GOETHE-TEST PRO: DEUTSCH FÜR DEN BERUF

UNTERSUCHUNG DER PSYCHOMETRISCHEN GÜTE

29.07.2020 - GOEHTE-INSTITUT E.V.

ARON FINK & ANDREAS FREY

GOETHE INSTITUT

Sprache. Kultur. Deutschland.

Inhalt

ABSTRACT

Der Online-Deutschtest *Goethe-Test PRO: Deutsch für den Beruf* wird seit April 2017 an den Prüfungszentren des Goethe-Instituts sowie bei Firmenkunden weltweit durchgeführt. Er ist als adaptiver Online-Test konzipiert und ermittelt die Hör- und Lesekompetenz am Arbeitsplatz auf den Stufen A1-C2 des Gemeinsamen europäischen Referenzrahmens für Sprachen. In dieser Studie wurde auf Grundlage der Testdaten von N = 5 636 Testpersonen aus 17 Staaten die psychometrische Güte des Goethe-Test PRO im Hinblick auf verschiedene Maße untersucht. Der Goethe-Test PRO liefert mit einer Reliabilität von > .9 zuverlässige Testergebnisse. Die psychometrischen Eigenschaften des Tests fallen in verschiedenen Staaten, in denen der Goethe-Test PRO durchgeführt wird, ähnlich aus.

1. EINLEITUNG

Der *Goethe-Test PRO: Deutsch für den Beruf* (GTP) ist ein vom Goethe-Institut e.V. angebotener computerbasierter Test, der die Hör- und Lesekompetenz am Arbeitsplatz basierend auf dem Gemeinsamen europäischen Referenzrahmen für Sprachen (GER) ermittelt (siehe z. B. https://www.goethe.de/Z/50/commeuro/101.htm). Der Test kann flexibel in einem der weltweit verteilten Standorte des Goethe-Instituts oder direkt im Unternehmen durchgeführt werden und soll als Entscheidungsgrundlage zum Beispiel für weitere Fortbildungsmaßnahmen oder Ähnliches dienen.

Der GTP nutzt die Methode des computerisierten adaptiven Testens (CAT; z. B. Frey, im Druck). Bei CAT werden nicht, wie bei den meisten Testverfahren, allen Testpersonen die gleiche vorab zusammengestellte Menge von Items präsentiert. Stattdessen orientiert sich die Itemauswahl am im Testverlauf gezeigten Antwortverhalten der Testperson. Vereinfacht dargestellt, wird bei CAT die Schwierigkeit der vorgegebenen Items an die Leistungsfähigkeit der getesteten Personen angepasst. Hiermit bekommen die Testpersonen nur diejenigen Items zur Beantwortung vorgelegt, die besonders viel diagnostische Information über die individuelle Merkmalsausprägung, im Fall des GTP also des Sprachniveaus, liefern. Dieses Vorgehen führt dazu, dass CAT im Vergleich zu traditionellen nicht-adaptiven Tests in der Regel präzisere Fähigkeitsschätzungen und/oder eine kürzere Testlänge liefert (z. B. Segall, 2005). Mit CAT kann es insbesondere ermöglicht werden, dass die Fähigkeiten der Testteilnehmerinnen und Testteilnehmer über das gesamte Fähigkeitsspektrum mit einem vergleichbaren Maß an Präzision gemessen werden (Frey & Ehmke, 2007). Das bedeutet, dass anders als bei traditionellen Testverfahren die Genauigkeit, mit der ein Testwert bestimmt werden kann, und hier im speziellem die Zuordnung einer Testperson zu einer Niveaustufe des GER, weitgehend unabhängig von der individuellen Fähigkeitsausprägung ist. Da bei CAT nicht alle Testpersonen die gleichen Items bearbeiten, kann die Einschätzung des Fähigkeitsniveaus einer Testperson nicht anhand der Anzahl korrekter Antworten erfolgen. Itemeigenschaften, beispielsweise wie schwierig die bearbeiteten Items waren, müssen in die Fähigkeitsschätzung miteinbezogen werden. Zur Berücksichtigung der Itemschwierigkeit werden bei CAT Modelle der Item Response Theory (IRT; z. B. van der Linden, 2016) genutzt. Diese beschreiben die Wahrscheinlichkeit, mit der ein Individuum mit einer bestimmten

Fähigkeit ein Item mit bestimmten Eigenschaften richtig löst. Der GTP nutzt als IRT-Modell das Rasch-Modell (Rasch, 1960).

Die Bearbeitung des GTP dauert 60 bis 90 Minuten. Alle Testpersonen bearbeiten zunächst 15 Tasks aus dem Bereich Lesen und anschließend 15 Tasks aus dem Bereich Hören. Als Tasks werden dabei Aufgaben bezeichnet, die einen Stamm haben und ein oder mehrere Items umfassen, die sich auf diesen beziehen. Der GTP umfasst dabei verschiedene innovative Itemformate. Direkt nach Beendigung des Tests werden der Testperson die Teilergebnisse für Lesen und Hören sowie das Gesamtergebnis inklusive der erreichten Niveaustufe des GER zurückgemeldet. Zusätzlich wird der Testperson ein digitales Zeugnis mit detaillierter Beschreibung des Sprachniveaus zur Verfügung gestellt.

2. FORSCHUNGSFRAGEN

Mit der vorliegenden Studie wird die psychometrische Güte des GTP untersucht und zwischen Saaten verglichen, in denen er angeboten wird. Es werden die folgenden Fragen untersucht und beantwortet:

1. Welche psychometrische Güte weist der Goethe-Test PRO auf?
    - 1.1 Sind die Items des GTP als eindimensional anzusehen?
    - 1.2 Wie zuverlässig sind die Testergebnisse des GTP?
    - 1.3 Wie hoch ist die Differenzierungsfähigkeit des GTP über das Fähigkeitsspektrum?
    - 1.4 Wie hoch ist der Grad an Adaptivität des GTP?
2. Differiert die psychometrische Güte zwischen Staaten?

3. VORGEHEN UND METHODE

Stichprobe

Die Datengrundlage bilden die aus der für die Testvorgabe genutzten Plattform Moodle extrahierten Testergebnisse von *N* = 5 636 Testpersonen, die seit April 2017 den Test bearbeitet haben. In Tabelle 1 ist die Verteilung der Testpersonen auf die untersuchten Staaten dargestellt. Ein Großteil der Personen (78.66 %) hat den Test in Deutschland, Frankreich oder den Niederlanden absolviert. In Tabelle 2 ist die Verteilung der Stichprobe auf die GER-Niveaustufen dargestellt.

Zu Fragestellung 1

Fragestellung 1.1: Sind die Items des GTP als eindimensional anzusehen?

Das Endergebnis des GTP wird durch Mittelwertsbildung der einzelnen Testergebnisse für die Teilbereiche Lesen und Hören gebildet. Dieses Vorgehen ist nur dann sinnvoll, wenn die beiden Teilbereiche eine gemeinsame Skala konstituieren, die Items des GTP also eindimensional sind. Daher soll in Bezug auf Fragestellung 1.1 geprüft werden, ob die Items des GTP als eindimensional angesehen werden können. Zu diesem Zweck werden auf Grundlage der Testdaten ein eindimensionales sowie ein zweidimensionales Modell mit den zwei korrelierten Dimensionen Lesen und Hören geschätzt. Die beiden

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
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 an Testpersonen pro GER-Niveaustufe für die Teilbereiche Lesen und Hören sowie für den Gesamttest.

		N_{Lesen}	$N_{Hören}$	N_{Gesamt}
GER-Niveaustufe	Vor-A1	5	21	7
	A1	131	78	50
	A2	2 786	819	1 600
	B1	1 750	2 646	2 665
	B2	605	1 345	875
	C1	218	538	343
	C2	141	189	96

Anmerkungen: GER = Gemeinsamer europäischer Referenzrahmen; Vor-A1 = Niveau unter A1.

Modelle werden mittels Likelihood-Ratio-Test miteinander verglichen. Der Likelihood-Ratio-Test erlaubt eine inferenzstatistische Aussage darüber, ob sich die Modellpassung zweier konkurrierender statistischer Modelle überzufällig unterscheidet oder nicht. Zusätzlich werden die Modelle mit dem Akaike Informationskriterium (engl.: Akaike information criterion, AIC; Akaike, 1974) und dem Bayesianischen Informationskriterium (engl.: Bayesian information criterion, BIC; Schwarz, 1978) miteinander verglichen. Bei beiden Koeffizienten ist die Modellpassung umso besser, je niedriger deren Wert ist.

Fragestellung 1.2: Wie zuverlässig sind die Testergebnisse des GTP?

Die Zuverlässigkeit von Testergebnissen kann durch das psychometrische Gütekriterium der Reliabilität quantifiziert werden. Sie gibt Auskunft über die Präzision der Testergebnisse. Um als reliabel zu gelten, sollte ein Test bei wiederholter Messung unter gleichen Rahmenbedingungen zu den gleichen Messergebnissen führen. Es existieren verschiedene Reliabilitätsmaße. Das in dieser Studie verwendete Reliabilitätsmaß leitet sich aus der Definition von Reliabilität als quadrierte Korrelation zwischen der wahren Fähigkeitsausprägung θ und der geschätzten Fähigkeit $\hat{\theta}$ ab (Squared-Correlation-Reliabilität; $p_{\theta\hat{\theta}}{}^2$; Kim, 2012), welche auch häufig zur Schätzung der Reliabilität in Simulationsstudien verwendet wird. Eine Schätzung aus empirischen Daten kann aus dem Quotienten aus der Varianz der Thetaschätzungen $\sigma^2_{\hat{\theta}_j}$ und der Summe aus der Varianz der Thetaschätzungen und dem mittleren quadrierten Standardfehler $SE^2_{\hat{\theta}_j}$ der individuellen Fähigkeitsschätzung $\hat{\theta}_j$ erfolgen:

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

Ein alternatives Reliabilitätsmaß basiert auf dem Ansatz, Reliabilität durch die Korrelation $p_{\hat{\theta}\hat{\theta}}$ der mit zwei parallelen Testformen geschätzten Fähigkeit $\hat{\theta}$ zu bestimmen. Die Analysen zeigten, dass mit beiden Reliabilitätsmaßen in ihrer Aussage vergleichbare Ergebnisse resultieren, weswegen sich in der Ergebnisdarstellung nur auf das in Formel 1 dargestellte Reliabilitätsmaß beschränkt wird.

Fragestellung 1.3: Wie hoch ist die Differenzierungsfähigkeit des GTP über das Fähigkeitsspektrum?

Um Aussagen über die Differenzierungsfähigkeit des GTP in Abhängigkeit von der Fähigkeitsausprägung treffen zu können, wird gemäß den Standards für pädagogische und psychologische Tests (AERA, APA & NCME, 2014) zur Beantwortung von Fragestellung 1.3 der auf die Fähigkeitsausprägung bedingte Standardfehler der Fähigkeitsschätzungen berechnet. Als relevanter Bereich wird jener Bereich der latenten Fähigkeitsskala betrachtet, der (a) alle Stufen des GER abdeckt und (b) mindestens 99 % der Testpersonen umfasst.

Fragestellung 1.4: Wie hoch ist der Grad an Adaptivität des GTP?

Zur Beantwortung von Fragestellung 1.4 wird der kürzlich für adaptive Tests auf Basis des Rasch-Modells vorgeschlagene Engineering Optimal Information Index (EOI; Kingsbury & Wise, 2020) zur Einschätzung des Grades an Adaptivität berechnet. Der EOI gibt den Anteil der tatsächlich durch den durchgeführten Test erhaltenen Information an der maximal erreichbaren Information auf Stufe der abschließenden Fähigkeitsschätzung an. Dieser Index setzt eine hypothetisch unter dem Rasch-Modell perfekte Itembank voraus, sodass den Testpersonen ausschließlich Items vorgelegt werden könnten, die ihrem geschätzten Fähigkeitsniveau exakt entsprechen. Er stellt somit einen theoretischen Wert dar, der den hypothetisch

informativsten Test quantifiziert, der durchgeführt werden könnte, und setzt ihn mit der tatsächlich beobachteten Testinformation ins Verhältnis. Für das Rasch-Modell berechnet sich der EOI wie folgt:

$$EOI = 100 \cdot \frac{\sum_{j=1}^{N}(IA_j/0.25K)}{N}, \qquad (2)$$

wobei IA_j die Testinformation von Person *j* für ihre aktuelle Fähigkeitsschätzung, *K* die Anzahl an beantworteten Items und 0.25 die maximal erreichbare Iteminformation eines Items unter dem Rasch-Modell ist. Der EOI hat ein Maximum von 100. Dieser Wert impliziert, dass gegeben der abschließenden Fähigkeitsschätzung die informativste Menge an Items für jedes Individuum in der getesteten Gruppe präsentiert wurde. Der EOI eignet sich darüber hinaus auch dazu, den Grad an Adaptivität eines Tests zwischen Gruppen von Testteilnehmern (z. B. GER-Niveaustufen, Staaten) zu vergleichen. So können etwaige Stärken oder Schwächen eines Testinstruments differenzierter untersucht werden.

Zu Fragestellung 2

Da der GTP in verschiedenen Staaten zur Anwendung kommt, ist es wichtig, dass der Test über Staaten hinweg in gleicher Weise funktioniert und verlässliche Ergebnisse liefert. Daher wurden zur Beantwortung von Fragestellung 2 das oben genannte Reliabilitätsmaß sowie der EOI auch für Einzelstaaten berechnet und miteinander verglichen. Hierfür wurden nur Staaten in die Analysen mit einbezogen, in denen mindestens 20 Testpersonen den GTP absolviert haben.

Alle Analysen wurden mit der Statistiksoftware R (R Core Team, 2020) durchgeführt. Zur Beantwortung von Fragestellung 1.1 wurde zudem das R-Paket „mirt“ (Chalmers, 2012) genutzt.

4. ERGEBNISSE UND ANALYSE

Deskriptive Statistiken

Die deskriptiven Statistiken für die beiden Itempools sind in Tabelle 3 dargestellt. In den Abbildungen 1 und 2 sind die absoluten Häufigkeiten der Items pro Schwierigkeitsbereich (linke y-Achse) sowie die relativen Häufigkeiten der Personen pro Fähigkeitsbereich (rechte y-Achse) in Logits zu sehen. Dabei wurden Items mit dicht beieinanderliegenden Schwierigkeitsparametern zusammengefasst. Hierbei ist für einen adaptiven Test wie den GTP erstrebenswert, dass für alle Bereiche der Logit-Skala, auf denen Personen lokalisiert sind, auch hinreichend Items vorhanden sind. Ein adaptiver Test kann sich dann optimal anpassen, wenn für jede Person mindestens so viele Items vorhanden sind, wie der Test lang ist. Dieses Optimum wird für Lesen und Hören unterschiedlich gut erzielt. Während für Lesen der Itempool eher um sehr leichte Items ergänzt werden könnte, mangelt es dem Itempool zum Hören noch an schweren und sehr schweren Items.

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

Anmerkungen. M = Mittelwert; *SD* = Standardabweichung.

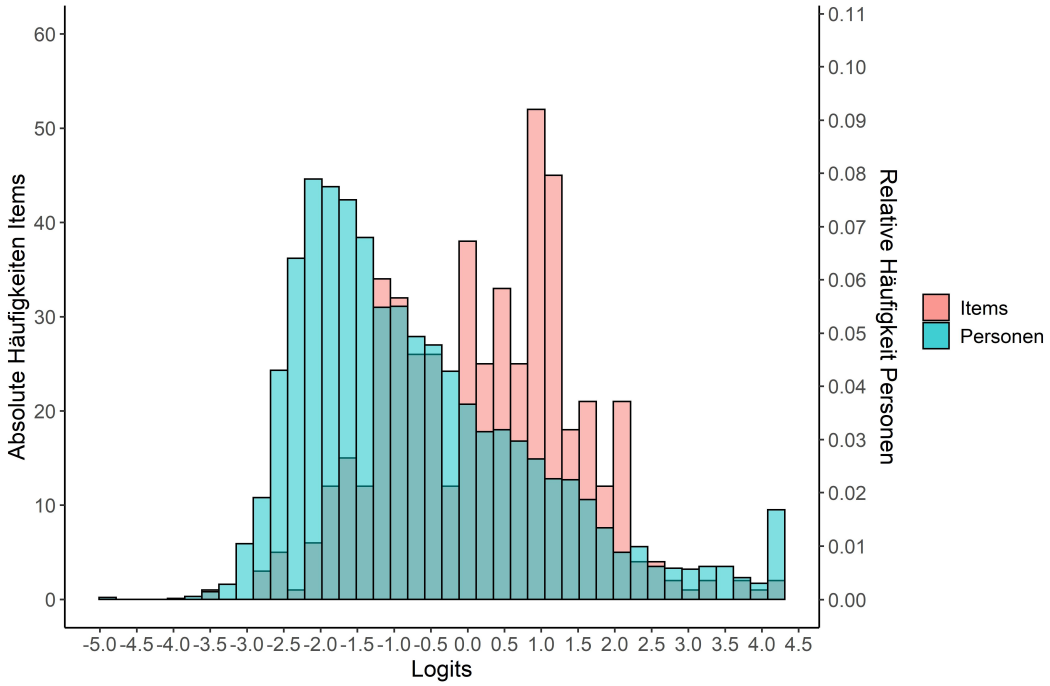

Abbildung 1. Absolute Häufigkeiten von Items (links) sowie relative Häufigkeiten der geschätzten Personenfähigkeiten (rechts) pro Schwierigkeitsbereich in Logits für die Domäne Lesen.

Zu Fragestellung 1 – Psychometrische Güte des GTP

Im Folgenden werden nun die Ergebnisse getrennt nach den einzelnen Fragestellungen dargestellt. Durch die standardisierte Durchführung, Auswertung, Bestimmung der Testergebnisse und deren Rückmeldung ist die Objektivität des GTP als gegeben anzusehen.

Zu Fragestellung 1.1 – Dimensionalität der Items

Tabelle 4 zeigt die Ergebnisse des Modellvergleichs zwischen dem eindimensionalen und dem zweidimensionalen Modell. Der Likelihood-Ratio-Test weist das zweidimensionale Modell als signifikant besser passend aus als das eindimensionale Modell. Dies ist aufgrund der großen Stichprobenanzahl und

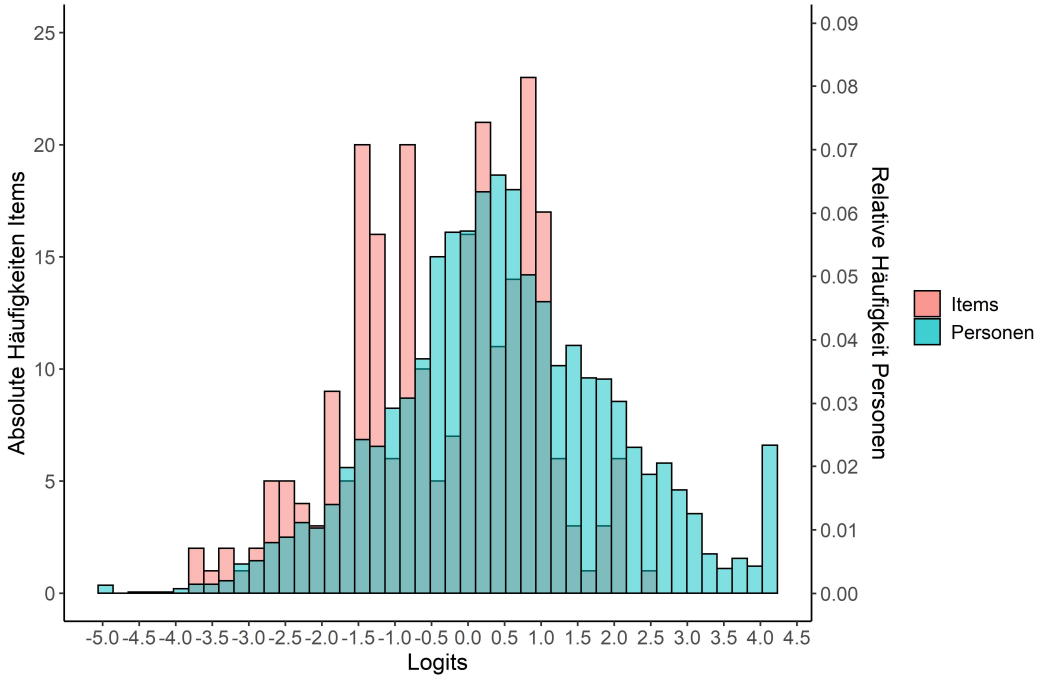

Abbildung 2. Absolute Häufigkeiten von Items (links) sowie relative Häufigkeiten der geschätzten Personenfähigkeiten (rechts) pro Schwierigkeitsbereich in Logits für die Domäne Hören.

Tabelle 4.
Ergebnisse des Modellvergleichs zwischen eindimensionalem und zweidimensionalem Modell.

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

Anmerkungen. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.

der damit verbundenen sehr hohen Teststärke dieses inferenzstatistischen Tests nicht verwunderlich. Die Informationskriterien AIC und BIC weisen das zweidimensionale Modell ebenfalls als knapp besser passend aus. Die beiden Dimensionen für Lesen und Hören korrelieren mit einer latenten Korrelation von .936 indes fast perfekt miteinander. Aufgrund dieser Befundlage ist sowohl die separate Ausweisung der beiden Dimensionen als auch die bei der Berichtlegung des GTP genutzte Zusammenfassung zu einer gemeinsamen Skala möglich.

Zu Fragestellung 1.2 – Reliabilität

In Tabelle 5 sind die Reliabilitätskoeffizienten für die Teilereiche Lesen und Hören sowie für den Gesamttest dargestellt. Hieraus ist zu erkennen, dass die Reliabilität für den Gesamttest mit über .9 als sehr hoch einzuschätzen ist. Beide Teilbereiche weisen eine gute Reliabilität auf. Die Reliabilität fällt dabei für den Teilbereich Lesen etwas höher aus als für den Teilbereich Hören.

Tabelle 5.
Deskriptive Statistiken sowie Reliabilitätskoeffizienten für den Goethe-Test PRO.

	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$
Lesen	-0.664	1.609	0.892
Hören	0.478	1.530	0.868
Gesamt	-0.093	1.445	0.923

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

Zu Fragestellung 1.3 – Differenzierungsfähigkeit

In Abbildung 3 sind die bedingten Standardfehler der Fähigkeitsschätzung für den Gesamttest sowie für die Teilbereiche Lesen und Hören dargestellt. Die Standardfehler für den Gesamttest liegen im Bereich von -3.0 bis 2.0 auf einem vergleichbar niedrigen Niveau. Der Test vermag es somit über einen breiten Bereich der Fähigkeitsskala mit vergleichbarer Präzision zu messen. Erwartungsgemäß werden für die kürzeren Tests zu den Teilbereichen Lesen und Hören im Niveau höhere Standardfehler und damit eine etwas niedrigere Präzision als beim Gesamttest erreicht. Die Standardfehler der Fähigkeitsschätzungen auf den Teilskalen sind im Bereich von -3.0 bis 1.5 weitgehend konstant. Selbst auf Ebene der Teilskalen kann somit über einen breiten Fähigkeitsbereich mit vergleichbarer Präzision gemessen werden. Es ist jedoch auch gut in Abbildung 3 sichtbar, dass der Mangel an schweren und sehr schweren Items für den Teilbereich Hören einen stärkeren Anstieg der Standardfehler im oberen Fähigkeitsbereich mit sich bringt.

Zu Fragestellung 1.4 – Adaptivität

Der EOI nähert sich sowohl für den Gesamttest mit 78.44 als auch für die Teilbereiche Lesen mit 86.55 und Hören mit 82.64 seinem theoretischen Maximum von 100 an und spricht somit für eine gute Adaptivität des GTP. Auch hier schneidet der Lesen-Test etwas besser ab als der Test für Hören. Das lässt sich zum Teil mit dem wesentlich kleineren Itempool für Hören begründen, dem es vor allem im oberen Bereich noch an sehr schweren Items mangelt. Dieser Umstand verdeutlicht sich ebenfalls in der Betrachtung der EOIs getrennt nach GER-Niveaustufen. Die Ergebnisse hierfür sind in Tabelle 6 dargestellt. Unter dem A1-Niveau eingestufte Testpersonen wurden aufgrund einer sehr kleinen Fallzahl (N = 7) nicht mit in die Analysen einbezogen. Die Ergebnisse lassen erkennen, dass für den Teilbereich Lesen von A1 bis B2 der EOI vergleichbar hoch ist. Ab C1 nimmt der EOI ab. Für den Teilbereich Hören liegen die EOIs ab Niveaustufe A2 unter den EOIs für Lesen. Der EOI für Niveaustufe C2 ist mit 24.04 als sehr niedrig einzustufen. Hier scheint der Test vergleichsweise wenig dem finalen Fähigkeitsniveau entsprechende Items zu administrieren (da diese im Itempool nicht enthalten sind). Dieser Umstand spiegelt sich auch im EOI für den Gesamttest wider, welcher für die GER-Niveaustufe C2 deutlich am niedrigsten ausfällt.

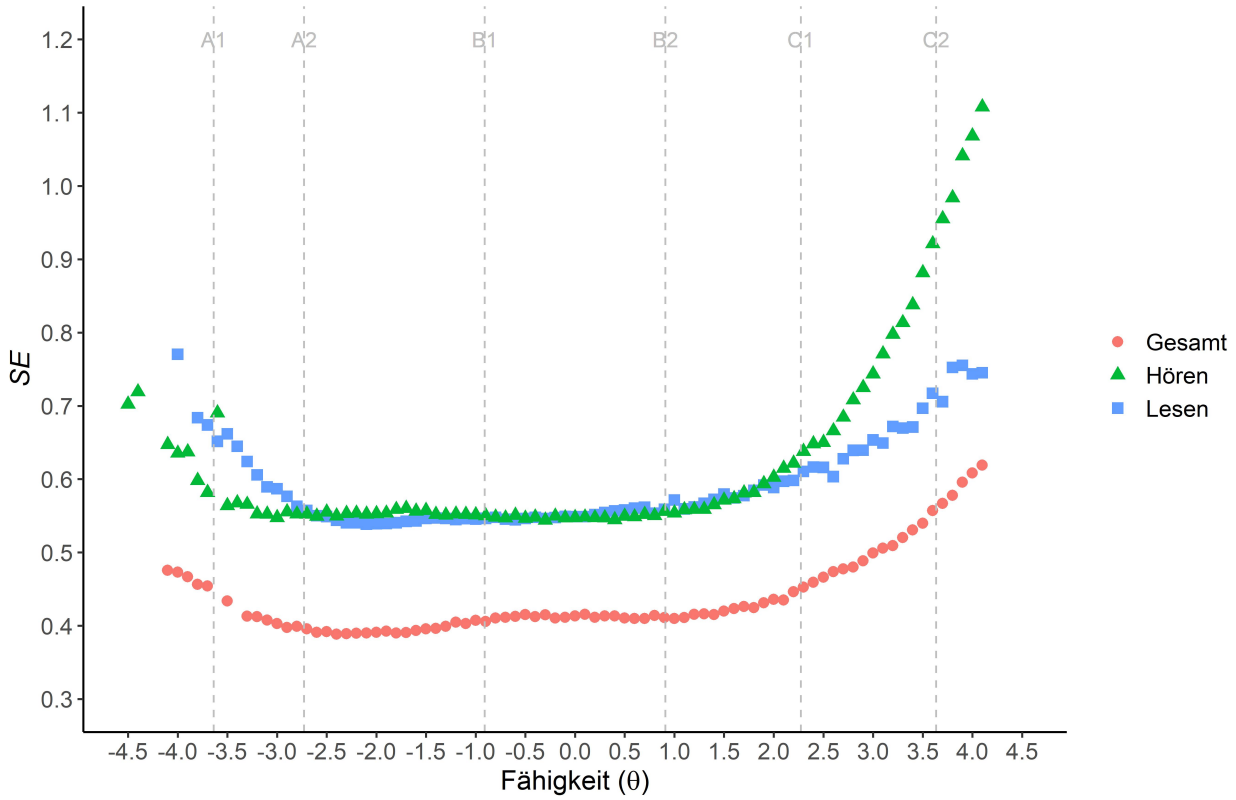

Abbildung 3. Bedingter Standardfehler (*SE*) der Fähigkeitsschätzungen für den Gesamttest sowie für die Teilbereiche Hören und Lesen. Die senkrechten gestrichelten Linien stellen die Grenzwerte der GER-Niveaustufen dar.

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

Anmerkungen. EOI = Engineering Optimal Information Index.

Zu Fragestellung 2 – Vergleich der psychometrischen Güte zwischen Staaten

In Tabelle 7 sind die Ergebnisse der Reliabilitätsanalysen sowie die EOIs getrennt nach Staaten jeweils für den Gesamttest sowie für die Teilbereiche Lesen und Hören dargestellt. Für die Gesamtskala wird in allen Staaten mit guter oder sehr guter Reliabilität gemessen. Somit ist trotz der Unterschiede in allen

Staaten eine hohe bis sehr hohe Messpräzision gewährleistet. Selbiges gilt für den Teilbereich Lesen. Hier liegen die Reliabilitäten zwischen .778 und .923 und somit für alle Staaten im guten bis sehr guten Bereich. Wie aufgrund der Ergebnisse über alle Staaten bereits zu erwarten, fallen die Reliabilitäten für Hören im Vergleich dazu etwas ab. Sie liegen aber dennoch in fast allen Staaten im guten bis sehr guten Bereich. In den Niederlanden, der Türkei, Russland und Argentinien sind sie als akzeptabel zu bezeichnen.

Mit Bezug auf den EOI zeigt sich in den staatenspezifischen Ergebnissen für nahezu alle Staaten ein vergleichbarer Grad an Adaptivität. Auch hier liegt der EOI für den Teilbereich Lesen über dem für den Teilbereich Hören. Der niedrigste EOI ergibt sich unabhängig vom Teilbereich für Russland. Allerdings liegt die durchschnittliche Leistung in Russland mit 1.691 für Lesen und 2.352 Hören deutlich über dem Gesamtdurchschnitt. Wie bereits aus Tabelle 6 erkennbar, ist der EOI für beide Teilbereiche im oberen Kompetenzbereich deutlich niedriger. Hierauf ist vermutlich der niedrigere EOI in Russland zurückzuführen. Insgesamt liegen die EOIs für den Gesamttest und die beiden Teilbereiche aber für alle Staaten auf einem vergleichbaren hohen Niveau.

Tabelle 7.

Staatenspezifische deskriptive Statistiken, Reliabilitätskoeffizienten und EOIs für den Goethe-Test PRO.

		Lesen				Hören				Gesamt			
Staat	N	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$	EOI	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$	EOI	$M(\hat{\theta})$	$SD(\hat{\theta})$	$p^2_{\theta\hat{\theta}}$	EOI
Niederlande	1 588	-1.234	1.026	.778	89.30	0.374	1.061	.780	86.25	-0.430	0.923	.833	79.40
Frankreich	1 572	-0.490	1.604	.892	86.60	0.622	1.596	.874	79.85	0.066	1.478	.926	78.50
Deutschland	1 273	-0.796	1.664	.897	86.01	0.004	1.560	.877	83.39	-0.396	1.474	.927	80.02
Schweiz	463	-0.398	1.829	.912	84.96	0.510	1.681	.882	79.27	0.056	1.640	.938	77.73
Polen	247	0.092	2.024	.923	81.29	1.152	1.934	.893	71.07	0.622	1.869	.947	72.59
Spanien	113	0.544	1.889	.914	81.24	1.417	1.810	.876	69.76	0.980	1.747	.941	72.71
Türkei	99	1.245	1.218	.815	81.33	2.305	0.972	.649	61.55	1.775	0.945	.818	70.06
Usbekistan	84	-0.778	1.467	.875	88.09	-0.432	1.200	.824	87.24	-0.605	1.211	.902	84.19
Großbritannien	56	-0.555	1.574	.886	85.42	1.109	1.444	.842	77.21	0.277	1.387	.912	74.47
Taiwan	34	-0.256	1.324	.850	86.79	0.289	1.322	.844	83.66	0.017	1.222	.901	82.05
Russland	23	1.691	1.767	.891	74.04	2.352	1.345	.747	57.72	2.022	1.436	.899	63.98
Griechenland	21	0.938	1.551	.875	80.65	1.431	1.274	.808	74.48	1.185	1.332	.906	75.96
Argentinien	20	-0.128	1.646	.896	86.57	1.210	1.043	.755	79.15	0.541	1.272	.901	76.15
Finnland	20	0.007	1.752	.906	85.60	1.217	1.499	.847	74.22	0.612	1.524	.927	75.93

Anmerkungen. $M(\hat{\theta})$ = Mittelwert der Fähigkeitsschätzungen; $SD(\hat{\theta})$ = Standardabweichung der Fähigkeitsschätzungen; $p^2_{\theta\hat{\theta}}$ = Reliabilität; EOI = Engineering Optimal Information Index.

5. FAZIT

Ziel von Fragestellung 1 war die Untersuchung der psychometrischen Güte des GTP. Dazu wurde neben deskriptivstatistischen Ergebnissen für Itempool und Testpersonen die Dimensionalität des Tests überprüft sowie Reliabilitäten, bedingte Standardfehler und der EOI berechnet. Aufgrund der Ergebnisse der Dimensionsanalyse erweist sich die bei der Berichtlegung des GTP genutzte Zusammenfassung zu einer gemeinsamen Skala als angemessen. Der verfügbare Itempool des GTP ist mit 738 Items sehr groß und erlaubt eine sehr feine Anpassung der Itemvorgabe an das Antwortverhalten der getesteten Personen. Entsprechend fallen die EOI-Werte hoch aus. Die Adaptivität nähert sich bei der Teilskala Lesen mit einem Wert von 86.55 der hypothetischen perfekten Adaptivität von 100.00 an. Entsprechend fällt die Reliabilität auch insgesamt sehr gut aus und variiert auf Gesamttestebene und auf Ebene der Teilskalen nur mäßig zwischen den Staaten. Die beobachteten Differenzen lassen sich mit hoher Wahrscheinlichkeit auf Unterschiede in den Fähigkeitsverteilungen zwischen den Staaten zurückführen. Allenfalls im oberen Extrembereich der Fähigkeitsverteilung zeigt sich im Bereich Hören noch Optimierungspotential, was sich auch im höheren Standardfehler der Fähigkeitsschätzung in diesem Bereich widerspiegelt. Dies kann durch die Erweiterung des Itempools durch schwere und sehr schwere Items behoben werden. Der EOI kann in zukünftigen Untersuchungen einfach interpretierbare Informationen darüber liefern, wie gut der GTP die Testinformation für die verschiedenen GER-Niveaustufen maximiert. Der Begriff „Engineering“ steht hier für das Potenzial des EOI, die Auswirkungen von Änderungen an Itempool, Testdesgin oder dem adaptiven Algorithmus auf die Adaptivität des Tests anzuzeigen.

Aufgrund der bei dieser Studie vorliegenden Daten können keine Aussagen hinsichtlich der Validität der abgeleiteten Testwertinterpretationen (z. B. Hartig, Frey & Jude, im Druck) gemacht werden. Untersuchungen zur Validität könnten deshalb Gegenstand künftiger Studien sein.

Zusammenfassend stellt der GTP ein computerbasiertes Testinstrument dar, mit dem in kurzer Zeit auf effiziente Weise staatenübergreifend präzise Messwerte zur Hör- und Lesekompetenz am Arbeitsplatz über einen breiten Fähigkeitsbereich ermittelt werden.

LITERATUR

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723. https://doi.org/10.1109/TAC.1974.1100705

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48(6),* 1-29. https://doi.org/10.18637/jss.v048.i06

Frey, A. (im Druck). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. aktualisierte und überarbeitete Auflage). Berlin, Heidelberg: Springer.

Frey, A. & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft*, *Sonderheft 8*, 169-184. https://doi.org/10.1007/978-3-531-90865-6_10

Hartig, J., Frey, A. & Jude, N. (im Druck). Validität von Testwertinterpretationen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (3. aktualisierte und überarbeitete Auflage). Berlin, Heidelberg: Springer.

Kingsbury, G. & Wise, S. L. (2020) Three measures of test adaptation based on optimal test information. *Journal of Computerized Adaptive Testing*. 8(1), 1–19. https://doi.org/10.7333/2002-0801001

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests. Studies in mathematical psychology. Copenhagen:* Danmarks Paedagogiske Institut.

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

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464, https://doi.org/10.1214/aos/1176344136

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429–438). Boston: Elsevier Academic. https://doi.org/10.1016/b0-12-369398-5/00444-8

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