g.a.s.t. gesellschaft für akademische studienvorbereitung und testentwicklung e.v.

Standard-Setting zum digitalen TestDaF: Ein Online-Verfahren für eine Online-Prüfung

Gabriele Kecker & Thomas Eckes (2021)

Der digitale TestDaF wurde im Oktober 2020 mit komplett neu konzipiertem Testformat eingeführt. Die Testaufgaben sind sowohl auf den Verwendungskontext als anerkannter Sprachnachweis für die Aufnahme eines Hochschulstudiums internationaler Studierender als auch auf den Einsatz in einer digitalen Testumgebung abgestimmt. Wie schon der papierbasierte TestDaF ist auch der neue digitale Test auf die Niveaus B2.1 bis C1.2 des Gemeinsamen europäischen Referenzrahmens für Sprachen (GER) ausgerichtet (vgl. Europarat, 2001, 2020). Der Nachweis der Zuordnung des digitalen TestDaF zu diesen GER-Niveaus stellt einen wichtigen Bestandteil seiner Validierung dar und wurde daher bereits in der Entwicklungsphase begonnen. In Erprobungen (Field-Tests) mit durchschnittlich 250 Teilnehmenden pro Testsatz wurde eine zufriedenstellende psychometrische Qualität der neuen Testaufgaben und Items erzielt. Mithilfe eines kalibrierten Ankertests (onSET) wurde zusätzlich ein konstanter Schwierigkeitsgrad über verschiedene Testversionen hinweg sichergestellt, der als Voraussetzung für eine Zuordnung zu einem externen Referenzsystem wie dem GER gilt (Council of Europe, 2009). Gleichzeitig wurden Testaufgaben und Items durch den Einsatz der Anker-Items bei den Erprobungen indirekt den GER-Niveaus zugeordnet. Ziel des Standard-Settings war es, diese indirekte Zuordnung bei einem High-Stakes-Test wie dem TestDaF durch ein direktes Verfahren der Zuordnung durch Experten-Panels zu ergänzen und zu überprüfen.

1. Testformat des digitalen TestDaF

Der digitale TestDaF besteht aus vier Prüfungsteilen (Lesen, Hören, Schreiben und Sprechen), die an einem Computer oder Laptop in einem Testzentrum online abgelegt werden (zu weiteren Einzelheiten siehe g.a.s.t., 2020). Das Konzept, das dem TestDaF zugrunde liegt, beinhaltet die Überprüfung kommunikativer Aufgaben und Sprachhandlungen sowie damit verbundener zentraler Kompetenzen, die für die Aufnahme eines Hochschulstudiums relevant sind. Ausgangspunkt bei der Auswahl der sprachlichen Anforderungen in den Testaufgaben bildeten die Ergebnisse der Bedarfsanalyse, der Erprobungen, Studien und Gutachten, die in der Planungs- und Designphase durchgeführt wurden (vgl. Kecker, Zimmermann & Eckes, 2022; Kecker, Althaus & Eckes, in Vorbereitung). Um einen möglichst engen Bezug zu den sprachlichen Anforderungen in einem Hochschulstudium herzustellen, wurden zusätzlich zu den auf vorwiegend eine Fertigkeit ausgerichteten Aufgabentypen im digitalen TestDaF auch fertigkeitsübergreifende, integrierte Aufgabentypen in das Testformat aufgenommen. Auf diese Weise konnten studienbezogene kommunikative Anforderungen, wie z.B. die Lektüre eines Textes und seine mündliche oder schriftliche Verarbeitung abgebildet werden.

Das neue Testformat und die Berücksichtigung neuer Skalen zur Mediation im Begleitband des GER (Europarat, 2020) ermöglichen eine stärkere inhaltliche Anbindung des digitalen TestDaF an den GER. Nicht nur die Skalen zur Verarbeitung von Texten, sondern beispielsweise auch zum Anfertigen von Notizen werden im Begleitband der Mediation zugeordnet. In dieser Hinsicht lassen sich verschiedene im digitalen TestDaF berücksichtigte sprachliche Aktivitäten gut auf den Skalen des Begleitbands (Europarat, 2020) verorten: die Rezeption eines schriftlichen bzw.

mündlichen Textes ggf. auch unter Einbeziehung von statistischen Daten, die Weiterverarbeitung der Inhalte in Form von Notizen oder einer Zusammenfassung und die schriftliche oder mündliche Weitergabe der Information an andere. Um diese Aktivitäten aus dem digitalen TestDaF auf den GER zu beziehen, wurden im Standard-Setting folgende Skalen herangezogen: *Spezifische Informationen weitergeben* (Europarat, 2020, S. 116–117), *Daten erklären* (Europarat 2020, S. 118–119), *Verarbeitung von Texten* (Europarat, 2020, S. 120–122), *Notizen anfertigen* (Europarat, 2020, S. 125–126).

2. Methoden des Standard-Setting

Beim Standard-Setting geht es im Kern darum, Grenzen auf der Testwertskala festzulegen, welche die Gruppe der Testteilnehmenden in mindestens zwei Kategorien unterteilen: Teilnehmende, die die erforderliche Kompetenz oder das Wissen nachgewiesen haben, und solche, die dieses Ziel nicht erreichen (Cizek, 2012). Bei der Zuordnung zum GER steht nicht die Einteilung in „bestanden“ oder „nicht bestanden“ im Vordergrund, sondern die Zuordnung von Sprachkompetenzen – hier als TestDaF-Niveaus 3, 4 und 5 beschrieben – zu den passenden GER-Niveaus, im vorliegenden Fall B2, B2+/C1 und C1 (Kecker & Eckes, 2010; Kecker, 2011).

Im vorliegenden Fall wurden leicht verständliche und anzuwendende Standard-Setting-Methoden ausgewählt: für die rezeptiven Fertigkeiten die Basket-Methode und für die produktiven Fertigkeiten die Benchmark-Methode. Nach der Basket-Methode (Council of Europe, 2009; Kaftandjieva, 2009; Kecker, 2011) müssen die Expert*innen für jedes Item in dem Prüfungsteil Lesen oder Hören die folgende Frage beantworten: „Auf welchem GER-Niveau kann ein Testteilnehmender das folgende Item schon richtig beantworten?“. Nach der Benchmarking-Methode (Council of Europe, 2009; Kecker, 2011) werden Leistungsbeispiele der Prüfungsteilnehmenden in den Prüfungsteilen Schreiben und Sprechen von den Expert*innen den GER-Niveaus zugeordnet, die im GER-Raster zur Beurteilung von Schreiben (Europarat, 2020, Anhang 4) und in den qualitativen Merkmalen gesprochener Sprache (Europarat, 2020, Anhang 3) beschrieben sind. Sie müssen dazu die Frage beantworten: „Welchem GER-Niveau lässt sich das Leistungsbeispiel zuordnen?“.

Die Bewältigung dieser Aufgabe erfordert von den beteiligten Expert*innen (a) eine genaue Kenntnis des Testformats, um die Schwierigkeit der Aufgaben und Items für die Zielgruppe der Prüfung realistisch einzuschätzen, und (b) ein hohes Maß an Vertrautheit mit den zuvor ausgewählten GER-Skalen, die dem Konstrukt der Testaufgaben möglichst nahekommen. Darüber hinaus müssen die Expert*innen in einer Trainingsphase mit der Anwendung der ausgewählten Standard-Setting-Methode vertraut gemacht werden, damit sie in der darauffolgenden Phase des Standard-Settings konsistent und zuverlässig urteilen können. Die Trainingsphase dient zusätzlich dem Ziel, die Deskriptoren der GER-Niveaus möglichst einheitlich zu interpretieren sowie ein gemeinsames Verständnis von der Schwierigkeit und den sprachlichen Anforderungen der TestDaF-Aufgaben und -Items zu erreichen. Diese Anforderungen an das Verfahren entsprechen internationalen Standards (AERA, APA & NCME, 2014), dem *Manual* des Europarats für die Zuordnung von Sprachprüfungen zum GER (Council of Europe, 2009) und der einschlägigen Referenzliteratur (Cizek, 2012).

3. Standard-Setting als Online-Veranstaltung

Für das Standard-Setting des digitalen TestDaF konnten 32 nationale und internationale Expert*innen gewonnen werden, die unterschiedlichen Stakeholder-Gruppen angehören: sieben Personen von vier Testanbietern in Deutschland und Österreich, 10 Testexpert*innen verschiedener Universitäten in Deutschland und Österreich sowie vom British Council in

Großbritannien, sieben Prüfungsbeauftragte von TestDaF-Testzentren aus dem In- und Ausland sowie acht zertifizierte Beurteiler*innen von g.a.s.t. für den digitalen TestDaF.

Die Online-Veranstaltung umfasste die im Manual (Council of Europe, 2009) vorgesehenen Phasen:

a) Familiarisierung mit den GER-Skalen,
b) Training/Standardisierung in der Anwendung der Standard-Setting-Methode und Zuordnung zum GER,
c) das Standard-Setting;

In Phase (c) ordnete jede*r Expert*in individuell Items aus einem kompletten Prüfungsteil im Lesen und Hören bzw. Leistungen von Teilnehmenden im Schreiben und Sprechen den GER-Niveaus zu.

Bereits eine Woche vor Beginn der eigentlichen Veranstaltung wurde ein interaktiver Modelltest (Demo-Version) für die Gruppe der Expert*innen freigeschaltet, damit sie sich mit dem Testformat und den Aufgabentypen vertraut machen konnten. Die Online-Veranstaltungen zum Standard-Setting umfassten drei halbe Tage à 3-4 Std. Der erste Tag wurde mit allen Expert*innen gemeinsam durchgeführt, an den zwei weiteren Tagen wurden die Teilnehmenden in zwei Gruppen aufgeteilt: Gruppe A absolvierte das Training und Standard-Setting für das Lesen und Hören, Gruppe B parallel dazu das Training und Standard-Setting für das Schreiben und Sprechen.

Am ersten Tag erhielten alle Teilnehmenden in einer Video-Konferenz eine Einführung in das Konzept des digitalen TestDaF und in das Standard-Setting. Im Anschluss daran sollten alle Expert*innen zur Familiarisierung mit den GER-Skalen die in der Lernplattform Moodle bereitgestellten Übungen individuell bearbeiten. Wiederum in einer Videokonferenz wurden dazu Fragen beantwortet und Probleme der GER-Skalen bzw. Unsicherheiten in der Niveaueinschätzung von Deskriptoren diskutiert.

Am zweiten Tag wurde in parallelen Gruppen das Training in der Basket-Methode mit Items aus dem Prüfungsteil Lesen absolviert (Gruppe A) bzw. das Training für die Anwendung der Benchmarking-Methode mit Leistungsbeispielen aus dem Prüfungsteil Schreiben (Gruppe B). Für die Abstimmung der Expert*innen über das GER-Niveau der Trainingsbeispiele wurden in Moodle die Beispiele und ausgewählten GER-Skalen zur Verfügung gestellt. Die Zuordnung zum GER erfolgte für jedes einzelne Trainingsbeispiel in zwei Runden mit Diskussion im Plenum über das Video-Konferenzsystem. Zusätzlich erhielten die Expert*innen im Plenum ein Feedback über ihre Abstimmungsergebnisse, die für die Diskussion aus Moodle exportiert wurden. Als zusätzliches Feedback nach der zweiten Runde wurden der Gruppe die psychometrischen Item-Kennwerte des Items bzw. das GER-Niveau des Leistungsbeispiels bekannt gegeben, um die empirische gemessene Schwierigkeit bzw. das kalibrierte GER-Niveau mit den im Training ermittelten urteilsbasierten Werten vergleichen zu können.

Nach Abschluss des Trainings folgte das eigentliche Standard-Setting in den zwei parallelen Gruppen: Jedes Gruppenmitglied bewertete individuell in Moodle die Items eines kompletten Prüfungsteils im Lesen bzw. eine größere Anzahl an Leistungsbeispielen zu allen Aufgaben im Schreiben mit den zur Verfügung gestellten GER-Skalen. Nach demselben Prinzip wurde am dritten halben Tag mit den Items des Prüfungsteils Hören und den Leistungsbeispielen des Prüfungsteils Sprechen verfahren.

Für das Training wurden exemplarische Items bzw. Leistungsbeispiele aus dem Modelltest (Demo-Version) verwendet, der mit 247 Personen erprobt worden war. In den Prüfungsteilen Lesen und Hören wurden jeweils vier Items aus unterschiedlichen Aufgabentypen nach dem zuvor beschriebenen Verfahren zugeordnet und diskutiert. In den produktiven Prüfungsteilen waren es vier Leistungsbeispiele im Schreiben und drei im Sprechen zu unterschiedlichen Aufgabentypen.
Für das Standard-Setting wurden Items und Leistungen aus dem ersten Testlauf im Oktober 2020 ausgewählt; diesen hatten insgesamt 129 Personen absolviert. In den rezeptiven Prüfungsteilen mussten 34 Items im Lesen (ein kompletter Prüfungsteil) und 26 Items im Hören (ein gesamter Prüfungsteil ohne die letzte Aufgabe 7 mit vier Items) dem GER zugeordnet werden. In den produktiven Prüfungsteilen wurden 10 Leistungsbeispiele im Schreiben und 12 Leistungsbeispiele im Sprechen mit den GER-Rastern bewertet. Für den Prüfungsteil Sprechen wurde eine höhere Anzahl an Beispielen ausgewählt, um die Bandbreite der sieben Aufgabentypen abdecken zu können. Im Prüfungsteil Schreiben, der lediglich zwei Aufgabentypen umfasst, war dies mit weniger Beispielen möglich.

4. Validität des Verfahrens

Die Expert*innen erhielten nach jedem Tag einen Fragebogen mit Fragen zur Validität des Verfahrens (Hambleton, 2001), der in Moodle eingestellt war. Nach den Angaben der Expert*innen fühlten sie sich durch die Einführung in das Standard-Setting und in die damit verbundene Aufgabe als Expert*in sehr gut vorbereitet. Die GER-Übungen und das Training mit dem Feedback wurden als sehr hilfreich bewertet. Die Zeit für die Diskussionen wurde überwiegend als ausreichend eingeschätzt. Die Umsetzung als Online-Veranstaltung wurde hinsichtlich Organisation und Technik als sehr gelungen bezeichnet und inhaltlich als anspruchsvoll angesehen. Da das digitale Format der Veranstaltung eine gewisse kognitive Belastung mit sich brachte und dadurch nur eine zeitlich begrenzte Ausdehnung der einzelnen Phasen möglich war, konnte in den Trainingsphasen lediglich die notwendige Mindestanzahl an Trainingsbeispielen behandelt werden. Auf diese Weise sollte für die einzelnen behandelten Beispiele genügend Zeit für die Diskussion und die zwei Abstimmungsrunden mit Feedback erhalten bleiben.

5. Auswertung und Ergebnisse

Die psychometrischen Analysen ergaben ein zufriedenstellendes Maß an Übereinstimmung innerhalb beider Gruppen nach Runde 2 der Trainingsphase. Im anschließenden Standard-Setting hatte Gruppe A (Lesen, Hören) Items aus der ersten digitalen TestDaF-Prüfung vom 22. Oktober 2020 wiederum nach der Basket-Methode zuzuordnen. Gruppe B (Schreiben, Sprechen) hatte aus dieser Prüfung stammende Leistungsbeispiele analog zum Training auf der GER-Skala einzuordnen.

Die Übereinstimmungs- und Reliabilitätswerte innerhalb beider Gruppen lagen wieder auf einem erfreulich hohen Niveau. Zudem stimmten die TDN-Einstufungen aus der ersten digitalen TestDaF-Prüfung mit den Einstufungen der Expert*innen weitgehend überein. Gelegentliche Abweichungen konnten auf die besonderen Merkmale des Online-Formats zurückgeführt werden.

6. Schlussbemerkung

Das Standard-Setting wurde relativ kurz nach der Einführung des digitalen TestDaF durchgeführt, um das indirekte Verfahren der GER-Zuordnung aus der Entwicklungsphase zeitnah durch ein direktes urteilsbasiertes Verfahren mit Expert*innen zu ergänzen. Pandemiebedingt konnte dies

nicht in Präsenzform stattfinden, sondern musste als Online-Veranstaltung umgesetzt werden. Die zu diesem Zweck eingeladenen 32 nationalen und internationalen Expert*innen sind insgesamt sehr gut mit dem Online-Verfahren zurechtgekommen. Die Evaluierung des Verfahrens mithilfe eines Online-Fragebogens und die psychometrische Auswertung des Abstimmungsverhaltens (Übereinstimmung und Konsistenz) ergaben in dieser Hinsicht erfreulich positive Rückmeldungen und Resultate.

Der Vorteil eines Online-Verfahrens liegt in der besseren Verfügbarkeit der Expert*innen, die an verschiedenen Standorten im In- und Ausland zu einer bestimmten Zeit gemeinsam zur Verfügung stehen. Des Weiteren ermöglicht ein Online-Verfahren für eine Online-Prüfung bessere logistische Möglichkeiten während des Verfahrens, um Testaufgaben an Computern interaktiv zu bearbeiten, da jede*r an dem eigenen Gerät arbeiten kann. Auch zeitliche Verzögerungen durch Expert*innen, die langsamer arbeiten, lassen sich in der Phase des Standard-Settings besser auffangen, da nicht alle die Bearbeitung gemeinsam beenden müssen. Ein Nachteil besteht jedoch darin, dass in Video-Konferenzsystemen bei Präsentationen nicht jeder Teilnehmende einer großen Gruppe für den/die Moderator*in sichtbar ist. Dadurch gehen nonverbale Signale aus der Gruppe verloren, die z.B. das Verständnis oder möglichen Erklärungsbedarf anzeigen. Durch das digitale Format der Veranstaltung ist es zudem schwieriger die Aufmerksamkeit der Teilnehmenden bei einer solch komplexen Aufgabe über einen relativ langen Zeitraum zu binden. Ablenkungen durch Ereignisse vor Ort können nicht ausgeschlossen werden.

Trotz der insgesamt positiven Ergebnisse sollte das Standard-Setting im Sinne einer weiteren Validierung des digitalen TestDaF wiederholt werden, wenn eine stabilere Datenbasis für die Ermittlung der Itemkennwerte vorliegt und eine größere Auswahl an Leistungsbeispielen zur Verfügung steht. Es bleibt zu hoffen, dass die Rahmenbedingungen dann Präsenzveranstaltungen erlauben, die mehr Möglichkeiten zur Diskussion und Moderation bieten.

Literatur:

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2014). *The Standards for Educational and Psychological Testing*. Washington, DC.

Cizek, G. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations* (2nd ed). Routledge.

Council of Europe. (2009). *Relating language examinations to the "Common European Framework of Reference for Languages: Learning, Teaching, Assessment" (CEFR): A manual.* Council of Europe/Language Policy Division.

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

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

g.a.s.t. (2020). *Der digitale TestDaF. Zielsetzung, Konzept und Testformat.* Gesellschaft für Akademische Testentwicklung und Studienvorbereitung, g.a.s.t. e.V.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, perspectives* (pp. 89–116). Erlbaum.

Kaftandjieva, F. (2009). Basket procedure: The breadbasket or the basket case of standard setting methods? In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 21–34). Cito/EALTA.

Kecker, G. (2011). *Validierung von Sprachprüfungen: Die Zuordnung des TestDaF zum Gemeinsamen europäischen Referenzrahmen für Sprachen*. Peter Lang.

Kecker, G., & Eckes, T. (2010). Putting the manual to the test: The TestDaF–CEFR linking project. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 50–79). Cambridge University Press.

Kecker, G., Zimmermann, S. & Eckes, T. (2022). Der Weg zum digitalen TestDaF: Konzeption, Entwicklung und Validierung. In P. Gretsch & N. Wulff (Hrsg.), *Deutsch als Zweit- und Fremdsprache in Schule und Beruf* (S. 393–410). Brill Schöningh.