Beitrag Themenheft „Testen und Prüfen“

Alice Friedland* und Milica Sabo*

Schwierigkeit und Trennschärfe von Aufgabentypen – evaluiert durch Ober- und Untergruppenanalysen in den DSH-Prüfungsteilen Leseverstehen und Hörverstehen

Quality criteria difficulty and discriminatory power within listening and reading tasks – evaluated through the so-called upper and lower group analysis for the DSH (German Language Examination for University Admission)

https://doi.org/10.1515/infodaf-2022-0066

Zusammenfassung: Die DSH-Prüfung (Deutsche Sprachprüfung für den Hochschulzugang) nimmt innerhalb der Prüfungen, welche in der Rahmenordnung für Deutsche Sprachprüfungen für den deutschen Hochschulzugang (kurz: RO-DT) gelistet sind, eine besondere Rolle ein. Sie stellt neben dem Prüfungsteil Deutsch der Feststellungsprüfung eines Studienkollegs die einzige Prüfung dar, die nicht grundsätzlich standardisiert ist und dementsprechend dezentral erstellt wird bzw. werden kann. Aus diesen Gründen ist eine ständige standortbezogene Evaluation der Prüfung notwendig. Am Beispiel des DSH-Standortes Friedrich-Schiller-Universität Jena soll aufgezeigt werden, wie die sogenannte Ober- und Untergruppenanalyse bzw. die vereinfachte Itemanalyse die Evaluation hinsichtlich der Gütekriterien *Schwierigkeit* und *Trennschärfe* unterstützen kann. Mithilfe der Ober- und Untergruppenanalyse ist auch eine Bestimmung des Schwierigkeitsgrads für jede Aufgabe möglich. Anhand von vier DSH-Prüfungssätzen soll dies für die Prüfungsteile der rezeptiven Fertigkeiten Leseverstehen und Hörver-

***Kontaktpersonen: Alice Friedland,** E-Mail: alice.friedland@uni-jena.de
Milica Sabo, E-Mail: milica.sabo@uni-jena.de

stehen aufgezeigt werden. Die zur Überprüfung der Fertigkeiten herangezogenen Aufgabentypen unterscheiden sich nicht wesentlich und sind deshalb für eine übergreifende Betrachtung der rezeptiven Prüfungsteile geeignet. Aus der Ober- und Untergruppenanalyse für die Prüfungsteile Lese- und Hörverstehen ergibt sich zum einen ein Vorschlag zur Testevaluation, welcher auch an anderen DSH-Standorten eingesetzt werden kann; zum anderen ergeben sich Hinweise für das Testerstellungsverfahren.

Schlüsselwörter: Gütekriterien in Sprachtests, Ober- und Untergruppenanalyse, rezeptive Fertigkeiten

Abstract: The DSH examination (German Language Examination for University Admission) plays a special role within the examinations listed in the Framework Regulations for German Language Examinations for German University Admission (RO-DT for short). Apart from the German examination part of the assessment test of a preparatory college, it is the only examination that is not standardized in principle and is or can therefore be created decentrally. For these reasons, a constant site-related evaluation of the examination is necessary. Using the example of the DSH site Friedrich-Schiller-University Jena it will be shown how the so-called upper and lower group analysis or the simplified item analysis can support the evaluation with regard to the quality criteria difficulty and discriminatory power. With the help of the upper and lower group analysis, it is also possible to determine the level of difficulty for each task. On the basis of four DSH examination sets, this will be shown for the examination parts of the receptive skills reading comprehension and listening comprehension. The evaluation of these skills does not differ significantly in the selection of tasks types and is therefore suitable for an overarching consideration of the receptive examination parts. The upper and lower group analysis for the reading and listening comprehension parts of the exam results in a proposal for test evaluation, which can also be used at other DSH sites, as well as hints for the test preparation procedure.

Keywords: quality criteria in language test examination, upper and lower group analysis, listening and reading comprehension

1 Einleitung

Mit der DSH-Prüfung (Deutsche Sprachprüfung für den Hochschulzugang) können potenzielle internationale Studierende ihre Sprachkenntnisse im Deutschen für den Hochschulzugang nachweisen (HRK und KMK 2019). Als eine dezentrale Prüfung wird sie an verschiedenen Standorten deutschlandweit erstellt, durch-

geführt und evaluiert. Diese Verfahren bilden sich im Testentwicklungszyklus ab (Eberharter/Kremmel/Zehentner 2018: 57–58). Mit der Bezeichnung Zyklus wird darauf hingewiesen, dass die Testerstellung kein linearer Vorgang ist, sondern jede Testerstellungsphase in die nächste eingreift und auf die vorherige zurückgeführt werden kann. Jede erstellte DSH-Prüfung sollte dementsprechend pilotiert und evaluiert werden.

Die vorgeschriebenen Grundsätze zum DSH-Testerstellungsverfahren sind durch das DSH-Handbuch (FaDaF 2012) festgelegt. In diesem sind sowohl die Vorgaben, die das DSH-Testkonstrukt abbilden, als auch die inhaltliche und organisatorische Ebene für die Konzipierung vorzufinden. Das DSH-Testkonstrukt prüft nach sprachlichen Kompetenzen auf dem Niveau B2/C1 nach dem GER. Die zu erreichenden Stufen sind als DSH 3, DSH 2 und DSH 1 ausgewiesen, wobei das Ergebnis DSH 1 die sogenannte Eingangsstufe darstellt und somit für den Hochschulzugang (an den meisten Hochschulstandorten) nicht ausreicht.

Der Nachweis der sprachlichen Studierfähigkeit wird nach dem DSH-Konstrukt in rezeptive und produktive Prüfungsteile aufgeteilt. Zu den produktiven Prüfungsteilen gehören das Verfassen eines argumentativen Sachtextes zu einem wissenschaftlichen Thema mit Alltagsbezug und eine mündliche monologische Vorstellung einer kurz dargestellten wissenschaftlichen Studie mit anschließendem dialogischem Gespräch. Zum Lese- und Hörverstehen als rezeptive Prüfungsteile gehört außerdem die Überprüfung der wissenschaftssprachlichen Strukturen, die auf der Umformulierung von wissenschaftssprachlichen Sätzen beruht und somit eher das explizite Wissen der sprachlichen Strukturen testet. Zur Überprüfung des Leseverstehens dienen ein komplexer wissenschaftlicher Text, der 6.000 Zeichen nicht übersteigen darf, und ein Aufgabenblatt zur Überprüfung des globalen, selektiven und detaillierten Leseverstehens. Zur Überprüfung des Hörverstehens hören die Prüfungsteilnehmenden einen wissenschaftlichen Vortrag, der in schriftlicher Form 7.000 Zeichen nicht übersteigen darf. Zudem erhalten sie ein Aufgabenblatt, welches das globale, selektive und detaillierte Hörverstehen dieses Vortrags überprüft (FaDaF 2012).

Um die Qualitätssicherung zu gewährleisten und insbesondere den Gütekriterien einer zumindest internen Validität und Reliabilität gerecht zu werden, werden im Erstellungsprozess (und nach der Durchführung) Prüfungssätze evaluiert. Hierzu eignen sich die Pilotierung eines Prüfungssatzes vor der Durchführung sowie eine Evaluation der pilotierten Prüfungssätze. Diese werden dann in einem Evaluationsverfahren weiterentwickelt und womöglich kalibriert.

Evaluationsverfahren kann man in qualitative und quantitative Verfahren aufteilen. Somit wäre eine Pilotierung mit geringer Teilnehmendenzahl und anschließendem Gespräch über die Prüfungsaufgaben eine qualitative Einschätzung der Schwierigkeit der einzelnen Aufgaben. Quantitative Evaluations-

verfahren erfordern größere Teilnehmendenzahlen und eine genaue Berechnung des Schwierigkeitsgrades p (p = probability) sowie des Trennschärfeindex D (D = discrimination). Eines der Verfahren zur Qualitätssicherung hinsichtlich der Trennschärfe und des Schwierigkeitsgrades in den Aufgabentypen stellen Burghoff und Leder (2011: 18–22) dar. In ihrer Handreichung zur DSH am Beispiel des Leseverstehens stellen sie eine Itemanalyse sowie eine Ober- und Untergruppenanalyse nach Schelten vor (1997: 132–136). Sie beschreiben, wie die Analyse für die Evaluation der DSH-Prüfung an der Freien Universität Berlin angepasst ist. In dem Registrierungsverfahren der DSH-Standorte spielt die Frage zur Ober- und Untergruppenanalyse zudem eine Rolle hinsichtlich der Qualitätssicherung.

Das Ziel dieses Beitrags ist zum einen, die Verfahren zur Ober- und Untergruppenanalyse darzustellen, und zum anderen, durch diese Analyse die Auswertung der Aufgaben hinsichtlich des Schwierigkeitsgrades und der Trennschärfe abzubilden. Der Schwierigkeitsgrad und die Trennschärfe werden außerdem in Bezug auf die Operatoren in Aufgabenformulierungen betrachtet. Erörtert wird daher in diesem Beitrag der testtheoretische Hintergrund zur Ober- und Untergruppenanalyse. Im Anschluss an die Darstellung der Ober- und Untergruppenanalyse wird in den nächsten Kapiteln eine Analyse der vier Prüfungssätze zu den Prüfungsteilen Leseverstehen und Hörverstehen gegeben. Außer der Darstellung dieses Evaluationsverfahrens werden in diesem Beitrag Hinweise und mögliche Verbesserungsvorschläge bezüglich der Aufgabenkonzipierung zu den rezeptiven Prüfungsteilen genannt.

2 Testtheoretische Einbettung der Ober- und Untergruppenanalyse

Validität, Reliabilität und Objektivität gelten als die drei Hauptgütekriterien im Testerstellungsverfahren und in der Testevaluation. Das Kriterium der Objektivität, das hauptsächlich der Durchführung und der Bewertung sowie der Darstellung der Aufgabenlösungen und der Punktevergabe gewidmet ist, wird nur am Rande betrachtet (Vigh 2018: 100). Die Reliabilität und die Validität eines Tests werden unter anderem durch weitere Kriterien beeinflusst – in diesem Falle den Schwierigkeitsgrad und die Trennschärfe in den Aufgaben – und stehen dementsprechend im Zentrum der Betrachtungen.

Die Ermittlung des Schwierigkeitsgrades und der Trennschärfe ist testtheoretisch in der deskriptiven Statistik angesiedelt. Dabei geht es hauptsächlich um eine Itemanalyse bzw. um eine Analyse der Itemschwierigkeiten und die Trennschärfeanalyse der Items (Pospeschill 2010: 72). Schelten (1997) wiederum passt

diese Testverfahren an die pädagogischen Bedürfnisse an und schlägt die einfache Itemanalyse (auch Ober- und Untergruppenanalyse) für die Anwendung in verschiedenen Testsituationen vor (Schelten 1997: 128–137).

Mit der Ober- und Untergruppenanalyse[1] (teils auch vereinfachte Itemanalyse genannt) können unter anderem der Schwierigkeitsgrad und die Trennschärfe[2] der einzelnen Aufgaben bzw. Items bestimmt und evaluiert werden (Schelten 1997: 128). Die Gütekriterien Schwierigkeit und Trennschärfe der einzelnen Testitems beeinflussen wiederum die Validität (insbesondere die Konstruktvalidität) sowie die Reliabilität des Gesamttests, das heißt, sie bestimmen direkt (neben anderen weiterführenden Gütekriterien) mit, ob ein Test tatsächlich das misst, was er vorgibt zu messen, und ob es sich bei diesem Test um ein zuverlässiges Messinstrument handelt (Dlaska/Krekeler 2009; Vigh 2018).

Der Schwierigkeitsgrad p jeder Aufgabe gibt an, von wie vielen Testteilnehmenden eine Aufgabe richtig bearbeitet wurde. Um einen ausgewogenen Test zu gestalten, gilt es, alle zu schweren und zu leichten Aufgaben – also Aufgaben, die nahezu von keinem Testteilnehmenden bzw. von fast allen Testteilnehmenden gelöst wurden – zunächst zu identifizieren und anschließend herauszufiltern bzw. durch andere, ausgewogenere Aufgabenformate zu ersetzen (Schelten 1997; Pospeschill 2010). Die finale Entscheidung, ob eine gegebenenfalls zu schwere oder zu leichte Aufgabe abgeändert wird, sollte jedoch nicht allein durch die Bestimmung des Schwierigkeitsgrads getroffen werden. Um eine fundiertere Entscheidung zu treffen, sollte zudem die Trennschärfe betrachtet werden. Diese bestimmt, inwieweit eine Aufgabe in der Lage ist, zwischen stärkeren und schwächeren Testteilnehmenden zu differenzieren (ebd.).

Mit Blick auf das Gesamttestkonstrukt der DSH-Prüfung ergibt sich, dass aufgrund einer komplexen Schreibaufgabe im Teil Textproduktion eine Ober- und Untergruppenanalyse lediglich für die rezeptiven Prüfungsteile Hörverstehen, Leseverstehen sowie wissenschaftssprachliche Strukturen zu empfehlen ist. Dementsprechend wird die Durchführung einer Ober- und Untergruppenanalyse als Teil der Qualitätssicherung im Fragebogen zur Re-Registrierung eines DSH-Standortes nur für die Fertigkeiten Leseverstehen und Hörverstehen durch den FaDaF erhoben (Koithan et al. o. J.: 2).

---

1 Wir orientieren uns in diesem Artikel am Begriff von Burghoff und Leder (2011), welche das Beiheft zum DSH-Handbuch für den Themenbereich Leseverstehen erstellt haben, und nutzen gleichsam den Terminus, der auch in der Checkliste für die Begutachtung und Erstellung eines DSH-Prüfungssatzes des FaDaF e.V. Verwendung findet.

2 Außerdem kann eine Distraktorenanalyse durchgeführt werden, auf die jedoch im Rahmen dieses Artikels nicht näher eingegangen wird, da Multiple-Choice-Aufgaben nur in geringem Umfang in den DSH-Prüfungsteilen zu den rezeptiven Fertigkeiten am Standort FSU Jena vorkommen.

3 Ablauf des Evaluationsverfahrens mithilfe der Ober- und Untergruppenanalyse am DSH-Standort Jena

Wie bereits einleitend beschrieben, ist es insbesondere an kleineren[3] DSH-Standorten teilweise nur bedingt möglich, direkt nach der Pilotierung bereits eine Ober- und Untergruppenanalyse durchzuführen. Dies begründet sich vor allem mit dem Aspekt, dass wenige Probandinnen und Probanden an der Pilotierung teilnehmen können und somit eine Ober- und Untergruppenanalyse zu wenig aussagekräftigen Ergebnissen kommen würde. Dieser Umstand entspricht nicht vollständig den Qualitätsvorgaben der Association of Language Testers in Europe (ALTE 2012: 39), ist jedoch teilweise auch in der dezentralen Organisation der DSH-Prüfung begründet und somit den geringen personellen Ressourcen (im Vergleich zu anderen Tests dieser Art) geschuldet. Um trotzdem bereits vor dem Einsatz des Prüfungssatzes eine Einschätzung hinsichtlich des Schwierigkeitsgrades und der Trennschärfe zu erhalten, wird eine qualitative Evaluation des Testsatzes im Rahmen einer Pilotierung durchgeführt und nach dem Testdurchlauf jede Aufgabe detailliert mit den Probandinnen bzw. Probanden besprochen (Weir 2005: 234). Anschließend werden die gewonnenen Erkenntnisse in die Überarbeitung des Prüfungssatzes sowie in die Weiterentwicklung eines Erwartungshorizontes aufgenommen.

Nach dem Einsatz wird der Prüfungssatz einer quantitativen Analyse (Ober- und Untergruppenanalyse) – insbesondere für die rezeptiven Prüfungsteile Leseverstehen und Hörverstehen – unterzogen. Zunächst werden alle Gesamtergebnisse der schriftlichen Prüfungsteile (Leseverstehen und wissenschaftssprachliche Strukturen, Hörverstehen sowie Textproduktion) absteigend sortiert und anschließend in zwei gleich große Gruppen – eine leistungsstärkere Gruppe und eine leistungsschwächere Gruppe – aufgeteilt. Burghoff und Leder (2011) empfehlen, die gesamte Menge der Testteilnehmenden in die Analyse miteinzubeziehen. Andere Autorinnen und Autoren (Schelten 1997; Pospeschill 2010) geben in ihren Darlegungen an, dass nur jeweils die oberen sowie die unteren 27 Prozent der Gesamtgruppe betrachtet werden sollten. Der DSH-Standort Jena folgt der Empfehlung von Burghoff und Leder (2011) und teilt entsprechend alle Testteilnehmenden eines Prüfungsdurchlaufs in eine Unter- und eine Obergruppe (bei ungerader Anzahl der Teilnehmenden wird die Person in der absoluten Mitte he-

3 Hier sind auch Standorte mitgedacht, die wenige Mitarbeiterinnen und Mitarbeiter für die DSH-Prüfung beschäftigen.

rausgelassen). Diese Entscheidung begründet sich unter anderem in der teilweise geringeren Teilnehmendenzahl zu einigen Testterminen (beispielsweise auch während der COVID-19-Pandemie). Für den Testtermin im Juli 2021 (vgl. Tab. 1) konnten jeweils 23 Personen der Ober- und Untergruppe zugeordnet werden. Ein Vorgehen nach Schelten (1997) und Pospeschill (2010) hätte lediglich zwölf Personen für die Ober- und Untergruppe zugelassen. Folglich konnten mehr Prüfungsdaten (nämlich im Beispiel Juli 2021 23 Personen je Gruppe) in die Analyse einbezogen werden und somit die Repräsentativität der Analyseergebnisse leicht erhöht werden.

Nach der Bestimmung der Ober- und Untergruppe wird für jede Aufgabe eine Tabelle erstellt, wie das Beispiel (vgl. Tab. 1 und Tab. 2) zum Prüfungsteil Hörverstehen aus Juli 2021 zeigt. Die Aufgabenformulierung lautete hierbei „Nennen Sie drei Funktionen von …". Für jede richtig genannte Funktion erhielten die Testteilnehmenden acht Punkte, für jedes halbe Item vier Punkte. In den Tabellen 1 und 2 kann jeweils in der oberen Zeile jede mögliche Punkteverteilung und in der unteren Zeile die Anzahl der Testteilnehmenden, die diese Punktzahl in der Aufgabe erreichten, abgelesen werden. Tabelle 1 zeigt zunächst die Ergebnisübersicht für die definierte Obergruppe, Tabelle 2 entsprechend die erreichten Punkte für die Untergruppe.

Tabelle 1: Übersicht Obergruppe (n = 23), HV-Aufgabe 7, Juli 2021

24	20	16	12	8	4	0	Ø = 17,56
8	3	9	-	1	-	2	0, 7317

Tabelle 2: Übersicht Untergruppe (n = 23), HV-Aufgabe 7, Juli 2021

24	20	16	12	8	4	0	Ø = 9,39
2	-	5	4	5	-	7	0, 3913

Der letzten Spalte der Tabellen kann zunächst der Durchschnitt der erreichten Punkte für die entsprechende Gruppe entnommen werden. Schelten (1997: 132) empfiehlt, für die Ermittlung des Schwierigkeitsgrades p den Anteil der Personen zu bestimmen, welche eine Testaufgabe richtig gelöst haben. Da in den Aufgaben der DSH-Prüfung jedoch eine reine Unterscheidung zwischen den beiden Polen *richtig gelöst* und *falsch gelöst* nicht möglich ist, erfolgt ein rechnerischer Umweg zu diesem Quotienten. Ausgehend von der durchschnittlich erreichten Punktzahl wird ein Mittelwert gebildet, der der Quotient aus der durchschnittlichen Punkt-

zahl und der maximal zu erreichenden Punktzahl einer jeden Aufgabe ist und somit p_{Hoch} und $p_{Niedrig}$ ersetzt. Aus den ermittelten Werten für p_{Hoch} und $p_{Niedrig}$ lässt sich nun der Schwierigkeitsgrad p wie folgt bestimmen (Schelten 1997: 132):

$$p = \frac{p_{Hoch} + p_{Niedrig}}{2}$$

Für die angegebene Beispielaufgabe ergibt sich nun ein Schwierigkeitsgrad p von 0,5615, wie in der folgenden Rechnung dargestellt:

$$0,5615 = \frac{0,7317 + 0,3913}{2}$$

Diese Aufgabe liegt somit im mittleren Schwierigkeitsbereich, da der Schwierigkeitsgrad p eine Spannweite von 0 (sehr schwer) und 1 (sehr leicht) abdeckt. ALTE (2012: 96) empfiehlt, nur Aufgaben innerhalb einer Spannweite von 0,25 und 0,8 in weitere statistische Auswertungen miteinzubeziehen. Außerdem sprechen Ergebnisse außerhalb dieses Spektrums wahrscheinlich dafür, dass die Aufgabe für die getestete Gruppe nicht geeignet war.

Nach der Ermittlung des Schwierigkeitsgrads für die Ober- und die Untergruppe, also p_{Hoch} und $p_{Niedrig}$, kann nun auch der Trennschärfeindex D wie folgt bestimmt werden (Schelten 1997: 132):

$$D = p_{Hoch} - p_{Niedrig}$$

Danach ergibt sich für die Beispielaufgabe ein Trennschärfeindex von 0,3404, wie der folgenden Rechnung zu entnehmen ist:

$$0,3404 = 0,7317 - 0,3913$$

Nach der Interpretation des Trennschärfeindex, die von Schelten (1997: 135) vorgeschlagen wird, kann diese Aufgabe als „brauchbare Testaufgabe“ gewertet werden, da D zwischen 0,30 und 0,39 liegt. Ab einem Trennschärfeindex D von 0,4 kann die Aufgabe als „ausgezeichnete Testaufgabe“ bezeichnet werden. Liegt der ermittelte Wert unter 0,29 sollte diese Aufgabe verbessert werden bzw. scheint sie weniger brauchbar. Durch die Entscheidung, die Gesamtgruppe in die Analyse einzubeziehen (und nicht nur die oberen und unteren 27 Prozent der Testsätze), ergibt sich jedoch, dass sich bei der Interpretation des Trennschärfeindex D nur im begrenzten Maße an den Vorgaben von Schelten (1997: 135) orientiert werden kann. Für Aufgaben, die in ihrer Schwierigkeit eher den Bereichen *mittel* und *schwer* zuzuordnen sind, ist eine Orientierung an den Interpretationen Scheltens noch gut möglich. Für Aufgaben des eher leichten Bereichs kann es eher zu Verzerrungen kommen, da die 46 Prozent der Mittelgruppe auch mitberücksichtigt

wurden. Nichtsdestotrotz kann ein DSH-Standort bereits zahlreiche Ergebnisse hinsichtlich des Schwierigkeitsgrades und der Trennschärfe der Testitems mithilfe der Ober- und Untergruppenanalyse gewinnen und diese für die stetige Qualitätssicherung am eigenen Standort nutzen. Im Folgenden werden die Ergebnisse der Ober- und Untergruppenanalysen von vier Testsätzen dargestellt und weitere Interpretationen vorgeschlagen.

4 Ergebnisse der Ober- und Untergruppenanalyse

Die Ober- und Untergruppenanalyse wurde an vier Prüfungssätzen bzw. an jeweils vier Prüfungsteilen zum Lese- und Hörverstehen vorgenommen. In der folgenden Tabelle (Tab. 3) befindet sich eine Übersicht zu den Prüfungszeiträumen und den Teilnehmendenzahlen für jeden Prüfungsdurchlauf, die in die folgenden Analysen einbezogen wurden. Alle Prüfungsteilnehmenden haben unmittelbar vor der Prüfung an einem 15-wöchigen DSH-Vorbereitungskurs in Jena teilgenommen. Die untere Zeile der Tabellen präsentiert jeweils die Anzahl der Prüfungen, die in die Ober- bzw. die Untergruppe einbezogen wurden (Burghoff/Leder 2011).

Tabelle 3: Übersicht der vier analysierten Prüfungssätze

DSH-Prüfung	Feb 2020	Juli 2020	Feb 2021	Juli 2021
TN-Zahl = N	69	53	29	46
Anzahl der Prüfungen in den Ober- und Untergruppen (nach Burghoff und Leder 2011) = n	34	26	14	23

Im Folgenden werden die Ergebnisse der Ober- und Untergruppenanalyse für die Prüfungsteile Leseverstehen und Hörverstehen getrennt voneinander betrachtet.

4.1 Evaluation der Aufgabentypen für den Prüfungsteil Leseverstehen

Der Text zur Überprüfung des Leseverstehens am Standort Jena erörtert aktuelle wissenschaftliche Themen mit Alltagsbezug. Die Struktur des jeweiligen Textes ist in den meisten Fällen ähnlich aufgebaut. Eine Einleitung in das Thema und gegebenenfalls die Vorstellung eines Forschungsteams oder eines Forschungszentrums mit besonderem wissenschaftlichem Schwerpunkt bilden den Texteinstieg. Anschließend folgt eine Erörterung der Fragestellung der im Text beschriebenen

Studie oder der Forschungsfrage. Ferner wird die im Text betreffende Studie beschrieben oder es werden zwei Studien verglichen. Abschließend werden der Mehrwert der Forschung bzw. die Ergebnisse dargestellt sowie ein passender Schlusspunkt zum Thema gegeben.

Die Aufgabengestaltung folgt der Struktur des Textes. Es wird globales, selektives und detailliertes Leseverstehen überprüft. Die Aufgabentypen sind in offene, halboffene und geschlossene Aufgaben im gleichen Verhältnis aufgeteilt. Für die Ober- und Untergruppenanalyse werden in diesem Beitrag die offenen Aufgaben, welche durch Operatoren zum Begründen und Benennen der verstandenen Aspekte, Facetten, Ursachen und Folgen aus dem Text repräsentiert werden, in Betracht gezogen.

In der folgenden Analyse ist demnach eine Evaluation der Aufgaben zu finden, die mit den Operatoren *erklären, erläutern* sowie *nennen* eingeleitet werden. Die Punktzahl in den erwarteten Lösungen ist nach den Items in der Aufgabe gestaffelt.

Schon aus dem Eintrag der Rohdaten in die Analysetabelle nach Ober- und Untergruppen können relevante Informationen für die Prüfungserstellenden gewonnen werden. Hier angegeben ist ein Beispiel einer Analysetabelle für eine Aufgabe aus einem Prüfungsteil Leseverstehen.

Tabelle 4: Frage 2: Erklären Sie ausgehend vom Text den Titel ... Gehen Sie auch darauf ein, wie der Forscher diesen Begriff versteht. (8 P. je Item – 2 Items) / Obergruppe (n = 23)

16	8	4	0	Ø = 11,13
13	6	0	4	p hoch = 0,70 (69,5 %)

Tabelle 5: Frage 2: Erklären Sie ausgehend vom Text den Titel ... Gehen Sie auch darauf ein, wie der Forscher diesen Begriff versteht. (8 P. je Item – 2 Items) / Untergruppe (n = 23)

16	8	4	0	Ø = 8,17
6	11	1	5	p niedrig = 0,51 (51,09 %)

Wie aus den Tabellen (Tab. 4 und 5) ersichtlich wird, beginnt die Aufgabenformulierung mit dem Operator „Erklären Sie ...“. Das Ziel der Aufgabe ist es, das globale Lesen zu überprüfen und eine Rückmeldung von den Prüfungsteilnehmenden zu bekommen, ob verstanden wurde, worum es in dem Text hauptsächlich geht. Die plausible Erklärung zum Thema des Textes wird als ein Item mit acht Punkten gewertet. Die weiteren acht Punkte bekommen die Prüfungsteilnehmenden, wenn sie außerdem auf den Forschungsbegriff aus der Sicht der/des Hauptforschenden eingehen. Für halbe Items wird jeweils die halbe Punktzahl ver-

geben. Aus der Tabelle ist zudem ersichtlich, dass 0, 4, 8, und 16 Punkte vergeben wurden, aber keine zwölf.

Hiermit ergibt sich ein Schwierigkeitsgrad p (siehe die Formel in Kapitel 3) von 0,60, was darauf hindeutet, dass sich die Aufgabe im mittleren Bereich des Schwierigkeitsgrades befindet. Ferner ergibt sich hier ein Trennschärfeindex D von 0,18. Gemäß der Interpretation der Trennschärfeindizes nach Schelten (1997: 135) ist diese Aufgabe wenig trennscharf und benötigt eine Revision.

Anhand dieses Beispiels wird auch sichtbar, dass trotz der Tatsache, dass die höchste Teilnehmendenzahl (13) aus der Obergruppe 16 Punkte erreicht hat und die höchste Teilnehmendenzahl (11) aus der Untergruppe 8 Punkte, die Aufgabe gemäß der Interpretation der Trennschärfe nach Schelten (1997: 135) nicht ausreichend trennscharf ist und eine Revision der Items notwendig scheint. Dies zeigt unter anderem auch, wie bereits oben erwähnt, dass einer Interpretation von Schelten nicht vollständig gefolgt werden kann. Eine Optimierung dieser Aufgabe kann nichtsdestotrotz in Betracht gezogen werden. Diese muss nicht unbedingt auf der inhaltlichen Ebene geschehen, denn man könnte die Punkteverteilung pro Item überdenken und entsprechend anpassen.

In der folgenden Tabelle wird zusammenfassend für alle vier analysierten Prüfungssätze zum Leseverstehen am Standort Jena aufgezeigt, wie der Schwierigkeitsgrad der Aufgaben hinsichtlich der Operatoren in den offenen Aufgaben in einen Zusammenhang gebracht werden kann. In der zusammenfassenden Analyse sind nur Operatoren in den offenen Aufgabenformaten dargestellt. Operatoren wie „Kreuzen Sie an“ würden eine Distraktorenanalyse nach sich ziehen, die in diesem Beitrag nicht betrachtet wird.

Für die Aufgabentypen zum Leseverstehen kann insgesamt gesagt werden, dass die Aufgaben, die eine stichpunktartige Benennung der im Text beschriebenen Prozesse, Facetten und Ergebnisse verlangen, nach dem Schwierigkeitsgrad p als eher leichte Aufgaben gewertet werden können. Die Aufgaben wiederum, die eine Erklärung der Sachverhalte, eine Begründung der Zusammenhänge aus dem Text sowie eine Beschreibung der Folgen und Ursachen erfordern, sind als mittlere Aufgaben nach dem Schwierigkeitsgrad p einzuschätzen.

Tabelle 6: Übersicht Operatoren Leseverstehen und Aufgabenformulierung nach Schwierigkeit

Operator und weitere Formulierungen	Einschätzung zur Schwierigkeit p
Nennen Sie Prozesse/Facetten/Ergebnisse ...	leicht (p im ø = 0,78)
Nennen Sie Kriterien/Aspekte ...	leicht (p im ø = 0,75)
Erläutern (Erklären) Sie, warum ...	mittel (p im ø = 0,64)
Nennen Sie Punkte/Gründe (wieso, wie ...)	mittel (p im ø = 0,64)

4.2 Evaluation der Aufgabentypen für den Prüfungsteil Hörverstehen

In jedem Hörtext am DSH-Standort Jena wird entweder eine Zusammenfassung eines aktuellen Forschungsprojekts zu einer eher alltagsnahen Forschungsfrage dargestellt oder es wird ein globaler Vergleich zwischen zwei Forschungsprojekten zu einer ähnlichen Thematik gegeben. Der jeweilige Vortrag folgt meist einer sehr ähnlichen Reihenfolge, die vergleichbar mit der Darstellung des klassischen Forschungskreislaufs ist: Darstellung des Forschungsinteresses bzw. der Forschungsfrage – gegebenenfalls Desiderate – Vorgehen der Forschenden – Ergebnisse – Ausblick.

Im zugeordneten Test zum Prüfungsteil Hörverstehen werden alle Hörverstehenstypen abgebildet, indem Aufgaben zum detaillierten Hören, globalen Hören und selektiven Hören präsentiert werden. Diese Aufgaben verlaufen chronologisch zum Vortrag.

Bereits in Kapitel 3 dieses Aufsatzes wurde anhand eines Beispiels aus dem Prüfungsteil Hörverstehen aufgezeigt, wie der Schwierigkeitsgrad p und der Trennschärfeindex D zu berechnen sind und welche Aussagen sich anschließend über die Ergebnisse treffen lassen. Dementsprechend soll im nächsten Schritt dargestellt werden, welche weiteren Ergebnisse mithilfe der Ober- und Untergruppenanalyse erzielt werden können, wenn man die Analysen von vier verschiedenen Prüfungssätzen miteinander vergleicht und in Beziehung setzt.

Aus der Ober- und Untergruppenanalyse von vier Prüfungssätzen für die Fertigkeit Hörverstehen kann beispielsweise auch abgeleitet werden, welche Operatoren sich eher dem leichten, dem mittleren und dem schwierigen Schwierigkeitsspektrum zuordnen lassen. Hierzu wurden zunächst alle Aufgaben der vier Prüfungssätze nach dem berechneten Schwierigkeitsgrad sortiert. Anschließend wurde festgestellt, dass sich einige (nicht alle) Operatoren und die entsprechende Aufgabenformulierung in einem ähnlichen Schwierigkeitsbereich bewegen. In der Tabelle 7 wird diese Hierarchie dargestellt. In der ersten Spalte sind der Operator sowie die nachfolgende Aufgabenformulierung angedeutet. Der zweiten Spalte kann der durchschnittliche Schwierigkeitsgrad für die formulierten Aufgaben entnommen werden.

Es sei an dieser Stelle darauf hingewiesen, dass es selbstverständlich auch darauf ankommt, wie explizit oder implizit die zu nennenden Fakten oder Erklärungsansätze im Hörtext präsentiert werden. Dementsprechend kann die nachfolgende Tabelle nicht generalisierend und nicht DSH-Standort-übergreifend verstanden werden. Die Ergebnisse in der Tabelle sollen lediglich verdeutlichen, wie DSH-Standort-intern nach einer Ober- und Untergruppenanalyse mit den Ergebnissen umgegangen werden kann und wie man diese Ergebnisse wiederum für die

Erstellung neuer Prüfungsteile nutzen kann – insbesondere, wenn immer eine ähnliche Personengruppe mit einem Prüfungsteil betraut ist.

Tabelle 7: Übersicht Operatoren Hörverstehen und Aufgabenformulierung nach Schwierigkeit

Operator und weitere Formulierungen	Einschätzung zur Schwierigkeit p
Nennen Sie Projektziele ...	leicht (p im ø = 0,796)
Geben Sie das Forschungsergebnis an.	leicht (p im ø = 0,8)
Geben Sie Informationen zur Probandengruppe an.	leicht (p im ø = 0,85)
Benennen Sie die Gliederungspunkte des Vortrags.	mittel (p im ø = 0,641)
Nennen Sie Gründe, weitere Ziele, kritische Punkte, Teilbereiche, Funktionen ...	mittel (p im ø = 0,56)
Erklären Sie anhand eines Beispiels einen Zusammenhang ...	mittel (p im ø = 0,65)
Beschreiben Sie das Vorgehen der Forschenden ...	mittel (p im ø = 0,42)
Vergleichen Sie das Vorgehen der Forschenden, die Ergebnisse der Forschenden ...	schwer (p im ø = 0,33)
Definieren Sie ...	schwer (p im ø = 0,3)

5 Fazit

Eine Besonderheit der DSH-Prüfung stellt ihre dezentrale Organisation, Durchführung und Evaluation dar – insbesondere im Vergleich zu anderen Prüfungen, welche für den Nachweis der sprachlichen Studierfähigkeit zugelassen sind (HRK und KMK 2019). Aufgrund der Dezentralisation sollte an jedem DSH-Standort eine Qualitätskontrolle jedes Prüfungssatzes erfolgen, um den Testgütekriterien Rechnung zu tragen. Im vorliegenden Beitrag wurde mit Blick auf die Gütekriterien *Schwierigkeit* und *Trennschärfe* eine Evaluation mithilfe der Ober- und Untergruppenanalyse beschrieben. Diese Darstellung soll anderen DSH-Standorten einen Einblick in die Evaluation am DSH-Standort Jena geben und aufzeigen, wie selbst an kleineren Einrichtungen eine Evaluation und stetige Weiterentwicklung betrieben und die eigene Prüfungserstellung fortwährend reflektiert werden kann. Selbstredend kann dies aufgrund einiger Rahmenbedingungen nicht immer im vollständigen und im Sinne der Testevaluation gewünschten Maße geschehen; jedoch bieten diese Ergebnisse viele Hinweise auf die Qualität der eigenen Prüfungssätze. Des Weiteren können diese Ergebnisse auch von den Anbieterinnen und Anbietern der DSH-Vorbereitungskurse als didaktische Implikationen Verwendung finden, beispielsweise für das Training der Operatoren.

Literaturverzeichnis

ALTE – Association of Language Testers in Europe (2012): *Handbuch zur Entwicklung und Durchführung von Sprachtests: Zur Verwendung mit dem GER*. Frankfurt am Main: telc GmbH.

Burghoff, Claudia; Leder, Gabriela (2011): *Handreichung zur DSH am Beispiel von Leseverstehen*. Göttingen: Klartext GmbH.

Dlaska, Andrea; Krekeler, Christian (2009): *Sprachtests: Leistungsbeurteilungen im Fremdsprachenunterricht evaluieren und verbessern*. Hohengehren: Schneider.

Eberharter, Kathrin; Kremmel, Benjamin; Zehentner, Matthias (2018): „Die Erstellung von Testaufgaben: Der Testentwicklungszyklus". In: Hinger, Barbara; Stadler, Wolfgang (Hrsg.): *Testen und Bewerten fremdsprachlicher Kompetenzen: Eine Einführung*. Tübingen: Narr Francke Attempto, 57–68.

Europarat, Goethe-Institut, Inter Nationes u.a. (Hrsg.) (2001): *Gemeinsamer europäischer Referenzrahmen für Sprachen: lernen, lehren, beurteilen*. Berlin: Langenscheidt.

FaDaF – Fachverband für Deutsch als Fremdsprache e.V. (2012): *DSH-Handbuch*. Göttingen: Klartext GmbH.

HRK und KMK – Hochschulrektorenkonferenz und Kultusministerkonferenz (2019): *Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen (RO-DT)*. Online: https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_06_25_RO_DT.pdf (30.09.2021).

Koithan, Ute; Leder, Gabriela; Wollert, Mattheus; Appel, Berit; Domes, Sonja (o. J.): *Checkliste zur Begutachtung und Erstellung eines DSH-Prüfungsexemplars*. Online: https://www.fadaf.de/de/rund_um_dsh/ordnung_zur_qualitaetssicherung_der_dsh/ (30.09.2021).

Pospeschill, Markus (2010): *Testtheorie, Testkonstruktion, Testevaluation*. München: Ernst Reinhardt.

Schelten, Andreas (1997): *Testbeurteilung und Testerstellung: Grundlagen der Teststatistik und Testtheorie für Pädagogen und Ausbilder in der Praxis*. 2. Auflage. Stuttgart: Franz Steiner.

Stadler, Wolfang (2018): „Rezeptive Fertigkeiten überprüfen und bewerten". In: Hinger, Barbara; Stadler, Wolfgang (Hrsg.): *Testen und Bewerten fremdsprachlicher Kompetenzen: Eine Einführung*. Tübingen: Narr Francke Attempto, 69–86.

Vigh, Tibor (2018): „Sprachtests im kommunikativen Fremdsprachenunterricht". In: Roche, Jörg; Einhorn, Ágnes; Suñer, Ferran (Hrsg.): *Unterrichtsmanagement*. Tübingen: Narr Francke Attempto, 85–121 (Kompendium DaF/DaZ 6).

Weir, Cyril J. (2005): *Language Testing and Validation: An Evidence-Based Approach*. London: Palgrave Macmillan.

Biographische Angaben

Alice Friedland

hat Deutsch als Fremd- und Zweitsprache sowie Erziehungswissenschaft studiert. Sie arbeitet seit Oktober 2019 als wissenschaftliche Mitarbeiterin am Institut für DaF/DaZ und interkulturelle Studien der Friedrich-Schiller-Universität Jena und ist unter anderem für die DSH-Prüfung mitverantwortlich. Ihre Forschungsschwerpunkte liegen in den Bereichen Testen und Prüfen, Wissenschaftssprache, Auslandsschulwesen und Didaktik/Methodik Deutsch als Fremd- und Zweitsprache.

Milica Sabo

hat Diplom auf Lehramt Anglistik und Germanistik sowie Master in Deutsch als Fremdsprache studiert. Sie ist promoviert im Fach Auslandsgermanistik und arbeitet seit Oktober 2018 als DSH-Prüfungsverantwortliche am Institut für DaF/DaZ und interkulturelle Studien der Friedrich-Schiller-Universität Jena. Ihre Forschungsschwerpunkte sind Testen und Prüfen und Evaluation, sprachenübergreifendes Lehren und Lernen sowie Didaktik/Methodik der Fremdsprachen.