedenen Versionen als unterschiedlich angesehen und aus der DIF-Prüfung ausgeschlossen. Hierbei handelt es sich um Aufgaben im komplexen Multiple-Choice-Format, wobei die einzelnen Entscheidungen nicht auf dieselben mathematischen Theoreme zurückführbar sind, obwohl der Inhalt der Aufgaben jeweils ähnlich ist. Ein Zusammenfassen allein auf Basis vergleichbarer Itemparameter erscheint nicht gerechtfertigt. Abbildung 2 zeigt ein Beispiel hierfür. Zwei weitere Items zeigen in jeweils einer Version eine deutliche Abweichung ihres Schwierigkeitsparameters vom mittleren Parameter und werden somit als DIF-auffällig klassifiziert. Diese Abweichung tritt in Geometrie-Aufgaben auf und kann wahrscheinlich auf unterschiedliche räumliche Komplexität oder unterschiedliche Komplexität des Lösungsweges oder des Lösungswertes begründet werden. Abbildung 3 zeigt ein Beispiel hierfür. Unter Berücksichtigung dieser Auffälligkeiten in der Modellierung resultieren die Kennwerte $h^2 = .99$ für alle drei Gruppen und MAD $\in [0.01{,}0.12]$. Bezüglich der Modi sind ebenfalls zwei Items auffällig, was bei Berücksichtigung zu den Kennwerten $h^2 = .98$ und MAD $\in [0.13, 0.19]$ führt. Hier stellte sich bei Sichtung der aktuellen Bildungspläne in Baden-Württemberg (KM BW & ZSL, 2016) heraus, dass der Logarithmus zwar noch Bestandteil schulischer Lehre ist, aber nicht mehr zwingend alle Rechenregeln hierzu. Somit scheint hier die unbeaufsichtigte Administration der Online-Form zu Vorteilen und somit einer Reduzierung des Schwierigkeitsparameters zu führen. Abbildung 4 zeigt ein Beispiel hierfür.

Trotz der hieraus resultierenden leicht unterschiedlichen itembezogenen Zusammensetzungen zwischen den Versionen sowie den Modi zeigt Tabelle 3, dass diese verschiedenen Zusammensetzungen keinen nennenswerten Einfluss auf die jeweilige mittlere Testschwierigkeit haben. Somit müssen diese Effekte im Folgenden nicht berücksichtigt werden, und es kann von einer gemeinsamen, interpretierbaren Skalierung ausgegangen werden.

Sei $x \neq -1$ eine reelle Zahl.
Welches ist eine richtige Umformung von $\frac{x^2-1}{x+1}$?

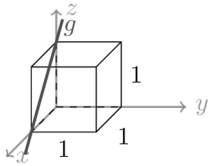
	richtig	falsch
$x-1$	☒	☐
$x+1$	☐	☒

Seien x, y reelle Zahlen, $x \neq 0$ und $y \neq 0$.
Welches ist eine richtige Umformung von $\frac{1}{\frac{1}{x}+\frac{1}{y}}$?

	richtig	falsch
$x+y$	☐	☒
$\frac{1}{x+y}$	☒	☐

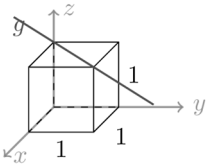
Abbildung 2: Teil einer Beispielaufgabe aus dem Bereich Terme mit unterschiedlichen zugrundeliegenden Theoremen

Die Gerade g verläuft durch die Punkte $(1,0,0)$ und $(0,0,1)$ (vgl. Skizze), die Gerade h verläuft durch die Punkte $(1,0,1)$ und $(1,1,1)$. Wie groß ist der Abstand der beiden Geraden g und h?



☐ 1 ☐ $1/2$ ☐ $\sqrt{2}$ ☒ $\sqrt{2}/2$

Die Gerade g verläuft durch die Punkte $(0,0,1)$ und $(1,1,1)$ (vgl. Skizze), die Gerade h verläuft durch die Punkte $(1,0,0)$ und $(0,1,0)$. Wie groß ist der Abstand der beiden Geraden g und h?



☒ 1 ☐ $1/2$ ☐ $\sqrt{2}$ ☐ $\sqrt{2}/2$

Abbildung 3: Beispielaufgabe aus dem Bereich Geometrie mit versionsabhängiger Auffälligkeit im *Differential Item Functioning*

Sei $x = \log 3$. Drücken Sie $\log 9$ mit Hilfe von x aus.

☐ x ☒ $2x$ ☐ $3x$ ☐ $4x$

Abbildung 4: Beispielaufgabe aus dem Bereich Funktionen mit modusabhängiger Auffälligkeit im *Differential Item Functioning*

Tabelle 3: Varianzanalyse der Itemschwierigkeit bezogen auf Version und Modus

Merkmal	Varianzanalyse	Effektgröße
Version	$F(2) = 0.19; p = .829$	$\eta^2_{\text{part}} = .002$
Modus	$F(1) = 0.10; p = .753$	$\eta^2_{\text{part}} = .001$
Version * Modus	$F(2) = 0.00; p = 1.000$	$\eta^2_{\text{part}} = .000$

Anmerkung: Die Anzahl der Items beträgt jeweils 28, außer für Version A lediglich 27.

Bezüglich des Geschlechts findet sich in den Kennwerten $h^2 = .98$ und $\text{MAD} \in [0.18, 0.25]$ kein Hinweis auf DIF. Somit können der Test und die Skala als fair bezüglich des Geschlechts angenommen werden.

3.3 Studiengangsunterschiede

Das Gruppieren der Studiengänge ergibt insgesamt 17 inhaltlich untereinander abgrenzbare Studiengangsgruppen. Dabei resultieren Gruppenmittelwerte im Intervall $\bar{\theta} \in [-0.65, 0.70]$ sowie ein Zwischengruppenvarianzanteil bezüglich der verschiedenen Gruppen an der Personenfähigkeit von $\eta^2 = .195$ $(F(16, 3440) = 52.1;\ p < .001)$. Innerhalb der Studiengangsgruppen liegen die Zwischengruppenvarianzanteile bezüglich der einzelnen Studiengänge im Bereich $\eta^2 \in [.000, .050]$. Die Signifikanz dieser Varianzanalysen liegt im Bereich $p \in [.087, .969]$.

Die Zwischengruppenvarianzanteile sprechen dafür, dass die Zusammenfassung in die Gruppen zu interpretierbaren Vergleichen führt. Die Zwischengruppenvarianzanteile in den einzelnen Studiengangsgruppen zeigen dabei keine messbaren Unterschiede. Die Skala kann also gut zwischen verschiedenen Leistungen differenzieren, wobei ein nicht unerheblicher Varianzanteil auf individueller Ebene verbleibt. Eine Analyse, inwieweit dies den Anforderungen der verschiedenen Studiengänge entspricht, ist noch ausstehend und wird im Ausblick aufgegriffen. Da gerade die Grundlagenveranstaltungen in Höherer Mathematik gerne über Studiengänge hinweg zusammengefasst werden, ergeben sich hieraus zu berücksichtigende Aspekte bezüglich der Leistungsheterogenität der Studierenden.

3.4 Niveaumodellierung

Es resultieren vier auf der Logit-Skala äquidistante Schwellen für das Niveaumodell bei $\theta_1 = -2.1$, $\theta_2 = -0.9$, $\theta_3 = 0.3$ sowie $\theta_4 = 1.5$. Einschließlich des nicht beschreibbaren Niveaus können somit fünf Niveaustufen generiert werden. Aus den Verteilungsparametern der Personenfähigkeit $(\theta \sim \mathcal{N}(0,1))$ folgen somit lediglich 1% der Personen auf dem nicht beschreibbaren Niveau X, 17% der Personen auf Niveau I, 43% der Personen auf Niveau II, 32% der Personen auf Niveau III sowie 7% der Personen auf Niveau IV. Die Beschreibung der Niveaus gestaltet sich wie folgt:

- Studierende auf Niveau I können ausreichend sicher …
  - Grundwissen der Sekundarstufe I wiedererkennen.
  - Grundtechniken der Sekundarstufe I anwenden.
- Studierende auf Niveau II können ausreichend sicher …
  - einschrittige Probleme lösen.
  - Algorithmen der Sekundarstufe I anwenden, wenn die Lösungsidee offensichtlich ist.
- Studierende auf Niveau III können ausreichend sicher …
  - mehrschrittige Probleme vollständig lösen.
  - Algorithmen der Sekundarstufe II anwenden, wenn die Lösungsidee trainiert wurde.
- Studierende auf Niveau IV können ausreichend sicher …
  - mehrschrittige Probleme vollständig lösen.
  - Algorithmen mehrerer Wissensbereiche verknüpfen, auch wenn die Lösungsidee unbekannt ist.

Diese Abstufungen zeigen Parallelen zum in den PISA-Studien identifizierten Niveaumodell (OECD, 2019, S. 115). Dort findet sich ab Stufe 3 die Fähigkeit, „sequentielle Entscheidungen“ zu berücksichtigen, was hier mit der Mehrschrittigkeit der Problemlösung aufgenommen wird. Allerdings erreichen im Jahr 2018 54% der PISA-Teilnehmenden diese Stufe, in dieser Modellierung lediglich 39% der Teilnehmenden. Die Verknüpfung mehrerer Wissens-

bereiche findet sich auf Stufe 6 („Sie können verschiedene Informationsquellen und Darstellungen miteinander verknüpfen“) wieder. Diese wird von lediglich 2% der PISA-Teilnehmenden, aber von 7% der Teilnehmenden in dieser Modellierung erreicht.

Insbesondere die unteren Niveaustufen scheinen problematisch. Es fehlt im Allgemeinen weniger das Wissen über die Mathematik der Sekundarstufe I, sondern das Anwenden und Verknüpfen dieses Wissens, um problemhaltige Situationen lösen zu können. Gerade dies ist aber grundlegender Bestandteil von WiMINT-Studiengängen.

3.5 Prognostische Validität

Das finale Modell zur Erklärung des Scheinerwerbs unter Einbezug aller relevanter Merkmale wird in Tabelle 4 dargestellt. Die einbezogene Stichprobengröße beträgt $n = 335$. Der Anteil erklärter Varianz liegt bei 51.3% ($\chi^2(4) = 155.3; p < .001$).

Tabelle 4: Modell der logistischen linearen Regression zur Erklärung des Scheinerwerbs

exogene Variable	Odds-Ratio	$\Delta R^2_{\text{Nagelkerke}}$	Modellvergleichstest
Gesamtnote	0.1	19.3%	$\chi^2(1) = 67.0; p < .001$
Testskala	2.5	4.5%	$\chi^2(1) = 17.0; p < .001$
Schulart	3.5	3.9%	$\chi^2(1) = 14.7; p < .001$
Zeitlicher Abstand	0.4	3.1%	$\chi^2(1) = 11.6; p < .001$

Anmerkung: Die beiden letzten Spalten beziehen sich auf einen Modellvergleich zwischen einem Modell mit allen vier exogenen Variablen sowie einem Modell mit den drei anderen verbleibenden exogenen Variablen. Somit ist hiermit die alleinige Erklärungskraft unter Berücksichtigung des Restmodells dargestellt.

Obwohl die Gesamtpunktzahl in Mathematik der HZB ebenfalls einen erheblichen univariaten Erklärungsanteil zum Erfolg aufweist (36%), bleibt sie im finalen Modell unberücksichtigt. Dies liegt darin begründet, dass die Gesamtnote der HZB allein eine Varianzaufklärung an der Punktzahl Mathematik in Höhe von $R^2 = .62$ sowie gemeinsam mit der Testskala in Höhe von $R^2 = .65$ liefert. Es liegt nahe, dass genau diese gemeinsame Varianz in die Aufklärung des Erfolgs eingeht und die alleinige Mathematikleistung der HZB ihre Vorhersagekraft damit verliert. Dass dabei aber die Testskala im Modell erhalten bleibt, verstärkt die Interpretation, dass hier Leistungen gemessen werden, die nur bedingt in die einbezogenen HZB-Maße Eingang finden.

In Abbildung 5 wird der Effekt der Testskala visualisiert. Hierbei wird statt dem intervallskalierten EAP-Schätzer ($R^2_{\text{Nagelkerke}} = 23.5\%; \chi^2(1) = 67.4; p < .001$) das ordinalskalierte erreichte Niveau verwendet, um die visuelle Aussagekraft, auch aus Sicht der betroffenen Interpretierenden, zu erhöhen. Es zeigen sich disjunkte Konfidenzintervalle auf dem 95%-Niveau.

Es zeigt sich somit, dass die Maße der Hochschulzugangsberechtigung ein bedeutsamer Prädiktor für den Studienerfolg sind, insbesondere da diese Leistungsmerkmale über einen längeren Zeitraum hinweg einbeziehen und weitere Merkmale, wie zum Beispiel Anstrengungsbereitschaft, integrieren. Das hier vorgestellte Instrument kann trotzdem mit geringem Aufwand einen ergänzenden Beitrag leisten und insbesondere im Hinblick auf die einfach zu interpretierende Skala Möglichkeiten zu individuellen Interventionen aufzeigen.

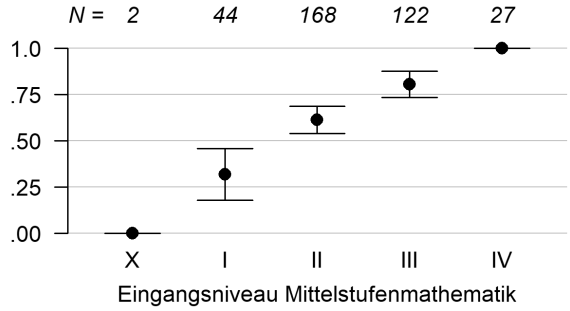


Abbildung 5: Wahrscheinlichkeit des Scheinerwerbs in Höhere Mathematik 1 nach erreichtem Eingangsniveau

Anmerkung: Die Fehlerbalken visualisieren das 95%-Konfidenzintervall der Mittelwertsschätzung.

4 Diskussion und Ausblick

4.1 Zusammenfassung der Ergebnisse

Mangels geeigneter Instrumente zur Überprüfung der mathematischen Basisfähigkeiten von Studierenden in der Studieneingangsphase wurde ein neues Instrument entwickelt, das sich insgesamt als reliabel und valide interpretierbar darstellt. Unter Verwendung eines 2PL-Bifactor-Modells mit vier Residualfaktoren (differenziert nach den inhaltlichen Testblöcken) kann die Gesamtfähigkeit der Studierenden modelliert werden, ein Rateparameter ist nicht notwendig (Fragestellung 1). Dieses kann mittels eines Niveaumodells auch für statistisch und psychometrisch Ungeübte einfach und verständlich interpretiert werden (Fragestellung 4).

Hinsichtlich der Darreichungsmodi sowie der Testvarianten ergeben sich nur marginale Unterschiede, welche in der Modellierung berücksichtigt werden. Im Vergleich der beiden Modi stellte sich ein Konstruktionsfehler heraus: Die Kennwerte des DIF identifizierten die beiden Logarithmus-Aufgaben als auffällig, was nach Betrachtung der Bildungspläne auf deren mangelnde curriculare Verankerung zurückzuführen ist. Darüber hinaus konnten aber keine Auffälligkeiten identifiziert werden, obwohl die Online-Bearbeitung ohne Aufsicht erfolgte. Die Zeitbegrenzung sowie die Freiwilligkeit der Teilnahme führen somit zu keinen Verzerrungen des gemessenen Konstrukts. Insgesamt sind alle Testformen auch in ihrer Schwierigkeit vergleichbar. Testfairness bezüglich des Geschlechts ist gegeben (Fragestellung 2).

Die Leistungsdifferenzierung nach Studiengängen sowie das Gruppieren in aussagekräftige Studiengangsgruppen zeigt erwartungskonforme und interpretierbare Ergebnisse. Eine differentielle Validität des Instruments kann somit impliziert werden (Fragestellung 3). Auch bezüglich der Vorhersagekraft zeigen sich eindeutige Effekte. Die Abiturnote hat zwar nach wie vor – als einziges der einbezogenen Merkmale – noch eine deutlich höhere Erklärungskraft an der erfolgreichen Teilnahme der Höheren Mathematik im ersten Studiensemester, trotzdem ergänzt der Test einen substanziellen zusätzlichen Anteil (Fragestellung 5). Dieser zusätzliche Anteil kann entscheidend sein für die Abwägung eines Studienabbruchs aufgrund fehlender Basisfähigkeiten, da er im Gegensatz zur allgemeineren Hochschulzugangsberech-

tigung ein gezielt zugeschnittenes Kompetenzspektrum erfasst und darüber hinaus mit zielführendem Feedback ergänzt.

Übergreifend zeigt sich, dass – obwohl das Instrument aus einer rein fachwissenschaftlichen Perspektive entwickelt wurde – eine hohe psychometrische Güte erreicht werden konnte. In einem nächsten Schritt sind nun, neben der Klärung der im Ausblick beschriebenen offenen Fragestellungen, konkrete praktische Implikationen abzuleiten. Dies betrifft insbesondere die Gestaltung der Hochschullehre in den Grundlagenveranstaltungen. Die identifizierten, individuellen Förderbedarfe könnten durch ergänzende, niedrigschwellige Zusatzangebote adressiert werden. Gleichzeitig könnte eine Diskussion über die gängige Praxis, viele Studiengänge in gemeinsamen Lehrveranstaltungen zur Höheren Mathematik zu bündeln, angestoßen werden.

4.2 Limitationen und Ausblick

Die Güte der Modellierung des Tests spricht für ein großes Potenzial dieses Instruments, wirksam gegen spätere Studienabbrüche zu sein. Dennoch sind einige Limitationen festzuhalten.

Erstens ist der Themenumfang aufgrund der begrenzten Testzeit klar eingeschränkt. Zwar scheinen die Themen adäquat ausgewählt, wofür die hohe prädiktive Kraft spricht. Eine Überinterpretation im Sinne einer allgemeinen mathematischen Fähigkeit ist jedoch keinesfalls zulässig. Rückmeldungen durch betroffene Dozierende zeigen zugleich, dass diese Interpretationseinschränkung der praktischen Bedeutung der Resultate für die Gestaltung der Lehrveranstaltungen nur wenig entgegenwirkt. Die durch die Technologiebasierung fehlende direkte Interaktion mit den Testteilnehmenden macht eine finale Bewertung dieser Facette insbesondere auf Ebene der Testteilnehmenden jedoch schwierig. Inwieweit die Testteilnehmenden also das Feedback auch im Sinne einer konkreten Nachholempfehlung annehmen und umsetzen, bleibt derzeit offen.

Zweitens sind in den Stichproben spezifische Selektionseffekte zu vermuten: Die Mathematik-Vorkurse sind an allen Standorten ein freiwilliges Zusatzangebot vor Studienbeginn. Zur Frage, wie sich die hier stattfindenden Selbstselektionen auf die Zusammensetzung dieser Gruppe bezüglich aller betroffenen Studienanfangenden auswirkt, gibt es nur wenig Evidenz. Karaponos und Pelz (2021) sehen hier lediglich geringe Effekte, eine Übertragbarkeit der Erkenntnisse auf die vorliegende Stichprobe ist aber zweifelhaft. Dies macht sich insbesondere auch bezüglich der im Folgenden aufgegriffenen ausstehenden Fragestellungen bemerkbar.

Nichtsdestotrotz können mit dem verwendeten technologie- und IRT-basierten Ansatz einige Vorteile gegenüber klassischem Testen erreicht werden: Es steht direkt ein Feedback für die Testteilnehmenden sowie für weitere betroffene Personengruppen (zum Beispiel Dozierende) zur Verfügung, welches nicht nur auf der Aufgabenebene oder einer summarischen Ebene, sondern direkt auf der Konstruktebene interpretiert werden kann. Dieser Ansatz befreit aber nicht von der Verpflichtung, ein Feedback zu entwickeln, welches qualitativ hochwertig ist. Im Gegenteil werden durch die fehlende Möglichkeit der Interaktion mit den Testteilnehmenden besonders hohe Anforderungen an die Test- und die Feedbackqualität notwendig. Die hier genannten Limitationen führen zu vier zentralen Implikationen für die Weiterentwicklung des Testinstruments und der analytischen Vorgehensweise:

(1) Die in diesem Beitrag berichteten Reliabilitätsmaße zeigen noch ein deutliches Verbesserungspotential. Insbesondere die Verwendung als individualdiagnostisches Instrument verlangt, dass hierauf weiterhin ein besonderes Augenmerkt gelegt wird. Es fanden bereits erfolgversprechende Ansätze Anwendung, die einerseits durch *Computerized Adaptive Testing* eine passgenauere Itemauswahl ermöglichen und andererseits durch die verwendeten Feedbackskalen die Bedeutsamkeit des Messfehlers für die Interpretation reduzieren. Die Erkenntnisse hierzu werden sich aus Umfangsgründen in einem weiteren Beitrag finden.

(2) In diesem Beitrag wurden die Ergebnisse erster Validitätsüberprüfungen berichtet. Allerdings sind weitere Prüfungen unerlässlich, um sicherstellen zu können, dass das entwickelte Leistungsmaß zuverlässig und zielgerichtet interpretiert werden kann. Einerseits sind Betrachtungen der inkrementellen Validität gegenüber alternativen Instrumenten und Skalen für eine bessere Einordnung notwendig. Andererseits muss eine repräsentative Stichprobe zur Prüfung der prognostischen Validität gefunden werden. Darüber hinaus muss der Test zu verschiedenen anderen Zeitpunkten, idealerweise längsschnittlich, eingesetzt werden, um einerseits eine Unabhängigkeit von den Selektionseffekten der Vorkurszusammensetzung und andererseits eine Eignung für die Studienwahlphase analysieren zu können.

(3) Inwieweit die differentiellen Unterschiede in den verschiedenen Studiengangsgruppen auch die tatsächlichen Anforderungen dieser Studiengänge widerspiegeln, ist noch nicht geklärt. Gleichzeitig scheint überaus interessant, welche Mindestniveaus für verschiedene Studiengänge oder Lehrveranstaltungen erreicht werden müssen. Hierzu ist eine Befragung der betroffenen Dozierenden auf Itemebene geplant, welche diese Informationen zuverlässig herausarbeiten kann. Gleichzeitig können hieraus noch Potentiale identifiziert werden, in welchen Skalenbereiche auf welche Art weitere Differenzierungen beziehungsweise Aufgaben integriert werden müssen, um die Aussagekraft der Interpretation weiter erhöhen zu können.

(4) Belastbare Aussagen zur Qualität und zu Optimierungsbedarfen des entwickelten Feedbacks sowie zu dessen praktischen Implikationen stehen ebenfalls noch aus. So sind jeweils Interviews geplant, welche sowohl die konkreten Fehlerursachen bei der Testbearbeitung als auch die konkreten Interpretationen der Studierenden bezüglich ihres Feedbackverständnisses erfassen sollen, um Informationen zu erhalten, welche Feedbackelemente zu Fehlinterpretationen führen. Dass die Ergebnisse bei den Dozierenden Reflektionsprozesse anstoßen, wurde von dieser Zielgruppe vielfach rückgemeldet. Inwieweit das Instrument auch Verhaltensänderungen bei den Testteilnehmenden und den Dozierenden auslöst sowie Studienabbrüchen entgegenwirkt, ist eine bislang offene empirische Frage.

Hinweis

Die hier vorgestellten Entwicklungen und Ergebnisse entstanden in einem Forschungsprojekt, welches vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg in Förderlinie 4 des Fonds *Erfolgreich Studieren* gefördert wurde.

Literatur

AERA, APA & NCME (2014). *Standards for educational and psychological testing.* AERA.

Bach, V. (2016). Kompetenzorientierung und Mindestanforderungen. *Mitteilungen der Deutschen Mathematiker-Vereinigung, 24*(1), 30–32. https://doi.org/10.1515/dmvm-2016-0015.

Bausch, I., Biehler, R., Bruder, R., Fischer, P. R., Hochmuth, R., Koepf, W., Schreiber, S. & Wassong, T. (Hrsg.). (2014a). *Mathematische Vor- und Brückenkurse. Konzepte, Probleme und Perspektiven*. Springer Spektrum. https://doi.org/10.1007/978-3-658-03065-0.

Bausch, I., Biehler, R., Bruder, R., Fischer, P. R., Hochmuth, R., Koepf, W. & Wassong, T. (2014b). VEMINT – Interaktives Lernmaterial für mathematische Vor- und Brückenkurse. In I. Bausch, R. Biehler, R. Bruder, P. R. Fischer, R. Hochmuth, W. Koepf, S. Schreiber & T. Wassong (Hrsg.), *Mathematische Vor- und Brückenkurse (Konzepte und Studien zur Hochschuldidaktik und Lehrerbildung Mathematik).* Springer Spektrum. https://doi.org/10.1007/978-3-658-03065-0_18.

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

Belcadhi, L. C. (2016). Personalized feedback for self assessment in lifelong learning environments based on semantic web. *Computers in Human Behavior, 55*(A), 562–570. https://doi.org/10.1016/j.chb.2015.07.042.

Bond, T. & Fox, C.M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9781315814698.

Brunner, S. (2017). *Online-Self-Assessments*. Koordinierungsstelle der Begleitforschung des Qualitätspaktes Lehre (KoBF).

Brunner, S., Ranft, A. & Wittig, W. (2015). Online-Self-Assessments: die Bedeutung von Feedback und Implikationen für die (Weiter-)Entwicklung von Verfahren für beruflich qualifizierte Studieninteressierte. In A. Hanft, O. Zawacki-Richter & W. B. Gierke (Hrsg.), *Herausforderung Heterogenität beim Übergang in die Hochschule* (S. 145–162). Waxmann.

Burnham, K. P. & Anderson, D. R. (2002). *Model selection and multimodel inference*. *A practical information-theoretic approach*. Springer. https://doi.org/10.1007/b97636.

Chen, W. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289. https://doi.org/10.2307/1165285.

CoSH-Gruppe (Hrsg.). (2021). *Mindestanforderungskatalog Mathematik*. Version 3.0. https://lehrerfortbildung-bw.de/u_matnatech/mathematik/bs/bk/cosh/katalog/makv3.0.pdf

Dürrschnabel, K., Dürr, R., Erben, W., Gercken, M., Lunde, K., Wurth, R. & Zimmermann, M. (2019). *So viel Mathe muss sein! Gut vorbereitet in ein WiMINT-Studium*. Springer Spektrum. https://doi.org/10.1007/978-3-662-57951-0.

Grisay, A. & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33*(1), 69–86. https://doi.org/10.1016/j.stueduc.2007.01.006.

Hanft, A., Zawacki-Richter, O. & Gierke, W. B. (Hrsg.). (2015). *Herausforderung Heterogenität beim Übergang in die Hochschule*. Waxmann.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487.

Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. & Woisch, A. (2017). *Motive und Ursachen des Studienabbruchs an baden-württembergischen Hochschulen und beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher*. DZHW Projektbericht.

Heublein, U., Richter, J. & Schmelzer, R. (2020). Die Entwicklung der Studienabbruchquoten in Deutschland. *DZHW BRIEF*, *3*, 1–12.

Jahnke, T., Klein, H. P., Kühnel, W., Sonar, T. & Spindler, M. (2014). Die Hamburger Abituraufgaben im Fach Mathematik. Entwicklung von 2005 bis 2013. *Mitteilungen der Deutschen Mathematiker-Vereinigung, 22*(2), 115–122. https://doi.org/10.1515/dmvm-2014-0046.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer. https://doi.org/10.1007/b98835.

Karapanos, M. & Pelz, R. (2021). Wer besucht Mathematikvorkurse? *Zeitschrift für Erziehungswissenschaft, 24*(5), 1231–1252. https://doi.org/10.1007/s11618-021-01035-2.

Karst, K., Ertelt, B.-J., Frey, A. & Dickhäuser, O. (2017). Studienorientierung durch Self-Assessments: Veränderung von Einstellungen zum Studienfach während der Bearbeitung eines Selbsttests. *Journal für Bildungsforschung Online, 9*(2), 205–227. https://doi.org/10.25656/01:14935.

Knorrenschild, M. (2004). *Vorkurs Mathematik. Ein Übungsbuch für Fachhochschulen.* Carl-Hanser.

Kubinger, K. D., Frebort, M. & Müller, C. (2012). Self-Assessment im Rahmen der Studienberatung: Möglichkeiten und Grenzen. In K. D. Kubinger, M. Frebort, L. Khorramdel & L. Weitensfelder (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 9–24). Pabst Science Publishers.

Krawitz, J. (2020). *Vorwissen als nötige Voraussetzung und potentieller Störfaktor beim mathematischen Modellieren*. Springer Spektrum. https://doi.org/10.1007/978-3-658-29715-2.

Krunke, S. O., Roegner, K., Schüler, L., Seiler, R. & Stens, R. L. (2012). Der Online-Mathematik-Brückenkurs OMB. Eine Chance zur Lösung der Probleme an der Schnittstelle Schule/Hochschule. *Mitteilungen der Deutschen Mathematiker-Vereinigung, 20*(2), 115–120. https://doi.org/10.1515/dmvm-2012-0048.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Praxis. *Psychologische Rundschau, 58*(2), 103–117. https://doi.org/10.1026/0033-3042.58.2.103.

Madsen, H. & Thyregod, P. (2010). *Introduction to General and Generalized Linear Models*. CRC Press.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*, 71–137. https://doi.org/10.1080/15366367.2013.831680.

Meyberg, K. & Vachenauer, P. (2001). *Höhere Mathematik 1*. Springer. https://doi.org/10.1007/978-3-642-56654-7.

Ministerium für Kultus, Jugend und Sport Baden-Württemberg (KM BW) & Zentrum für Schulqualität und Lehrerbildung (ZSL) (Hrsg.). (2016). *Bildungspläne Baden-Württemberg. Gymnasium – Mathematik*. https://www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GYM_M.pdf

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691–692. https://doi.org/10.1093/biomet/78.3.691.

Neubrand, M., Biehler, R., Blum, W., Cohors-Fresenborg, E., Flade, L., Knoche, N., Lind, D., Löding, W., Möller, G. & Wynands, A. (2004). Eine systematische und kommentierte Auswahl von Beispielaufgaben des Mathematiktests in PISA 2000. In M. Neubrand (Hrsg.), *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland* (Bd. 23) (S. 259–270). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-80661-1_13.

Neugebauer, M., Heublein, U. & Daniel, A. (2019). Studienabbruch in Deutschland: Ausmaß, Ursachen, Folgen, Präventionsmöglichkeiten. *Zeitschrift für Erziehungswissenschaft, 22*, 1025–1046. https://doi.org/10.1007/s11618-019-00904-1.

Neumann, I., Pigge, C. & Heinze, A. (2017). *Welche mathematischen Lernvoraussetzungen erwarten Hochschullehrende für ein MINT-Studium?* IPN.

Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) (2019). *PISA 2018 Ergebnisse*. Band I. Was Schülerinnen und Schüler wissen und können. wbv Media. https://doi.org/10.3278/6004763w.

Petri, P. S. (2020). *Ein Prozessmodell des Studieneinstiegs. Differentielle Aspekte studiumsbezogener Kognitionen und deren Effekte auf Studienerfolg und Studienabbruch*. Dissertation, Justus-Liebig-Universität Gießen.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555.

Reise, S. P. (Hrsg.). (2015). *Handbook of item response theory modeling. Applications to typical performance assessment*. Routledge.

Reiss, K., Weis, M., Klieme, E. & Köller, O. (2019). *PISA 2018. Grundbildung im internationalen Vergleich*. Waxmann. https://doi.org/10.31244/9783830991007.

Robitzsch, A. & Steinfeld, J. (2018). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling, 60*(1), 101–139.

Schäfer, W., Georgi, K. & Trippler, G. (1997). *Mathematik-Vorkurs. Übungs- und Arbeitsbuch für Studienanfänger.* Teubner. https://doi.org/10.1007/978-3-322-97616-1.

Schwippert, K., Kasper, D., Köller, O., McElvany, N., Selter, C., Steffensky, M. & Wendt, H. (2020). *TIMSS 2019. Mathematische und naturwissenschaftliche Kompetenzen von Grundschulkindern in Deutschland im internationalen Vergleich*. Waxmann.

Tripp, A. & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? *Journal of Nursing Education, 24*(3), 92–98. https://doi.org/10.3928/0148-4834-19850301-04.

TU9 (Allianz führender Technischer Universitäten in Deutschland) (2020). *Starthilfe fürs Studium: Online-Brückenkurs Physik. TU9-zertifiziertes kostenfreies digitales Lernangebot.*

van der Linden, W. J. & Ren, H. (2019). A fast and simple algorithm for Bayesian adaptive testing. *Journal of Educational and Behavioral Statistics, 45*(1), 58–85. https://doi.org/10.3102/1076998619858970.

Wilson, M. (2005). *Constructing measures. An item response modeling approach*. Routledge.

Kontakt

Stefan Behrendt
Schwenninger Str. 21
78083 Dauchingen
E-Mail: behrendt@bestetistics.de

Jan Köllner
Universität Stuttgart
Institut für Analysis, Dynamik und Modellierung
Pfaffenwaldring 57
70569 Stuttgart
E-Mail: jan.koellner@mathematik.uni-stuttgart.de

Prof. Dr. Kristina Kögler
Universität Stuttgart
Institut für Erziehungswissenschaft
Geschwister-Scholl-Str. 24D
70174 Stuttgart
E-Mail: koegler@bwt.uni-stuttgart.de

Prof. Dr. Christine Sälzer · Andreas Just
Universität Stuttgart
Institut für Erziehungswissenschaft
Azenbergstr. 16
70174 Stuttgart
E-Mail: christine.saelzer@ife.uni-stuttgart.de
E-Mail: just@bwt.uni-stuttgart.de