**Beitrag Themenheft „Testen und Prüfen“**

Aron Fink, Andreas Frey und Katharina Klein*

# Deutsch im Beruf – eine Untersuchung der psychometrischen Güte des Goethe-Test PRO

## German for professionals – an evaluation of the psychometric quality of the Goethe-Test PRO

https://doi.org/10.1515/infodaf-2022-0065

**Zusammenfassung:** Deutsch im Beruf ist mit zunehmender Migration und Mobilität ein zentrales Thema im DaF/DaZ-Bereich. Berufsbezogene Sprachkenntnisse können mithilfe von verschiedenen Tests für den Beruf diagnostiziert werden. Ein Test zum Nachweis berufsbezogener Sprachkenntnisse ist der Goethe-Test PRO: Deutsch für den Beruf (GTP). Der vorliegende Beitrag widmet sich zunächst der berufsorientierten Ausrichtung, dem Testformat sowie dem adaptiven Testverfahren des GTP. Der zweite Teil dieses Beitrags beschäftigt sich mit der Beschreibung seiner psychometrischen Güte. Abschließend werden die Ergebnisse mit Blick auf die Relevanz von berufsbezogenen Tests für Deutsch als Fremdsprache diskutiert.

**Schlüsselwörter:** Berufsorientierung, Computerisiertes Adaptives Testen, psychometrische Güte, Testen

**Abstract:** With increasing migration and mobility, German at the workplace has become an important topic in the field of German as a foreign Language/German as a second language. Vocational language skills can be measured with the help of various tests for the profession. The Goethe-Test PRO: German for professionals (GTP) is one of those tests that provides evidence of vocational language skills. This article first deals with the vocational orientation, the test format and the

**Aron Fink,** E-Mail: a.fink@psych.uni-frankfurt.de
**Andreas Frey,** E-Mail: frey@psych.uni-frankfurt.de
***Kontaktperson: Katharina Klein,** E-Mail: katharina.klein@goethe.de

adaptive test procedure of the GTP. The second part of this paper is devoted to the description of its psychometric quality. Finally, the results are discussed with regard to the relevance of vocational tests for German as a foreign language.

**Keywords:** language for specific purposes, computerized adaptive testing, psychometric quality, testing

# 1 Einleitung

Weltweit lernen aktuell über 15 Millionen Menschen Deutsch als Fremdsprache (Auswärtiges Amt 2020: 9). Dabei ist für diese Lernenden Deutschland nicht nur als Studien-, sondern auch als Arbeitsort attraktiv (ebd.: 6).

Die seit 2015 relativ gleichbleibend hohe Deutschlernendenzahl weltweit (ebd.: 9), aber auch die Arbeitswelt beeinflussende Faktoren unserer Zeit wie Digitalisierung, Globalisierung sowie demografischer, kultureller und gesellschaftlicher Wandel (BMAS 2016: 18–39) tragen dazu bei, dass berufliche Mobilität thematisch aus dem DaF-Unterricht – insbesondere im Ausland – nicht mehr wegzudenken ist (Dimitrijević 2020: 14). Mit Blick auf den 2021 veröffentlichten Bericht der Internationalen Arbeitsorganisation (ILO) ist der Fokus auf berufsbedingte Mobilität im Rahmen der sprachlichen Qualifizierung von Sprachkursteilnehmenden nicht verwunderlich: Im Jahre 2019 sind von 272 Millionen Migrantinnen und Migranten weltweit allein 169 Millionen aus beruflichen Gründen migriert (ILO 2021: 11).

Wie wichtig neben der fachlichen Qualifikation auch der Nachweis von Sprachkenntnissen auf dem deutschen Arbeitsmarkt ist, verdeutlicht eine bereits vor einigen Jahren von der OECD und dem DIHK gemeinsam durchgeführte Befragung deutscher Unternehmen. Von diesen Unternehmen schätzten über 60 Prozent Deutschkenntnisse als Kriterium für die Auswahl von ausländischen Arbeitskräften als sehr wichtig ein (OECD 2013: 129–130).

Die Zielgruppe der Erwerbsmigrantinnen und -migranten wird zudem durch das am 1. März 2020 in Kraft getretene Fachkräfteeinwanderungsgesetz (online: https://fachkraefteeinwanderungsgesetz.de/) in den Fokus gerückt. Das Gesetz geht auf den aus vielen Branchen beklagten Fachkräftemangel (online: https://de.statista.com/themen/887/fachkraeftemangel/) mit einem vereinfachten Anwerbeverfahren von qualifizierten Fachkräften aus Drittländern, das heißt Nicht-EU-Ländern, ein. Auch im Fachkräfteeinwanderungsgesetz ist der Nachweis von Deutschkenntnissen verankert. Gefordert wird in der Regel ein B1- oder B2-Niveau nach dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (Europarat 2001: 33–35).

Zum Nachweis der Deutschkenntnisse bedarf es standardisierter Testverfahren, die in der Erwachsenenbildung im Lernprozess unterschiedliche Funktionen erfüllen. Neben dem Einstufungs-, Diagnostik- und Lernfortschrittstest *(achievement test)* gibt ein Sprachstandstest *(proficiency test)* Auskunft über die sprachliche Performanz der getesteten Person zum Zeitpunkt der Prüfungsabnahme (Handt 2002: 187). Das Ergebnis von Sprachstandstests kann durch ein Sprachenzertifikat bescheinigt werden. Ein solches Sprachenzertifikat „wird als Voraussage über die sprachliche Leistungsfähigkeit einer geprüften Person in der realen Welt interpretiert" (Perlmann-Balme 2016: 420).

Die eingangs beschriebene wachsende Zahl der Erwerbsmigrantinnen und -migranten hat nicht nur Auswirkungen auf die sprachliche Qualifizierung im DaF-Unterricht (Dimitrijević 2020: 16–18), sondern zeigt auch im Kontext des Testens und Prüfens den Bedarf an speziellen Testverfahren auf, die auf die Bedürfnisse dieser Zielgruppe abgestimmt sind. Insbesondere fallen hierunter berufsorientierte Prüfungen, die – im Gegensatz zu allgemeinsprachlichen Prüfungen – Charakteristika von Berufssprache in den Aufgaben berücksichtigen.

Ein auf die oben genannte Zielgruppe abgestimmter Test ist der Goethe-Test PRO: Deutsch für den Beruf (GTP). Der GTP ist ein psychometrisch konstruierter, adaptiver Deutschtest, der die Hör- und Lesekompetenz am Arbeitsplatz auf den Stufen A1–C2 des Gemeinsamen Europäischen Referenzrahmens für Sprachen misst und vom Goethe-Institut weltweit angeboten wird. Ziele des hier vorliegenden Beitrags sind die Vorstellung des GTP sowie die Beschreibung seiner psychometrischen Güte. Hierfür werden in Kapitel 2 zunächst das berufsrelevante Register des GTP, sein Testformat und die weltweite Form der Durchführung beschrieben. Nach dieser einführenden Darstellung wird in Kapitel 3 der dem GTP zugrunde liegende psychometrische Ansatz, das Computerisierte Adaptive Testen (CAT), erläutert. Anschließend werden in Kapitel 4 die Fragestellungen der Studie dargestellt. Kapitel 5 beschreibt die zur Beantwortung der Fragestellungen genutzten Methoden. Schließlich werden in Kapitel 6 die Ergebnisse präsentiert und in Kapitel 7 diskutiert.

## 2 Der Sprachtest Goethe-Test PRO: Deutsch für den Beruf

Der GTP ist ein adaptiver Online-Deutschtest, der am Prüfungszentrum auf der Goethe-eigenen Testplattform (Moodle) unter standardisierten Testbedingungen abgelegt wird (siehe Abschnitt 2.3). Die Bearbeitung des Tests dauert ca. 60–90 Minuten. Als Mitglied der Association of Language Testers in Europe (ALTE;

online: https://www.alte.org/) verpflichtet sich das Goethe-Institut, bei der Entwicklung von Sprachprüfungen alle Schritte der Prüfungsentwicklung umzusetzen (ALTE 2012). Zum Nachweis der Qualität von Sprachprüfungen setzt ALTE 17 Mindeststandards an. Durch die Umsetzung dieser Mindeststandards werden die Gütekriterien von Tests und deren Nutzung, also Objektivität, Reliabilität, Validität und Fairness sowie deren Einfluss auf Gesellschaft und Unterricht, umgesetzt (Perlmann-Balme 2016: 423). Diese Standards finden sowohl bei allgemeinsprachlichen Prüfungen im Prüfungsportfolio des Goethe-Instituts als auch beim berufsorientierten Sprachtest GTP Anwendung.

## 2.1 Berufssprachliche Orientierung im Goethe-Test PRO

Deutschprüfungen für den Beruf werden von verschiedenen Testanbietern vertrieben (Fromme/Korb 2018). Die verfügbaren Tests unterscheiden sich unter anderem durch ihren Fokus auf bestimmte Berufsfelder – so zum Beispiel Medizin oder Wirtschaft – und somit durch den Grad der Spezifizierung (van Gorp/Vîlcu 2018: 3–4). Der GTP ist eine dieser Prüfungen für den Beruf und wird seit 2017 als Nachfolger des Business Language Testing Service, kurz BULATS (online: https://www.cambridgeenglish.org/bulats), vom Goethe-Institut angeboten.

Der GTP kann als berufsorientierter Test im weiteren Sinne charakterisiert werden, also im Gegensatz zu berufs- bzw. fachsprachlichen Tests im engeren Sinne (van Gorp/Vîlcu 2018: 3–4). Berufsorientierte Tests weisen im Spannungsfeld berufsrelevanter Register zwischen Allgemein- und Fachsprache eine größere Nähe zur Allgemeinsprache auf (Efing 2014: 420). In der beruflichen Kommunikation kann beobachtet werden, dass ein Großteil des sprachlichen Handelns am Arbeitsplatz nicht mit dem Register der Fachsprache bewältigt wird. In Abgrenzung zur Fachsprache entstand so der Begriff der Berufssprache (ebd.: 419). Im Berufsalltag werden zum Großteil kommunikative Handlungen verlangt, die zu allgemeinen Kommunikationsformen gezählt werden können, zum Beispiel systematisches Verarbeiten von Informationen sowohl schriftlicher als auch mündlicher Quellen (Kuhn 2019: 54).

Solche kommunikativen Handlungen werden bei der Itemerstellung unter Berücksichtigung der individuellen Ausprägungen einbezogen und im GTP durch unterschiedliche Aufgaben operationalisiert. Diese werden im weiteren Verlauf Items genannt. Dem GTP liegt eine umfangreiche Itembank zugrunde, die Items für alle GER-Niveaustufen beinhaltet (siehe Abschnitt 6.1). Beim Aufbau dieser Itembank wurde darauf geachtet, dass der Fokus auf berufsorientiertem Deutsch liegt, um den branchenübergreifenden Einsatz als Sprachnachweis zu gewährleisten.

Zielgruppe des GTP sind Erwachsene im Berufsleben, die ihre Sprachkenntnisse am Arbeitsplatz möglichst schnell, beispielsweise im Bewerbungsprozess, nachweisen möchten. Der Test kann auch direkt im Unternehmen mit Mitarbeiterinnen und Mitarbeitern oder mit Bewerberinnen und Bewerbern durchgeführt werden. Hierbei kann er als Entscheidungsgrundlage unter anderem für Fortbildungsmaßnahmen dienen, zum Beispiel berufsbegleitende Sprachkurse. In den Niederlanden wird der GTP auch an berufsbegleitenden Schulen (online: https://duitsmbo.nl/) von Berufsschülerinnen und -schülern abgelegt. Testteilnehmende erhalten direkt nach Testabschluss ein digitales Zeugnis mit Beschreibung des Sprachniveaus, das online zum Beispiel von (zukünftigen) Arbeitgebern auf Echtheit überprüft werden kann (online: https://goethe.de/pro/relaunch/prf/de/Pruefungsordnung_GTP.pdf).

## 2.2 Beschreibungen des Testformats

Der GTP testet in zwei Teilbereichen kommunikative Sprachaktivitäten (Europarat 2020: 57), die in Kombination abgelegt werden. Der erste Teilbereich testet die Fertigkeit Lesen, der zweite die Fertigkeit Hören. Dem Test liegen in diesen beiden rezeptiven Fertigkeiten die Kompetenzniveaus und die Kann-Beschreibung des Gemeinsamen Europäischen Referenzrahmens für Sprachen (Europarat 2001) zugrunde. Bei der Konstruktion der Testitems des GTP wurde das Prinzip der Handlungsorientierung berücksichtigt, welches Fremdsprachenlernende als „sozial Handelnde betrachtet, das heißt als Mitglieder einer Gesellschaft, die unter bestimmten Umständen und in spezifischen Umgebungen und Handlungsfeldern kommunikative Aufgaben bewältigen müssen […]“ (ebd.: 21). Die Testitems des GTP sind so konstruiert, dass sie die (zukünftige) Lebenswelt der Lernenden im beruflichen Kontext möglichst authentisch widerspiegeln, das heißt für die Lernenden eine unmittelbare Relevanz haben.

Getestet wird in beiden Teilbereichen anhand verschiedener, vorwiegend geschlossener Itemformate, die eine hohe Auswertungsobjektivität (Rost 2004: 39) gewährleisten. Der GTP verfügt über insgesamt acht Itemtypen, die von kurzen Einzelsätzen und Lücken bis hin zu längeren Lese- und Hörtexten reichen (Demo-Version online: www.goethe.de/gtpro). Die Auswahl der Items wird durch das adaptive Testverfahren individuell an die Testperson angepasst (siehe Kapitel 3).

Bei der Bewältigung der Lese- und Höritems rezipieren die Teilnehmenden Textsorten, die verschiedene Gesprächspartnerinnen und -partner in den Fokus rücken. Im Bereich Lesen treten kurze (in-)formelle Korrespondenzen wie zum Beispiel E-Mails, SMS oder andere Kurztexte aus sozialen Netzwerken, (Stellen-)Anzeigen, Mitteilungen oder Informationsschreiben, aber auch längere Artikel zu

beruflichen Themen auf. Der Teilbereich Hören beinhaltet kürzere Hörtexte, darunter (in-)formelle Gespräche am Arbeitsplatz, zum Beispiel Terminvereinbarungen oder Pausengespräche sowie Durchsagen und Ansagen. Bei längeren Hörtexten handelt es sich beispielsweise um Interviews oder Kundengespräche.

### 2.3 Durchführung weltweit

Der GTP wird seit 2017 an Prüfungszentren des Goethe-Instituts sowie bei Firmenkunden weltweit einheitlich durchgeführt. Um eine standardisierte Testdurchführung an allen Prüfungszentren zu garantieren, werden Durchführungsbestimmungen in der Prüfungsordnung zum GTP festgelegt, so zum Beispiel die Beaufsichtigung der Testdurchführung durch eine qualifizierte Aufsichtsperson. Die ordnungsgemäße Durchführung wird unter anderem durch regelmäßige Kontrollen der Testdurchführung sichergestellt (online: https://goethe.de/pro/relaunch/prf/de/Pruefungsordnung_GTP.pdf). Durch diese Art der standardisierten Testdurchführung, die automatisierte Auswertung und Bestimmung der Testergebnisse sowie deren Rückmeldung an die Testteilnehmenden ist das Gütekriterium der Objektivität (ALTE 2012: 57; Rost 2004: 39) für den GTP als gegeben anzusehen.

Prüfungszentren des GTP im In- und Ausland sind Goethe-Institute, Prüfungskooperationspartner und Institutionen, die vertraglich zur Durchführung berechtigt sind, unter anderen Firmenkunden (online: https://goethe.de/pro/relaunch/prf/de/Pruefungsordnung_GTP.pdf). Die Prüfungszentren mit dem höchsten GTP-Testaufkommen spiegeln sich in der Verteilung der Stichprobe in Tabelle 1 wider.

## 3 Computerisiertes Adaptives Testen

Beim Computerisierten Adaptiven Testen (CAT) (Frey 2020) orientiert sich die Auswahl der einer Testperson im Verlauf des Tests zur Bearbeitung vorgelegten Items an ihrem vorherigen Antwortverhalten. Üblicherweise führt ein solches Vorgehen dazu, dass Personen mit einer hohen Fähigkeitsausprägung schwierigere Items präsentiert werden als Personen mit einer niedrigeren Fähigkeitsausprägung. Vereinfacht gesprochen passt sich der Test bei CAT also in seinem Schwierigkeitsniveau dem Kompetenzstand der Testperson maßgeschneidert an. Dieses Vorgehen ähnelt dem klassischer mündlicher Prüfungen – den individuellen Verlauf des Tests bestimmt für jede Testperson allerdings anstatt der Prüferin beziehungsweise des Prüfers der Computer mittels eines adaptiven Algorithmus. Das Ziel dieser Vorgehensweise ist es, den Testpersonen keine individuell zu

schwierigen oder zu leichten Items zur Bearbeitung vorzulegen, um so das individuelle Kompetenzniveau der Testpersonen möglichst genau (d. h. mit möglichst geringem Messfehler) zu bestimmen. Da bei CAT Testpersonen unterschiedlicher Fähigkeitsausprägung auch unterschiedlich schwierige Items bearbeiten, ist ein interindividueller Vergleich durch simples Aufsummieren der richtig gelösten Items nicht möglich. Vielmehr müssen Itemmerkmale (z. B. die Itemschwierigkeit) in die Schätzung des Fähigkeitsniveaus miteinbezogen werden. Hierfür werden bei CAT Modelle der Item Response Theory (IRT) (van der Linden 2016) genutzt. IRT-Modelle erlauben die separate Bestimmung von Itemcharakteristika (Itemparameter) und den individuellen Testergebnissen. Dabei werden Itemparameter und Testergebnisse mathematisch auf der gleichen Dimension lokalisiert. Dies ermöglicht es, Testergebnisse zwischen Testpersonen zu vergleichen, auch wenn diese unterschiedliche Items bearbeitet haben.

Für den Einsatz von CAT ist es notwendig, dass vor der Testanwendung die Itemparameter für die komplette Menge an Items (Itempool), aus denen der Algorithmus Items auswählen kann, mithilfe der IRT bestimmt wurden. Dies erfolgt im Rahmen von Kalibrierungsstudien mit großen Stichproben.

Der grundlegende Ablauf eines Computerisierten Adaptiven Tests lässt sich wie folgt beschreiben: Der Test wählt nach vorher definierten Regeln ein Item aus dem Itempool aus und legt dieses zur Bearbeitung vor (Schritt 1). Anschließend registriert und bewertet das Computerprogramm die Antwort der Testperson auf dieses Item (Schritt 2). Mithilfe der bewerteten Antwort kann die individuelle Merkmalsausprägung geschätzt werden (Schritt 3). Anschließend wird geprüft, ob ein oder mehrere vorab definierte Abbruchkriterien (z. B. Anzahl vorgelegter Items, Testzeit) erreicht sind. Ist dies nicht der Fall, werden die Schritte 1–3 erneut durchlaufen. Die Itemauswahl erfolgt in jeder weiteren Runde auf Basis der jeweils aktuellen Schätzung der Merkmalsausprägung aus Schritt 3. Diese wird durch das Vorhandensein von zunehmend mehr Antworten mit jedem Durchgang präziser. Neben einem rein statistischen Optimalitätskriterium werden bei der Itemauswahl üblicherweise auch weitere, nicht statistische Einschränkungen miteinbezogen. Eine typische, nicht statistische Einschränkung wäre beispielsweise, dass für jede Testperson sichergestellt ist, dass jeder Inhaltsbereich des Tests gleichmäßig durch Items abgedeckt ist. Die Schritte 1–3 werden bis zum Erreichen aller Abbruchkriterien wiederholt. Anschließend wird der Test mit der endgültigen Schätzung der individuellen Merkmalsausprägung beendet. Aus diesem kurz skizzierten Ablauf lässt sich ableiten, dass bei der Konstruktion Computerisierter Adaptiver Tests Festlegungen bezüglich sechs elementarer Bausteine von CAT zu treffen sind (für detaillierte Informationen hierzu siehe Frey 2020):

- Itempool
- Itemauswahl zu Beginn des Tests

- Schätzung der individuellen Merkmalsausprägung
- Itemauswahl während des Tests
- Umgang mit Einschränkungen bei der Itemauswahl
- Kriterien zur Beendigung des Tests

# 4 Forschungsfragen

Aufgrund der hohen Relevanz der Testergebnisse des GTP sowohl für das getestete Individuum (z. B. als Einstellungskriterium) als auch Unternehmen und Organisationen, in denen der Test angewendet wird (z. B. als Basis für die Identifikation von kostenintensiven Weiterbildungsmaßnahmen), ist es wichtig sicherzustellen, dass der GTP eine hohe psychometrische Güte ausweist. Mit der vorliegenden Studie wird deshalb die psychometrische Güte des GTP differenziert untersucht. Dabei werden die folgenden Forschungsfragen betrachtet:

1 Sind die Items des GTP als eindimensional anzusehen?
2 Wie reliabel sind die Testergebnisse des GTP?
3 Wie hoch ist die Differenzierungsfähigkeit des GTP über das komplette Fähigkeitsspektrum?
4 Wie hoch ist der Grad an Adaptivität des GTP?

# 5 Methode

## 5.1 Stichprobe

Zu Beantwortung der Fragestellungen wurden die Testergebnisse von $N = 5.626$ Testpersonen herangezogen, welche seit April 2017 den GTP durchgeführt haben. Tabelle 1 zeigt die Verteilung der Testpersonen auf die verschiedenen Staaten, in denen der Test durchgeführt wurde. Zudem ist in Tabelle 2 die Verteilung der Testpersonen auf die GER-Niveaustufen dargestellt.

**Tabelle 1:** Anzahl und prozentualer Anteil an Testpersonen pro Standort

| | *N* | % |
|---|---|---|
| Argentinien | 20 | 0,35 |
| Brasilien | 6 | 0,11 |
| China | 6 | 0,11 |
| Deutschland | 1.273 | 22,59 |

**Tabelle 1:** (fortgesetzt)

| | *N* | % |
|---|---|---|
| Finnland | 20 | 0,36 |
| Frankreich | 1.572 | 27,89 |
| Großbritannien | 56 | 0,99 |
| Griechenland | 21 | 0,37 |
| Italien | 11 | 0,20 |
| Niederlande | 1.588 | 28,18 |
| Polen | 247 | 4,38 |
| Russland | 23 | 0,41 |
| Schweiz | 463 | 8,21 |
| Spanien | 113 | 2,00 |
| Taiwan | 34 | 0,60 |
| Türkei | 99 | 1,76 |
| Usbekistan | 84 | 1,49 |
| Gesamt | 5.636 | 100,00 |

**Tabelle 2:** Anzahl und prozentualer Anteil an Testpersonen pro GER-Niveaustufe für die Teilbereiche Lesen und Hören sowie für den Gesamttest

| GER-Stufe | $N_{Lesen}$ (%) | $N_{Hören}$ (%) | $N_{Gesamt}$ (%) |
|---|---|---|---|
| Vor-A1 | 5 (0,1 %) | 21 (0,3 %) | 7 (0,1 %) |
| A1 | 131 (2,3 %) | 78 (1,4 %) | 50 (0,9 %) |
| A2 | 2.786 (49,4 %) | 819 (14,5 %) | 1.600 (28,4 %) |
| B1 | 1.750 (31,1 %) | 2.646 (46,9 %) | 2.665 (47,3 %) |
| B2 | 605 (10,7 %) | 1.345 (23,9 %) | 875 (15,5 %) |
| C1 | 218 (3,9 %) | 538 (9,5 %) | 343 (6,1 %) |
| C2 | 141 (2,5 %) | 189 (3,4 %) | 96 (1,7 %) |

Anm.: GER = Gemeinsamer Europäischer Referenzrahmen; Vor-A1 = Niveau unter A1

## 5.2 Zu Fragestellung 1

Die bei der Ergebnisrückmeldung des GTP genutzte Mittelwertsbildung der beiden Teilergebnisse für die Kompetenzbereiche Lesen und Hören ist nur dann psychometrisch vertretbar, wenn die beiden Teilbereiche eine gemeinsame Skala

bilden und somit als eindimensional angesehen werden können. In Bezug auf Fragestellung 1 wurden daher ein eindimensionales sowie ein zweidimensionales Modell mit den zwei korrelierten Dimensionen Lesen und Hören geschätzt und anschließend mittels Likelihood-Ratio-Test hinsichtlich ihrer Modellpassung inferenzstatistisch verglichen. Als weitere Kriterien für den Modellvergleich wurden das Akaike-Informationskriterium (Akaike Information Criterion, AIC) (Akaike 1974) sowie das Bayesianische Informationskriterium (Bayesian Information Criterion, BIC) (Schwarz 1978) genutzt.

## 5.3 Zu Fragestellung 2

Zur Beantwortung von Fragestellung 2 wurde die Reliabilität des GTP bestimmt. Das in dieser Studie genutzte Reliabilitätsmaß ist definiert als quadrierte Korrelation zwischen der wahren Fähigkeitsausprägung θ und der geschätzten Fähigkeit $\hat{\theta}$ (Kim 2012). Eine Schätzung aus empirischen Daten kann aus dem Quotienten aus der Varianz der Fähigkeitsschätzungen $\sigma^2_{\hat{\theta}_j}$ und der Summe aus $\sigma^2_{\hat{\theta}_j}$ und dem mittleren quadrierten Standardfehler $SE^2_{\hat{\theta}_j}$ der individuellen Fähigkeitsschätzung $\hat{\theta}_j$ von Person *j* erfolgen:

$$p^2_{\theta\hat{\theta}} = \frac{\sigma^2_{\hat{\theta}_j}}{\sigma^2_{\hat{\theta}_j} + \frac{1}{N}\sum_{j=1}^{N} SE^2_{\hat{\theta}_j}} \quad (1)$$

## 5.4 Zu Fragestellung 3

Zur Beantwortung von Fragestellung 3 wird den Empfehlungen der Standards for Educational and Psychological Testing (AERA/APA/NCME 2014) folgend der auf die Fähigkeitsausprägung bedingte Standardfehler der Fähigkeitsschätzung berechnet. Dieser gibt Aufschluss darüber, inwiefern sich die Messgenauigkeit des GTP in Abhängigkeit der Merkmalsausprägung unterscheidet, und somit auch, wie gut der Test in der Lage ist, zwischen verschiedenen Kompetenzniveaus zu differenzieren.

## 5.5 Zu Fragestellung 4

Als Maß für den Grad an Adaptivität des GTP wurde der Engineering Optimal Information Index (EOI) (Kingsbury/Wise 2020) genutzt. Der EOI gibt den Anteil der

tatsächlich durch den Test realisierten Testinformation an der theoretisch maximal erreichbaren Testinformation auf Stufe der finalen Fähigkeitsschätzung an. Der realisierte Test wird hierbei mit einem Test auf Grundlage eines hypothetisch perfekten Itempools verglichen, bei dem den Testpersonen ausschließlich Items vorgelegt werden, die ihrem Fähigkeitsniveau exakt entsprechen. Der EOI ist ein theoretischer Wert, der den hypothetisch informativsten Test quantifiziert und ihn mit der tatsächlich beobachteten Testinformation ins Verhältnis setzt. Er hat sein Maximum bei 100 und berechnet sich für das beim GTP genutzte IRT-Modell (eindimensionales 1PL-Modell) wie folgt:

$$EOI = 100 \cdot \frac{\sum_{j=1}^{N}(IA_j/0{,}25K)}{N} \tag{2}$$

Hierbei ist $IA_j$ die Testinformation von Person *j* gemäß ihrer aktuellen Fähigkeitsschätzung, *K* die Anzahl beantworteter Items und 0,25 die maximal erreichbare Iteminformation eines Items unter dem eindimensionalen 1PL-Modell. Um etwaige Stärken und Schwächen des EOI differenzierter in den Blick zu nehmen, wurden neben einem globalen EOI für den Gesamttest und die beiden Teilbereiche Lesen und Hören auch die jeweiligen EOIs getrennt nach GER-Niveaustufen berechnet.

Alle Analysen wurden mit der Statistiksoftware R (R Core Team 2020) unter Verwendung des R-Pakets *mirt* durchgeführt (Chalmers 2012).

# 6 Ergebnisse

## 6.1 Deskriptive Statistiken

Tabelle 3 zeigt deskriptivstatistische Ergebnisse für die Itempools für Lesen und Hören. Ein Computerisierter Adaptiver Test wie der GTP kann sich im Testverlauf nur dann optimal dem individuellen Fähigkeitsniveau der Testpersonen anpassen, wenn für alle Teilbereiche genügend Items vorhanden sind. Um dies differenziert zu betrachten, sind in den Abbildungen 1 und 2 (siehe Abschnitt 6.3) die absoluten Häufigkeiten der Items pro Schwierigkeitsbereich sowie die relativen Häufigkeiten der Personen pro Fähigkeitsbereich dargestellt. Die Ergebnisse zeigen eine breite Überlappung und damit Passung von Itemschwierigkeiten und Fähigkeiten. Eine weitergehende Optimierung der Passung kann durch die Ergänzung leichter Items für den Fähigkeitsbereich Lesen und schwieriger Items für den Fähigkeitsbereich Hören erzielt werden.

**Tabelle 3:** Deskriptive Statistiken der Itempools für die Teilbereiche Lesen und Hören

| | Lesen | Hören |
|---|---|---|
| Anzahl Tasks | 420 | 168 |
| Anzahl Items | 449 | 289 |
| *M* Itemschwierigkeit | 0,035 | 0,043 |
| *SD* Itemschwierigkeit | 1,339 | 1,245 |

Anm.: *M* = Mittelwert; *SD* = Standardabweichung

## 6.2 Dimensionalität der GTP-Items

In Tabelle 4 sind die Ergebnisse des Modellvergleichs zwischen dem eindimensionalen und dem zweidimensionalen Modell dargestellt. Der Likelihood-Ratio-Test identifiziert das zweidimensionale Modell als das signifikant besser passende Modell. Auch der AIC und der BIC weisen das zweidimensionale Modell als knapp besser passend aus. Die Unterschiede in den Informationskriterien sind allerdings sehr gering. Aufgrund dieser Ergebnisse sowie der nahezu perfekten latenten Korrelation von 0,936 zwischen den Dimensionen Lesen und Hören sind somit die separate Ausweisung der beiden Dimensionen sowie die Zusammenfassung zu einer gemeinsamen Skala möglich.

**Tabelle 4:** Ergebnisse des Modellvergleichs zwischen eindimensionalem und zweidimensionalem Modell

| Modell | Log-Likelihood | $x^2$ | *df* | *p* | AIC | BIC |
|---|---|---|---|---|---|---|
| eindimensional | -132905 | - | - | - | 267268 | 272107 |
| zweidimensional | -132802 | 206.479 | 2 | < .001 | 267066 | 271918 |

Anm.: AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion

## 6.3 Reliabilität des GTP

Die Reliabilitätskoeffizienten getrennt für die Teilbereiche Lesen und Hören sowie für den Gesamttest sind in Tabelle 5 dargestellt. Die Reliabilität des Gesamttests fällt mit einem Reliabilitätskoeffizienten von über 0,9 sehr hoch aus. Die beiden Teilbereiche Lesen und Hören weisen jeweils gute Reliabilitäten auf.

**Tabelle 5:** Deskriptive Statistiken sowie Reliabilitätskoeffizienten für den GTP

| | $M(\hat{\theta})$ | $SD(\hat{\theta})$ | $p^2_{\theta\hat{\theta}}$ |
|---|---|---|---|
| Lesen | -0,664 | 1,609 | 0,892 |
| Hören | 0,478 | 1,530 | 0,868 |
| Gesamt | -0,093 | 1,445 | 0,923 |

Anm.: $M(\hat{\theta})$ = Mittelwert der Fähigkeitsschätzungen; $SD(\hat{\theta})$ = Standardabweichung der Fähigkeitsschätzungen; $p^2_{\theta\hat{\theta}}$ = Reliabilität

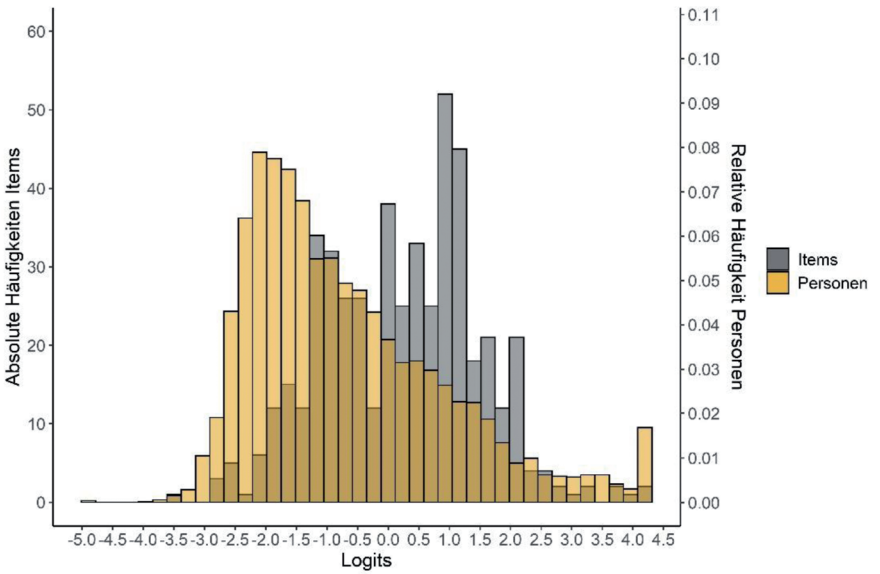


**Abbildung 1:** Absolute Häufigkeiten von Items (linke Y-Achse, helle Balken) und relative Häufigkeiten der Personenfähigkeiten (rechte Y-Achse, dunkle Balken) in Abhängigkeit der IRT-Skala θ für Lesen. Niedrige Werte auf der X-Achse bedeuten niedrige Itemschwierigkeiten sowie niedrige Fähigkeiten und hohe Werte hohe Itemschwierigkeiten sowie hohe Fähigkeiten.

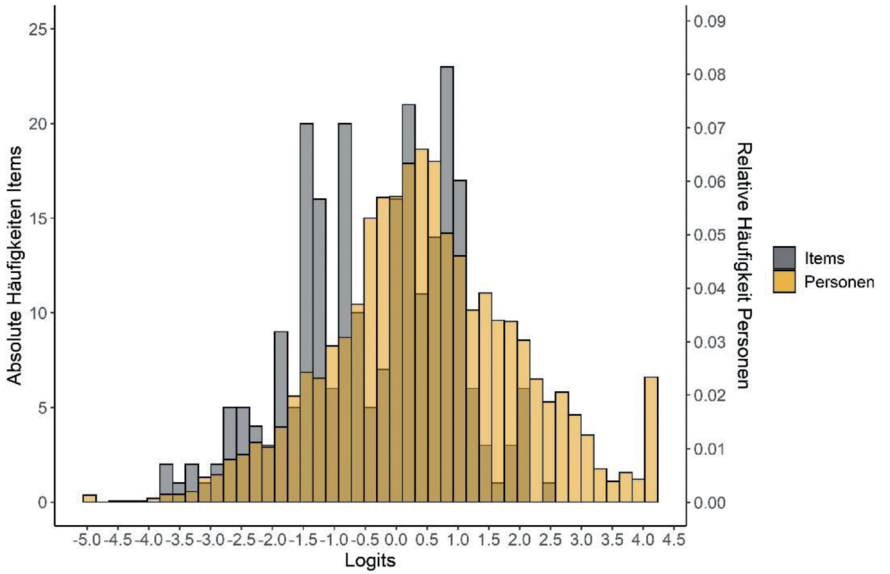


**Abbildung 2:** Absolute Häufigkeiten von Items (linke Y-Achse, helle Balken) und relative Häufigkeiten der Personenfähigkeiten (rechte Y-Achse, dunkle Balken) in Abhängigkeit der IRT-Skala θ für Hören. Niedrige Werte auf der X-Achse bedeuten niedrige Itemschwierigkeiten sowie niedrige Fähigkeiten und hohe Werte hohe Itemschwierigkeiten sowie hohe Fähigkeiten.

## 6.4 Differenzierungsfähigkeit des GTP

Zur Einschätzung der Differenzierungsfähigkeit des GTP und somit zur Beantwortung von Fragestellung 3 wurden die bedingten Standardfehler für den Gesamttest sowie für die beiden Teilbereiche Lesen und Hören berechnet. Die Ergebnisse sind in Abbildung 3 dargestellt. Für den Gesamttest liegen die Standardfehler in einem sehr breiten Bereich von –3,0 bis 2,0 auf einem vergleichbar niedrigen Niveau. Das heißt, dass für alle Personen mit Fähigkeiten zwischen –3,0 und 2,0 (beim vorliegenden Datensatz 90,2 %) ähnlich präzise Testergebnisse ermittelt werden können. Die Standardfehler für die beiden Teilbereiche fallen aufgrund der kürzeren Testlängen im Vergleich zum Gesamttest erwartungsgemäß höher aus. Auch auf Ebene der Teilskalen zeigt sich, dass der Test über einen breiten Fähigkeitsbereich (–3,0 bis 1,5) in der Lage ist, die Hör- und Lesekompetenz mit vergleichbarer Messpräzision zu messen. Die geringere Anzahl von schwierigen Items im Bereich Hören spiegelt sich in höheren Standardfehlern im oberen Fähigkeitsbereich wider.

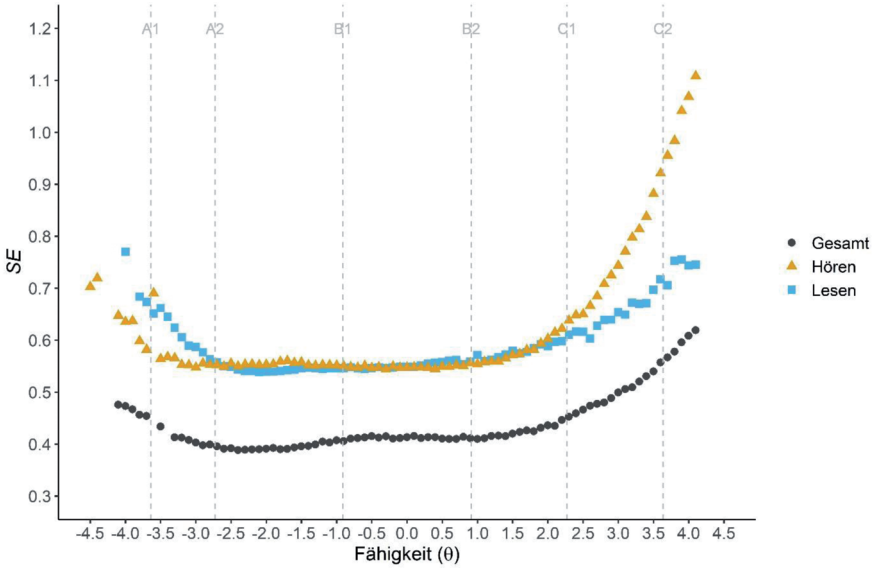


**Abbildung 3:** Bedingter Standardfehler (SE) der Fähigkeitsschätzungen für den Gesamttest sowie für die Teilbereiche Hören und Lesen. Die senkrechten, gestrichelten Linien stellen die Grenzwerte der GER-Niveaustufen dar.

## 6.5 Adaptivität des GTP

Mit einem EOI von 78,44 für den Gesamttest sowie 86,55 für den Teilbereich Lesen bzw. 82,64 für den Teilbereich Hören ist die Adaptivität des GTP als gut einzuschätzen. In Tabelle 6 sind die EOIs getrennt nach GER-Niveaustufen dargestellt. Dabei wurden unter dem A1-Niveau eingestufte Testpersonen aufgrund einer sehr kleinen Fallzahl ($N = 7$) nicht mit in die Analysen einbezogen. Für den Teilbereich Lesen sind die EOIs auf den GER-Niveaustufen A1 bis B2 auf einem vergleichbaren Niveau. Ab C1 nimmt der EOI deutlich ab.

Ab Niveaustufe A2 liegen die EOIs für den Teilbereich Hören unter den EOIs für Lesen. Der EOI für Hören auf der Niveaustufe C2 ist mit 24,04 als sehr niedrig einzustufen. Eine Erhöhung der Adaptivität auch in diesem Bereich kann durch die Ergänzung von schweren und sehr schweren Items für Hören erreicht werden. Der adaptive Algorithmus kann sich im oberen Fähigkeitsbereich nicht so gut an das individuelle Fähigkeitsniveau anpassen. Aus demselben Grund fällt auch der EOI für den Gesamttest für die GER-Niveaustufe C2 deutlich am niedrigsten aus.

**Tabelle 6:** EOI getrennt nach GER-Niveaustufen für den Goethe-Test PRO

| GER-Niveaustufen | *N* | $EOI_{Lesen}$ | $EOI_{Hören}$ | $EOI_{Gesamt}$ |
|---|---|---|---|---|
| A1 | 50 | 82,32 | 81,55 | 82,19 |
| A2 | 1.600 | 89,34 | 88,08 | 85,00 |
| B1 | 2.665 | 89,51 | 87,12 | 79,39 |
| B2 | 875 | 83,88 | 71,79 | 76,22 |
| C1 | 343 | 68,35 | 46,96 | 57,28 |
| C2 | 96 | 52,92 | 24,04 | 37,77 |

Anm.: EOI = Engineering Optimal Information Index

# 7 Diskussion

Mit zunehmender berufsorientierter Ausrichtung in der sprachlichen Qualifizierung von Deutschlernenden im In- und Ausland wird auch die Notwendigkeit einer berufssprachlichen Kompetenz und deren Diagnostik immer deutlicher. Prüfungen, die diese bisher ermitteln und ausweisen können, existieren in verschiedenen Formaten auf dem Markt. Ziel dieses Beitrags war es einerseits, einen dieser auf dem Markt befindlichen Tests für den Beruf und sein adaptives Testverfahren vorzustellen.

Aufgrund der hohen Relevanz der Testergebnisse des GTP für das getestete Individuum wie auch das Unternehmen oder die Organisation, in denen der Test zum Einsatz kommt, ist es wichtig, dass der GTP als zentrale Grundlage eine hohe psychometrische Güte aufweist, um verlässliche Rückschlüsse auf das Kompetenzniveau der Testpersonen zu ermöglichen. Daher war zweites Ziel des hier vorliegenden Beitrags, die psychometrische Güte des GTP zu beleuchten. Hierfür wurde neben deskriptivstatistischen Ergebnissen die Dimensionalität des Tests überprüft sowie Reliabilitäten, bedingte Standardfehler und der EOI analysiert. Die Ergebnisse der Dimensionsanalyse verdeutlichen, dass die bei der Berichtlegung des GTP genutzte Zusammenfassung der Teilergebnisse für Hören und Lesen zu einer gemeinsamen Skala angemessen ist. Sowohl der Gesamttest als auch die beiden Teilbereiche weisen hohe Werte von ≥ 0,87 für die Reliabilität auf. Mit seinen 738 Items, die ein breites Spektrum der in der Realität beobachtbaren Fähigkeitsverteilung abdecken, ist die Anpassungsfähigkeit des GTP an das individuelle Fähigkeitsniveau der getesteten Personen sehr hoch, was sich auch in den hohen EOI-Werten widerspiegelt. Optimierungsmöglichkeiten bestehen im oberen Extrembereich der Fähigkeitsverteilung für den Teilbereich Hören. Aufgrund

fehlender schwerer und sehr schwerer Items in diesem Bereich zeigen sich hier niedrigere Werte für die SEs sowie den EOI. Diesem Umstand kann durch die Erweiterung des Itempools um schwere und sehr schwere Items im Bereich Hören begegnet werden.

Zusammenfassend lässt sich sagen, dass mit dem Goethe-Test PRO ein adaptiver Online-Deutschtest zur Messung von berufsorientierten Deutschkenntnissen vorliegt, mit dem in einer relativ kurzen Testzeit effizient und messgenau die Hör- und Lesekompetenz am Arbeitsplatz über ein breites Spektrum der GER-Niveaustufen ermittelt werden kann.

Zukunftsweisende Impulse für das Lernen, Lehren und Beurteilen von Deutsch im Beruf bietet abschließend der in 2020 veröffentlichte Begleitband des GER. Die Aktualisierungen und Ergänzungen, die bei der Neuerscheinung vorgenommen wurden, rücken unter anderem den Fokus verstärkt auf die Domäne Beruf und haben Potenzial, die berufssprachliche Kompetenzdiagnostik, beispielsweise in Bezug auf neue Aufgabenformate, nachhaltig zu beeinflussen (Bärenfänger 2021: 240–244).

## Literaturverzeichnis

AERA/APA/NCME (2014): *Standards for Educational and Psychological Testing*. Washington: AERA Publication Sales.

Akaike, Hirotsugo (1974): „A new look at the statistical model identification". In: IEEE Transactions on Automatic Control 19, 716–723. Online: https://doi.org/10.1109/TAC.1974.1100705 (11.08.2021).

ALTE (2012): *Handbuch zur Entwicklung und Durchführung von Sprachtests: Zur Verwendung mit dem GER, erstellt von ALTE im Auftrag des Europarats/Abteilung für Sprachenpolitik*. Frankfurt am Main: telc gGmbH. Online: https://www.telc.net/fileadmin/user_upload/handbuch_zur_entwicklung_und_durchfuehrung_von_sprachtests.pdf (04.08.2021).

Auswärtiges Amt (2020): *Deutsch als Fremdsprache weltweit: Datenerhebung 2020*. Online: https://www.auswaertigesamt.de/blob/2344738/b2a4e47fdb9e8e2739bab2565f8fe7c2/deutsch-als-fremdsprache-data.pdf (11.08.2021).

Bärenfänger, Olaf (2021): „Berufssprachliche Kompetenzdiagnostik: Welche Ansatzpunkte bietet der Begleitband?". In: Vogt, Karin; Quetz, Jürgen (Hrsg.): *Der neue Begleitband zum Gemeinsamen europäischen Referenzrahmen für Sprachen: Kolloquium Fremdsprachenunterricht.* Berlin: Peter Lang, 225–245.

BMAS – Bundesministerium für Arbeit und Soziales (2016): *Arbeit weiter denken – Arbeiten 4.0*. Online: https://www.bmas.de/DE/Service/Publikationen/Broschueren/a883-weissbuch.html (18.08.2021).

Chalmers, R. Philip (2012): „mirt: A multidimensional item response theory package for the R environment". In: *Journal of Statistical Software* 48 (6), 1–29. Online: https://doi.org/10.18637/jss.v048.i06 (02.09.2021).

Dimitrijević, Anna (2020): „Neue Perspektiven für Deutsch: Fachkräfteausbildung und Qualifikation der Lehrenden". In: *IDV-Magazin: Deutsch für Fachkräfte: Herausforderungen und Er-*

*fahrungen*. Nr. 97, Juni 2020, 14–20. Online: https://idvnetz.org/wp-content/uploads/2020/06/IDV-Magazin-JUNI-2020.pdf (09.08.2021).

Efing, Christian (2014): „Berufssprache & Co: Berufsrelevante Register in der Fremdsprache. Ein varietätenlinguistischer Zugang zum berufsbezogenen DaF-Unterricht". In: *Information Deutsch als Fremdsprache* 41 (4), 415–441.

Europarat (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

Europarat (2020). *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Begleitband. Stuttgart: Ernst Klett Sprachen.

Frey, Andreas (2020): „Computerisiertes adaptives Testen". In: Moosbrugger, Helfried; Kelava, Augustin (Hrsg.): *Testtheorie und Fragebogenkonstruktion*. 3. Auflage. Berlin: Springer, 501–524. Online: https://doi.org/10.1007/978-3-662-61532-4_20 (08.08.2021).

Fromme, Linda; Korb, Eva (2018): „Aktuelle Tests und Prüfungen Deutsch für den Beruf". In: *Fremdsprache Deutsch. Zeitschrift für die Praxis des Deutschunterrichts* 59, 49–53.

Handt, Gerhard von der (2002): „Sprachtests und Zertifikate". In: Quetz, Jürgen; Handt, Gerhard von der (Hrsg.): *Neue Sprachen lehren und lernen: Fremdsprachenunterricht in der Weiterbildung*. Bielefeld: Bertelsmann, 187–193.

ILO – International Labour Organization (2021): *ILO Global Estimates on International Migrant Workers: Results and Methodology*. 3. Auflage. Genf. Online: https://www.ilo.org/global/topics/labour-migration/publications/WCMS_808935/lang–en/index.htm (18.08.2021).

Kim, Seonghoon (2012): „A note on the reliability coefficients for item response model-based ability estimates". In: *Psychometrika* 77, 153–162. Online: https://doi.org/10.1007/S11336-011-9238-0 (12.08.2021).

Kingsbury, G. Gage; Wise, Steven L. (2020): „Three measures of test adaptation based on optimal test information". In: *Journal of Computerized Adaptive Testing* 8 (1), 1–19. Online: https://doi.org/10.7333/2002-0801001 (11.08.2021).

Kuhn, Christina (2019): „Jenseits von Fachsprache? Eine Studie zur Kommunikation im Betrieb und ihre Implikationen für den berufsorientierten Fremdsprachenunterricht und die Materialgestaltung". In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht (ZIF)* 24 (1), 49–60. Online: https://zif.tujournals.ulb.tu-darmstadt.de/article/id/3174/ (18.08.2021).

OECD (2013): *Zuwanderung ausländischer Arbeitskräfte: Deutschland (German Version)*. OECD Publishing. Online: http://dx.doi.org/10.1787/9789264191747-de (16.08.2021).

Perlmann-Balme, Michaela (2016): „Sprachzertifikate". In: Burwitz-Melzer, Eva; Mehlhorn, Grit; Riemer, Claudia; Bausch, Karl-Richard; Krumm, Hans-Jürgen (Hrsg.): *Handbuch Fremdsprachenunterricht*. Tübingen: Francke, 420–423.

R Core Team (2020): *R: A language and environment for statistical computing* [Software]. R Foundation for Statistical Computing. Online: www.r-project.org (06.08.2021).

Rost, Jürgen (2004): *Lehrbuch Testtheorie – Testkonstruktion*. Zweite, vollständig überarbeitete und erweiterte Auflage. Bern: Hans Huber.

Schwarz, Gideon (1978): „Estimating the dimension of a model". In: *Annals of Statistics* 6, 461–464. Online: https://doi.org/10.1214/aos/1176344136 (28.09.2021).

van der Linden, Wim J. (Hrsg.) (2016): *Handbook of item response theory. Band 1: Models*. Boca Raton: Chapman & Hall/CRC.

van Gorp, Koen; Vîlcu, Dina (2018): *Guidelines for the Development of Language for Specific Purposes Tests*. Cambridge: ALTE.

## Internetquellen

Association of Language Testers in Europe (ALTE): Online: https://www.alte.org/ (18.08.2021).
Cambridge Assessement English: Online: https://www.cambridgeenglish.org/de/exams-and-tests/bulats/ (28.09.2021).
Demo-Version Goethe-Test PRO: Online: www.goethe.de/gtpro (28.09.2021).
Fachkräfteeinwanderungsgesetz: Online: https://fachkraefteeinwanderungsgesetz.de/ (09.08.2021).
Initiative „Duits in de beroepscontext“: Online: https://duitsmbo.nl/ (18.08.2021).
Prüfungsordnung Goethe-Test PRO: Deutsch für den Beruf: Online: https://www.goethe.de/pro/relaunch/prf/de/Pruefungsordnung_GTP.pdf (17.08.2021).
Statista Research Department (2019): Statistiken zum Fachkräftemangel. Online: https://de.statista.com/themen/887/fachkraeftemangel/ (12.08.2021).
Zentrale Webseite Goethe-Institut e.V.: Online: https://www.goethe.de/de/wwt.html (28.09.2021).

# Biographische Angaben

**Aron Fink**

erhielt seinen Master of Science in Psychologie von der Universität Erfurt. Er ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Pädagogische Psychologie mit Schwerpunkt Beratung, Diagnostik und Evaluation an der Goethe-Universität Frankfurt. Zuvor arbeitete er am Lehrstuhl für Empirische Methoden der erziehungswissenschaftlichen Forschung an der Friedrich-Schiller-Universität Jena. Seine Forschungsinteressen liegen in den Bereichen empirische Bildungsforschung und Psychometrie – insbesondere Computerisiertes Adaptives Testen (CAT) sowie Einsatz Künstlicher Intelligenz in der empirischen Bildungsforschung.

**Andreas Frey**

ist Professor für Pädagogische Psychologie mit Schwerpunkt Beratung, Diagnostik und Evaluation und wissenschaftlicher Leiter der Psychologischen Beratungsstelle MAINKIND an der Goethe-Universität Frankfurt. Zuvor war er Professor für Empirische Methoden der erziehungswissenschaftlichen Forschung an der Friedrich-Schiller-Universität Jena und parallel von 2016–2021 Professor II für Educational Measurement an der Universität Oslo. Er ist Mitglied in zahlreichen methodischen Expertengremien und Mitglied des nationalen IGLU-2021-Konsortiums. Von 2012 bis 2014 war er Vorstandsmitglied der Deutschen Gesellschaft für Psychologie (DGPs). Seine Forschungsinteressen sind aktuelle methodische Entwicklungen, computerbasiertes Testen, Kompetenzdiagnostik, automatisiertes Feedback und Methoden von Large-Scale Assessments.

**Katharina Klein**

studierte Interkulturelle Germanistik / Deutsch als Fremdsprache an der Georg-August-Universität Göttingen. Danach war sie als DAAD-Sprachassistentin am Indian Institute of Technology in Mumbai und als freiberufliche DaF-Lehrkraft tätig. Seit 2016 arbeitet sie beim Goethe-Institut e.V. in der Zentrale München als Referentin im Bereich Prüfungen der Abteilung Sprache. Zu ihren Arbeitsgebieten zählen die Prüfungsentwicklung sowie die Digitalisierung von Prüfungen.