**Beitrag Themenheft „Testen und Prüfen“**

Gabriele Kecker* und Thomas Eckes*

# Der digitale TestDaF: Aufbruch in neue Dimensionen des Sprachtestens

## The digital TestDaF: Entering new dimensions of language testing

https://doi.org/10.1515/infodaf-2022-0057

**Zusammenfassung:** Der vorliegende Beitrag beschreibt die Entwicklungsphasen des digitalen TestDaF und zeigt, wie die neuen Aufgabentypen zusammen mit einer komplett digitalisierten Durchführung den Bezug zum Hochschulkontext und zum GER stärken. Der Beitrag geht im Detail auf das Testkonzept, das Testformat und die im Lesen, Hören, Schreiben und Sprechen verwendeten Aufgabentypen ein. Innovative Ansätze und Methoden der Leistungsbewertung und Ergebnisermittlung kommen ebenso zur Sprache wie ihre Umsetzung in der Praxis. Abschließend werden Perspektiven der Forschung zum digitalen TestDaF aufgezeigt und Bezüge zur aktuellen Sprachtestforschung diskutiert.

**Schlüsselwörter:** Testentwicklung, digitales Testen, integrierte Aufgaben, Sprachtestforschung

**Abstract:** This paper describes the developmental phases of the digital TestDaF and shows how the new task types and the digital test administration strengthen the relationship to the context of higher education institutions in Germany and to the CEFR. The contribution deals extensively with the test construct, the test format, and the different types of tasks used in the reading, listening, writing, and speaking sections. Innovative approaches and methods of performance assessment and scoring are discussed as much as their practical implementation. Finally, the paper presents some perspectives of future studies on the digital TestDaF and elaborates on relations to contemporary language testing research.

**Keywords:** test development, web-based testing, integrated tasks, language testing research

***Kontaktpersonen: Dr. Gabriele Kecker,** E-Mail: kontakt@gast.de
**Dr. Thomas Eckes,** E-Mail: Thomas.Eckes@gast.de

# 1 Einleitung

Seit Jahren ist die Nutzung digitaler Medien ein selbstverständlicher Bestandteil der wissenschaftlichen Arbeit und der Kommunikation an Hochschulen. Studierende müssen in einem zunehmend digitalen Studium über sprachliche Kompetenzen verfügen, die den Umgang mit diesen Formen der Kommunikation einschließen. Mündliche Kommunikation erfolgt im Studium nicht nur in Präsenz, sondern häufig über Videokonferenzsysteme. Sprecherwechsel werden durch Moderation eingeleitet, Statements sind oft länger und weniger spontan als in einer Seminardiskussion. Texte werden am Computer geschrieben, Vorträge am Bildschirm und über Lautsprecher verfolgt und später in eigenen Texten am Computer verarbeitet. Es gibt gute Gründe, solche Entwicklungen bei der Überprüfung von Sprachkompetenzen, die für die Studienzulassung internationaler Studierender obligatorisch sind, zu berücksichtigen. Zum einen können Prüfungsaufgaben, die möglichst viele Merkmale aus der (auch digitalen) Sprachverwendungssituation abbilden, als authentischer angesehen werden; zum anderen bietet eine digitale Prüfungsdurchführung mehr Möglichkeiten, Prüfungsaufgaben dem spezifischen Kontext entsprechend und vielseitig zu präsentieren. Hinzu kommen Vorteile in der Ergebnisermittlung und Prüfungsverwaltung.

Im vorliegenden Beitrag stellen wir den im Oktober 2020 eingeführten digitalen TestDaF und seine Anwendung im deutschen Hochschulkontext vor. Im Mittelpunkt stehen die Besonderheiten seines in einem insgesamt zehnjährigen Arbeitsprozess entwickelten Formats und die für die praktische Umsetzung erforderliche Testumgebung. Der folgende Abschnitt gibt zunächst einen Überblick über den Ausgangspunkt der Neuentwicklung und die verschiedenen Phasen der Gestaltung und Validierung des Testformats. In einem weiteren Abschnitt stellen wir das Konzept des digitalen TestDaF sowie die vier Prüfungsteile mit ihren Aufgabentypen dar. Dabei berücksichtigen wir auch die Form der Präsentation der Testaufgaben auf dem Bildschirm. Anschließend gehen wir auf die Verbindung zwischen dem digitalen Test und der Testumgebung ein, die für eine sichere Online-Durchführung und damit einhergehende Verwaltungsprozesse wie Registrierung oder Ähnliches erforderlich ist. In den letzten beiden Abschnitten zeigen wir einige Forschungsperspektiven auf und geben einen Ausblick auf die Rolle des digitalen TestDaF in der deutschen Bildungslandschaft.

# 2 Entwicklungsphasen des digitalen TestDaF

Ein Nachweis von sprachlichen Kompetenzen, die für die Aufnahme eines Studiums an einer deutschen Hochschule erforderlich sind, sollte präzise Auskunft da-

rüber geben, wie gut die kommunikativen Aufgaben in einem Hochschulstudium von Studienbewerbern und -bewerberinnen bewältigt werden können. Prüfungsteilnehmende, ihre Eltern, zulassende Einrichtungen, Lehrkräfte und andere Stakeholder verlassen sich auf die Zuverlässigkeit und prognostische Validität von Prüfungsergebnissen und eine entsprechende sprachliche Kompetenz von Absolventen und Absolventinnen im Studienalltag. In einer standardisierten Sprachprüfung wird dieses Ziel durch die Berücksichtigung international akzeptierter Qualitätsstandards bei der Prüfungsentwicklung und Validierung erreicht (vgl. Bachman/Palmer 2010; Kecker/Zimmermann/Eckes 2022; Lane/Raymond/Haladyna 2016). Dazu gehört eine möglichst genaue Abbildung von kommunikativen Aufgaben und den durch sie ausgelösten kognitiven Verarbeitungsprozessen, wie sie für Situationen der Sprachverwendung als charakteristisch gelten können. Im Folgenden soll aufgezeigt werden, in welchen Schritten die Entwicklung des Testformats, der Benutzeroberfläche und der Durchführungssoftware vollzogen wurde und welche Maßnahmen zur Validierung damit verbunden waren. Die einzelnen Entwicklungsphasen sind Abbildung 1 zu entnehmen.

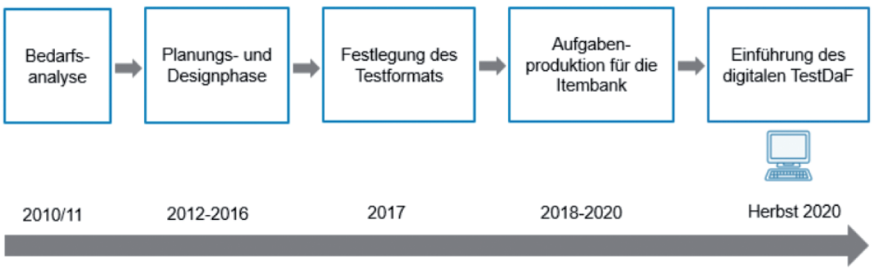


**Abb. 1:** Entwicklungsphasen des digitalen TestDaF

## 2.1 Bedarfsanalyse

Die Entwicklung des papierbasierten TestDaF erfolgte in den späten 1990er Jahren. Daher wurde zu Beginn der Formatentwicklung des digitalen TestDaF in den Jahren 2010/2011 zunächst mithilfe einer Bedarfsanalyse überprüft, ob und wie sich die sprachlichen Anforderungen im Hochschulstudium seither verändert hatten (vgl. Arras 2012; Marks 2015).

Für die Bedarfsanalyse wurde ein Mehrmethoden-Design angewendet, das sowohl qualitative Methoden (1. Phase) als auch quantitative Methoden (2. Phase) umfasste. In der ersten Phase wurden sprachliche Anforderungen für ein Studium an deutschen Hochschulen in Zusammenarbeit mit Lehrkräften aus dem Hochschulbereich zusammengestellt. Diese wurden in leitfadenorientierten Interviews

mit insgesamt 38 internationalen und deutschen Studierenden sowie Hochschuldozenten und -dozentinnen an verschiedenen Hochschulen und Fachbereichen in mehreren Bundesländern überprüft. In der zweiten, quantitativen Phase wurden ausgehend von den ermittelten sprachlichen Anforderungen, das heißt kommunikativen Aufgaben und sprachlichen Aktivitäten (vgl. Europarat 2001) im Hochschulstudium, Fragebögen für internationale Studierende und Hochschullehrkräfte entwickelt. Mithilfe der Fragebögen wurde erhoben, welche sprachlichen Aktivitäten an den Hochschulen in Deutschland aus Sicht der Befragten relevant sind, häufig vorkommen und welche sprachlichen Schwierigkeiten die Befragten im akademischen Kontext damit verbinden. Es gingen 1327 vollständig ausgefüllte Fragebögen von 9493 befragten Studierenden und 120 Rückläufe von 1045 befragten Dozenten bzw. Dozentinnen in die Auswertung ein.

Als Ergebnis wurden die nachstehend aufgeführten kommunikativen Aufgaben als Ausgangspunkt für das Testformat des digitalen TestDaF festgelegt:

a) Gespräche zur Studienorganisation und zu Alltagsfragen führen
b) *fachbezogene Gespräche* mit Kommilitonen und Kommilitoninnen und Lehrkräften führen, *dabei auch auf Lektüre Bezug nehmen*
c) in Seminaren Kurzpräsentationen halten
d) *in Seminaren auf andere Beiträge reagieren* und/oder einen eigenen Beitrag leisten
e) Diskussionen folgen und gegebenenfalls Stellung nehmen
f) Vorlesungen/Vorträge hören, dabei Handout/Präsentationsfolien lesen, Notizen machen
g) *Lektüre verarbeiten*, Notizen machen, *Texte dazu schreiben*
h) schriftliche Texte *wie Forumsbeiträge* oder Textpassagen für eine Hausarbeit verfassen

Die in der Auflistung kursiv gesetzten Textpassagen bezeichnen Anteile der kommunikativen Aufgaben, die im digitalen TestDaF abgebildet werden und in der papierbasierten Version nicht enthalten sind.

Im folgenden Abschnitt wird erläutert, wie der nächste Entwicklungsschritt von den oben angegebenen kommunikativen Aufgaben zu den Testaufgaben im digitalen TestDaF vollzogen wurde.

## 2.2 Planungs- und Designphase

### 2.2.1 Berücksichtigung von zentralen Kompetenzen für ein Hochschulstudium

Werden Testaufgaben aus einer Bedarfsanalyse abgeleitet und sollen wesentliche Elemente der kontextuellen Sprachverwendung dabei nicht verloren gehen, muss eine genaue Analyse der kommunikativen Situation und der durch die kommunikativen Aufgaben angestoßenen kognitiven Verarbeitungsprozesse erfolgen; dies schließt die dabei verwendeten Input-Materialien ein. Die nähere Betrachtung der Verarbeitungsprozesse, die für die Bewältigung der kommunikativen Aufgaben a) bis h) (siehe 2.1) als notwendig gelten können, ergab Cluster von Kompetenzen, die für diese Prozesse relevant sind. Die Kompetenzen weisen sowohl rezeptive als auch produktive Anteile auf; sie spielen daher für die Überprüfung aller vier Fertigkeiten oder Teilkompetenzen eine erhebliche Rolle:

a) Positionen/Einstellungen anderer erkennen, wiedergeben und (gegebenenfalls eigenen) gegenüberstellen
b) persönliche Meinung und Sachargument unterscheiden
c) mündlich oder schriftlich Stellung nehmen
d) Unterschiede bzw. Übereinstimmungen erkennen und wiedergeben
e) kausale Zusammenhänge erkennen und ausdrücken
f) Grafiken erfassen und Informationen daraus versprachlichen
g) Notizen anfertigen und verarbeiten
h) Informationen aus Zusammenfassungen verarbeiten und Zusammenfassungen produzieren

Diese Liste ließe sich fortsetzen. Jedoch ist dabei zu beachten, dass die Auswahl möglichst allgemeine studienrelevante Kompetenzen erfasst, also solche, die für viele Studienfächer passen, da der TestDaF die Sprachkompetenz nicht studienfachbezogen überprüft. Grundlegend für die Auswahl ist darüber hinaus, dass Wissensvermittlung – ganz gleich, ob durch eigene Lektüre, Seminare oder Vorlesungen – in fast allen Fällen auf statistische Daten, Grafiken, Abbildungen, Fotos, Videos und ähnliche, häufig visuelle Materialien zurückgreift, um Inhalte (z. B. Theorien, Fakten, Positionen) zu veranschaulichen oder zu ergänzen (vgl. Ockey 2007: 532–533). Dabei kommt es in vielen Fällen darauf an, Textinhalte (auch auditive in Seminaren oder Vorlesungen) mit visuell präsentierten Inhalten oder anderen Texten abzugleichen, Zusammenhänge zu erkennen, Sachverhalte kurz festzuhalten und weiterzuverarbeiten. Des Weiteren bestehen wesentliche Schritte in der wissenschaftlichen Arbeit darin, Argumente, Strukturen, Positionen bzw. Einstellungen zu erkennen, um sie zum Beispiel einander gegenüberzustellen und gegebenenfalls schriftlich oder mündlich dazu Stellung nehmen zu können.

Daher wurden bei der Entwicklung der Testaufgaben visuelle Medien in die Aufgabenstellungen eingebunden. Diese Medien dienen sowohl der Veranschaulichung und dem besseren Verständnis von Fachterminologie und Sachverhalten als auch der Illustration kommunikativer Situationen, indem sie zum Beispiel Gesprächspartner und -partnerinnen, Seminarteilnehmende oder Vortragende auf Fotos zeigen (Prüfungsteile Hören und Sprechen). Sie werden somit nach Ockey (2007) sowohl inhaltsbezogen als auch kontextbezogen eingesetzt.

### 2.2.2 Integrierte Aufgaben

Bei den kommunikativen TestDaF-Aufgaben (Abschnitt 2.1) wird deutlich, dass bei der Bearbeitung der Aufgaben häufig ein schriftlicher oder mündlicher Text-Input rezipiert und verstanden sowie für eine eigene Äußerung weiterverarbeitet werden muss. Somit sind Kompetenzen in mehr als einer Fertigkeit angesprochen. Diesen Sachverhalt in Testaufgaben zu übertragen, bedeutet, auch fertigkeitsübergreifende Testaufgaben, sogenannte integrierte Aufgaben (*integrated tasks*; vgl. Plakans 2013; Knoch/Sitajalabhorn 2013), einzuführen. Dabei sind zwei Gesichtspunkte ausschlaggebend: Einerseits soll das Testkonstrukt durch die Einbindung integrierter Aufgaben erweitert und somit auch authentischer umgesetzt werden, andererseits muss sichergestellt sein, dass eine Fertigkeit durch zusätzliche unabhängige oder isolierte Testaufgaben zuverlässig und ohne Beeinträchtigung durch eine andere Fertigkeit gemessen werden kann. Aus diesem Grund besteht der überwiegende Anteil der 23 Testaufgaben im digitalen TestDaF aus Aufgaben, die vorwiegend eine Fertigkeit erfordern, und lediglich vier sind integrierte oder fertigkeitsübergreifende Aufgaben, die zwei Fertigkeiten einbinden. Aufgaben aus dem Hörverstehen, die mit Kurzantworten zu lösen sind (Notizen anfertigen), gelten noch nicht als fertigkeitsübergreifend. Vielmehr sollten nach Plakans (2013) integrierte Aufgaben einen substanziellen Input-Text und bei produktiven Aufgaben als Antwort der Prüfungskandidaten und -kandidatinnen eine umfangreichere Textproduktion aufweisen, um als integriert gelten zu können. Klare Hinweise zur Festlegung der Länge solcher Input-Texte oder der erwarteten produktiven Äußerung sind in der Forschungsliteratur allerdings noch nicht zu finden. Daher folgt die Festlegung im digitalen TestDaF den Erfahrungswerten aus den Erprobungen.

### 2.2.3 Anforderungen an die Testaufgaben

Bei einer derart umfangreichen Formatrevision wie im vorliegenden Fall spielt nicht nur die Anpassung bzw. Optimierung des Testkonstrukts eine Rolle. Vielmehr kommen auch psychometrische Überlegungen und Erfordernisse einer digitalen Testdarbietung ins Spiel. Im Sinne einer Konstrukterweiterung sollten in die Prüfungsteile Lesen, Hören und Schreiben des digitalen TestDaF mehr Aufgabentypen als im papierbasierten Format aufgenommen werden. Des Weiteren sollte in den neuen Prüfungsteilen Lesen und Hören die Gesamtzahl der Items zwar weitgehend gleichbleiben, die Anzahl der Items pro Lese- oder Hörtext jedoch reduziert werden. Dies dient einer besseren Kombinationsmöglichkeit der einzelnen Items im jeweiligen Prüfungsteil.

Bei der Formatentwicklung wurden Aufgabenformate, die eine Auswertung mit möglichst hoher Reliabilität ermöglichen, bevorzugt. Dennoch wurden Kurzantworten (im Hören) und offene Formate (im Schreiben und Sprechen) im Sinne einer möglichst weitgehenden Abbildung von sprachlichen Anforderungen im Studium beibehalten. Im Hinblick auf die Testökonomie war der Aufwand für die Erstellung einer Testaufgabe in ein akzeptables Verhältnis zur Anzahl der Items und somit zum Anteil am Testergebnis zu bringen.

### 2.2.4 Erprobungen von Aufgabentypen (Try-Outs)

Erste Aufgabentypen, die die kommunikativen Aufgaben und zentralen Kompetenzen (vgl. 2.1 und 2.2) so weit als möglich berücksichtigten, wurden in der Planungs- und Designphase mit kleinen Gruppen der Zielpopulation von 20 bis 60 Personen in Deutschland erprobt, ausgewertet und revidiert. Von 2012 bis 2016 wurden vier Try-Outs mit insgesamt 534 Probanden und Probandinnen durchgeführt. Der lange Zeitraum geht unter anderem darauf zurück, dass die Software zur Darbietung der schrittweise veränderten neuen Aufgabentypen stetig weiterentwickelt werden musste.

Folgende Kriterien wurden bei der Analyse der Try-Out-Ergebnisse und der Entscheidung über eine Revision der Aufgaben zugrunde gelegt:

- Beantwortung der Items/Aufgaben mit den intendierten Lösungen bzw. Texten
- Angemessenes Schwierigkeitsniveau der Items/Aufgaben
- Psychometrische Qualität der Items/Aufgaben
- GER-Niveau der Probandengruppe (Einsatz des onSET als Ankertest)
- Verständlichkeit und Handhabbarkeit der Aufgabendarstellung auf dem Bildschirm

Als weiteres Kriterium für die Eignung einer Aufgabe wurden Ergebnisse der Online-Befragungen der Probanden und Probandinnen herangezogen. Diese gaben insbesondere Aufschluss über wahrgenommene Bearbeitungszeiten, Verständlichkeit der Anleitungen, kognitive Belastung und Benutzerfreundlichkeit (vgl. 2.4.2).

## 2.3 Aufgabenproduktion für die Itembank

### 2.3.1 Festlegung des Testformats

Eine der Voraussetzungen für den Aufbau einer Itembank ist, dass das Testformat feststeht, bevor mit der Aufgabenproduktion in größerem Umfang begonnen wird. Dazu gehören Erprobungen mit einem größeren Sample an internationalen Probanden und Probandinnen und entsprechende psychometrische Analysen. Dementsprechend wurden die Aufgabentypen, die in den Try-Outs bei den qualitativen und quantitativen Auswertungen gut abgeschnitten hatten, zu den vier Prüfungsteilen Lesen, Hören, Schreiben und Sprechen zusammengestellt und umfassend erprobt. Die erste Erprobung (Field-Test 1) fand 2017 mit 247 Teilnehmenden an den Testzentren von g.a.s.t. statt. Die psychometrischen Analysen ergaben zufriedenstellende Ergebnisse. Ausgehend von der empirisch ermittelten Qualität der Aufgabentypen und Items wurden zur Validierung des Testformats in den vier Prüfungsteilen zusätzlich Gutachten von vier ausgewiesenen Experten und Expertinnen der Sprachtestforschung eingeholt (vgl. 2.4.2). Diese erste Testversion wurde vor Einführung des digitalen TestDaF als Demo-Version bzw. Modelltest auf der Webseite des TestDaF-Instituts (www.testdaf.de) veröffentlicht.

### 2.3.2 Field-Tests

Das begutachtete Testformat aus Field-Test 1 wurde nach geringfügigen Anpassungen als Grundlage für alle weiteren Field-Tests festgelegt. Es wurden Aufgaben nach diesem Format für den sukzessiven Aufbau der Itembank fortlaufend entwickelt und erprobt. Von 2018 bis Ende 2020 wurden auf diese Weise sechs Field-Tests mit insgesamt 2924 Probanden bzw. Probandinnen und circa 250 Probanden bzw. Probandinnen pro Testsatz in den Testzentren von g.a.s.t. durchgeführt. Ziel der Erprobungen ist es, mithilfe einer gut ausgestatteten Itembank Testsätze automatisiert generieren und Testzentren Prüfungstermine bedarfsgerecht (*on demand*) anbieten zu können. Zu diesem Zweck werden während der Prüfungen zusätzlich sogenannte Einstreuaufgaben eingesetzt. Diese Aufgaben

sind von den Prüfungsteilnehmenden zusammen mit dem eigentlich für die Prüfung vorgesehenen Testmaterial zu bearbeiten. Einstreuaufgaben werden psychometrisch analysiert, bleiben jedoch von der Ergebnisermittlung ausgeschlossen (siehe auch Abschnitt 4.1).

## 2.4 Validierung

Bereits während der Entwicklungsphase wurden quantitative und qualitative Forschungsmethoden zur Validierung des digitalen TestDaF eingesetzt.

### 2.4.1 Quantitative Analysen

Quantitative Methoden wie Itemanalysen nach der klassischen Testtheorie und Rasch-Analysen dienten der Absicherung der Item- und Aufgabenqualität. Auch die Eignung der neu entwickelten Beurteilungsskalen für die Prüfungsteile Schreiben und Sprechen und die Leistung der Beurteilenden selbst wurden mithilfe psychometrischer Analysen, insbesondere Multifacetten-Rasch-Analysen (Eckes 2015a; Eckes/Jin 2021), evaluiert.

Darüber hinaus wurde eine erste Studie mit Teilnehmenden an Erprobungen des digitalen TestDaF durchgeführt, die zeitnah auch den papierbasierten TestDaF abgelegt hatten. Gegenstand der Studie war der Vergleich der Ergebnisse, die diese Probandengruppe in beiden Testformaten erreicht hatten. Dabei ist zu beachten, dass der digitale TestDaF keine bloße Übertragung des papierbasierten Tests in ein Online-Format darstellt. Wie zu Beginn dieses Beitrags erläutert, wurde für die Entwicklung des digitalen TestDaF das Ziel gesetzt, die aktuellen sprachlichen Anforderungen und charakteristischen kommunikativen Aufgaben eines Hochschulstudiums so authentisch wie möglich abzubilden. Dies führte zu einer deutlich erweiterten Operationalisierung des Testkonstrukts.

In der Forschung zur Vergleichbarkeit bzw. Äquivalenz papierbasierter und digitaler Testversionen werden üblicherweise drei methodische Zugänge verfolgt: (a) Score-Äquivalenz, (b) kognitive Äquivalenz (oder Konstrukt-Äquivalenz) und (c) psychometrische Äquivalenz (Chan/Bax/Weir 2018; Dadey/Lyons/DePascale 2018).

Bei der Score-Äquivalenz geht es um die Frage, ob Teilnehmende, die beim papierbasierten TestDaF gut (oder schlecht) abgeschnitten haben, ähnlich gut (oder schlecht) beim digitalen TestDaF abschneiden. Auswertungen von Daten aus einem Field-Test und drei papierbasierten TestDaF-Prüfungen sprechen da-

für, dass dies der Fall ist. Die (durchweg positiven) Korrelationen erwiesen sich in allen vier Prüfungsteilen trotz relativ niedriger Fallzahlen (durchschnittlich 59.7 Teilnehmende) bis auf wenige Ausnahmen als statistisch hochsignifikant; die Werte bewegten sich in der Mehrzahl der Korrelationen zwischen .50 und .70.

Kognitive Äquivalenz bezieht sich auf die Frage nach den kognitiven Prozessen (Sprachverarbeitungsprozessen), die mit der Aufgabenbearbeitung einhergehen. Quantitative und qualitative Studien (s. Abschnitt 2.4.1 und 2.4.2) belegen, dass die beim papierbasierten TestDaF intendierten kognitiven Prozesse auch beim digitalen TestDaF angestoßen werden. Verarbeitungsprozesse für die kompetenzübergreifenden, integrierten Aufgaben des digitalen TestDaF gehen allerdings darüber hinaus.

Psychometrische Äquivalenz betrifft die Vergleichbarkeit hinsichtlich Messgenauigkeit (Reliabilität), Validität und Fairness. Reliabilitätswerte der digitalen Prüfungsteile um .80 oder deutlich darüber verweisen auf eine mindestens so hohe Genauigkeit wie beim papierbasierten TestDaF. Bislang vorliegende Analysen zur Validität (z. B. Funktionalität der neu entwickelten Aufgabenformate und Beurteilungsverfahren) sowie Fairness (Beurteilungseffekte) sprechen für eine hohe Messqualität auch beim digitalen TestDaF (Eckes/Jin 2021). Mit steigenden Zahlen von Teilnehmenden an digitalen TestDaF-Prüfungen wird es künftig darum gehen, die breitere Datenbasis für weitere, eingehende Analysen zur Frage der Vergleichbarkeit zu nutzen.

### 2.4.2 Qualitative Analysen

Qualitative Methoden in Form von Gutachten, Befragungen und retrospektiven Interviews wurden für die Überprüfung spezifischer Aspekte des Testkonstrukts verwendet (z. B. Authentizität der Aufgaben im Hörverstehen, Schwierigkeiten bei der Bearbeitung der integrierten Aufgaben).

Die Eignung des Testkonstrukts und dessen Operationalisierung in Form der Aufgaben wurden von vier internationalen Experten und Expertinnen der Sprachtestforschung begutachtet: Prof. Ute Knoch (Universität Melbourne), Peter Lenz (Universität Fribourg), Prof. Henning Rossa (Universität Trier) und Dr. Carol Spöttl (Universität Innsbruck); dabei waren der intendierte Verwendungszweck der Prüfung und ihr digitaler Einsatz zu berücksichtigen.

Gegenstand der Gutachten waren folgende Fragestellungen:

- Sind die Aufgabentypen geeignet, um das in den Testspezifikationen dargestellte Testkonstrukt zu überprüfen?
- Ist das Konstrukt durch das Aufgabenmaterial in dem jeweiligen Prüfungsteil ausreichend abgedeckt?

- Sind die Aufgaben klar und verständlich dargestellt, auch hinsichtlich der Präsentation auf dem Bildschirm? Könnte sich durch die Art der Darstellung eine Verzerrung der intendierten Messung ergeben?
- Erscheint das Schwierigkeitsniveau der Aufgaben angemessen im Hinblick auf die sprachlichen Anforderungen an der Hochschule und die angestrebten GER-Niveaus B2 und C1?
- Sind die Bearbeitungszeiten für die einzelnen Aufgaben im Hinblick auf die Sprachkompetenzen der Zielgruppe angemessen?

In den Gutachten wurde das Format des digitalen TestDaF aus dem ersten Field-Test als geeignet angesehen, den Nachweis der deutschen Sprachkenntnisse für die Studienzulassung erbringen zu können. Anregungen der Gutachter und Gutachterinnen zur Optimierung der Darstellung auf dem Bildschirm (vgl. 3.4) oder zur Formulierung von Aufgaben wurden so weit wie möglich umgesetzt.

Ergänzend zu den Gutachten wurde die subjektive Einschätzung von Teilnehmenden an den Erprobungen mithilfe von Online-Befragungen direkt nach der Prüfungsdurchführung erhoben. Folgende Aspekte des Konstrukts wurden thematisiert: Bezug der Aufgaben zum Studium, Aufgabenschwierigkeit und Bearbeitungszeit der Aufgaben, Hörtexte und Hörfrequenz, Verständlichkeit der Anleitungen, kognitive Belastung und Vertrautheit mit dem Computer sowie Benutzerfreundlichkeit der digitalen Testumgebung. Die Befragungen ergaben erwartungsgemäß Hinweise auf Probleme bei der Verwendung der deutschen Tastatur und beim Tippen. Sie wurden für die Konzeption von Vorbereitungsmaterialien aufgegriffen (siehe Depner/Peters in diesem Band). Weitere Angaben in den Umfragen führten dazu, dass kurze Pausen zwischen den einzelnen Aufgaben in jedem Prüfungsteil eingeführt wurden, um die kognitive Belastung zu reduzieren.

Spezifische Fragestellungen wie die kognitiven Anforderungen durch den Einsatz von Videos (Prüfungsteil Hören) bzw. durch integrierte Aufgabentypen (Prüfungsteil Schreiben) wurden mithilfe von Eye-Tracking untersucht. Dazu wurden die Augenbewegungen der Probanden bzw. Probandinnen bei der Bearbeitung der Aufgaben im Hören und Schreiben festgehalten; zusätzlich haben die Teilnehmenden das eigene Vorgehen in einem retrospektiven Interview kommentiert (vgl. Kecker/Zimmermann/Eckes 2022: 400–401; Zimmermann 2020a; 2020b).

# 3 Das Konzept des digitalen TestDaF

Im Folgenden gehen wir darauf ein, zu welchen Skalen des GER-Begleitbands (Europarat 2020) ein Bezug hergestellt werden kann und in welcher Weise die in

diesen Skalen beschriebene Sprachkompetenz im Testformat des digitalen TestDaF berücksichtigt wurde. Des Weiteren wird das Format beschrieben und die Beurteilung in den produktiven Prüfungsteilen sowie die Ergebnisermittlung in allen Prüfungsteilen dargestellt. Abschließend werden Aspekte der benutzerfreundlichen Darstellung am Bildschirm erörtert.

## 3.1 Bezug des digitalen TestDaF zum GER

Im Begleitband des GER (Europarat 2020) wurde unter anderem der Mediation große Bedeutung zugemessen. Mediation im Sinne von Textverarbeitung gehört zu den Schlüsselkompetenzen wissenschaftlichen Arbeitens an der Hochschule und wird in Kombination mit anderen kommunikativen Sprachaktivitäten in rezeptiver oder produktiver Form zu einer wesentlichen Voraussetzung für die sprachliche Bewältigung akademischer Arbeitsprozesse. Anders als im GER von 2001 wird Mediation im Begleitband von 2020 nicht allein als Übertragungsleistung von einer Sprache in die andere angesehen, sondern berücksichtigt stärker die Informationsübermittlung innerhalb einer Sprache zwischen zwei oder mehreren Personen, wie es beispielsweise im akademischen Kontext üblich ist. Auch das Anfertigen von Notizen wird als Teil von Mediation angesehen. In dieser Hinsicht lassen sich verschiedene im digitalen TestDaF abgebildete sprachliche Aktivitäten gut an die Skalen des GER (Europarat 2020) anbinden, die zu der Mediation von Texten gehören: die Rezeption eines schriftlichen bzw. mündlichen Textes, gegebenenfalls 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 im GER zu verorten, können die folgenden Skalen herangezogen werden: „Spezifische Informationen weitergeben“ (Europarat 2020: 116–117), „Daten erklären“ (Europarat 2020: 118–119), „Verarbeitung von Texten“ (Europarat 2020: 120–122), „Notizen anfertigen“ (Europarat 2020: 125–126).

Der TestDaF ist auf den GER-Niveaus B2.1 bis C1.2 angesiedelt. Diese GER-Niveaus sind den TestDaF-Niveaus (TDN) 3, 4 und 5 zugeordnet. Diese Zuordnung ist für den papierbasierten und den digitalen TestDaF identisch (Abb. 2).

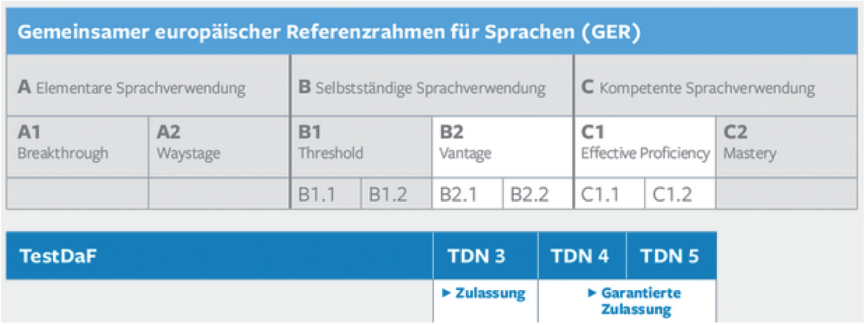


**Abb. 2:** Niveauvergleich TestDaF – GER (TestDaF o. D.)

In der Planungs- und Entwicklungsphase des digitalen TestDaF wurde eine indirekte Zuordnung zum GER vorgenommen, indem der onSET (C-Test mit den GER-Niveaus A2–C1; vgl. Eckes 2010; https://www.onset.de/) bei den Try-Outs und Field-Tests von der Probandengruppe zeitnah absolviert werden musste. Auf diese Weise konnten die neuen Items und Testaufgaben des digitalen TestDaF auf einer gemeinsamen Skala mithilfe der Anker-Items des onSET kalibriert werden. Gleichzeitig wurde bei der Entwicklung der Aufgabenformate eine intuitive Form der Zuordnung durch die Mitarbeitenden des TestDaF-Instituts vorgenommen.

Nach den ersten Testläufen des digitalen TestDaF wurde Ende Juni 2021 eine empirische, direkte Zuordnung der Testaufgaben zum GER mithilfe von Experten-Panels (Standard-Setting) durchgeführt. Insgesamt 32 Personen, die als Testexperten und -expertinnen, Lehrkräfte, TestDaF-Prüfungsbeauftragte sowie als TestDaF-Beurteilende tätig waren, haben in einer mehrtägigen Online-Veranstaltung die Testitems sowie die schriftlichen und mündlichen Leistungen von Teilnehmenden des digitalen TestDaF ausgewählten Skalen des GER zugeordnet.

Für die rezeptiven Fertigkeiten wurde die Basket-Methode (vgl. Council of Europe 2009; Kaftandjieva 2009), für die produktiven Fertigkeiten die Benchmarking-Methode verwendet (vgl. Council of Europe 2009; Kecker 2011). Während der Veranstaltung wurden die Experten und Expertinnen zunächst mit den Skalen des GER vertraut gemacht. Danach erhielten sie eine Einführung in das Konzept des digitalen TestDaF sowie in die Methoden des Standard-Settings. Anschließend konnten sie in einem Training mit Diskussion und Feedback die notwendige Routine in der Zuordnung von Testmaterial und Prüfungsleistungen zum GER erwerben. In einer letzten Phase ordneten die Experten und Expertinnen individuell Testitems und Teilnehmendenleistungen aus der ersten digitalen TestDaF-Prüfung vom Oktober 2020 dem GER zu. Analysen der Interrater-Reliabilität und Multifacetten-Rasch-Analysen (Eckes 2015a) belegten, dass die Experten und Expertinnen in hohem Maße vom Training profitierten, das heißt zufriedenstellende

Konsens- und Konsistenzwerte erreichten, und ihre Einstufungen im Standard-Setting mit den TDN-Einstufungen aus der ersten digitalen TestDaF-Prüfung weitgehend übereinstimmten (für Einzelheiten siehe Kecker/Eckes 2022).

## 3.2 Beschreibung des Testformats

Wie bereits ausgeführt verwendet der digitale TestDaF neben unabhängigen auch integrierte Testaufgaben. Im Prüfungsteil Lesen wird allerdings auf integrierte Aufgaben verzichtet, da die Textlektüre in Kombination mit ihrer Weiterverarbeitung in die anderen drei Prüfungsteile eingebunden ist.

Die Themen stammen aus den Geistes- und Gesellschaftswissenschaften, Naturwissenschaften, Ingenieurwissenschaften und Technik sowie Wirtschaftswissenschaften, Medizin und Humanwissenschaften. Sie sind für die Zielgruppe, die sich für unterschiedliche Studienfächer bewirbt und in den meisten Fällen über kein spezifisches Vorwissen verfügt, verständlich aufbereitet.

**Tab. 1:** Testformat des digitalen TestDaF

| Prüfungsteil | Aufgabentypen/ Items | Davon integrierte Aufgabentypen | Dauer |
|---|---|---|---|
| Lesen | 7 Aufgabentypen<br>34 Items | – | ca. 55 Min. |
| Hören | 7 Aufgabentypen<br>30 Items | 1 | ca. 40 Min. |
| Schreiben | 2 Aufgabentypen | 1 | ca. 60 Min. |
| Sprechen | 7 Aufgabentypen | 2 | ca. 35 Min. |

Die Dauer der Prüfungsteile ist als ungefähre Zeit angegeben, da in der Prüfung zusätzlich Testaufgaben zu bearbeiten sind, die für Erprobungszwecke eingestreut werden. Die Bearbeitungsdauer der Einstreuaufgaben beträgt insgesamt für jeden Teilnehmenden maximal ca. 30 Minuten (vgl. 4.1).

Die folgenden Ausführungen zu den vier Prüfungsteilen orientieren sich an den aktuell erprobten Handbüchern für Autoren und Autorinnen zu den jeweiligen Prüfungsteilen. Wir danken den Kollegen und Kolleginnen Günther Depner, Daniela Marks, Leska Schwarz und Sonja Zimmermann für ihre Unterstützung.

### 3.2.1 Prüfungsteil Lesen

Das Konstrukt der Lesekompetenz wird vom Sprachmodell kommunikativer Kompetenz im GER (vgl. 3.1 zu Mediation) und von kognitiven Prozessen bestimmt, die von hochschulbezogenen Leseabsichten typischerweise hervorgerufen werden. Zu diesen gehören die folgenden: *Informationen identifizieren und verstehen*, *Textverständnis entwickeln*, *von Texten lernen*, *Informationen zusammenführen*. Ausgehend von den Leseabsichten und den zugrundeliegenden Verarbeitungsprozessen wird in der Prüfung evaluiert, ob Studienbewerber und -bewerberinnen die Struktur eines Textes erfassen und die Hauptideen (explizit/implizit) erkennen können. Es wird überprüft, ob zukünftige Studierende bei der Lektüre kausale Bezüge (z. B. Forderung und Argument, Ursache und Folge; vgl. Tabelle 2, Aufgabentyp 5 und 6) erkennen und Zusammenhänge herstellen können. Sie müssen Informationen sowie Inhalte aus verschiedenen Quellen (Texte und Grafik) zu einem Thema abgleichen, zusammenführen und inhaltliche Abweichungen erkennen können (vgl. Tabelle 2, Aufgabentyp 7). Darüber hinaus werden als Voraussetzung für diese weiterführende Lesekompetenz grundlegende linguistische und pragmatische Kenntnisse erfasst, die für eine erfolgreiche Bewältigung der studienbezogenen Kommunikationsaufgaben notwendig sind. Zu diesen Grundlagen gehören rezeptive und produktive Wortschatzkenntnisse, pragmatische Textkompetenzen (z. B. Einstellungen, Haltungen erkennen) sowie ein detailliertes Sprachverstehen (vgl. Tabelle 2, Aufgabentypen 1, 2 und 4).

Die in der Prüfung verwendeten Textausschnitte stammen aus journalistischen Texten zu wissenschaftlichen Themen oder aus wissenschaftlichen Texten, die einführenden Charakter haben bzw. für Nicht-Spezialisten konzipiert wurden, zum Beispiel Wissenschaftsmagazine, Lehrbücher für Studienanfänger und -anfängerinnen und Forschungsberichte. Die Texte weisen eine beschreibende, erklärende oder argumentative Diskursart auf und unterscheiden sich im Grad der Abstraktheit und Informationsdichte.

**Tab. 2:** Prüfungsteil Lesen

| | Aufgabentyp | Kognitive Prozesse | Textsorte und -länge |
|---|---|---|---|
| 1 | Lückentext ergänzen<br>5 Items mit je 4 Optionen | Verstehen von Einzelinformationen, Schlüsselbegriffen und Hauptideen, Aktivierung von Wortschatz | 150–200 Wörter<br>Beschreibende Texte |
| 2 | Textabschnitte ordnen<br>4 Items (4 Übergänge zwischen 5 Textteilen) | Kohärenz/Kohäsion erkennen, Verbindungen von Inhalten über Abschnitte hinweg nachvollziehen | 100–150 Wörter<br>Beschreibende Texte über gesellschaftlich relevante Themen, Kurzberichte |

**Tab. 2:** (fortgesetzt)

| | Aufgabentyp | Kognitive Prozesse | Textsorte und -länge |
|---|---|---|---|
| 3 | Multiple Choice (MC)<br>7 Items mit je 4 Optionen | Verstehen von Details, Hauptideen, impliziten Informationen; Inferieren; Einstellungen/Haltungen verstehen | 600–680 Wörter<br>Forschungs- und Projektberichte |
| 4 | Sprachhandlungen zuordnen<br>4 Items aus 8 Optionen | Sprachhandlungen (auch implizite) im Kontext verstehen | 220–270 Wörter<br>Meinungsäußernde Kommentare |
| 5 | Aussagen Kategorien zuordnen<br>7 Items mit je 4 Optionen | Informationen einordnen, gewichten und verknüpfen; kausale, temporale Zusammenhänge erkennen | 320–370 Wörter<br>Argumentative, erklärende Texte, in denen zwei Konzepte vergleichend vorgestellt werden |
| 6 | Aussagen einem Begriffspaar zuordnen<br>4 Items aus 8 Optionen | Beziehungen zwischen Argumenten herstellen; kausale Zusammenhänge erkennen | 280–320 Wörter<br>Argumentative, erklärende Texte; Beschreibung und Bewertung kausaler Zusammenhänge |
| 7 | Fehler in Zusammenfassung erkennen<br>3 Items | Verstehen von Hauptideen, Verstehen von impliziten Informationen, Verknüpfen von Informationen, Abgleich von Text, Grafik und Zusammenfassung | 200–250 Wörter<br>Erklärende, beschreibende Texte, z. B. Einführungstexte zu einem Thema aus einem Lehrbuch |

### 3.2.2 Prüfungsteil Hören

Die Sprachkompetenz im Hörverstehen wird (mehr als im Leseverstehen) wesentlich von den Kommunikationssituationen in den Lehrveranstaltungen und von digitalen Medien bestimmt, die häufig zur Unterstützung eingesetzt werden. So unterscheidet sich das Hörverstehen in Vorlesungen, Vorträgen oder Seminaren, die mithilfe von Präsentationstools veranschaulicht werden können, von dem Hörverstehen in Sprechstunden, Diskussionen unter Studierenden – etwa in Arbeitsgruppen – oder in Alltagsgesprächen auf dem Campus. Daher werden im digitalen TestDaF möglichst viele dieser Hör- oder Gesprächssituationen abgebildet, um zu überprüfen, ob die Sprachkompetenz der Studienbewerber und -bewerberinnen diesen Anforderungen genügt. Dazu gehört auch, dass beispielsweise in Diskussionen oder Vorlesungen Notizen angefertigt werden müssen, um kausale Verbindungen anzugeben oder Kerninformationen festzuhalten (siehe Tabelle 3, Aufgabentyp 2 und 5).

Eine Besonderheit stellt Aufgabentyp 3 dar, der einzige mit fertigkeitsübergreifendem Format in diesem Prüfungsteil, bei dem die Prüfungsteilnehmenden die schriftliche Zusammenfassung eines gehörten Vortrags lesen und fehlerhafte Informationen darin identifizieren müssen. Damit die Zusammenfassung nicht beim Hören des Vortrags mitgelesen wird, erscheint sie erst im Anschluss an den Hörtext auf dem Bildschirm. Dies bedeutet, dass die Prüfungsteilnehmenden die inhaltlichen Fehler in der Zusammenfassung ausschließlich auf der Grundlage ihres Textverständnisses und/oder ihrer Notizen identifizieren können. Die Teilnehmenden müssen zu diesem Zweck eine mentale Repräsentation des Textes erarbeiten. Im Gegensatz zu den anderen Aufgabentypen wird das Hörverstehen hier nicht durch Items in Form von Fragen oder Aussagen bzw. bei halboffenen Formaten durch Gliederungspunkte für die eigenen Notizen unterstützt (vgl. Field 2012; Song 2012).

In zwei Aufgabentypen (4 und 5) sind Videos eingebunden, die den Kontext (in Aufgabe 4 die in einer Diskussion agierenden Personen) bzw. zusätzlich den Inhalt des Hörtextes veranschaulichen (in Aufgabe 5 durch Präsentationsfolien zum Thema einer Vorlesung).

Die Hörtexte werden im Prüfungsteil Hören in allen sieben Aufgabentypen einmal gehört. Ein bis vier Sprecher bzw. Sprecherinnen kommen in den Hörtexten vor.

In den Aufgabentypen werden verschiedene Hörabsichten operationalisiert, die einer Progression folgen: *Selektives Verstehen, Detailverstehen* und *Globalverstehen* gehören genau wie phonetische Dekodierungsprozesse zu den grundlegenden Anforderungen. *Implizite Bedeutungen verstehen*, *eine mentale Repräsentation eines Textes erarbeiten* oder die *Analyse und Synthese von Informationen aus verschiedenen Quellen* zu den anspruchsvolleren.

Phonetische Dekodierungsprozesse werden im digitalen TestDaF überprüft, obwohl diese zum Hörverstehensprozess allgemein gehören und keine Besonderheit des Hörverstehens im akademischen Kontext darstellen. Da diese Prozesse jedoch die weitere Informationsverarbeitung und die Erschließung der Lexik und Semantik sowie darauf aufbauende Schlussfolgerungen erst ermöglichen und somit wichtige Informationen über die Kompetenz im Hörverstehen liefern, werden sie im Aufgabentyp 7 (siehe Tabelle 3) explizit erfasst.

**Tab. 3:** Prüfungsteil Hören

| | Aufgabentyp | Kognitive Prozesse | Textsorte und -länge |
|---|---|---|---|
| 1 | Kurzantwort:<br>Übersicht ergänzen<br>5 halboffene Items | Selektives Verstehen, Details verstehen, Notizen anfertigen | 400–500 Wörter<br>Halb- oder informelles Gespräch aus dem Studienalltag |
| 2 | Kurzantwort: Textstellen zu Begriffspaar notieren<br>4 halboffene Items | Hauptaussagen erkennen, Verbindung zwischen Teilen des Hörtextes und kausale Zusammenhänge erkennen, Textstellen notieren | 300–400 Wörter<br>Diskussion |
| 3 | Fehler in Zusammenfassung erkennen<br>2 Items | Hauptaussagen erkennen, mentale Repräsentation des Textes erarbeiten, mit Lesetext vergleichen | 400–500 Wörter (Hörtext) für Vortrag/Vorlesung und 100–150 Wörter für die Zusammenfassung |
| 4 | Aussagen Personen zuordnen<br>6 Items mit je 4 Optionen | Hauptideen erfassen,<br>Pragmatik: Einstellung erkennen | 400–550 Wörter<br>Diskussion |
| 5 | Kurzantwort: Gliederungspunkte zu Vortrag ergänzen<br>4 halboffene Items | Hauptaussagen erkennen, gegebenenfalls mit Video abgleichen, Detailverstehen, Notizen anfertigen | 300–450 Wörter<br>Vorlesung/Vortrag |
| 6 | Multiple-Choice<br>5 Items mit je 4 Optionen | Textaufbau nachvollziehen, Detail-/Globalverstehen, Einstellung/Haltung erkennen | 400–500 Wörter<br>Vorlesung/Vortrag/Präsentation |
| 7 | Laut- und Schriftbild abgleichen<br>4 Items | Fehler beim Abgleich von Laut- und Schriftbild erkennen, phonetische Dekodierungsprozesse | 80–120 Wörter<br>Populärwissenschaftlicher Text ohne Fremdwörter |

### 3.2.3 Prüfungsteil Schreiben

Hier geht es darum zu überprüfen, ob Studienbewerber und -bewerberinnen vorgegebene, gesellschaftlich relevante Themen oder Fragestellungen schriftlich bearbeiten und sich dazu zusammenhängend und klar strukturiert äußern können. Dabei nutzen sie unterschiedliches Quellenmaterial und verfolgen wie im Studium unterschiedliche Ziele beim Schreiben: *die Gliederung und logische Struktur der eigenen Texte festlegen*; *eigene Ideen und eine eigene Position adressatengerecht schriftlich formulieren, aber auch fremde Meinungen und Gedanken in eigenen Texten adäquat wiedergeben*; *Informationen aus verschiedenen Quellen (Text, Grafik) für den Text nutzen und dazu Übereinstimmungen sowie Unterschiede erken-*

*nen*; *Informationen aus verschiedenen Quellen im Hinblick auf eine Fragestellung zusammenfassen*. Diese Schreibkompetenz wird in dem Prüfungsteil mithilfe von zwei Schreibaufgaben überprüft, einer isolierten und einer integrierten (vgl. Tab. 4). Die in den beiden Schreibaufgaben zu behandelnden Fragestellungen umfassen Themen aus Bildungs- und Gesellschaftspolitik sowie geistes-, umwelt- und naturwissenschaftliche Themen.

**Tab. 4:** Prüfungsteil Schreiben

| Aufgabentyp | Bearbeitungszeit und erwartete Textlänge | Input | Sprachhandlung |
|---|---|---|---|
| **1. Argumentativen Text schreiben**<br><br>**Isolierte Schreibaufgabe** | 30 Min.,<br>mindestens<br>200 Wörter | – | Vor-/Nachteile bzw. positive/ negative Aspekte eines Themas abwägen, Begründungen geben und Beispiele nennen, dabei gegebenenfalls vorgegebene kurze Statements berücksichtigen und Stellung nehmen |
| **2. Informationen aus Lesetext und Grafik zusammenfassen**<br><br>**Integrierte Schreibaufgabe** | 30 Min.,<br>ca. 100–150 Wörter | Lesetext<br>(ca. 250–300 Wörter)<br>und grafischer Input | Informationen aus Text und Grafik im Hinblick auf eine konkrete Fragestellung zusammenfassen |

Die kommunikative Einbettung der zwei Schreibaufgaben erleichtert den Prüfungsteilnehmenden die Zuordnung zum akademischen Kontext und erhöht die Authentizität. In der isolierten Schreibaufgabe 1 geht es darum, einen argumentativen Text zu einem bildungs- oder gesellschaftspolitischen Thema zu verfassen, der auf einer Lernplattform an der Hochschule eingestellt wird. In die Themenstellung können auch sehr kurze Statements eingebunden werden, die bei der Bearbeitung berücksichtigt werden müssen, dennoch bleibt das Input-Material sehr reduziert, damit Planung und Strukturierung gänzlich als Leistung der Teilnehmenden angesehen werden können.

Die integrierte Schreibaufgabe 2 ist in einem Seminar angesiedelt, für das eine Hausarbeit angefertigt werden muss. Die Aufgabenstellung sieht vor, dass dafür ein Textabschnitt produziert werden soll, der die wichtigsten Informationen aus einem Lesetext und einer Grafik im Hinblick auf das Thema und die Fragestellung zusammenfasst. Lesetext und Grafik können sich inhaltlich zu dem angegebenen Thema ergänzen oder alternativ redundante oder auch gegensätz-

liche Informationen enthalten. Eine solche Verarbeitung in eigenen Worten – im Hinblick auf ein bestimmtes Erkenntnisinteresse – ist ein wichtiger Schritt im Verarbeitungsprozess von Informationen wie er im Hochschulkontext üblich ist. Er erfordert das Verständnis und den Abgleich der beiden Quellen, das Erfassen der kausalen Zusammenhänge und die Fähigkeit, die Inhalte eigenständig zu formulieren.

### 3.2.4 Prüfungsteil Sprechen

Studienanfänger und -anfängerinnen müssen ihre Sprachkompetenz im Sprechen und im mündlichen Ausdruck in einer Vielzahl von Kommunikationssituationen an der Hochschule unter Beweis stellen. Diese Situationen variieren nach Sprechhandlung (z. B. Vorteile/Nachteile abwägen, widersprechen, Kritik formulieren) und anderen situativen Merkmalen wie zum Beispiel Anzahl, Status und Rolle der Gesprächsteilnehmer und -teilnehmerinnen. Des Weiteren spielen für eine erfolgreiche Kommunikation Kompetenzen eine Rolle, die über einzelne Sprechhandlungen hinausgehen: Studierende müssen *schriftliche oder mündliche Beiträge anderer adäquat mündlich zusammenfassen*, *den eigenen Standpunkt dazu formulieren und sachliche Argumentation von persönlicher Meinung unterscheiden können*. Sie müssen *Informationen aus verschiedenen Quellen schnell erfassen*, *gegebenenfalls Abweichungen erkennen und versprachlichen*. Diese Facetten der Sprechkompetenz abzubilden, erfordert eine größere Anzahl von Aufgabentypen, damit eine möglichst große Bandbreite der genannten Kompetenzen erfasst werden kann. Im Prüfungsteil Sprechen des digitalen TestDaF wird die Sprechkompetenz daher mithilfe von sieben Aufgabentypen überprüft, von denen jeder eine andere Sprechhandlung fokussiert und situative Merkmale berücksichtigt, wie zum Beispiel die Zahl der Gesprächsteilnehmenden (Einzelperson vs. Seminargruppe), die Rollen und Beziehungen der Gesprächsteilnehmenden zum/zur Prüfungsteilnehmenden (Freunde bzw. Freundinnen, Diskussionsteilnehmende; vertraut, unbekannt), die Themen (Studienalltag, Inhalte aus Vorlesungen) sowie das Setting (Lehrräume, Lernorte, Service-Einrichtungen der Hochschule). Die Aufgaben beinhalten überwiegend Themen aus Geistes- und Gesellschaftswissenschaften, Wirtschaftswissenschaften, Humanwissenschaften/Medizin oder Themen aus Hochschule und Forschung allgemein.

Von den sieben Aufgabentypen sind fünf isolierte, die das Sprechen möglichst als einzige Fertigkeit erfassen, und zwei integrierte, die zusätzlich eine andere Fertigkeit berücksichtigen (siehe Tabelle 5). In den isolierten Aufgabentypen wird das Input-Material eher kurz gehalten; in den zwei integrierten Aufgabentypen werden dagegen zwei längere Texte verwendet: ein schriftlicher Text, der mündlich zusammengefasst werden muss (Aufgabentyp 3), und ein längerer

mündlicher Beitrag eines/einer Seminarteilnehmenden, zu dem mündlich Stellung zu nehmen ist (Aufgabentyp 6). Die integrierten Aufgabentypen erhöhen die Authentizität der Testaufgaben, da sie durch die Einbindung von einem schriftlichen oder mündlichen Textinput kognitive Anforderungen abbilden, die im wissenschaftlichen Diskurs häufig vorkommen.

**Tab. 5:** Prüfungsteil Sprechen

| | Aufgabentyp | Sprechzeit | Input-Material | Sprechhandlung |
|---|---|---|---|---|
| 1 | Rat geben | Isoliert (0:45 Min.) | – | Rat/Tipps geben |
| 2 | Optionen abwägen | Isoliert (1:30 Min.) | – | Vorteile/Nachteile bzw. positive/negative Folgen abwägen, begründen |
| 3 | Text zusammenfassen | Integriert (2:00 Min.) Fertigkeit Lesen | Kurzer Lesetext (250–300 Wörter) | Zusammenfassen |
| 4 | Informationen abgleichen, Stellung nehmen | Isoliert (1:30 Min.) | Grafik-Input + kurzer Diskussionsbeitrag | Informationen aus unterschiedlichen Quellen erfassen, Stellung nehmen |
| 5 | Thema präsentieren | Isoliert (2:30 Min.) | Präsentationsfolie/ Übersicht/Handout | Sachverhalt beschreiben und präsentieren |
| 6 | Argumente wiedergeben, Stellung nehmen | Integriert (2:00 Min.) Fertigkeit Hören | Längerer Seminarbeitrag eines/einer Studierenden | Argumente wiedergeben, Stellung nehmen, begründen |
| 7 | Maßnahmen kritisieren | Isoliert (1:30 Min.) | Programm der Veranstaltung oder Aushang der Hochschulverwaltung | Kritik äußern und begründen, gegebenenfalls Alternativvorschlag machen |

Die Sprachkompetenz im Prüfungsteil Sprechen wird auch im digitalen TestDaF im semidirekten Format als adaptierte Form des *Simulated Oral Proficiency Interviews* (SOPI) überprüft (vgl. Kenyon 2000; Kecker 2011). Das bereits im papierbasierten TestDaF eingesetzte Format hat sich bewährt und wurde hinsichtlich der Durchführung für den digitalen TestDaF insofern verändert, als die Testaufgaben nicht mehr in einem Testheft, sondern auf dem Bildschirm abgebildet werden. Darüber hinaus müssen Testzentren und Teilnehmende keine zusätzliche Software für die Aufnahme ihrer Äußerungen bedienen, denn Antworten werden automatisch aufgenommen und gespeichert. Somit können

sich Teilnehmende ausschließlich auf die Bearbeitung der Aufgaben konzentrieren.

## 3.3 Leistungsbewertung und Ergebnisermittlung

### 3.3.1 Leistungsbewertung

Die Antworten und Leistungen der Prüfungsteilnehmenden werden in den vier Prüfungsteilen separat ausgewertet. Im Lesen und Hören wird pro Prüfungsteil mithilfe der hinterlegten Lösungsschlüssel automatisiert anhand der richtig gelösten Items ein Gesamtpunktwert ermittelt (1 Punkt pro richtige Antwort). Im Prüfungsteil Hören werden für die drei Aufgabentypen mit Kurzantworten (Aufgabentyp 1, 2 und 5) zuvor geschulte und zertifizierte Beurteiler und Beurteilerinnen eingesetzt, die die Antworten nach Vorgaben als richtig oder falsch codieren. Dabei werden vor Beginn der Bewertungsphase eindeutig richtige oder falsche Antworten aus dem Pool der zu bewertenden Antworten herausgefiltert und automatisch bewertet. Derzeit wird geprüft, inwieweit die Bewertung durch Beurteiler und Beurteilerinnen noch weitergehend automatisiert unterstützt werden kann. Erste Untersuchungen dazu wurden in Kooperation mit dem *Language Technology Lab* der Universität Duisburg/Essen (Prof. Torsten Zesch, Dr. Andrea Horbach, Dr. Ronja Laarmann-Quante) durchgeführt.

In den Prüfungsteilen Schreiben und Sprechen werden die in der Datenbank gespeicherten schriftlichen Texte oder mündlichen Äußerungen der Teilnehmenden ebenfalls von geschulten und zertifizierten Beurteilenden ausgewertet. Die Leistungen zu jeder einzelnen Aufgabe werden in beiden Prüfungsteilen auf sechsstufigen holistischen Beurteilungsskalen mit 0 bis 5 Punkten bewertet. In beiden Prüfungsteilen werden die für jede Aufgabe vergebenen Punkte zu einem Gesamtpunktwert pro Prüfungsteil addiert.

Die Beurteilungsskalen wurden für diesen Zweck im Laufe der Entwicklungsphase des digitalen TestDaF in Zusammenarbeit mit Experten und Expertinnen aus der Sprachtestforschung, erfahrenen TestDaF-Beurteilenden und DaF-Lehrkräften neu konzipiert und in den Field-Tests erprobt (Zimmermann/Marks 2018). In den Skalen wurden die spezifischen Merkmale der jeweiligen Aufgabenstellung im Schreiben und Sprechen berücksichtigt. Dies gilt insbesondere für die Beurteilung der integrierten oder kompetenzübergreifenden Aufgabentypen, bei denen der Textinput in den Beurteilungskriterien beachtet werden muss (vgl. Kecker/Zimmermann/Eckes 2022: 402–403).

### 3.3.2 Ergebnisermittlung

Bei der Ermittlung der Testergebnisse gilt es zunächst zu berücksichtigen, dass die in den vier Prüfungsteilen jeweils erreichbaren Punktzahlen (die beobachteten Werte) abhängig von der Anzahl der Items bzw. Aufgaben variieren. In den rezeptiven Prüfungsteilen bestimmen sich die Punktzahlen als Anzahl der korrekt beantworteten Items; im Lesen sind dies maximal 34 Punkte, im Hören maximal 30 Punkte. In den produktiven Prüfungsteilen ergeben sich die Punktzahlen als Summe der Einzelbewertungen; im Schreiben sind maximal 10 Punkte, im Sprechen maximal 30 Punkte erreichbar. Hinzu kommt, dass auch bei einem Höchstmaß an Standardisierung der Aufgabenerstellung zwei Testsätze eines Prüfungsteils unterschiedlich schwer sein können. Ist zum Beispiel im Lesen ein Testsatz A etwas leichter als ein Testsatz B, so würde (unter sonst gleichen Bedingungen) eine Punktzahl von 21 im Testsatz A einer geringeren Leistung entsprechen als dieselbe Punktzahl in Testsatz B. Punktzahlen sind demnach stets an einen bestimmten Testsatz gebunden und nicht direkt zwischen Testsätzen desselben Prüfungsteils vergleichbar.

Um format- und schwierigkeitsbedingte Unterschiede zwischen Prüfungsteilen bzw. Testsätzen auszugleichen, werden beim digitalen TestDaF die erreichten Punktzahlen je Prüfungsteil in Werte einer einheitlichen 20-Punkte-Skala überführt (Kolen/Brennan 2014). Die Werte auf dieser Skala, die so genannten skalierten Werte, sind das Ergebnis psychometrischer Analysen auf der Grundlage von Rasch-Modellen (Bond/Yan/Heene 2020; Eckes 2015a). Im Lesen und Hören kommt das dichotome Rasch-Modell, im Schreiben und Sprechen das Multifacetten-Rasch-Modell zum Einsatz. Die Analysen liefern für alle Teilnehmenden Schätzungen ihres Sprachstands in den vier Teilkompetenzen, ausgedrückt in Einheiten der Logitskala. Die Logitwerte werden schließlich in geeigneter Weise linear transformiert, sodass in jedem Prüfungsteil skalierte Werte zwischen 0 und 20 resultieren. Der Schwierigkeitsausgleich zwischen Testsätzen eines Prüfungsteils erfolgt nach gängigen Verfahren der Rasch-basierten Testangleichung (*test equating*).

Als Ergebnis von Skalierung und Testangleichung sind die skalierten Werte unabhängig vom Prüfungsteil und von dessen Schwierigkeit gültig. Die skalierten Werte besitzen im Unterschied zu den beobachteten Werten die Eigenschaft der Vergleichbarkeit; sie sind in gleicher Weise hinsichtlich des jeweils erreichten Sprachstands interpretierbar.

Neben den TestDaF-Niveaus weist das TestDaF-Zertifikat für jeden Prüfungsteil die skalierten Werte aus. Weiter ist ein Gesamtwert zwischen 0 und 80 aufgeführt, welcher sich als Summe der vier skalierten Werte berechnet. Anhand dieser Informationen können zulassende Stellen zum Beispiel erkennen, ob Prü-

fungsteilnehmende ein TestDaF-Niveau nur knapp erreicht oder das nächsthöhere Niveau nur knapp verfehlt haben. Sie können darüber hinaus Zulassungsentscheidungen vom Erreichen einer Mindestanzahl von skalierten Werten (Mindestgesamtwert) abhängig machen (Eckes/Althaus 2020).

### 3.3.3 Monitoring und Qualitätssicherung der Beurteilung

Die Qualitätssicherung in der Beurteilung von Teilnehmendenleistungen im TestDaF basiert auf drei Komponenten: Der Qualifizierung und Zertifizierung der aktiven Beurteilenden in umfassenden Schulungsmaßnahmen, der Verlinkung der an der Beurteilung in einem Prüfungsteil Beteiligten und dem regelmäßigen Monitoring ihrer Beurteilungsleistung.

Alle für die Beurteilung im TestDaF zugelassenen Personen müssen zuvor für den ausgewählten Prüfungsteil (Hören, Schreiben oder Sprechen) ein eintägiges Schulungsprogramm absolvieren und zusätzlich ihre Kompetenz durch eine Probebeurteilung unter Beweis stellen. Sofern sie diese Qualifizierungsmaßnahmen erfolgreich bewältigt haben, werden sie vom TestDaF-Institut für einen Zeitraum von zwei Jahren zertifiziert und müssen nach Ablauf dieser Zeit erneut an einem Training zur Erneuerung ihres Zertifikats teilnehmen. Voraussetzung für die Beurteilung von TestDaF-Leistungen sind zudem ein abgeschlossenes einschlägiges Hochschulstudium und Unterrichtserfahrung.

Die Beurteilung der Prüfungsleistungen werden in den Prüfungsteilen Hören, Schreiben und Sprechen den an einem Prüfungsereignis beteiligten Beurteilenden online über ein Portal zur Verfügung gestellt. Die Verteilung der Prüfungsleistungen erfolgt in allen drei Prüfungsteilen geordnet nach Aufgabentypen und nicht nach Teilnehmenden. Dies bedeutet, dass mehrere Beurteilende an der Beurteilung der Gesamtleistung eines/einer Teilnehmenden in einem Prüfungsteil beteiligt sind. Im Prüfungsteil Hören sind dafür drei Aufgabentypen (Aufgabentyp 1, 2 und 5) zu bewerten, im Schreiben zwei und im Sprechen sechs (die erste Aufgabe wird zum Warming-up eingesetzt und nicht bewertet). Hinzu kommen gegebenenfalls die Einstreuaufgaben (siehe Abschnitt 4.1). Alle Beurteilenden erhalten in ihrem Portal automatisiert zusammengestellte virtuelle Beurteilungspakete, in denen Leistungen zu allen Aufgaben des von ihnen bearbeiteten Prüfungsteils enthalten sind.

In den Prüfungsteilen Schreiben und Sprechen werden die Beurteilenden in dem von ihnen bearbeiteten Prüfungsteil durch die Verteilung der Leistungen nach den Testaufgaben und durch den Einsatz mehrerer Beurteilender für die Gesamtleistung eines Teilnehmenden in diesem Prüfungsteil miteinander verlinkt. Diese Verteilung hat den Vorteil, dass Beurteilungseffekte (Halo- und Kontext-

effekte) minimiert werden, die durch die festgelegte Reihenfolge der Aufgaben entstehen können. Ferner kann die individuelle Beurteilungsqualität durch eingestreute, bereits beurteilte Leistungen regelmäßig überprüft werden. Durch die systematische Verlinkung von Beurteilenden und Teilnehmenden kann die Beurteilungsqualität aller an den Beurteilungen beteiligten Personen in einem gemeinsamen Bezugsrahmen untersucht werden.

Im Prüfungsteil Hören werden für die Leistungen bei den Aufgabentypen 1, 2 und 5 je unterschiedliche Bewertende eingesetzt. Damit wird die Gesamtleistung eines Teilnehmenden auch im Hören immer von mehr als einer Person bewertet. Die Qualität der Bewertungen wird durch Stichproben von Mitarbeitern und Mitarbeiterinnen des TestDaF-Instituts überprüft.

## 3.4 Benutzeroberfläche und Interaktionskonzept

Die Benutzeroberfläche des digitalen TestDaF, das sogenannte *Graphical User Interface* (GUI), wurde in Kooperation mit dem Studiengang Informationsdesign, Fachrichtung Interaktionsdesign (Prof. Ralph Tille) der Hochschule der Medien (HdM) in Stuttgart entwickelt. Dabei wurde ein Drei-Phasen-Modell zugrunde gelegt (Fulcher 2003), das neben einer Planungs- und Designphase auch erste Erprobungen mit der Zielgruppe sowie weitere Pilotierungen zur Feinabstimmung vorsieht. Maßgebliches Ziel der Entwicklung war es, eine Benutzeroberfläche zu gestalten, die eine klare, eindeutige Benutzerführung beinhaltet, einen möglichst zurückhaltenden Rahmen für die dargestellten Testaufgaben bietet und eine Fokussierung auf die Aufgabenbearbeitung unterstützt. Das Design der Testaufgaben wurde diesem Ziel untergeordnet. Die Verständlichkeit und Handhabbarkeit der neuen Benutzeroberfläche wurde zunächst mit einer kleinen Probandengruppe im *User Experience Laboratory* der HdM mithilfe einer Laut-Denken-Studie getestet. Die vorläufige Endversion wurde in zahlreichen Erprobungen überprüft und, sofern notwendig, optimiert. Grundlage dafür bildeten Angaben, die Teilnehmende nach den Erprobungen in einem Fragebogen machten.

# 4 Durchführung und Testumgebung des digitalen TestDaF

## 4.1 Durchführung der Prüfung

Der digitale TestDaF wird in den Testzentren von g.a.s.t. unter Aufsicht online am Computer durchgeführt. Die Testzentren stellen dafür die notwendige Ausstattung bereit, die aus PCs oder Laptops mit einer Mindestbildschirmgröße von 15 Zoll und Headsets besteht. Zudem muss eine stabile Internetverbindung vorhanden sein. Für die Durchführung der Prüfung ist der Download eines sicheren Browsers, des sogenannten *Safe Exam Browsers* (SEB), auf allen verwendeten Geräten erforderlich. Durch die Verwendung des SEB läuft die Prüfung im sogenannten Kiosk-Modus, das heißt, während der Prüfung können keine weiteren Software-Anwendungen benutzt werden außer der Software für die Prüfungsdurchführung. Auf diese Weise wird die Grundlage für die notwendige Testsicherheit während der Prüfung geschaffen.

Als Erstes bearbeiten Teilnehmende in der Prüfung die Aufgaben im Lesen, dann folgen nach entsprechenden Pausen Hören, Schreiben und Sprechen. Kurzantworten im Prüfungsteil Hören sowie Texte im Prüfungsteil Schreiben werden auf der Tastatur getippt. Das Sprechen wird in einer adaptierten Version des SOPI (vgl. 3.2.4) ebenfalls am Computer durchgeführt. Während der Prüfung bearbeiten Teilnehmende zusätzlich zu der in jedem Prüfungsteil vorgesehenen Anzahl von Aufgabentypen weitere Aufgaben, die erprobt werden (sogenannte Einstreuaufgaben). Diese Aufgaben werden auf verschiedene Prüfungsteile aufgeteilt und dürfen insgesamt eine Bearbeitungszeit von 30 Minuten nicht überschreiten. Die zusätzliche Zeit ist für alle Teilnehmenden gleich, jedoch unterscheiden sich die eingestreuten Aufgaben und Prüfungsteile. Sofern im Prüfungsteil Schreiben eine Einstreuaufgabe eingesetzt wird, können in den anderen Prüfungsteilen keine weiteren Aufgaben erprobt werden, da damit die Zeitspanne von 30 Minuten ausgeschöpft wird. Die Ergebnisse der Teilnehmenden in den eingestreuten Aufgaben werden bei der Ergebnisermittlung nicht berücksichtigt; sie werden für psychometrische Analysen und zur Qualitätssicherung der Prüfungsaufgaben verwendet.

Das für die Prüfungsdurchführung verantwortliche Aufsichtspersonal lässt die angemeldeten Prüfungsteilnehmenden, die die Identitätskontrollen im Testzentrum ohne Beanstandungen absolviert haben, in ihrem Online-Portal zu, gibt den Beginn der Prüfung für alle frei und verfolgt den weiteren Ablauf am Bildschirm und im Prüfungsraum. Am Ende der Prüfung wird mittels einer entsprechenden Anzeige die erfolgreiche Speicherung der Prüfungsantworten aller Teil-

nehmenden kontrolliert. Sofern durch technische Defekte an einem Gerät die Bearbeitung der Prüfungsaufgaben unterbrochen wird, werden die bis zu diesem Zeitpunkt gegebenen Antworten gespeichert und die Prüfung nach einem Re-Login an der entsprechenden Stelle fortgesetzt.

## 4.2 Testumgebung des digitalen TestDaF

Für den digitalen TestDaF steht eine Plattform mit einer Vielzahl an Funktionen bereit. Mit dieser Plattform kann der gesamte Prüfungsprozess digital abgebildet werden. Aus Sicht der Teilnehmenden betrifft dies zum Beispiel die Prüfungsanmeldung und -bezahlung sowie die Vorbereitung mithilfe einer interaktiven Demo-Version der Prüfung und einem Einführungsvideo, das die Bildschirmoberfläche und die Funktionsbuttons erklärt. Des Weiteren erhalten die Teilnehmenden technische Informationen zur Prüfungsdurchführung und können nach der Prüfung ihre Ergebnisse abrufen, ihr Prüfungszertifikat digital speichern, ausdrucken oder gegebenenfalls das Ergebnis beim Prüfungsausschuss reklamieren. Alle genannten Funktionen stehen den Teilnehmenden nach ihrer Prüfungsanmeldung in einem Online-Portal zur Verfügung.

Die Testzentren können ihrerseits über ein eigenes Online-Portal die Prüfungstermine verwalten, alle für die Identitätskontrolle und die Prüfungsdurchführung notwendigen Funktionen und Handreichungen nutzen sowie Vorbereitungsmaterialien herunterladen oder Werbematerialien bestellen. Die Kommunikation mit Teilnehmenden und Testzentren wird ebenfalls über Portale für diese Zielgruppen abgewickelt.

Die Anwendungen der g.a.s.t.-Plattform ermöglichen darüber hinaus den digitalen Zugriff und die digitale Erfassung aller für die interne Prüfungsorganisation erforderlichen Daten. Die darin abgebildeten Arbeitsprozesse und Funktionen beinhalten die Erfassung und Verwaltung der Prüfungsaufgaben in einer Itembank, die automatisierte Zusammenstellung der Testversionen für Prüfungen sowie die Online-Beurteilung der Leistungen aus den produktiven Prüfungsteilen und die Ergebnisermittlung. Diese Funktionen erlauben eine stärkere Flexibilisierung der Prüfungstermine für den Bedarf der Testzentren und eine wesentlich verkürzte Auswertungszeit und damit raschere Ergebnisbekanntgabe. Die Zusammenführung der Daten aus den Portalen der Teilnehmenden und Testzentren erleichtert zudem eine fristgerechte Bereitstellung der Prüfungsaufgaben für alle Teilnehmenden einschließlich eines Nachteilsausgleichs für Teilnehmende mit Beeinträchtigung sowie die Einbindung einer ausreichenden Anzahl von Honorarkräften für die Beurteilung nach einem Prüfungslauf. Auch Verwaltungsabläufe wie Ergebniseinsprüche oder die Abrechnung von Prüfungsentgelten und Ho-

noraren für Beurteilungen werden über die Testumgebung der Plattform organisiert.

## 5 Forschungsperspektiven

Die Einführung des digitalen TestDaF eröffnet eine Fülle an Perspektiven für die künftige Forschung. Diese Perspektiven betreffen einerseits Fragen, die bei neuen oder grundlegend revidierten Sprachtests zur Absicherung ihrer Reliabilität, Validität und Fairness stets zu untersuchen sind, unabhängig davon, ob es sich um digitale oder papierbasierte Formate handelt. Andererseits versteht es sich von selbst, dass digitale Sprachtests eine ganze Reihe von Merkmalen aufweisen, die sie nicht mit papierbasierten Testformaten teilen. Diese formatspezifischen Merkmale werfen ganz eigene, zum Teil neuartige Fragen auf, die entsprechend innovative Untersuchungsansätze erfordern. Im Folgenden greifen wir aus beiden Kategorien einige Fragestellungen heraus, die als Anregungen für Forschungsarbeiten der nächsten Jahre aufgefasst werden können.

In den beiden rezeptiven Prüfungsteilen Lesen und Hören kommen ausschließlich Aufgaben zum Einsatz, die aus zwei Teilen bestehen: aus einem Aufgabenstamm, Stimulus oder Input (z. B. einem kurzen Text, einer Grafik) und einer Reihe von Fragen oder Items (z. B. Multiple-Choice-Items, Zuordnungsitems), die auf den Input Bezug nehmen. Es handelt sich um jeweils sieben Aufgaben mit 34 Items (Lesen) bzw. 30 Items (Hören).

Diese weit verbreiteten Aufgabenformate werden *Aufgabenbündel* (Rosenbaum 1988; Wilson/Adams 1995) oder auch *Testlets* (Wainer/Kiely 1987) genannt. Testlets bieten eine Reihe von Vorteilen:

a) effiziente Aufgaben- bzw. Testentwicklung (ein und derselbe Input ermöglicht die Formulierung mehrerer Items),
b) zeitsparende Testdurchführung (ein konstanter Input bedeutet weniger Bearbeitungszeit pro Item) und
c) gezielte Erfassung höherer kognitiver Fähigkeiten (die Aufgabenstellung lässt sich in einen größeren, realitätsnahen Kontext einbetten).

Mit der Verwendung von Testlets können aber auch unerwünschte Effekte verbunden sein. So könnten zum Beispiel einzelne Teilnehmende trotz hoher Sprachkompetenz aufgrund fehlenden Vorwissens besondere Schwierigkeiten mit einzelnen Sätzen oder Ausdrücken einer fremdsprachlichen Textpassage haben. Die Folge wäre eine Minderung der Validität bzw. Fairness der Interpretation ihrer Testergebnisse. Für einzelne Aufgaben des papierbasierten TestDaF konnte Eckes

(2014; 2015b) Testlet-Effekte nachweisen. Es bleibt zu untersuchen, ob dies auch für den digitalen TestDaF gilt.

Im Prüfungsteil Hören des digitalen TestDaF kommen unter anderem Kurzantwort-Aufgaben zum Einsatz. Speziell geschulte Beurteilende bewerten die Antworten als richtig oder falsch anhand einer Liste von zutreffenden Antworten. Eine Frage, die schon Gegenstand erster Untersuchungen war (Laarmann-Quante/Schwarz 2021), lautet: Lassen sich die Beurteilenden durch technologiegestützte Identifikation gleicher Antworten oder auch durch die automatische Bündelung ähnlicher Antworten bei ihrer Arbeit unterstützen, ohne dass die Qualität der Bewertungen gemindert wird? Im positiven Fall könnten eine kognitive Entlastung der Beurteilenden und eine deutliche Zeitersparnis des gesamten Beurteilungsprozesses resultieren. Eine andere, formatunabhängige Frage zielt auf mögliche Beurteilungseffekte, insbesondere Unterschiede zwischen den Beurteilenden hinsichtlich ihrer Strenge oder Milde bei der Bewertung der Antworten auf Kurzantwort-Aufgaben. Liegen Einstufungen der Antworten als richtig oder falsch durch eine Gruppe von Experten und Expertinnen als eine Art Referenz vor, kann zusätzlich die Genauigkeit der Bewertungen untersucht werden (vgl. Eckes 2020).

Die Untersuchung von Beurteilungseffekten gehört zum Standardrepertoire der psychometrischen Analyse von anerkannten Sprachprüfungen, insbesondere in den produktiven Prüfungsteilen (Eckes 2015a; Engelhard/Wind 2018). Häufig untersuchte Effekte betreffen Unterschiede zwischen den Beurteilenden in ihrer Strenge oder Milde, in ihrer Tendenz, die mittleren Kategorien einer Ratingskala bevorzugt zu verwenden (Zentraltendenz), oder in ihrer Abhängigkeit von kontextspezifischen Merkmalen der Beurteilung (Halo-Effekte; differenzielle Beurteilungsfunktionen). Mit geeigneten psychometrischen Modellen lassen sich diese Effekte quantifizieren und gegebenenfalls in ihren Auswirkungen auf die Qualität der Beurteilungen kontrollieren (Eckes 2005; Eckes/Jin 2022; Jin/Eckes 2021). Erste Untersuchungen mit Erprobungsdaten des digitalen TestDaF haben wie erwartet gezeigt, dass Beurteilende auch dann, wenn sie Leistungen von Teilnehmenden online bewerten, Strenge-Effekten und Zentraltendenzen unterliegen (Eckes/Jin 2021). Es hat sich zudem herausgestellt, dass ein neu entwickeltes Multifacetten-Rasch-Modell (Jin/Wang 2018) diese unerwünschten Tendenzen zu korrigieren imstande ist. Weitere Untersuchungen mit Daten aus Echtprüfungen müssen folgen.

Die Identifikation unterschiedlicher Arten von Beurteilungseffekten erlaubt eine detaillierte Rückmeldung an Beurteilende. Schulungen, aber auch die Zertifizierung von Beurteilenden, lassen sich damit auf eine sichere Basis stellen. Künftige Untersuchungen sollten verschiedene Formen der Schulung (z. B. online oder Präsenz, individuell oder in Gruppen) in ihren Auswirkungen auf die Qualität der Beurteilungen, insbesondere im Hinblick auf die Minderung der oben

beschriebenen Beurteilungseffekte betrachten (Knoch/Read/von Randow 2007; Raczynski et al. 2015). Die beim digitalen TestDaF online durchgeführte Beurteilung von Teilnehmendenleistungen ermöglicht zudem eine fortlaufende Qualitätskontrolle. Dazu werden von Experten und Expertinnen vorbewertete Leistungen eingestreut und Abweichungen von diesen Referenzbewertungen registriert. Bei Häufung größerer Abweichungen können die betroffenen Beurteilenden nachgeschult oder gegebenenfalls aus der Gruppe der aktiven Beurteilenden ausgeschlossen werden. Es sollte untersucht werden, wie sich dieser Prozess des Online-Monitorings gestalten lässt, damit das Ziel einer kontinuierlich hohen Qualität der Beurteilungen bei gleichzeitig hoher Akzeptanz aller Beteiligten erreicht wird. Hierfür kommen auch qualitative Methoden, etwa in Form von strukturierten Interviews, in Betracht.

Wie zuvor ausgeführt, fand im Juni 2021 ein erstes Standard-Setting zum digitalen TestDaF statt (Abschnitt 3.1). Bedingt durch die Corona-Pandemie wurde es komplett online durchgeführt. Mit wenigen Ausnahmen lieferten die psychometrische Analysen Belege dafür, dass es die beim digitalen TestDaF verwendeten Aufgaben erlauben, Leistungen auf den Niveaus TDN 3 bis TDN 5 abzubilden und den GER-Niveaus B2 bis C1 zuzuordnen. Allerdings stützte sich dieses Standard-Setting auf Leistungsdaten und Kennwerte aus der ersten digitalen TestDaF-Prüfung im Oktober 2020 mit einer relativ geringen Zahl von Teilnehmenden. Es stellt sich daher die Frage, inwieweit sich die Ergebnisse bei Verfügbarkeit einer größeren Datenbasis (wichtig für die präzise Kalibrierung der Items und die Auswahl geeigneter Leistungsbeispiele) und unter anderen Bedingungen (Präsenzveranstaltung) bestätigen lassen. Sobald ausreichend viele Kurse zur Vorbereitung auf den digitalen TestDaF mit einer großen Zahl an Teilnehmenden stattfinden, könnte neben den bislang verwendeten testzentrierten Methoden ergänzend ein personenzentrierter Ansatz wie die Prototypgruppenmethode (Eckes 2017), die sich beim Spracheinstufungstest onSET bewährt hat, zum Einsatz kommen.

Eine ausreichend große Datenbasis vorausgesetzt, lassen sich quantitative Analysen zur faktoriellen Struktur des digitalen TestDaF durchführen, um seine Konstruktvalidität abzusichern. Mittels konfirmatorischer Faktorenanalysen (Dunn/McCray 2020; Eckes 2022; Sawaki/Stricker/Oranje 2009) sind Fragen wie die folgenden zu beantworten:

a) Wie gut lassen sich die vier Teilkompetenzen Lesen, Hören, Schreiben und Sprechen voneinander abgrenzen?
b) Bilden diese Teilkompetenzen gleichgewichtige Komponenten einer übergeordneten Sprachkompetenz?
c) Tragen die integrierten Aufgaben primär zur Messung der jeweils intendierten Zielkompetenz bei (z. B. Lesen bei der integrierten Schreibaufgabe zum Schreiben) oder gibt es relevante Beziehungen zu Kompetenzen, die in den

jeweils anderen Prüfungsteilen angezielt werden (z. B. Lesen bei der integrierten Schreibaufgabe zum Lesen)?

# 6 Schlussbemerkung

Im Jahr 2018 erschien in der renommierten Fachzeitschrift *Language Testing* eine ausführliche Rezension zum TestDaF in seiner papierbasierten Form (Norris/Drackert 2018). Es war in dieser Zeitschrift die erste Rezension eines Sprachtests, der nicht auf die Erfassung englischer Sprachkenntnisse zielt. In ihrem Fazit beschrieben John Norris und Anastasia Drackert den TestDaF mit Blick auf sein hohes Maß an Standardisierung, die umfangreichen Arbeiten zur Sicherung seiner psychometrischen Qualität und die eingehenden Analysen zum Nachweis einer reliablen und fairen Leistungsmessung als das „‚go-to' assessment" zur Ermittlung hochschulbezogener Sprachkompetenz in Deutsch als Fremdsprache.

Norris und Drackert (2018) verwiesen aber auch auf einige kritische Punkte oder Schwachstellen, die sich in drei Fragen bündeln lassen:

a) Wie gut repräsentieren die im (papierbasierten) TestDaF verwendeten Aufgabentypen die sprachlichen Anforderungen, die Hochschulen in Deutschland heutzutage an internationale Studierende stellen?
b) Wie eng ist die Relation der TestDaF-Niveaus zu den Skalen des GER (Europarat 2001)?
c) Was sind die empirischen Belege dafür, dass der TestDaF den primär intendierten Zweck erfüllt, also den Nachweis liefert, dass die für die Aufnahme eines Studiums in Deutschland erforderliche Sprachkompetenz in ausreichendem Maße vorliegt?

Der digitale TestDaF erlaubt Antworten auf jede dieser Fragen. Im vorliegenden Beitrag haben wir aufgezeigt, dass die neu konzipierten Aufgabentypen dazu dienen, die aktuellen kommunikativen Anforderungen an Hochschulen besser abzubilden. Hierzu zählen digitale, mediengestützte Sprachhandlungen und die Verwendung von integrierten Aufgaben. Im Hinblick auf die zweite Frage hat ein Standard-Setting erste Erkenntnisse darüber geliefert, dass sich verschiedene, im digitalen TestDaF abgebildete sprachliche Aktivitäten gut auf die im Begleitband des GER (Europarat 2020) erschienenen Skalen zur Mediation von Texten beziehen lassen. Zweifelsohne sind zur komplexen Frage des GER-Bezugs weitere Studien erforderlich. Diese Studien werden sich in einigen Jahren auf eine deutlich stabilere Datenbasis stützen können. Auch sollten dabei andere Verfahren des Standard-Settings angewendet werden, um eine höhere Generalisierbarkeit der Ergebnisse zu gewährleisten. Was die dritte Frage betrifft, so ermöglicht der digitale TestDaF

dank seiner gegenüber dem papierbasierten Format signifikant verbesserten Abbildung der über die Jahre veränderten Sprachanforderungen eine detaillierte, aussagekräftige Untersuchung seiner prognostischen Validität hinsichtlich des Studienerfolgs.

Im Kern betreffen die drei oben genannten Fragen die Validität der Schlussfolgerungen aus den Ergebnissen (den skalierten Werten), die Teilnehmende erzielen, wenn sie den digitalen TestDaF ablegen. Folgt man dem argumentbasierten Ansatz der Testvalidierung (Kane 2013), dann besteht ein zentraler Aspekt der Validität darin, von der Höhe der skalierten Werte der Teilnehmenden auf ihre Chancen, ein Studium an einer deutschen Hochschule erfolgreich zu bewältigen, zu schließen. Hierbei handelt es sich nach Kane um eine Extrapolation, das heißt um einen Schluss von der Leistung im Test auf eine Leistung in einer Nicht-Testsituation, also den Erfolg oder Misserfolg im Hochschulstudium. Die Annahme lautet, dass Teilnehmende mit hohen skalierten Werten, die sich in TestDaF-Niveaus von mindestens 4 übersetzen, gut darauf vorbereitet sind, die sprachlichen Anforderungen ihres Studiums zu erfüllen, und damit Chancen haben, zufriedenstellende akademische Leistungen zu erbringen (Eckes/Althaus 2020). Besser denn je ist der TestDaF in seiner nunmehr vorliegenden digitalen Version geeignet, der Gültigkeit dieser Annahme nachzugehen.

## Literatur

Arras, Ulrike (2012): „Im Rahmen eines Hochschulstudiums in Deutschland erforderliche sprachliche Kompetenzen – Ergebnisse einer empirischen Bedarfsanalyse". In: Tinnefeld, Thomas (Hrsg.): *Hochschulischer Fremdsprachenunterricht: Anforderungen – Ausrichtung – Spezifik*. Online: https://hochschulfremdsprachenunterricht.blogspot.com/search/label/41%20Arras (24.11.2021).

Bachman, Lyle; Palmer, Adrian (2010): *Language assessment in practice*. Oxford, UK: Oxford University Press.

Bond, Trevor G.; Yan, Zi; Heene, Moritz (2020): *Applying the Rasch model: Fundamental measurement in the human sciences*. 4th ed. New York: Routledge.

Chan, Sathena; Bax, Stephen; Weir, Cyril (2018): „Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test". In: *Assessing Writing* 36, 32–48. DOI: https://doi.org/10.1016/j.asw.2018.03.008.

Council of Europe (2009): *Manual for relating language examinations to the Common European Framework of Reference*. Strasbourg: Language Policy Division.

Dadey, Nathan; Lyons, Susan; DePascale, Charles (2018): „The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice". In: *Applied Measurement in Education* 31(1), 30–50. DOI: https://doi.org/10.1080/08957347.2017.1391262.

Depner, Günther; Peters, Anja (2022): „Sprachkompetenzen entwickeln und trainieren: Ein Konzept für eine kompetenzorientierte Prüfungsvorbereitung". In: *Informationen Deutsch als Fremdsprache* 49 (4), 325–345.

Dunn, Karen J.; McCray, Gareth (2020): „The place of the bifactor model in confirmatory factor analysis investigations into construct dimensionality in language testing". In: *Frontiers in Psychology* 11:1357. DOI: https://doi.org/ 10.3389/fpsyg.2020.01357.

Eckes, Thomas (2005): „Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis". In: *Language Assessment Quarterly* 2(3), 197–221. DOI: https://doi.org/10.1207/s15434311laq0203_2.

Eckes, Thomas (2010): „Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung". In: Grotjahn, Rüdiger (Hrsg.): *C-Test: Beiträge aus der aktuellen Forschung/The C-test: Contributions from current research*. Frankfurt am Main: Peter Lang, 125–192.

Eckes, Thomas (2014): „Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach". In: *Language Testing* 31(1), 39–61. DOI: https://doi.org/10.1177/0265532213492969.

Eckes, Thomas (2015a): *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* 2nd ed. Frankfurt am Main: Peter Lang. DOI: https://doi: 10.3726/978-3-653-04844-5.

Eckes, Thomas (2015b): „Lokale Abhängigkeit von Items im TestDaF-Leseverstehen: Eine Testlet-Response-Analyse". In: *Diagnostica* 61(2), 93–106. DOI: https://doi.org/10.1026/0012-1924/a000118.

Eckes, Thomas (2017): „Setting cut scores on an EFL placement test using the prototype group method: A receiver operating characteristic (ROC) analysis". In: *Language Testing* 34(3), 383–411. DOI: https://doi.org/10.1177/0265532216672703.

Eckes, Thomas (2020): „Rater-mediated listening assessment: A facets modeling approach to the analysis of raters' severity and accuracy when scoring responses to short-answer questions". In: *Psychological Test and Assessment Modeling* 62(4), 449–471. Online: https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2020/PTAM-4-2020_ebook_2_eckes.pdf (26.04.2022).

Eckes, Thomas (2022): „Exploratorische und konfirmatorische Faktorenanalysen". In: Caspari, Daniela; Klippel, Friederike; Legutke, Michael K.; Schramm, Karen (Hrsg.): *Forschungsmethoden in der Fremdsprachendidaktik: Ein Handbuch*. 2. Aufl. Tübingen: Narr Francke Attempto, 378–392.

Eckes, Thomas; Althaus, Hans-Joachim (2020): „Language proficiency assessments in higher education admissions". In: Oliveri, Maria E.; Wendler, Cathy (Hrsg.): *Higher education admissions practices: An international perspective.* Cambridge, UK: Cambridge University Press, 256–275. DOI: https://doi.org/10.1017/9781108559607.

Eckes, Thomas; Jin, Kuan-Yu (2021): „Examining severity and centrality effects in TestDaF writing and speaking assessments: An extended Bayesian many-facet Rasch analysis". In: *International Journal of Testing* 21(3–4), 131–153. DOI: https://doi.org/10.1080/15305058.2021.1963260.

Eckes, Thomas; Jin, Kuan-Yu (2022): „Detecting illusory halo effects in rater-mediated assessment: A mixture Rasch facets modeling approach". In: *Psychological Test and Assessment Modeling* 64(1), 87–111. Online: https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam_2022-1/PTAM__1-2022_5_kor.pdf (26.04.2022).

Engelhard, George; Wind, Stefanie A. (2018): *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. New York: Routledge. DOI: https://doi.org/10.4324/9781315766829.

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

Europarat (2020): *Gemeinsamer europäischer Referenzrahmen für Sprachen – Begleitband.* Stuttgart: Klett.

Field, John (2012): „The cognitive validity of the lecture-based question in the IELTS listening paper". In: Taylor, Linda; Weir, Cyril (Hrsg.): *IELTS collected papers 2: Research in reading and listening assessment.* Cambridge: Cambridge University Press, 391–453.

Fulcher, Glenn (2003): „Interface design in computer-based language testing". In: *Language Testing* 20(4), 384–408. DOI: https://doi.org/10.1191/0265532203lt265oa.

Jin, Kuan-Yu; Eckes, Thomas (2021): „Detecting differential rater functioning in severity and centrality: The dual DRF facets model". In: *Educational and Psychological Measurement.* DOI: https://doi.org/10.1177/00131644211043207.

Jin, Kuan-Yu; Wang, Wen-Chung (2018): „A new facets model for rater's centrality/extremity response style". In: *Journal of Educational Measurement* 55(4), 543–563. DOI: https://doi.org/10.1111/jedm.12191.

Kaftandjieva, Feljanka (2009): „Basket procedure: The breadbasket or the basket case of standard setting methods?". In: Figueras, Neus; Noijons, José (Hrsg.): *Linking to the CEFR levels: Research perspectives*. Arnhem: Cito/EALTA, 21–34.

Kane, Michael T. (2013): „Validating the interpretations and uses of test scores". In: *Journal of Educational Measurement* 50(1), 1–73. DOI: https://doi.org/10.1111/jedm.12000.

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

Kecker, Gabriele; Eckes, Thomas (2022): *Standard-Setting zum digitalen TestDaF: Ein Online-Verfahren für eine Online-Prüfung*. Online: https://www.testdaf.de/fileadmin/testdaf/downloads/Publikationen/Standard-Setting-2021.pdf (27.04.2022).

Kecker, Gabriele; Zimmermann, Sonja; Eckes, Thomas (2022): „Der Weg zum digitalen TestDaF: Konzeption, Entwicklung und Validierung". In: Gretsch, Petra; Wulff, Nadja (Hrsg.): *Deutsch als Zweit- und Fremdsprache in Schule und Beruf*. Paderborn: Brill Schöningh, 393–410.

Kenyon, Dorry (2000): „Tape-mediated oral proficiency testing: Considerations in developing Simulated Oral Proficiency Interviews (SOPIs)". In: Bolton, Sibylle (Hrsg.): *TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. Beiträge aus einem Expertenseminar*. Köln: VUB Gilde, 87–106.

Knoch, Ute; Read, John; von Randow, Janet (2007): „Re-training writing raters online: How does it compare with face-to-face training?". In: *Assessing Writing* 12(1), 26–43. DOI: https://doi.org/10.1016/j.asw.2007.04.001.

Knoch, Ute; Sitajalabhorn, Woranon (2013): „A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes". In: *Assessing Writing* 18(4), 300–308. DOI: https://doi.org/10.1016/j.asw.2013.09.003.

Kolen, Michael J.; Brennan, Robert L. (2014): *Test equating, scaling, and linking: Methods and practices*. New York: Springer.

Laarmann-Quante, Ronja; Schwarz, Leska (2021): *‚Meet me at the ribary' – Analyzing and categorizing spelling variants in free-text answers to listening comprehension prompts*. Unveröffentlichtes Manuskript.

Lane, Suzanne; Raymond, Mark R.; Haladyna, Thomas M. (Hrsg.) (2016): *Handbook of test development*. 2nd ed. New York: Routledge.

Marks, Daniela (2015): „Prüfen sprachlicher Kompetenzen internationaler Studienanfänger an deutschen Hochschulen – Was leistet der TestDaF?". In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 20(1), 21–39.

Norris, John; Drackert, Anastasia (2018): „Test review: TestDaF". In: *Language Testing* 35(1), 149–157. DOI: https://doi.org/10.1177/0265532217715848.

Ockey, Gary J. (2007): „Construct implications of including still image or video in computer-based listening tests". In: *Language Testing* 24(4), 517–537. DOI: https://doi.org/10.1177/0265532207080771.

Plakans, Lia (2013): „Assessment of integrated skills". In: Chapelle, Carol (Hrsg.): *The Encyclopedia of Applied Linguistics*. Malden, MA: Wiley-Blackwell, 205–212.

Raczynski, Kevin R.; Cohen, Allan S.; Engelhard, George; Lu, Zhenqiu (2015): „Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment". In: *Journal of Educational Measurement* 52(3), 301–318. DOI: https://doi.org/10.1111/jedm.12079.

Rosenbaum, Paul R. (1988): „Item bundles". In: *Psychometrika* 53(3), 349–359. DOI: https://doi.org/10.1007/BF02294217.

Sawaki, Yasuyo; Stricker, Lawrence J.; Oranje, Andreas H. (2009): „Factor structure of the TOEFL internet-based test". In: *Language Testing* 26(1), 5–30. DOI: https://doi.org/10.1177/0265532208097335.

Song, Min-Young (2012): „Note-taking quality and performance on an L2 academic listening test". In: *Language Testing* 29(1), 67–89. DOI https://doi.org/10.1177/0265532211415379.

TestDaF (o.D.): *Das müssen Sie können. Anforderungen im TestDaF*. Online: https://www.testdaf.de/de/teilnehmende/warum-testdaf/das-muessen-sie-koennen/ (27.04.2022).

Wainer, Howard; Kiely, Gerard L. (1987): „Item clusters and computerized adaptive testing: A case for testlets". In: *Journal of Educational Measurement* 24(3), 185–201. DOI: https://doi.org/10.1111/j.1745–3984.1987.tb00274.x.

Wilson, Mark; Adams, Raymond J. (1995): „Rasch models for item bundles". In: *Psychometrika* 60(2), 181–198. DOI: https://doi.org/10.1007/BF02301412.

Zimmermann, Sonja (2020a): „„Das ist doch Leseverstehen!" – Eine empirische Untersuchung zum Konstrukt von integrierten Schreibaufgaben". In: Drackert, Anastasia; Mainzer-Murrenhof, Mirka; Soltyska, Anna; Timukova, Anna (Hrsg.): *Testen bildungssprachlicher Kompetenzen und akademischer Sprachkompetenzen. Zugänge für Schule und Hochschule*. Frankfurt am Main: Peter Lang, 187–213.

Zimmermann, Sonja (2020b): „„Das ist Zeitverlust für mich, den Text wieder lesen" – Einblicke in das Schreiben von Zusammenfassungen in der Fremdsprache Deutsch". In: *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 25(2), 111–133.

Zimmermann, Sonja; Marks, Daniela (2018): „Integrierte Prüfungsleistungen mit dem GER beurteilen: Was die Kann-Beschreibungen nicht können". In: Brandt, Aniko; Buschmann-Göbels, Astrid; Harsch, Claudia (Hrsg.): *Der Gemeinsame Europäische Referenzrahmen für Sprachen und seine Adaption im Hochschulkontext*. Bochum: AKS, 158–166.

# Biographische Angaben

**Gabriele Kecker**

war bis Ende 2021 stellvertretende Institutsleiterin des TestDaF-Instituts in Bochum und leitete die Abteilung Testentwicklung und Qualitätssicherung. Sie wurde 2011 an der Ruhr-Universität Bochum promoviert und hat ihre Dissertation zum Thema Validierung von Sprachprüfungen verfasst. Ihr Forschungsinteresse gilt Validierungsmodellen für Sprachprüfungen und Fragen der Niveau-Zuordnung zu Bezugssystemen wie dem GER sowie Methoden des Standard-Settings. Arbeitsschwerpunkte waren neben der Entwicklung und Validierung des digitalen TestDaF die Beratung von Institutionen zu Fragen der Testentwicklung. Gabriele Kecker verfügt über langjährige Erfahrung in der Fortbildung von Lehrkräften im In- und Ausland zu Themen der Leistungsmessung und zum GER.

**Thomas Eckes**

war bis Ende 2021 stellvertretender Institutsleiter des TestDaF-Instituts in Bochum und leitete die Abteilung Psychometrie und Sprachtestforschung. Promotion (1981) und Habilitation (1989) erfolgten an der Universität des Saarlandes, Saarbrücken. Sein Forschungsinteresse gilt der psychometrischen Modellierung sprachlicher Kompetenzen, der Weiterentwicklung von Multifacetten-Rasch-Modellen und der Analyse von Beurteilereffekten in der Leistungsmessung. Er war Mitglied der Editorial Boards der Zeitschriften *Assessing Writing* (2015 bis 2020) und *Language Testing* (2017 bis 2021). Thomas Eckes erhielt 2019 den erstmals vergebenen „Language Testing Reviewer of the Year Award“ und 2021 den „British Council International Assessment Award“.