2

Kompetenzmessungen
im österreichischen Schulsystem:
Analysen, Methoden & Perspektiven

Antonia Bachinger, Michael Krelle, Maria Engelbert-Kocher,
Gabriele von Eichhorn (Hrsg.)

Zuhörkompetenzen messen

Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe



WAXMANN

Institut des Bundes
für Qualitätssicherung im
österreichischen Schulwesen

Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven

herausgegeben vom
Institut des Bundes für Qualitätssicherung
im österreichischen Schulsystem (IQS)

Band 2

Antonia Bachinger, Michael Krelle,
Maria Engelbert-Kocher, Gabriele von Eichhorn (Hrsg.)

Zuhörkompetenzen messen

Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe

Waxmann 2022
Münster · New York

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://dnb.dnb.de abrufbar.

Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven, Bd. 2

ISSN 2628-9598
Print-ISBN 978-3-8309-4483-6
E-Book-ISBN 978-3-8309-9483-1
DOI 10.31244/9783830994831

© Waxmann Verlag GmbH, 2022
www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Anne Breitenbach, Münster
Cover-Bild: Andreas Paar
Satz: Andreas Kamenik, IQS

Dieses Werk ist unter der Lizenz CC BY-NC-SA 4.0 veröffentlicht:
Namensnennung – Nicht-kommerziell –
Weitergabe unter gleichen Bedingungen 4.0 International (CC BY-NC-SA 4.0)



Inhalt

Vorwort des IQS-Direktors ... 7
Danksagung der Herausgeber/innen ... 9

Gabriele von Eichhorn, Maria Engelbert-Kocher, Michael Krelle, Antonia Bachinger
Kapitel 1 Überblick ... **11**

Michael Krelle, Antonia Bachinger
Kapitel 2 Ein Konstrukt zum Testen von Zuhörkompetenz ... **25**

Maria Engelbert-Kocher, Antonia Bachinger
Kapitel 3 Konstruktoperationalisierung: Testaufgabenentwicklung und -durchführung ... **47**

Antonia Bachinger, Michael Krelle, Maria Engelbert-Kocher
Kapitel 4 Beispieltestaufgaben mit fachdidaktischer Kommentierung ... **73**

Benedikt O. J. Winter, Manuel Wick, Stefan Gugerell, Gloria Schneeweiß
Kapitel 5 Studiendesign und Durchführung der Pilotierung ... **105**

Michael Ober, Matthias Trendtel
Kapitel 6 Verlinkung zur ersten Bildungsstandardüberprüfung und Bestimmung des Administrationseffekts ... **123**

Antonia Bachinger, Roman Freunberger, Michael Krelle
Kapitel 7 Ergebnisse der Schülerinnen und Schüler im Zuhören ... **139**

Gabriele von Eichhorn, Matthias Trendtel
Kapitel 8 Aspekte psychometrischer Konstruktäquivalenz – Untersuchung zweier Operationalisierungen von Zuhörkompetenz ... **161**

Autorinnen und Autoren ... 181

Vorwort des IQS-Direktors

Verstehend zuhören zu können, ist eine zentrale Kompetenz im Deutschunterricht, in der Schule allgemein sowie für die außerschulische Entwicklung von Kindern; das wird auch mit Blick auf die Bildungsstandards für die 4. Schulstufe deutlich. Aus diesem Grund wurden im Jahr 2015 im Rahmen der flächendeckenden Bildungsstandardüberprüfung erstmals für Österreich umfassende Daten zu diesem Themenbereich erhoben. Mit dem darauf aufbauenden Berichtsband „Zuhörkompetenzen messen. Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe" wird dem System der evidenzinformierten Qualitätsentwicklung im österreichischen Schulwesen ein weiterer Meilenstein hinzugefügt.

Dieser Sammelband wird als zweiter Band der vom Waxmann Verlag herausgegebenen Reihe des IQS (ehemals BIFIE) „Kompetenzmessung im österreichischen Schulsystem: Analysen, Methoden & Perspektiven" veröffentlicht. Primäres Ziel der Reihe ist es, allen Interessierten und Beteiligten im deutschsprachigen wie im internationalen Raum vertiefende Analysen auf Basis der am IQS erhobenen und analysierten Daten zu präsentieren und damit gleichzeitig die wissenschaftliche Expertise des IQS verstärkt in den Fokus zu rücken. Zusätzlich dient die Reihe dazu, die Bildungsforschung vermehrt in den Blick zu nehmen und gleichzeitig den Wissenstransfer in die schulische Praxis zu unterstützen.

Im Berichtsband werden die unterschiedlichen Perspektiven von Bildungswissenschaft, Fachdidaktik sowie Bildungspolitik eingenommen und zusammengeführt. So werden die bildungspolitischen Vorzeichen, unter denen die Studie entstand, ebenso thematisiert wie die fachdidaktischen und fachwissenschaftlichen Hintergründe der Untersuchung. Sowohl die technische Umsetzung als auch dahinterliegende Entscheidungen sind ein wesentlicher Teil dieser Publikation. Zudem werden zentrale Evidenzen etwa zu Schülerleistungen, aber auch zur Konstruktäquivalenz erörtert.

Der Fokus des vorliegenden Berichtsbands richtet sich demnach darauf, wichtige Ergebnisse aus der D4-Studie 2018 zum Zuhören einer breiten Öffentlichkeit zu präsentieren. Die Untersuchung fand in Vorbereitung auf die Bildungsstandardüberprüfung statt, die Ergebnisse sind jedoch über den Kontext der Erhebung hinaus von Interesse. Insofern bildet dieser Sammelband auch eine Brücke zwischen der Bildungsstandardüberprüfung (BIST-Ü) und der Informellen Kompetenzmessung (IKM) auf der einen Seite und der individuellen Kompetenzmessung PLUS (iKM^{PLUS}) auf der anderen Seite.

Eine weitere Besonderheit des Berichtsbands liegt in dem Umstand, dass zentrale theoretische Informationen aus der Bildungsforschung und Fachdidaktik in das System

Schule eingebracht werden sollen. Die Daten können von Lehrerinnen und Lehrern, den entsprechenden Fachwissenschaften, aber auch der Bildungspolitik genutzt werden, um die evidenzinformierte Schul- und Unterrichtsentwicklung weiter voranzutreiben, etwa indem sprachsensiblem Unterricht eine höhere Bedeutung in der täglichen Schulpraxis beigemessen wird.

Die Publikation entstand in Kooperation mit der Technischen Universität Chemnitz, mit der das IQS schon seit Jahren in enger und fruchtbarer Zusammenarbeit steht. Die Expertise von Univ.-Prof. Dr. Michael Krelle und Maria Engelbert-Kocher im Bereich der standardisierten Erhebung von Kompetenzen im Fach Deutsch stellt eine große Bereicherung und Aufwertung der Untersuchung dar. Die Zusammenarbeit mit Expertinnen und Experten führt dabei auch zum fachwissenschaftlichen Austausch und zur weiteren Professionalisierung am IQS.

Die Herausgabe eines Berichtsbands erfordert sowohl hohe wissenschaftliche und organisatorische Kompetenz als auch – über die Kernaufgaben des IQS hinausgehende – Eigeninitiative; an dieser Stelle darf ich daher die Leistungen von Antonia Bachinger und Gabriele von Eichhorn nicht unerwähnt lassen. Mein besonderer Dank gilt sämtlichen Autorinnen und Autoren dieses Berichtsbands sowie allen Mitarbeiterinnen und Mitarbeitern des IQS, die an dessen Entstehung beteiligt waren, und im Besonderen den Kooperationspartnerinnen und -partnern der Technischen Universität Chemnitz.



Robert Klinglmair
Salzburg, im Jänner 2022

Danksagung der Herausgeber/innen

Zur Erstellung dieses Berichtsbands „Zuhörkompetenzen messen. Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe“ haben viele Personen beigetragen. Verschiedene Forschungsfragen können nur im Verbund aus Bildungsforschung und Fachdidaktik beantwortet werden. Insofern ist auch der vorliegende Berichtsband in enger Zusammenarbeit entstanden. Deshalb möchten wir uns an dieser Stelle bei allen bedanken, die uns bei diesem umfangreichen Projekt unterstützt haben.

Bedanken möchten wir uns insbesondere bei unseren Kolleginnen und Kollegen am IQS, die an der Entwicklung und Durchführung der D4-Studie 2018 beteiligt waren. Zudem möchten wir uns bei externen Partnerinnen und Partnern bedanken, die maßgeblich an der Entwicklung der Testaufgaben mitgearbeitet haben.

Bedanken möchten wir uns auch bei den Rechteinhaberinnen und -inhabern und bei den an der Entwicklung der Hörtexte Beteiligten sowie bei den Vertreterinnen und Vertretern der Hörbuch-Company und des Gerstenberg Verlags. Außerdem danken wir den Vertreterinnen und Vertretern des Waxmann Verlags herzlich für die gute Kooperation.

Wir danken den Autorinnen und Autoren für die gute Zusammenarbeit. Für die hilfreichen Rückmeldungen und Anregungen zu einzelnen Kapiteln gebührt ein besonderes Dankeschön auch den Reviewerinnen und Reviewern der wissenschaftlichen Artikel in diesem Band, insbesondere Johann Bacher, Ulrike Behrens, Katrin Brandmair, Marcel Illetschko, Ulrike Kipman, Konrad Oberwimmer, Veronika Österbauer, Daniel Paasch, Jutta Ransmayr, Alexander Robitzsch, Elisabeth Rothe, Rebecca Schneider und Sebastian Weirich.

In besonderem Maße bedanken wir uns außerdem bei den studentischen Hilfskräften bzw. Praktikantinnen und Praktikanten an der Technischen Universität Chemnitz und am IQS: Luisa Gronek, Tobias Pietz, Joel Scherzer und Thomas Wimmer.

Wir richten ganz herzlichen Dank an Andreas Kamenik, der uns beim Satz und bei der Organisation eine große Unterstützung war, und an Ann-Cathrice George, Claudia Griesmayr, Saya Gurtner-Reinthaler, Thomas Kiefer, Veronika Österbauer, Elisabeth Rothe, Benedikt Winter und Ramona Zintl, die uns stets mit guten Ratschlägen zur Seite standen. Unser Dank geht auch an Martin Schreiner (Lektorat) und an Gunther Griesenauer für die Erstellung des Webauftritts unseres Berichtsbands. Auch viele andere Mitarbeiter/innen des IQS waren direkt oder indirekt an den Arbeiten beteiligt, sie waren im Rahmen der Erstellung des Berichtsbands „Zuhörkompetenzen messen“ stets hilfsbereit und sind bereit-

willig eingesprungen, wenn zusätzliche Unterstützung benötigt wurde. Auch bei ihnen möchten wir uns sehr herzlich bedanken.

Zu guter Letzt gilt unser besonderer Dank auch der Führungsebene des IQS, die uns bei diesem Vorhaben stets unterstützt hat, insbesondere Michael Bruneforth, Roman Freunberger, Marcel Illetschko, Maria Neubacher und Sylvia Opriessnig, sowie dem Direktor Robert Klinglmair und der ehemaligen Direktorin Angela Weilguny.

Antonia Bachinger, Maria Engelbert-Kocher,
Michael Krelle und Gabriele von Eichhorn

Gabriele von Eichhorn, Maria Engelbert-Kocher, Michael Krelle, Antonia Bachinger

Kapitel 1
Überblick

In diesem Berichtsband sind zentrale Ergebnisse der Pilotierungsstudie zu den Bildungsstandardüberprüfungen im Fach Deutsch veröffentlicht. Die hier vorgestellte Studie (im Folgenden D4-Studie 2018[1]) bezieht sich auf Zuhörkompetenzen, eine Teilkompetenz im Kompetenzbereich *Hören, Sprechen und Miteinander-Reden.* Damit erscheint dieser Berichtsband im Kontext nationaler österreichischer Kompetenzmessungen. Der Fokus liegt auf der Volksschule. Zukünftig dienen die hier vorgestellten Erkenntnisse als Grundlage für die Erhebung von Zuhörkompetenzen in der individuellen Kompetenzmessung PLUS (iKM^{PLUS}), die ab dem Schuljahr 2021/22 stufenweise zur Anwendung gebracht wird und die bisherigen Bildungsstandardüberprüfungen (BIST-Ü) sowie die Informelle Kompetenzmessung (IKM) ablöst. Zuhörkompetenzen werden im Zuge der sogenannten Zyklusmodule aufgegriffen und ab 2024 flächendeckend erhoben (https://iqs.gv.at/themen/nationale-kompetenzerhebung/ikm-plus/allgemeine-informationen-und-zielsetzung).

Vor diesem Hintergrund werden im Folgenden die inhaltlichen Entscheidungen und die Ergebnisse der Studie der interessierten Öffentlichkeit vorgestellt. Die Reihe „Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven" bildet deshalb einen idealen Ort für die Veröffentlichung.

1.1 Bildungsstandards und ihre Überprüfung in Österreich

Die Bildungsstandards wurden im Schuljahr 2008/2009 eingeführt, um kompetenzorientierten Unterricht in Österreich zu etablieren (Schreiner & Breit, 2016; Wiesner & Schreiner, 2019, S. 14). Insgesamt geht es bei diesem Vorhaben darum, Aspekte evidenzbasierter Qualitätsentwicklung auszubauen und zu entwickeln (Haider, Eder, Specht, Spiel & Wimmer, 2005). Mit den Bildungsstandards in Österreich wird einerseits definiert, welche Kompetenzen Schüler/innen in bestimmten Fächern erreichen sollen, und andererseits, welche Qualitätskriterien das Bildungssystem erfüllen soll (Klieme et al., 2003). Dabei liegt den Bildungsstandards ein bestimmter Kompetenzbegriff zugrunde, der sich auf solche

1 Die Abkürzung bezeichnet die Pilotierungsstudie in Deutsch in der 4. Schulstufe im Jahr 2018. In dieser Studie wurden neben den nachfolgend beschriebenen Teilstudien in *Zuhören* auch Pilotierungen neuer Testaufgaben in den Kompetenzbereichen *Lesen, Rechtschreiben, Sprachbetrachtung* und *Verfassen von Texten* durchgeführt.

Fähigkeiten und Fertigkeiten bezieht, die in konkreten Anforderungssituationen gezeigt werden (Pant, Böhme, Stanat, Schipolowski & Köller, 2017, S. 45 f.). Im Sinne von Weinert (2001) geht es also um erlernbare kognitive Fähigkeiten und Fertigkeiten, um bestimmte Probleme zu lösen. Vor diesem Hintergrund sind die Bildungsstandards in Österreich als *Könnensbeschreibungen* (Can-Do-Statements bzw. Deskriptoren) gesetzlich fixiert (BGBl. II Nr. 1/2009 i. d. F. BGBl. Nr. 548/2020). In den Can-Do-Statements wird also beschrieben, welche Kompetenzen österreichische Schüler/innen am Ende der 4. respektive der 8. Schulstufe erreichen sollen.

Im Zuge der gesetzlich verordneten Bildungsstandards sollte außerdem eine an Daten orientierte Unterrichtsentwicklung etabliert werden. Aus diesem Grund erfolgte die Einführung der Bildungsstandardüberprüfung für alle Schüler/innen in Österreich in Deutsch, Englisch und Mathematik. Mit der Überprüfung der Bildungsstandards wurde das BIFIE, das Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens[2], betraut.

Mit diesem Instrument sollten die Kompetenzen von Schülerinnen und Schülern am Ende der 4. Schulstufe und am Ende der 8. Schulstufe flächendeckend erhoben werden. Die erste flächendeckende Überprüfung im Fach Deutsch/Lesen/Schreiben auf der 4. Schulstufe fand 2015 statt. Ein Überblick über den Ablauf der Studien in den Jahren 2012–2019 lässt sich anhand der Abbildung 1 gewinnen. Eine Zusammenschau der Ergebnisse aus dem ersten Zyklus dieser Erhebungen von 2012 bis 2016 ist in George, Schreiner, Wiesner, Pointinger und Pacher (2019) zu finden.

Fast jährlich wurden die Kompetenzen der Schüler/innen in Österreich erfasst. Dabei wurden die Untersuchungen jeweils als Vollerhebung – also als Erhebung eines kompletten Schülerjahrgangs (mit einzelnen Ausnahmen) – durchgeführt.[3] Ein sogenannter Bildungsstandard-Zyklus setzte sich dabei aus Deutsch und Mathematik in der 4. Schulstufe sowie Deutsch, Mathematik und Englisch in der 8. Schulstufe zusammen. Aus den Ergebnissen dieser Testungen wurden umfangreiche Rückmeldungen an unterschiedliche Adressatinnen und Adressaten generiert, z. B. an Lehrpersonen, an Schulleitungen, an die Schulaufsicht und an Schüler/innen. Die Ergebnisse der BIST-Ü wurden primär für die Systembeobachtung (Systemmonitoring) bzw. das Qualitätsmanagement an Schulen genutzt (Itzlinger-Bruneforth, Kuhn & Kiefer, 2016, S. 22). Der Rückmeldung an Schüler/innen kam weniger Bedeutung zu: Diese konnte aufgrund aufwändiger Logistik- und Analyseprozesse erst einige Zeit nach der Testung abgerufen werden, wenn die Schüler/innen bereits eine Schulstufe aufgestiegen waren, und wurde dementsprechend nur wenig genutzt.

2 Aus dem BIFIE wurde am 1. Juli 2020 das Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS).

3 Eine Ausnahme bildet hier zum Beispiel die Erhebung in Sprechen: Die Überprüfung des Bereichs fand aus organisatorischen Gründen nur an einer Stichprobe von 2.354 Schülerinnen und Schülern statt (Breit, Bruneforth & Schreiner, 2016, S. 91).

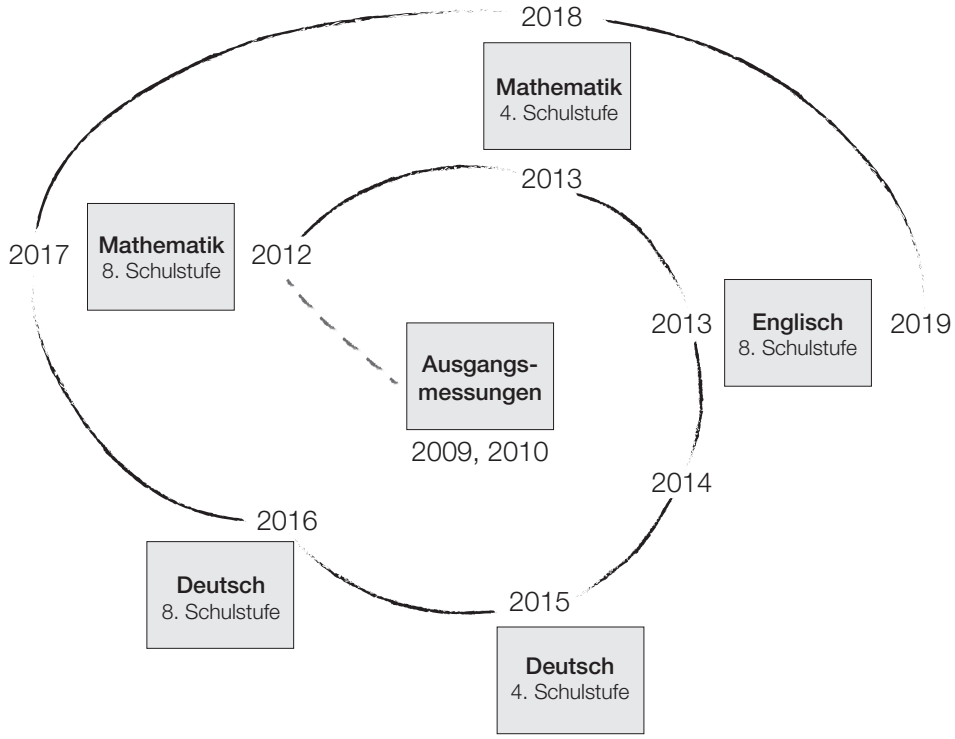


Abbildung 1: Bildungsstandard-Zyklus (George et al., 2019)

Neben der BIST-Ü stand den Lehrpersonen noch ein zweites Instrument zur Verfügung: die Informelle Kompetenzmessung (IKM). Während das Studien- und Berichtsdesign der BIST-Ü auf das Systemmonitoring ausgerichtet waren, standen bei der Informellen Kompetenzmessung Förderung und Unterrichtsevaluierung im Vordergrund. Die vorwiegend itembasierte Rückmeldung an Lehrer/innen erfolgte direkt nach der Durchführung der IKM und konnte somit gut für die Förderung von Schülerinnen und Schülern bzw. Klassen genutzt werden. In der Volksschule wurden im Rahmen der IKM vier Aufgabenpakete angeboten: *Mathematik* sowie in Deutsch *Lesen, Sprachbetrachtung* und *Verfassen von Texten.*

Im Rahmen des 2018 eingeführten Pädagogik-Pakets wurde eine Neugestaltung der Bildungsstandardüberprüfungen in Österreich beschlossen (BGBl. I Nr. 101/2018; BMBWF, 2021). Beginnend mit dem Schuljahr 2019/20 erfolgte eine stufenweise Weiterentwicklung der bestehenden Instrumente der Bildungsstandardüberprüfung und der IKM hin zum Instrument der individuellen Kompetenzmessung PLUS (iKM^{PLUS}) (siehe Abbildung 2), das schließlich ab dem Schuljahr 2021/22 implementiert wird. Bei diesem Verfahren sollen die Vorteile von BIST-Ü und IKM kombiniert werden. Die Grundlage der Erhebung bleiben im ersten Zyklus der iKM^{PLUS} die gesetzlich geregelten Bildungsstandards. Perspektivisch werden die aktuell in Entwicklung befindlichen Kompetenzraster (BMBWF, 2021) voraussichtlich zu wichtigen Referenzpunkten für künftige Erhebungen. Zuhörkompetenzen werden dort auch weiterhin eine zentrale Rolle spielen.

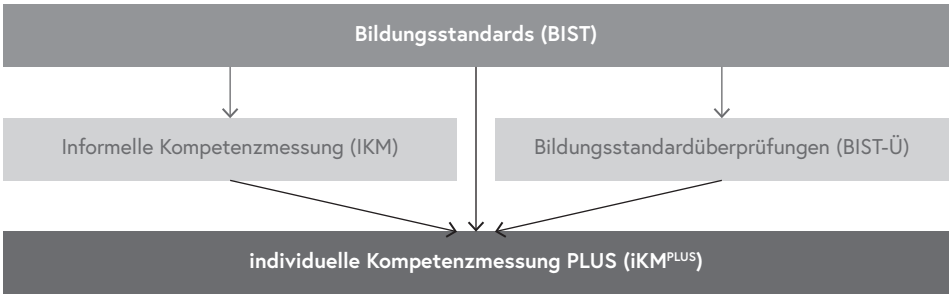


Abbildung 2: Verlaufsschema von den BIST zu IKM und BIST-Ü und zu iKM^{PLUS}

1.2 Ablauf und Grundlagen der vorliegenden Studie

Die Vorbereitung für und die Durchführung von Bildungsstandardüberprüfungen folgt jeweils einem ähnlichen Schema, das auch in Abbildung 3 verdeutlicht wird (siehe dazu auch Breit & Schreiner, 2016). Der Testkonstruktion, bestehend aus Konstruktspezifikation und Itementwicklung, folgen Überlegungen, welche Fragen durch die nachfolgende Pilotierungsstudie beantwortet werden sollen. Zumeist steht hier die empirische Prüfung von neuen Testaufgaben im Vordergrund. Wurden jedoch weitere Änderungen am Testinstrument vorgenommen, sollten auch diese empirisch untersucht werden. Als Vorbereitungsschritte folgen die Ziehung der Schulen- und Klassenstichprobe, der Schulkontakt, die Erstellung des Testdesigns und die Zuteilung der Testhefte zu den Schülerinnen und Schülern sowie Druck- und Logistikschritte zur Distribution der Testhefte. Die umfangreichen Vorbereitungsarbeiten münden schließlich in der Durchführung der Studie und der nachfolgenden Datenerfassung und -aufbereitung. Ab diesem Zeitpunkt können die im Studienkonzept angezielten Fragestellungen empirisch untersucht werden. Die Testaufgaben werden psychometrisch auf ihre Eignung hin überprüft. Auffällige Testaufgaben werden weiterhin fachdidaktisch analysiert und können aus dem Einsatz in künftigen Studien ausgeschlossen werden. Weitere Analyseschritte beinhalten (je nach Studienzweck) eine Skalierung der Daten, eine Verlinkung mit einer bereits bestehenden Metrik und die Auswertung etwaiger Teilstudien. In der Regel steht nach diesem Prozess ein geprüfter Itempool für den Einsatz in künftigen Studien zur Verfügung. Im vorliegenden Fall entsteht aus der Pilotierungsstudie zudem ein kurzer Ergebnisbericht.

Die Can-Do-Statements der Bildungsstandards bilden die zentrale Grundlage der Erhebungen: Aus ihnen werden – in Kombination mit der fachdidaktischen und psychometrischen Forschung – sowohl Kompetenzstufenmodelle als auch Testkonstrukte abgeleitet. Dabei werden zumeist nicht alle Can-Do-Statements verwendet, sondern lediglich diejenigen ausgewählt, die im Kontext von Large-Scale Assessments überprüfbar sind.

Bei Kompetenzstufenmodellen handelt es sich um sogenannte Niveaumodelle, die empirisch erfasste Kompetenzen konkreter inhaltlich beschreiben und somit auch eine kriteriums-

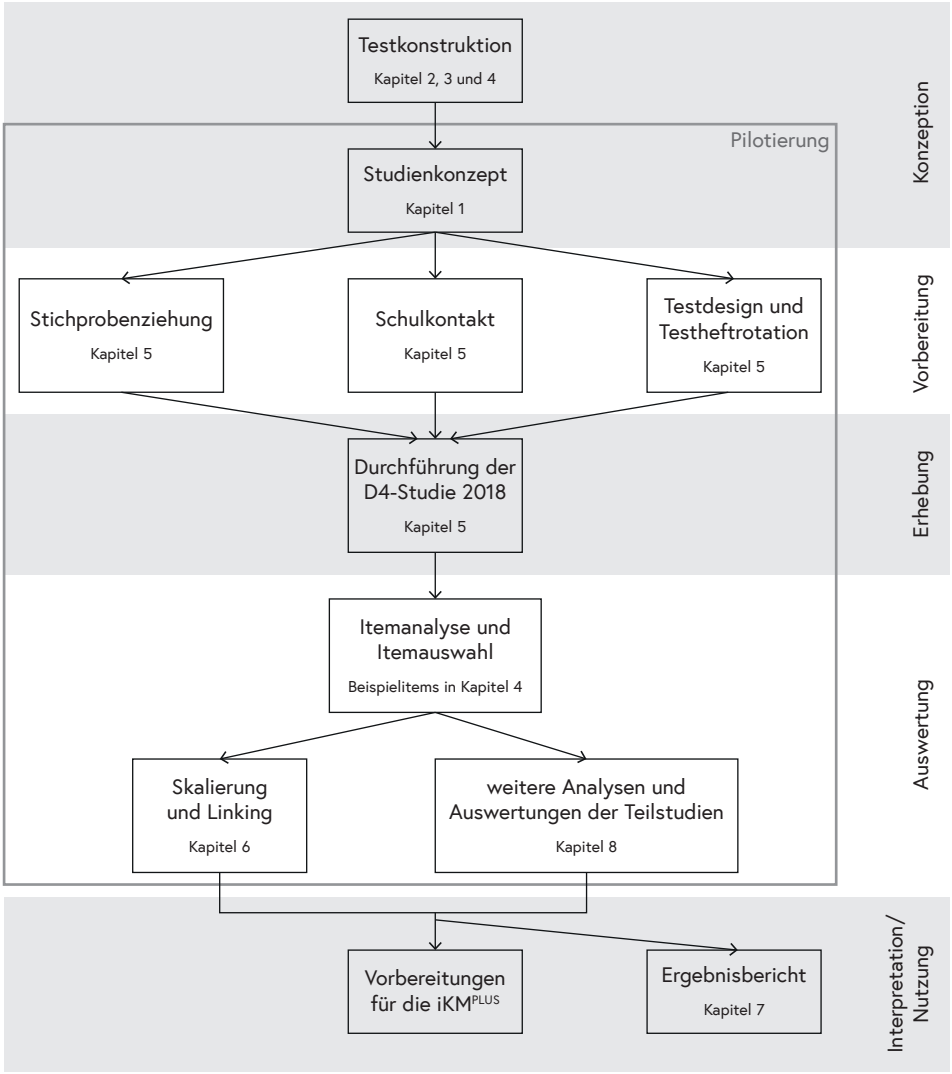


Abbildung 3: Ablauf einer Studie

orientierte Interpretation der Leistungswerte möglich machen (Hartig & Klieme, 2006). Sie stellen eine Momentaufnahme im Bildungsverlauf dar, sagen jedoch nichts darüber aus, in welchen Etappen Schüler/innen Kompetenzen erwerben (Entwicklungsmodell) oder aus welchen Facetten sich die Kompetenz zusammensetzt (Strukturmodell, siehe dazu auch Kapitel 2 – Konstrukt).

Die Testkonstrukte wiederum sind die Basis für die Entwicklung konkreter Testaufgaben, die Schüler/innen im Rahmen der Überprüfung vorgelegt bekommen, damit aus den korrekten und falschen Antworten ihr Ausmaß an Kompetenz abgeleitet werden kann.

1.3 Überarbeitung der Grundlagen der Testung

Im Rahmen der BIST-Ü 2015 wurde auch das *Zuhören* in Österreich erstmals flächendeckend überprüft. Die Ergebnisse für Österreich können dem Bundesergebnisbericht (Breit et al., 2016) entnommen werden. Im Anschluss wurde eine Weiterentwicklung des Tests beschlossen. Es sollten sowohl die zugrundeliegenden Konstrukte als auch die Testaufgaben adaptiert und verbessert werden: Die Neuerungen umfassen Anpassungen an die aktuelle Forschung und Änderungen in der Auslegung der zugrundeliegenden Deskriptoren der Bildungsstandards, in den Testaufgaben insbesondere bei den Stimuli (z. B. beim verstärkten Einsatz von literarischen Texten) und bei der Administration. Die Änderungen wurden im Rahmen einer Kooperation zwischen dem BIFIE und der Fachdidaktik Deutsch der Technischen Universität Chemnitz umgesetzt.

Ziel dieser Forschungskooperation war es, die Messung von Zuhörkompetenz einerseits unterrichtsvalider zu gestalten und andererseits mit Blick auf einzelne Aspekte Verbesserungen vorzunehmen. Die Umsetzung war ab 2016 geplant. Um weiterhin die Verknüpfung der Bildungsstandardüberprüfung 2015 und der (geplanten) Bildungsstandardüberprüfung 2020 zu ermöglichen, wurde für die D4-Studie 2018 ein komplexes Studiendesign geplant, das aus vier Teilstudien bestand.

Pilotierungsstudien wie die D4-Studie 2018 dienen üblicherweise dazu, Erhebungsinstrumente und Administrationsformen zu erproben. Neben der Pilotierung von Testaufgaben erfüllte die Studie, die die Datengrundlage dieses Berichtsbands bildet, weitere Funktionen. Das Studiendesign folgt der Logik, dass in jeder Teilstudie ein Aspekt der Testung verändert wurde, um Änderungen in den Testleistungen auf diesen Aspekt zurückführen zu können. Genauere Informationen, insbesondere zu den beiden Teilstudien Brückenstudie und Administrationseffektstudie, können in Kapitel 6 – Verlinkung – nachgelesen werden.

Es wurde angenommen, dass die Schüler/innen in den Stichproben, die diese Teilstudien bearbeiteten, der gleichen Population entstammten und diese repräsentativ vertraten. Diese verschiedenen Teilstudien werden im Folgenden erläutert.

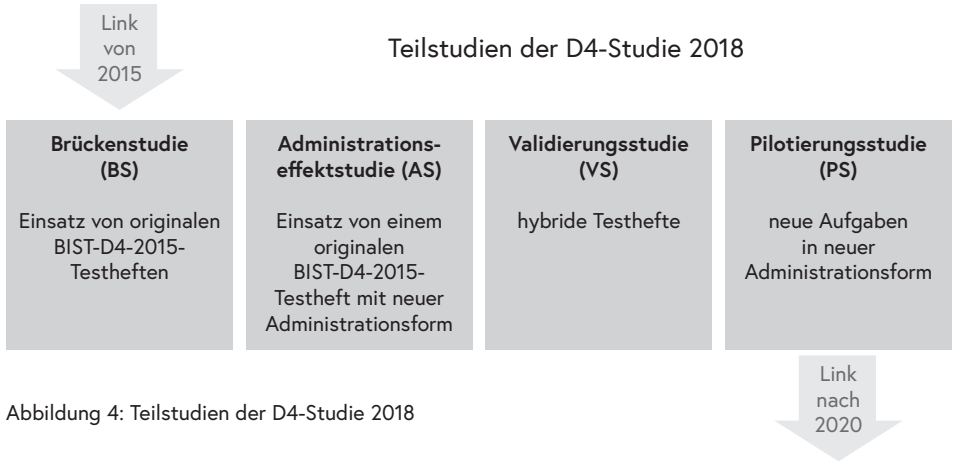


Abbildung 4: Teilstudien der D4-Studie 2018

1.3.1 Teilstudien

Brückenstudie

Um eine stabile Trendberichterstattung der Bildungsstandardüberprüfung zu künftigen Testungen zu gewährleisten, wurde eine Brückenstudie eingesetzt. In dieser wurden vier Testhefte aus der Bildungsstandardüberprüfung von 2015 im gleichen Ablauf und der gleichen Administrationsform von einer repräsentativen Teilstichprobe an Schülerinnen und Schülern bearbeitet. So ist sichergestellt, dass sich Leistungsunterschiede auf die Fähigkeit der Kohorte zurückführen lassen. Durch den Einsatz üblicher Linkingverfahren können die Leistungsscores dieser Stichprobe auf die Metrik der Bildungsstandardüberprüfung von 2015 transformiert werden. Das bedeutet, dass die Leistungen vergleichbar sind und Leistungsveränderungen der Schülerpopulation zwischen 2015 und 2018 festgestellt werden können. Die tatsächliche Verlinkung der Studien soll in diesem Berichtsband allerdings nicht im Mittelpunkt stehen, da die Gleichsetzung der Metrik auf die primären Fragestellungen dieses Berichtsbands keinen Einfluss hat (siehe Kapitel 6 – Verlinkung).

Administrationseffektstudie

Im Vergleich zur Bildungsstandardüberprüfung von 2015 wurden Änderungen in der Testadministration vorgenommen. So wurde z. B. eine kurze Instruktion vor jedem Stimulus angeboten, um den Kindern beim Zuhören eine Kontextualisierung anzubieten (siehe Kapitel 3 – Operationalisierung). Solche Änderungen können die Schwierigkeit der Aufgaben beeinflussen. Daher wurde eine Administrationseffektstudie durchgeführt, in der einer Teilstichprobe eines der originalen Testhefte aus der Bildungsstandardüberprüfung von 2015 mit neuer Administrationsform vorgegeben wurde. Über die Annahme, dass die Stichproben aus der Brückenstudie und aus der Administrationseffektstudie äquivalente Fähigkeiten aufweisen, können Unterschiede in der Testleistung auf die Änderungen in der Administrationsform zurückgeführt werden (siehe Kapitel 6 – Verlinkung).

Validierungsstudie

Neben den Änderungen in der Administration gab es noch einige wesentliche Änderungen im Testkonstrukt, die in Kapitel 2 – Konstrukt – näher erläutert werden und deren Auswirkungen auf die Testleistung der Schüler/innen von besonderem Interesse sind. Um die Unterschiede zwischen dem Konstrukt von 2015 und dem Konstrukt von 2018 genauer untersuchen zu können, wurde ein Testdesign gewählt, in dem Schüler/innen ein Hybrid-Testheft bearbeiteten, in dem zu gleichen Teilen Items aus beiden Konstrukten enthalten waren. Die Reihenfolge der Items aus beiden Konstrukten wurde über die Testhefte rotiert. So kann davon ausgegangen werden, dass Unterschiede in den Leistungen zwischen beiden Testheft-Teilen im Mittel nicht auf die Personenfähigkeit oder Ermüdungseffekte, sondern auf die Unterschiede zwischen den Konstrukten zurückzuführen sind. In Kapitel 8 – Konstruktäquivalenz – werden die Ergebnisse dieser Validierungsstudie berichtet.

Pilotierungsstudie

Der größte Anteil an der Stichprobe entfiel auf die Testaufgabenpilotierung. Diese hat zum Ziel, die neu entwickelten Aufgaben empirisch auf ihre Eignung hin zu überprüfen.

Anschließend werden die Items selektiert und für ihren Einsatz in der nächsten Erhebung bereitgestellt. In diesem Fall dient die Pilotierungsstudie zusätzlich dem Zweck, Zwischenergebnisse über die Leistungen der österreichischen Schüler/innen für Hörverstehen zu erheben und zu berichten. Dabei muss allerdings beachtet werden, dass Pilotierungsstudien üblicherweise nicht für die Ergebnisrückmeldung entwickelt werden und somit gewisse Nachteile mit sich bringen.[4] Dennoch zeigt sich, dass solche Ergebnisse wichtig sind, wenn Konstrukte neu oder weiterentwickelt werden (Granzer et al., 2009). Zudem liegen aufgrund der Weiterentwicklung der Bildungsstandardüberprüfung hin zur iKM^{PLUS} aktuell keine Ergebnisse zu den Kompetenzen von Schülerinnen und Schülern im Fach Deutsch vor. Vor diesem Hintergrund stehen also auch deskriptive Analysen von Zuhörkompetenzen im Mittelpunkt des Berichtsbands (siehe Kapitel 7 – Ergebnisse). Allerdings können keine verlässlichen Rückschlüsse darüber getroffen werden, welchen Fortschritt Schülerinnen und Schüler in Österreich von 2015 auf 2018 erzielt haben. Diese Ergebnisse werden erst in den Zyklusberichten der iKM^{PLUS} berichtet.

1.4 Weiterentwicklung zur iKM^{PLUS}

Im Zuge des Pädagogik-Pakets wurde 2018 entschieden, dass die BIST-Ü und die IKM in einem neuen Instrument verbunden werden (siehe oben). Damit sollen die Vorteile beider Erhebungen und deren unterschiedliche Testzwecke in einem Instrument mit sowohl freiwilligen als auch verpflichtend durchzuführenden Modulen kombiniert werden. Beispielsweise sollen durch eine zeitnahe Ergebnisrückmeldung die Daten im Unterricht besser verwertet werden können. Auf dieser Basis können mit Blick auf einzelne Standards, Kompetenzen und Fächer und auf unterschiedlichen Ebenen im System (z. B. Bundesland, Schule, Klasse, Schüler/innen) rechtzeitig gezielte Maßnahmen eingeleitet werden, wenn Defizite erkannt werden. Durch die flächendeckende verpflichtende Durchführung sollen aber dennoch weiterhin Daten für das Systemmonitoring zur Verfügung stehen.

Im Zuge der Entwicklung der iKM^{PLUS} wurde beschlossen, *Zuhören* zukünftig dreijährlich flächendeckend im Rahmen der sogenannten Zyklusmodule zu erheben, wodurch die Wichtigkeit dieser Kompetenz hervorgehoben wird. Von dieser politischen Entscheidung wurden auch die Arbeiten des BIFIE (jetzt IQS) und der Technischen Universität Chemnitz zur Änderung der Messung in *Zuhören* geprägt.

4 Ergebnisse aus Pilotierungsstudien sind mit einem Stichprobenfehler belegt. Weiterhin würden bei der iKM^{PLUS} oder der BIST-Ü nur Aufgaben zum Einsatz kommen, die sich bei der Pilotierung als besonders geeignet herausgestellt haben. Auch die Testdesigns für Erhebungen werden auf Basis der Informationen aus der Pilotierung optimiert zusammengestellt, um das Konstrukt angemessen abzudecken.

1.5 Der Berichtsband „Zuhörkompetenzen messen. Ergebnisse der Bildungsstandard-Pilotierung in der 4. Schulstufe“

Mit dem hier vorliegenden Berichtsband legen wir verschiedene Ergebnisse vor, die im Rahmen der Überarbeitung nach der BIST-Ü initiiert wurden und die auf Daten basieren, die in der D4-Studie 2018 erhoben wurden.

Dabei wird diese Progression auch als Chance begriffen, neuere fachwissenschaftliche und fachdidaktische Erwägungen in die iKM^{PLUS} aufzunehmen und wichtige Änderungen am Testkonstrukt und der Testadministration vorzunehmen. Diese Änderungen erleichtern auch, die Rückmeldungen für die jeweiligen Akteure passgenau zu gestalten, wie dies auch in der iKM^{PLUS} geplant ist (Behrens & Krelle, 2014).

Im vorliegenden Berichtsband wird in Kapitel 2 zunächst das theoretische Konstrukt diskutiert, das der D4-Studie 2018 zugrunde lag und zukünftig auch der iKM^{PLUS} zugrunde liegen wird. Die Operationalisierung des Konstrukts wird darauf aufbauend in den Kapiteln 3 und 4 dargestellt. In Kapitel 3 – Operationalisierung – wird zunächst beschrieben, wie die Entwicklung der Testaufgaben durchgeführt wurde. Die fachdidaktischen und testmethodischen Grundlagen werden entlang des Entwicklungsprozesses der Testaufgaben diskutiert. In Kapitel 4 – Beispielaufgaben – werden einzelne Testaufgaben der Pilotierungsstudie diskutiert: Sowohl die Stimuli als auch die dazugehörigen Testaufgaben (Items) werden ausführlich didaktisch thematisiert. In Kapitel 5 – Studiendesign – folgt schließlich die Beschreibung der Vorbereitung und Durchführung der Studie.

In Kapitel 6 – Verlinkung – werden primär Ergebnisse aus der Administrationseffektstudie vorgestellt und mögliche Gründe für diese Ergebnisse erörtert. Weiterhin werden die Ergebnisse in einen Kontext mit weiteren linkerschwerenden Aspekten und der Linkingstrategie gesetzt.

In Kapitel 7 – Ergebnisse – erfolgt die deskriptive Darstellung der Kompetenzen österreichischer Schüler/innen in der Pilotierungsstudie. Dabei werden verschiedene Kriterien wie Migrationsstatus, Geschlecht etc. für Subgruppen-Analysen verwendet.

Abschließend werden in Kapitel 8 – Konstruktäquivalenz – die beiden Konstrukte gegenübergestellt und mit der Frage nach Konstruktäquivalenz aus verschiedenen Perspektiven und mit verschiedenen empirischen Evidenzen diskutiert. Thema sind die Zusammenhänge zwischen den beiden Konstrukten und mit weiteren leistungsrelevanten Hintergrundmerkmalen, das Ausmaß von lokalen Abhängigkeiten innerhalb der beiden Konstrukte sowie die Unterrichtsnähe.

1.6 Ausblick – zukünftige Forschung in *Zuhören*

Eine Reihe zentraler Fragestellungen wird in diesem Berichtsband nicht aufgegriffen, soll aber im Rahmen weiterer Forschung angegangen werden, die hier knapp skizziert wird. Dabei wird nachfolgend primär auf drei Aspekte eingegangen: weitere Nutzungsmöglichkeiten der vorliegenden Daten, Zukunftsaussichten in der Messung von Zuhörkompetenzen sowie Fördermöglichkeiten von Zuhörkompetenzen. Abschließend soll noch dementsprechend ein Ausblick auf solche zukünftigen Untersuchungen gegeben werden.

1.6.1 Weitere Nutzungsmöglichkeiten der vorliegenden Daten

Trotz ausführlicher Analysen in diesem Berichtsband bleiben noch Fragen unbeantwortet, die in zukünftigen Studien untersucht werden sollen.

So ist etwa eine genauere Prüfung zur Dimensionalität von Zuhörkompetenz von Interesse. Dabei kann die Frage gestellt werden, ob und in welcher Weise sich theoretisch plausible und unterscheidbare Teilkompetenzen auch empirisch wiederfinden. Diese Frage ist aus verschiedenen Perspektiven interessant und relevant. Erstens können empirische Untersuchungen zur Faktorstruktur eines Tests eine Evidenz für die Konstruktvalidität bzw. strukturelle Validität nach Messick (1995) liefern. Die Frage nach der empirischen Dimensionalität kann jedoch zweitens in Überlegungen zur Rückmeldung einfließen: angesichts wachsender Forderungen nach mehr Information über die Testteilnehmerin/den Testteilnehmer ist abzuwägen, welche und wie viele Subscores rückgemeldet werden und ob diese genügend Reliabilität aufweisen (Sinharay, Puhan & Haberman, 2011). Drittens sind Dimensionalitätsuntersuchungen auch für die iterative Weiterentwicklung von Tests nützlich: Sind die Teilkompetenzen empirisch gut trennbar oder muss in der Itementwicklung mehr Fokus auf die Unterschiedlichkeit und Passgenauigkeit der Testaufgaben in verschiedenen Teilkompetenzen gelegt werden? Für Zuhören konkret können Teilkompetenzen aus Verstehensprozessen (hierarchieniedrig vs. hierarchiehoch) oder auch aus dem Umgang mit verschiedenen Texten (expositorisch vs. literarisch) abgeleitet werden. Zukünftig ist außerdem zu bedenken, dass klangliche Merkmale gesprochener Sprache selbst auch Träger von Bedeutung sind, die als eigenständige Teilkompetenz modelliert werden könnte (Behrens, Käser-Leisibach, Krelle, Weirich & Zingg Stamm, 2020).

Ebenfalls denkbar sind Untersuchungen zu narratologischen Kriterien als Einflussfaktoren auf die Schwierigkeit von Testaufgaben. Die Kenntnis schwierigkeitsbestimmender Merkmale ist hilfreich für das Verständnis der Kompetenz und des Antwortverhaltens der Schüler/innen und kann – ebenso wie Dimensionalitätsuntersuchungen – zur Konstruktvalidierung beitragen. Entsprechende Studien können zudem zukünftige Itementwicklungsprozesse informieren, um sie gezielter zu gestalten und den Ertrag an geeigneten Items zu erhöhen (Hartig & Frey, 2012). In der DESI-Studie wurden die Informationen aus den Untersuchungen zur Itemschwierigkeit zusätzlich für die Skalenbeschreibung in der Berichtslegung verwendet (Hartig, 2007). Böhme, Robitzsch und Busè (2010) untersuchten bereits

schwierigkeitsbestimmende Merkmale auf Item- und Textebene für Zuhören mit besonderer Hervorhebung der Merkmale, die hörspezifisch sind.

Zuletzt sind Vergleiche zwischen Lesen und Zuhören denkbar und mit den vorliegenden Daten möglich. Da der Fokus empirischer Forschung zur Diagnostik rezeptiver Kompetenzen auf dem Leseverstehen liegt, ist die Frage, ob und wie weit zwischen Lese- und Hörverstehen Unterschiede bestehen, nach wie vor interessant. Böhme et al. (2010) beschäftigten sich bereits aus einer theoretischen Perspektive mit den Spezifika von Zuhören. Behrens et al. (2020) untersuchten das Verhältnis von Hör- und Leseverstehen sowie verschiedene Präsentationsmodi der Items auch empirisch. In der vorliegenden Untersuchung wurden zu Forschungszwecken acht Testaufgaben in Lesen und Zuhören eingesetzt. Dabei wurden die Stimuli einmal in gedruckter Form und einmal als Hörtext dargeboten. Damit wäre es – ähnlich Behrens et al. (2020) – möglich, zu untersuchen, ob beiden Kompetenzbereichen die gleichen Verstehensprozesse zugrunde liegen oder ob sich Positionseffekte der zu den Stimuli gehörenden Items unterscheiden, da beim einmaligen Hören die Erinnerung an den Text möglicherweise schneller schwindet als beim Lesen, wo es möglich ist, den Stimulus erneut anzusehen.

1.6.2 Die Zukunft der Messung von Zuhörkompetenzen

Zukunftsweisende Technologien zur Testung von Kompetenzen gelangen mehr und mehr in den Einsatz und eröffnen neue Möglichkeiten, Konstrukte zu operationalisieren. Besonders für Zuhören ergeben sich einige interessante Aussichten.

So kann in Zukunft beispielsweise verstärkt auf multimediale und multimodale Texte und Testinstrumente gesetzt werden. Das bedeutet, dass Hörverstehen etwa auch mit dem Sehverstehen (z. B. beim Einsatz von audiovisuellen Medien) oder Leseverstehen (z. B. bei der Einbettung von Hörfiles in multimedialen Texten) kombiniert werden kann (Behrens & Krelle, 2014). Vereinfacht wird diese Möglichkeit dadurch, dass viele Large-Scale Assessments in den letzten Jahren den Schritt in die digitale Testung wagen, wodurch die Interaktivität der Testung steigen kann. Es ergeben sich neue Möglichkeiten zur Testung, aber auch neue Herausforderungen bei Tests und bei der Vermittlung von Zuhörkompetenzen.

Die digitale Testdurchführung ermöglicht es auch, Verfahren, die bisher mit fixierten Testheften arbeiten, in Zukunft adaptiv zu gestalten. Das bedeutet, dass an die geschätzte Fähigkeit der Testteilnehmerin/des Testteilnehmers angepasste Testaufgaben präsentiert werden und somit weniger Testaufgaben pro Schüler/in eingesetzt werden müssen, um hinreichend zuverlässige Ergebnisse zu erhalten. Darüber hinaus erhöhen computerisierte Testverfahren die Effizienz in Bezug auf den administrativen, personellen und finanziellen Aufwand (Asseburg, 2011). Hinsichtlich der häufig geäußerten Annahme, dass adaptives Testen die Motivation zur Testbearbeitung steigert, konnten bislang keine fundierten empirischen Ergebnisse vorgelegt werden. Untersuchungen deuten eher auf keinen (Ling, Attali, Finn & Stone, 2017) oder einen differenziellen Effekt zugunsten von Testteilnehmerinnen und

Testteilnehmern mit niedrig ausgeprägter Fähigkeit hin (Ortner, Weißkopf & Koch, 2014; Wise, 2014).

Vor diesem Hintergrund rücken Fragen zum Zuhören im Gespräch in den Mittelpunkt, das in bisherigen Testverfahren aus testmethodischen Gründen ausgeklammert worden war. In Österreich wurde der in den Bildungsstandards definierte Kompetenzbereich *Hören, Sprechen und Miteinander-Reden* beispielsweise in die Bereiche Zuhören und Sprechen aufgeteilt und somit getrennt erfasst. Die Vernetzung des Sprechens und Zuhörens könnte in Zukunft auch in der Operationalisierung der Bereiche abgebildet werden. So rückt die Möglichkeit, Gesprächssituationen zu simulieren, in den Bereich des Machbaren. Auf diese Weise könnten bisher noch nicht testfähige Deskriptoren der Bildungsstandards überprüft werden (siehe dazu auch Kapitel 3 – Operationalisierung) (Behrens, 2010; Nold & Rossa, 2017; Weirich, Bachinger, Trendtel & Krelle, 2019).

1.6.3 Fördermöglichkeiten von Zuhörkompetenzen

Neben der Erfassung ist die Förderung von Zuhörkompetenzen ein zentraler Bereich, in dem noch Forschungsdesiderate bestehen. Formativem Assessment kommt hier eine große Bedeutung zu. Mit Blick auf Aspekte der Unterrichtsentwicklung wäre es für Lehrkräfte sinnvoll, dieselben Schüler/innen mit vergleichbaren Aufgaben zu mehreren Zeitpunkten zu testen, um mögliche Resultate von Fördermaßnahmen in den Blick zu nehmen.

Letztlich soll beim Zuhören noch weiter erforscht werden, wie Kompetenzen von Schülerinnen und Schülern gut gefördert werden können. Für den deutschsprachigen Raum liegen zwar Förderkonzepte vor (Hagen, 2006; Krelle, 2011), allerdings wurden diese bisher selten empirisch validiert. Wünschenswert wäre hier eine Orientierung an den Kompetenzen der Schüler/innen, für die bestimmte Fördermethoden evidenzbasiert empfohlen werden können, vor allem in Hinblick auf die iKM^{PLUS}. Ein besonderer Blick sollte nach Ansicht der Autorinnen und Autoren dieses Kapitels auf die Erforschung digitaler Fördermethoden im Zuhören gelegt werden.

Diese Forschungsfragen sollten in zukünftigen Studien bearbeitet werden. Die Vernetzung von Forschung und Anwendung zur Erfassung und Förderung von Zuhörkompetenzen ist ein Anliegen der Herausgeber/innen dieses Berichtsbands. Der vorliegende Berichtsband richtet sich deshalb gleichermaßen an Bildungsforscher/innen, Fachdidaktiker/innen, Sprachwissenschaftler/innen sowie interessierte Lehrkräfte, die sich intensiv mit den Zuhörkompetenzen von Schülerinnen und Schülern in der Volksschule auseinandersetzen möchten.

Literatur

Asseburg, R. (2011). *Leistungsbereitschaft in Testsituationen. Motivation zur Bearbeitung adaptiver und nicht-adaptiver Leistungstests.* Marburg: Tectum-Verlag.

Behrens, U. (2010). Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 31–50). Göttingen: Vandenhoeck & Ruprecht.

Behrens, U., Käser-Leisibach, U., Krelle, M., Weirich, S. & Zingg Stamm, C. (2020). *stim·mig – Ergebnisse einer Studie zum Zuhören in der dritten Jahrgangsstufe.* Münster: Waxmann.

Behrens, U. & Krelle, M. (2014). Hörverstehen – Ein Forschungsüberblick. *Didaktik Deutsch, 20* (36), 86–107. doi:10.25656/01:17208

BMBWF (2021). *Das Pädagogik-Paket.* Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/zrp/pp.html

Böhme, K., Robitzsch, A. & Busè, A.-K. (2010). Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung der Testaufgaben. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule – Beiträge aus Wissenschaft und Praxis.* Göttingen: Vandenhoeck & Ruprecht.

Breit, S., Bruneforth, M. & Schreiner, C. (2016). *Standardüberprüfung 2015, Deutsch/Lesen/Schreiben, 4. Schulstufe – Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/0d879bba1b6fb81469e09bc99aab7e4916f88bef/BiSt_UE_D4_2015_Bundesergebnisbericht.pdf

Breit, S. & Schreiner, C. (2016). Vorwort. In C. Schreiner & S. Breit (Hrsg.), *Large-Scale Assessment mit R – Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung.* Wien: facultas.

George, A. C., Schreiner, C., Wiesner, C., Pointinger, M. & Pacher, K. (Hrsg.). (2019). *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich – Vertiefende Analysen zum Zyklus 2012 bis 2016* (Bd. 1). Münster: Waxmann.

Granzer, D., Köller, O., Bremerich-Vos, A., van den Heuvel-Panhuizen, M., Reiss, K. & Walther, G. (2009). *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule.* Weinheim: Beltz.

Hagen, M. (2006). *Förderung des Hörens und Zuhörens in der Schule* (Bd. 6). Göttingen: Vandenhoeck & Ruprecht.

Haider, G., Eder, F., Specht, W., Spiel, C. & Wimmer, M. (2005). *Abschlussbericht der Zukunftskommission an Frau Bundesministerin Elisabeth Gehrer.* Wien: BMBWK.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus. In E. Klieme & B. Beck (Hrsg.), *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International* (S. 83–99). Weinheim: Beltz.

Hartig, J. & Frey, A. (2012). Konstruktvalidierung und Skalenbeschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten. *Psychologische Rundschau, 63* (1), 43–49. doi:10.1026/0033-3042/a000109

Hartig, J. & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127–143). Berlin: Springer. doi:10.1007/3-540-33020-8_9

Itzlinger-Bruneforth, U., Kuhn, J.-T. & Kiefer, T. (2016). Testkonstruktion. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R – Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung.* Wien: facultas.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M. et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bonn: BMBF.

Krelle, M. (2011). Zuhördidaktik. Anmerkungen zur Förderung rezeptiver Fähigkeiten des mündlichen Sprachgebrauchs im Deutschunterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 51–69). Göttingen: Vandenhoeck & Ruprecht.

Ling, G., Attali, Y., Finn, B. & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement, 41* (7), 495–511. doi:10.1177/0146621617707556

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50* (9), 741–749. doi:10.1037/0003-066X.50.9.741

Nold, G. & Rossa, H. (2017). Hörsehverstehen. In B. Tesch, X. von Hammerstein, P. Stanat & H. Rossa (Hrsg.), *Bildungsstandards aktuell: Englisch/Französisch in der Sekundarstufe II* (S. 100–119). Braunschweig: Diesterweg.

Ortner, T. M., Weißkopf, E. & Koch, T. (2014). I Will Probably Fail: Higher Ability Students' Motivational Experiences During Adaptive Achievement Testing. *European Journal of Psychological Assessment, 30* (1), 48–56. doi:10.1027/1015-5759/a000168

Pant, H. A., Böhme, K., Stanat, P., Schipolowski, S. & Köller, O. (2017). Kompetenzstufenmodelle für den Primarbereich. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 45–52). Münster: Waxmann.

Schreiner, C. & Breit, S. (2016). Konzeption der Überprüfung der Bildungsstandards in Österreich. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R – Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung.* Wien: facultas.

Sinharay, S., Puhan, G. & Haberman, S. J. (2011). An NCME Instructional Module on Subscores. *Educational Measurement: Issues and Practice, 30* (3), 29–40. doi:10.1111/j.1745-3992.2011.00208.x

Weinert, F. E. (2001). Leistungsmessung in Schulen – eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Hrsg.), *Leistungsmessung in Schulen.* Weinheim: Beltz.

Weirich, S., Bachinger, A., Trendtel, M. & Krelle, M. (2019). Listening comprehension tests in Germany and Austria – Research report and critical review. *L1 Educational Studies in Language and Literature, 19,* 1–26. doi:10.17239/l1esll-2019.19.03.05

Wiesner, C. & Schreiner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – eine gelungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *Fünf Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich – Vertiefende Analysen zum Zyklus 2012 bis 2016* (Bd. 1). Münster: Waxmann.

Wise, S. L. (2014). The Utility of Adaptive Testing in Addressing the Problem of Unmotivated Examinees. *Journal of Computerized Adaptive Testing, 2* (3), 1–17.

Michael Krelle, Antonia Bachinger

Kapitel 2
Ein Konstrukt zum Testen von Zuhörkompetenz

2.1 Einleitung

Die Bildungsstandardüberprüfung und die Informelle Kompetenzmessung (IKM) sind standardisierte Instrumente zur Diagnose der Kompetenzen von Schülerinnen und Schülern zu bestimmten Zeitpunkten in Österreich. Mit der Bildungsstandardüberprüfung ist eine Datenbasis für das Schulqualitätsmanagement und das Monitoring des Bildungssystems geschaffen. Zudem sollen mit der IKM Pädagoginnen und Pädagogen förderdiagnostische Rückmeldungen zu ihren Schülerinnen und Schülern erhalten.

Im Folgenden wird für die nationale Leistungsmessung ein Konzept entworfen, mit dem solche Aspekte des Zuhörens der Schüler/innen überprüft werden, die im Rahmen von Tests ermittelt werden können. Dazu wird zunächst Zuhörkompetenz in fachwissenschaftlicher und fachdidaktischer Perspektive diskutiert. Anschließend wird der Bezug zu den Bildungsstandards in Österreich hergestellt und es werden bisherige Umsetzungen von Zuhörtests diskutiert, um Konstruktionsprinzipien für die Messung von Zuhörkompetenz im Rahmen der Bildungsstandardüberprüfung abzuleiten. Die Bildungsstandardüberprüfung und die IKM werden in Zukunft zusammengeführt, weiterentwickelt und in Form der individuellen Kompetenzmessung PLUS (iKM^{PLUS}) ausgebaut (siehe Kapitel 1 – Überblick). Grundlage der iKM^{PLUS} bleiben die per Verordnung definierten Bildungsstandards (BGBl. II Nr. 1/2009).

2.2 Zuhörkompetenz in fachwissenschaftlicher und -didaktischer Perspektive

In der Regel wird zwischen (physiologischem) Hören und (psychologischem) Zuhören unterschieden (Ackermann, 2006; Behrens, Käser-Leisibach, Krelle, Weirich & Zingg Stamm, 2021; Worthington & Bodie, 2017): Unter den Begriff *Hören* fallen dabei in erster Linie Aspekte der Schalldetektion und -verarbeitung durch das Hörsystem (Hellbrück & Ellermeier, 2004, S. 91). Solche grundlegenden Fähigkeiten werden nach Behrens, Böhme und Krelle (2009) bei Schuleintritt vorausgesetzt oder sie werden im inklusiven Unterricht als Beeinträchtigungen bzw. Einschränkungen aufseiten der Schüler/innen angesehen (Lipkowski & Schüller, 2017, S. 19). Unter dem Begriff *Zuhören* werden hingegen unter-

schiedliche kognitive bzw. psychische Prozesse verstanden (Imhof, 2010). Mitunter werden auch sprachliche Handlungen als *Zuhören* bezeichnet. Bei einer solchen umfassenden Definition geht es dann beispielsweise um *Back-Channel-Behavior,* also die Verwendung von Feedbacksignalen zu den Gesprächsschritten (Brinker & Sager, 2001, S. 59), sowie um *aktives Zuhören* als Kommunikationsstrategie (Rogers, 1985). Solche Handlungsformen und -muster sollen im Laufe der Schullaufbahn gezielt durch schulische Förderung ausgebildet werden, sie sind aber in der Regel nicht Gegenstand von Testungen (Krelle, 2010).

Vor diesem Hintergrund werden im Folgenden *Zuhören* und *Hören* terminologisch unterschieden und eng geführt: Zuhören wird hier als mentaler Prozess der Informationsverarbeitung angesehen. Das damit in der Forschung verbundene Konstrukt wird in diesem Sinne als *verstehendes Hören* bzw. als *Hörverstehen* bezeichnet.

2.2.1 Hörverstehen als Konstrukt

Das Hörverstehen wird als komplexer Prozess der aktiven, rezeptiven Sprachverarbeitung angesehen, bei dem im Zusammenspiel von Vorwissen und auditivem Input kognitive Repräsentationen des Gehörten entstehen (Schwarz, 1992, S. 133). Ein prominentes Modell solcher kognitiven Prozesse stammt von Imhof (2003, 2010, S. 15). Dort spiegelt sich auch die oben diskutierte Unterscheidung von Hören und Zuhören wider: Grundlage sogenannter *auditiver Prozesse* ist nach Imhof die Unterscheidung einer akustischen und einer auditorischen Seite der (parallelen) Reizverarbeitung im Gehirn. Die akustische Verarbeitung von Lautreizen (i. S. v. Hören) wird als die Wahrnehmung und Verarbeitung von Sprachreizen oder anderen akustischen Signalreizen aufgrund physikalischer Kriterien angesehen. Auditorische Prozesse umfassen hingegen die kognitive Verarbeitung und differenzierte Analyse von Reizen über das Ohr (i. S. v. Zuhören). Vor diesem Hintergrund modelliert Imhof (2010) das Hörverstehen als mehrstufigen Prozess der Informationsverarbeitung (vgl. Abbildung 1).

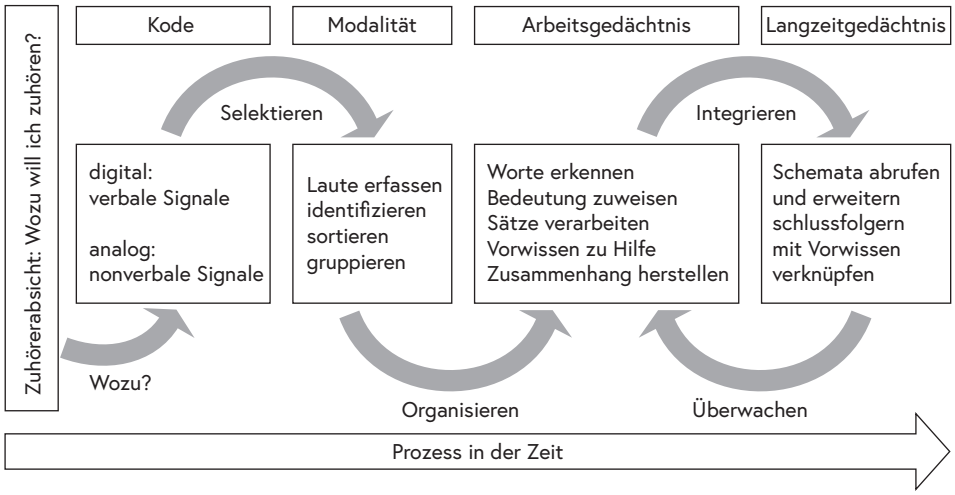


Abbildung 1: Zuhörmodell (nach Imhof, 2010, S. 19)

Grundlage aller auditiven Prozesse ist eine Intention bzw. eine sogenannte *Zuhörabsicht*. Das heißt hier, dass Zuhören immer mit bewussten oder unbewussten Zielen verbunden ist: Zuhörer/innen selektieren akustische Signale aus dem Lautstrom, um auf Basis von Sprach- und Weltwissen Informationen zu organisieren bzw. in die eigene Repräsentation des Gehörten zu integrieren. Diese grundlegenden Fähigkeiten sind notwendige Voraussetzungen für Kommunikations- und Verstehensprozesse. Weil für die genannten Aspekte kognitive Ressourcen benötigt werden, spielen Arbeits- und Langzeitgedächtnis eine wichtige Rolle beim Zuhören. Schließlich prüft man auf der Grundlage des eigenen Vorwissens und neuer Informationen das bisher Verstandene hinsichtlich Plausibilität, wobei durchaus Revisionen vorgenommen werden. Das Hörverstehen kann also als ein kognitiver Verarbeitungsprozess angesehen werden, in dem die Zuhörer/innen aktiv verschiedene interpretatorische Leistungen auf Basis des Vorwissens vollbringen. Krelle und Prengel (2014, S. 213) schreiben dazu:

> Um solche Leistungen zu beschreiben, wird von verschiedenen Formen der mentalen Repräsentation von Bedeutungen ausgegangen. Man unterscheidet hier eine hierarchieniedrige Ebene des kognitiven Systems, auf der es um das mehr oder weniger sinngemäße Erinnern einer oder mehrerer Formulierungen geht. Daneben wird von einer hierarchiehohen Ebene gesprochen, in der Repräsentationen angesiedelt sind, die sich auf größere Textteile oder den gesamten Text und seine Strukturen beziehen, z. B. in Form eines mentalen Modells.

Prozesse auf hierarchieniedriger und hierarchiehoher Ebene erfordern jeweils kognitive Ressourcen, etwa wenn Informationen zu Einheiten („chunks“) verdichtet werden und sich eben sinngemäß oder in ihrer Bedeutung für das Gesamte erinnert wird. In der Folge resultiert ein Verstehen des Gehörten.

Mitunter werden für die Unterscheidung von kognitiven Leistungen unter Rückgriff auf Textverstehenstheorien und vergleichbar zum Lesen auch eine *propositionale Struktur* und eine *Makrostruktur* unterschieden (Kintsch, 1998). Eine Einordnung der Begriffe liefert Imhof (2003, S. 140): Von propositional spricht man, wenn bei der Informationsgewinnung Elemente oder Einheiten in sogenannte Propositionen (im Sinne von Satzaussagen und ihren Relationen) zerlegt werden. Diese sind dann kognitiv im Verstehensprozess so geordnet, dass sie (makrostrukturell) zu einer mehr oder weniger sinnvollen Gesamtaussage integriert werden kann. Als Verarbeitungsprozesse werden noch weitere Prozesse angenommen, u. a. solche der Reduktion von Informationen (z. B. Zusammenfassen, Generalisieren, Auslassen), einfache, automatische Schlussprozesse bzw. Inferenzen (so genannte *bridging inferences* im Sinne von Kintsch (1998, S. 189), Prozesse der lokalen Kohärenzbildung, elaborative Prozesse (z. B. Assoziation) und ggf. die Hinzunahme nichtsprachlicher Signale:

> Auf diese Weise mag es dem Zuhörer gelingen, den sprachlichen Äußerungen eine Bedeutung zuzuweisen, die durch eine gewisse Kohärenz im Sinne von subjektiv empfundener Stimmigkeit und Plausibilität ausgezeichnet ist. (Imhof, 2003, S. 140)

Generieren Menschen beim Zuhören dann also eine solche Makrostruktur, überführen sie die Propositionen mit Hilfe von Selektions-, Reduktions- und Gewichtungsprozessen in eine übergeordnete Struktur. Dabei wird auch Wichtiges von Unwichtigem unterschieden, Bekanntes wiedererkannt, Neues integriert etc. Zuhören heißt in dieser Perspektive dementsprechend auch selektieren, gewichten, auslassen, generalisieren, schließen, konstruieren und integrieren zu können (Imhof, 2003, S. 140).

2.2.2 Einflussfaktoren auf das Hörverstehen und Bezug zu anderen Kompetenzen

Über die Entwicklung des Hörverstehens ist bisher wenig bekannt: So wird zwar z. B. die Sprechfähigkeit biografisch eindeutig nach dem Zuhören ausgebildet, dennoch liegen fast ausschließlich elaborierte Modelle zum sprachproduktiven Erwerb vor. Allerdings sind Befunde über das Hören und Zuhören im frühen Säuglingsalter einschlägig (Behrens, 2010, S. 31; zusammenfassend Hennon, Hirsh-Pasek & Golinkoff, 2000, S. 51). Weitere Untersuchungen finden sich im Kontext des Erwerbs des Schriftsystems, etwa wenn Aspekte *phonologischer Bewusstheit* als Faktor für den Schriftspracherwerb insgesamt modelliert werden (Pröscholdt et al., 2013). Dort geht es aber weniger um den Erwerb von Zuhörfähigkeiten als mehr um ihre *dienende Funktion* für den Erwerb des Schriftsystems.

Gut dokumentiert sind die Einflussfaktoren auf das Hörverstehen. In einem Forschungsüberblick fassen Behrens und Krelle (2014) etwa folgende Aspekte als bedeutsam zusammen:

(1) Das Arbeitsgedächtnis und seine Leistungsfähigkeit sind für das Hörverstehen zentral (s. o.), was u. a. mit der Multimodalität alltäglicher Zuhörsituationen zu erklären ist. Es müssen nämlich sehr viele (Arten von) Informationen prozessiert werden:

> Zuhörer müssen die bislang gehörte flüchtige Information so lange präsent halten, bis sie mit der prinzipiell nicht vorhersehbaren noch folgenden Information zu einem sinnvollen Ganzen integriert werden kann. (Behrens und Krelle, 2014, S. 88; siehe auch Hagen, 2006; Kahlert, 2000)

(2) Akustische Faktoren beeinflussen das Zuhören. Das betrifft z. B. die Lautstärke, die Art der Artikulation, die Geschwindigkeit, mögliche Störgeräusche u. v. m. (siehe dazu auch Buck, 2001; Kapitel 3 – Operationalisierung).
(3) Auditive Faktoren sind durch die Signalwahrnehmung der Zuhörerinnen und Zuhörer bedingt. Bereits minimale Hörverluste zwischen 10–15 dB, die normalerweise bei entsprechenden Screenings nicht entdeckt werden, können zu beträchtlichen Lücken im Hörverstehen führen (Coninx, 2009).
(4) Weitere kognitive Faktoren betreffen die Verarbeitung des ankommenden Signals, etwa den Zusammenhang von Sprechtempo und Sprachkenntnissen des Rezipienten.
(5) Motivationale Faktoren stehen in Zusammenhang mit Aspekten der situativen Interessiertheit sowie mit einem relativ überdauernden Interesse.
(6) Zu den kommunikativen Faktoren zählen etwa:

> Aspekte der sozialen Nähe oder Distanz zwischen Gesprächspartnern, der Reziprozität der Kommunikation (etwa die Möglichkeit zur Nachfrage oder Beeinflussung des Sprechers

durch Hörersignale), der persönlichen Bereitschaft, sich auf eine Situation „einzulassen" (Behrens & Krelle, 2014, S. 90).

(7) Als weiterer Faktor kommen noch Fähigkeiten der Aufmerksamkeitssteuerung und der Konzentration ins Spiel. Insbesondere die Konzentration gilt als wichtigster Prädiktor von Hörverstehensleistungen in Abhängigkeit der Variablen sprachlicher Hintergrund (Deutsch als Erst- oder Zweitsprache) und Selbstkonzept (Gschwend, 2014).

Vor diesem Hintergrund lassen sich auch bedeutsame Unterschiede des Hörverstehens zu anderen sprachlichen Kompetenzen ausmachen, z. B. zum Leseverstehen. Die größten Unterschiede finden sich auf der hierarchieniedrigen Ebene des kognitiven Systems und ergeben sich aus der jeweils akustischen oder visuellen Darbietung der Informationen (Grotjahn, 2005), z. B. mit Blick darauf, dass Texte beim Lesen selbstgesteuert gescannt werden können, wenn eine Information wiedergefunden werden muss. Die größten Gemeinsamkeiten zwischen Lesen und Zuhören werden hingegen auf der Ebene der Inferenzen und der Bildung mentaler Modelle ausgemacht (Hagtvet, 2003; Kürschner, Schnotz & Eid, 2006; Kürschner & Schnotz, 2008). Im Rahmen von Vergleichsstudien korrelieren Lese- und Hörverstehen mit .74 auf latenter Ebene (Behrens, Böhme & Krelle, 2009, S. 370), was allerdings auch mit der Gestaltung der Testitems zusammenhängen könnte (Behrens et al., 2021). Mitunter wird auch von einer Zunahme solcher Korrelationen zwischen Lese- und Hörverstehen bei steigender Leseerfahrung berichtet (Marx & Jungmann, 2000). Solche Ergebnisse sind bisher allerdings noch uneindeutig (Knechtel, 2011). Weitere Untersuchungen liegen auch zum Einfluss von Lehrmethoden zum Hörverstehen auf das Lesenlernen im Grundschulalter vor (Marx, 2000). Zudem dürfte das Leseverstehen von Grundschulkindern auch von der Fähigkeit des Hörverstehens profitieren (Knoepke, Richter, Isberner, Neeb & Naumann, 2013, S. 274). Bremerich-Vos, Böhme und Robitzsch (2009) weisen in einer explorativen Faktorenanalyse (mit zwei Faktoren: sprachproduktiv und sprachrezeptiv) nach, dass nur das Hörverstehen eine hohe Ladung auf den „rezeptiven Faktor" hat. Das Lesen hingegen lädt nicht auf diesen rezeptiven Faktor. Hier ist allerdings unklar, ob ggf. die medialen Besonderheiten des Zuhörens einen eigenen Faktor bilden (Bremerich-Vos et al., 2009, S. 212).

Mit Blick auf weitere (auch sprachproduktive) Kompetenzen hält Behrens (2010, S. 34) fest, dass das Zuhören „[...] offenkundig in ein Netz von Kompetenzen eingewoben [ist], die über die Spezifika dieses Bereichs hinausweisen". Zentral ist hier neben anderen Aspekten insbesondere das Sehverstehen. Vor diesem Hintergrund hat sich auch der Begriff „Hör-Sehverstehen" etabliert, zu dem eine Reihe von Modellierungen in den Bereichen Deutsch als Fremd- und Zweitsprache vorliegen (Porsch, Grotjahn & Tesch, 2010; Rösch, 2011). Das Hör-Sehverstehen ist im Alltag ebenfalls von den oben genannten Faktoren beeinflusst: Je nach Situation ist z. B. das Arbeitsgedächtnis mehr beansprucht, wenn widersprüchliche Signale unterschiedlicher Quellen interpretiert und für die Informationsgewinnung genutzt werden müssen; die Belastung des Arbeitsgedächtnisses kann sich jedoch auch als weniger anspruchsvoll erweisen, wenn die Datenlage weniger widersprüchlich ist (Grotjahn, 2000, S. 8). Wie sich bewegte Bilder in Hörverstehenstests auswirken, wird bisher ausschließlich in der Fremdsprachendidaktik diskutiert (Wagner, 2007; Behrens & Krelle, 2014, S. 90).

Vor dem Hintergrund des Hör-Sehverstehens gibt es zudem noch enge Verbindungen zum Sprechen. Hier liegen linguistische Ansätze vor, in denen z. B. Gesprächsstörungen aus Perspektive der Zuhörenden untersucht werden (Fiehler, 1998), in denen Verstehensleistungen auf der para- und nonverbalen Ebene im Gespräch analysiert werden (Hartung, 1998) oder das Potenzial von *Back-Channel-Behaviors* gesprächslinguistisch in den Blick genommen wird (Spiegel, 2006, 2009). Dabei geht es in der Regel um komplexe Gesprächsfähigkeiten bzw. um Aspekte von Gesprächskompetenz (Hartung, 2004) sowie mündlicher Kommunikationskompetenz (Becker-Mrotzek, 2009), also um einen weiteren Begriff von Zuhören als hier (siehe oben, Abschnitt 2.1 in diesem Kapitel).

Was Kinder zu bestimmten Zeitpunkten des schulischen Erwerbs können sollen, wird in den deutschsprachigen Ländern im Rahmen von Bildungsstandards festgelegt (Breit, Bruneforth & Schreiner, 2016, 2017; Stanat, Pant, Böhme & Richter, 2012). Für Österreich sind vor allem die Formulierungen in den Bildungsstandards für die vierte und achte Schulstufe maßgeblich, wie sie in der Verordnung des Bundesministeriums für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen in Österreich zu finden sind (BGBl. II Nr. 1/2009). Die Bildungsstandards für die vierte Schulstufe bilden auch die Grundlage, wenn hier ein Konstrukt zum Testen von Zuhörkompetenz entwickelt wird.

2.3 Bildungsstandards zum Zuhören

2.3.1 Bildungsstandards zum Zuhören

Die Bildungsstandards im Schulwesen für die 4. Schulstufe der Volksschule in Österreich weisen „Hören" zusammen mit „Sprechen" und „Miteinander-Reden" als einen eigenen Kompetenzbereich aus. Die wesentlichen Kompetenzen sind nachstehend abgedruckt.

Tabelle 1: Auswahl an Bildungsstandards zum Zuhören (BGBl. II Nr. 1/2009)

Kompetenzbereich: *Hören, Sprechen und Miteinander-Reden*	
Verständlich erzählen und anderen verstehend zuhören	Die Schüler/innen können anderen aufmerksam zuhören.
Informationen einholen und sie an andere weitergeben	Die Schüler/innen können Informationen über Lebewesen, Gegenstände sowie Sachzusammenhänge einholen.
In Gesprächen Techniken und Regeln anwenden	Die Schüler/innen können sich an Gesprächsregeln halten, anderen respektvoll zuhören und sich fair mit deren Meinungen auseinandersetzen [...].
Kompetenzbereich: *Lesen – Umgang mit Texten und Medien*	
Die Lesemotivation bzw. das Leseinteresse festigen und vertiefen	Die Schüler/innen können beim Vorlesen interessiert und bewusst zuhören [...].

Auch wenn hier von „Hören" gesprochen wird, zeigt sich auf Ebene der Standards, dass „Zuhören" bzw. „Hörverstehen" im oben genannten Sinne gemeint ist. Unter dem Standard „Verständlich erzählen und anderen verstehend zuhören" finden sich nämlich solche sprachrezeptiven Kompetenzen, die sich in narrativen (ggf. fiktionalen) Kontexten ergeben. Es geht darum, anderen aufmerksam zuhören zu können, wenn diese z. B. von Begebenheiten und Erfahrungen erzählen. Gleichwohl spielt das Zuhören auch bei dem Standard „Informationen einholen und sie an andere weitergeben" eine zentrale Rolle. Beim Zuhören in Gesprächen sollen sich die Schüler/innen außerdem „an Gesprächsregeln halten, anderen respektvoll zuhören und sich fair mit deren Meinungen auseinandersetzen". Darüber hinaus werden die Kompetenzbereiche im Sinne eines integrativen Deutschunterrichts aufeinander bezogen und miteinander in Beziehung gesetzt. Mit Blick auf das Zuhören wird das z. B. im Kompetenzbereich „Lesen – Umgang mit Texten und Medien" deutlich: Die Schülerinnen und Schüler sollen „beim Vorlesen interessiert und bewusst zuhören [können]".

Hinzu kommt eine Reihe von Bildungsstandards aus anderen Bereichen wie „Einsicht in Sprache durch Sprachbetrachtung", die sich ebenfalls auf das Zuhören beziehen lassen, auch wenn sie oben nicht explizit ausgewiesen sind, z. B. die Kompetenz, Merkmale von Regionalsprache und Standardsprache unterscheiden zu können.

2.3.2 Zur Lesart der Standards im Bereich Zuhören

Die Standardformulierungen aus Abschnitt 2.3.1 lassen sich folgendermaßen auf die fachwissenschaftliche und fachdidaktische Diskussion (Abschnitt 2.2) beziehen: In den Bildungsstandards sind durchweg sprachrezeptive (mündliche) Kompetenzen versammelt, die man für aufmerksamkeitsgesteuerte Bedeutungsrekonstruktion und -sicherung in variablen Kommunikationssituationen benötigt (Becker-Mrotzek, 2008, S. 66). Dabei ist folgende Unterscheidung leitend: Es kann um Situationen gehen, in denen man mit anderen interagiert und in denen das Zuhören von umfangreichen Gesprächstätigkeiten begleitet ist. Gleichwohl sind aber auch solche Situationen gemeint, *in denen man zuhört, ohne dass man mit anderen in Interaktion tritt,* etwa wenn man Radiosendungen, Podcasts, Hörspiele etc. hört. Damit nimmt das Zuhören eine besondere Rolle im Rahmen der Bildungsstandards ein (siehe dazu auch Krelle & Neumann, 2014, S. 14): Anders als bei schriftsprachlichen Kompetenzen geht es – erstens – auch um Kompetenzen, die im Gesprächsprozess in gegenseitiger Abhängigkeit mit anderen gezeigt werden. Teils werden solche Kompetenzen – zweitens – zusammen mit dem Sprechen bereits weit vor der Schulzeit und auch darüber hinaus außerhalb des Unterrichts erworben. Die in den Bildungsstandards formulierten Kompetenzen werden im Unterricht dann mit Blick auf grundlegende Bildung und kulturelle Teilhabe weiterentwickelt. Mit Blick auf den Beitrag des Deutschunterrichts zur Bildung heißt es dementsprechend an zentraler Stelle in einer Erläuterung zu den Standards:

> Die Grundschule ist ein pädagogischer Raum für die Grundlegung der Bildung. Sprachliche Grundbildung erfolgt dann einerseits als Persönlichkeitsbildung, andererseits weckt, erhält und fördert sie die Fähigkeit zur Kulturaneignung. Dabei leistet sie einen wesentlichen Beitrag für die Sinnperspektive des Lernens und Handelns. (BIFIE, 2011, S. 7)

Das Zuhören ist in dieser Perspektive wesentlicher Bestandteil kommunikativer Kompetenz. Insofern geht es um mehr als Grundfertigkeiten, nämlich um komplexe Fähigkeiten, die für die Lösung wiederkehrender Kommunikationsaufgaben dienlich sein können (Krelle & Neumann, 2014, S. 15). Dabei ist dann auch zu bedenken, dass Aspekte mündlicher Kommunikation nicht nur als Lernziel, sondern auch als Lernmedium für andere Kompetenzen und andere Fächer bedeutsam sind.

Ausgehend von diesen Standardformulierungen und ihrer Interpretation werden im Folgenden Konstruktionsprinzipien und Einschränkungen bei der Gestaltung von Zuhörtests zu den Bildungsstandards diskutiert.

2.4 Konstruktionsprinzipien und Einschränkungen bei der Gestaltung von Zuhörtests

An prominenter Stelle heißt es in der Expertise zur Entwicklung nationaler Bildungsstandards, dass diese in ihrer Formulierung so konkret beschrieben sein sollen, „dass sie in Aufgabenstellungen umgesetzt und prinzipiell mit Hilfe von Testverfahren erfasst werden können" (Klieme et al., 2003, S. 9). Dabei ist man sich allerdings einig, dass die Entwicklung von Testaufgaben zu den Standards kein mechanischer, sondern ein kreativer Prozess ist, denn die Standards sind mehr oder weniger auslegungsbedürftig (Bremerich-Vos, 2009, S. 21). Dabei kommt es auch auf eine enge Zusammenarbeit von Pädagogik, Psychologie und Fachdidaktik an (Klieme et al., 2003, S. 22). Dass diese Umsetzung von Standards in Testaufgaben im Bereich Zuhören von einer Reihe von Einschränkungen bestimmt ist, wurde in der Vergangenheit schon häufig betont (Behrens, Böhme & Krelle, 2009): Das Zuhören kann nach dem derzeitigen Stand der Forschung und unter vertretbarem Ressourcenaufwand flächendeckend nur in solchen Situationen getestet werden, in denen die Schüler/innen nicht mit anderen interagieren bzw. sprechen.[1] In solchen Fällen könnte man schließlich nicht sicherstellen, dass alle Beteiligten unter den gleichen Bedingungen getestet werden (Krelle & Prengel, 2014, S. 213). Dazu schreiben Behrens und Krelle (2014, S. 100):

> Obwohl es unstrittig ist, dass das Hörverstehen im Alltag und in der Schule zumeist nur zusammen mit Sprech- und Gesprächstätigkeiten relevant ist, gibt es dazu nur sehr wenige empirische Untersuchungen. Teils finden sich Arbeiten in der Gesprächslinguistik und -didaktik. [...] Einstweilen bleibt es aber außerordentlich schwierig, das komplexe Gesprächsganze sachangemessen und gleichzeitig methodisch kontrolliert zu erfassen, um zu belastbaren Aussagen zu kommen.

Ob perspektivisch im Rahmen digitaler Testungen auch Gesprächssituationen (mit anderen) simuliert werden können, ist noch nicht hinreichend geklärt. Vor diesem Hintergrund können die für den Unterricht wichtigen Kompetenzen zum Standard „In Gesprächen Techniken und Regeln anwenden" aktuell nicht getestet werden.

1 Zur Administration des Bereichs *Sprechen* siehe Breit et al. (2016, S. 91–96).

Deshalb werden ausschließlich Aufgaben zu Hörtexten auf CD gestellt, die der ganzen Klasse (Volksschule) vorgespielt werden. In der Sekundarstufe werden die Hörtexte künftig als Einzeltestung am Computer durchgeführt. Die Hörsituationen sind in diesem Sinne auf rezeptive mündliche Fähigkeiten beschränkt, wie sie in den folgenden Standards für die Volksschule formuliert sind (siehe Abschnitt 2.3):

- (beim Erzählen) anderen verstehend bzw. aufmerksam zuhören,
- Informationen einholen (und sie an andere weitergeben),
- beim Vorlesen interessiert und bewusst zuhören, [...].

Allerdings ist auch der letztgenannte Standard schwierig in ein Testformat zu überführen, insbesondere wenn es um Fragen des „Interesses" geht.

Vor dem Hintergrund der genannten Standards und fachwissenschaftlichen bzw. -didaktischen Grundlagen wird im Folgenden das Testkonstrukt dargestellt.

2.4.1 Vorarbeiten und Bezugnahmen für die Testentwicklung

Mit Blick auf bereits vorhandene Operationalisierungen, also Umsetzung von Kompetenzbeschreibungen in konkrete Testaufgaben, liegen im deutschsprachigen Raum drei Erhebungsinstrumente vor: Die Arbeiten des Bundesinstituts für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens[2], (siehe Kapitel 1 – Überblick), des Instituts zur Qualitätsentwicklung im Bildungswesen (IQB, Deutschland, Bremerich-Vos et al., 2010; Bremerich-Vos, Böhme, Krelle, Weirich & Köller, 2012) sowie der Schweizerischen Konferenz der kantonalen Erziehungsdirektoren (EDK, 2011, S. 46).

Die Arbeiten des BIFIE beziehen sich auf ein Testkonstrukt von Thomas, Latschen, Frank, Weishaupt und Weishaupt (2011a; siehe auch Thomas et al., 2011b), in dem die folgenden Subkompetenzen für die Testentwicklung unterschieden werden:

Tabelle 2: Subkompetenzen für die Itementwicklung (Thomas et al., 2011a; Thomas et al., 2011b)

K1	Die Bedeutung ähnlich klingender Wörter unterscheiden können.
K2	Aus Texten Detailinformationen entnehmen.
K3	Aus Texten die wesentlichen Informationen/Abläufe/Zusammenhänge entnehmen.
K4	Aus Texten die Grundgedanken erfassen und weiterführen und/oder Schlussfolgerungen ziehen.
K5	Die Bedeutung parasprachlicher Informationen erkennen und die Mitteilungsabsicht derselben einschätzen können.

2 Aus dem Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) wurde am 1. Juli 2020 das Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS).

K6	Die Inhalte eines Textes auch bei Auftreten von Stör- bzw. Hintergrundgeräuschen erfassen können.
K7	Die eigene Verstehensleistung kontrollieren (Comprehension Monitoring).
K8	Die Angemessenheit von Sprachhandlungen (Sprachkonventionen) erkennen und einschätzen können.
K9	Einen altersadäquaten Wortschatz passiv verstehen und die Bedeutung der Wörter erfassen können.

Diese Subkompetenzen von Thomas et al. (2011a, 2011b) wurden hier nicht für die Testentwicklung berücksichtigt. Dafür gibt es gleich mehrere Gründe: Die Bedeutung ähnlich klingender Wörter unterscheiden zu können (K1), lässt sich zwar auf basale Fähigkeiten beziehen („Phonlogische Bewusstheit"), aber nur schwierig auf einschlägige fachwissenschaftliche und -didaktische Modelle zum Hörverstehen. Die Subkompetenzen K2–K4 sowie K8 lassen sich hingegen eindeutig auf kognitionsbasierte Theorien zum Hörverstehen beziehen, sie sind aber teils schwierig zu operationalisieren. So ist z. B. die Abgrenzung von „Detailinformationen" (K2) und „wesentlichen Informationen" (K3) nicht immer leicht zu treffen, insbesondere, wenn die Texte zeitlich begrenzt sind oder die Menge der Informationen überschaubar ist. Zu der Subkompetenz K5 liegen zudem ebisher kaum Testverfahren vor. Zuletzt wurde zwar im „Stimmig-Projekt" (Zingg Stamm, Käser-Leisibach, Behrens, Krelle & Weirich, 2016) die Fähigkeit getestet, zu verstehen, was gesagt wird, selbst dann, wenn der Wortlaut einer Äußerung mehrere Interpretationen zulässt (Behrens et al., 2021, S. 9). Allerdings fehlt es noch an weitergehenden bzw. umfangreicheren Grundlagenforschungen in diesem Bereich. Bei der Subkompetenz K6 sind – anders als bei den anderen Subkompetenzen – schwierigkeitsbestimmende Merkmale („Störgeräusche") genuiner Bestandteil der Kompetenzbeschreibung. Aufgrund dieser Einschränkungen wurde K9 schon in der Bildungsstandardüberprüfung des BIFIE im Jahr 2015 nicht operationalisiert (siehe dazu auch Kapitel 1 – Überblick). Zudem ist festzuhalten: Die Subkompetenz K7 dürfte schon im Alltag bzw. im Unterricht schwierig in Aufgaben umzusetzen sein; in Testsituationen kann die Kontrolle „eigener Verstehensleistungen" nicht zuverlässig beobachtet werden. Manche Subkompetenzen (z. B. K9) lassen sich zudem auch auf Fragen des Sprachgebrauchs bzw. auf Aspekte allgemeiner Sprachkompetenz zurückführen. Insofern ergeben sich auch Fragen zur Validität der Subkompetenzen.

Vor diesem Hintergrund erfolgte eine Orientierung an einer Reihe von nationalen und internationalen Vorarbeiten für die Primar- und Sekundarstufe. Das betrifft auch die Arbeiten zu den KMK-Bildungsstandards in Deutschland (Behrens, Böhme & Krelle, 2009; Bremerich-Vos et al., 2010; Bremerich-Vos et al., 2012; Krelle & Prengel, 2014) sowie die Arbeiten zu den Nationalen Bildungsstandards der EDK in der Schweiz (Eriksson, Lindauer & Sieber, 2008; Eriksson, 2009; Waibel & Eriksson, 2010; EDK, 2011). Die vorliegenden Tests weisen eine Reihe von Gemeinsamkeiten auf (Behrens & Krelle, 2014), die auch hier die Grundlage der Testentwicklung bilden:

- *Die theoretischen Grundlagen sind vergleichbar:* Sie beziehen sich durchwegs auf Theorien zum Hörverstehen, wie sie weiter oben dargestellt sind (siehe Abschnitt 2.2).

- *Das Setting der Tests ist vergleichbar:* Zu Audioquellen (u. a. von CD) werden schriftliche Testaufgaben gestellt, um Vergleichbarkeit zwischen großen Gruppen von Schülerinnen und Schülern herzustellen. Die Hörstimuli werden einmal abgespielt, um eine möglichst authentische Hörsituation zu evozieren (siehe Kapitel 3 – Operationalisierung). Es gibt kein Itempreview (Neumann, 2012, S. 115), d. h., die Schüler/innen sehen die Aufgaben erst, nachdem sie den Hörtext zu Ende gehört haben.
- *Die Grundlagen der Aufgabenentwicklung sind vergleichbar:* So werden schwierigkeitsbestimmende Merkmale auf Hörtext- und Itemebene zugrunde gelegt, die ausführlich in Kapitel 3 – Operationalisierung diskutiert werden.
- *Die Hörtexte sind ähnlich bzw. lassen sich auf vergleichbare Textsorten beziehen:* Es wird jeweils darauf geachtet, verschiedene Arten von Hörtexten zu berücksichtigen, insbesondere Varianten expositorischer und literarischer Texte. Bei der Auswahl wird darauf geachtet, dass die Texte den Hörerfahrungen von Kindern entsprechen (z. B. Hörspiele, Radiosendungen). Es werden Hörtexte eingesetzt, in denen konzeptionell schriftliche Sprache im Medium der Mündlichkeit zum Tragen kommt (Koch & Österreicher, 1985). Mitunter sind auch Interview-Passagen Teil der Testung, in denen konzeptionell mündliche Sprache dominant ist, z. B., wenn in Features oder Radiosendungen „O-Töne" eingespielt werden. Auf alltagsnahe Gespräche oder Gesprächsmitschnitte wird verzichtet, weil diese normalerweise Partizipation am Gespräch voraussetzen und nicht von CD konsumiert werden. In der Überprüfung der Bildungsstandards 2015 durch das BIFIE (vgl. Kapitel 1 – Überblick) wurden so bezeichnete semi-authentische Texte eingesetzt, die zwar auf authentischen Texten beruhen, aber für die Erhebung adaptiert und aufgenommen wurden (Thomas et al., 2011b, S. 13). Es handelte sich jedoch letztlich um eine nichtauthentische Hörsituation. Die Rezeption von *gescripteten* Texten, die den Eindruck von Authentizität erwecken sollen, ist vor allem im muttersprachlichen Unterricht ungewöhnlich, weshalb auf diese Texte hier verzichtet wird.
- *Die Grundlagen der Testaufgabenerstellung sind vergleichbar:* Man unterscheidet z. B. gängige geschlossene und offene Formate. Durch die Einbeziehung von offenen Aufgaben ergeben sich Möglichkeiten zur sprachproduktiven Auseinandersetzung mit den Hörtexten. Weitere Informationen zur Entwicklung der Items werden ebenfalls in Kapitel 3 – Operationalisierung dieses Berichtsbands besprochen.
- *Die Funktion der Modelle und ihr Entstehungskontext sind vergleichbar:* Die Operationalisierungen sind dazu gedacht, Kompetenzstufenmodelle zu Bildungsstandards in den jeweiligen Ländern vorzulegen. Die Kompetenzstufen ergeben sich dabei aus der Analyse der nach Lösungshäufigkeit sortierten Testitems (Standard-Setting) durch Expertinnen und Experten. Dabei werden jeweils ähnlich schwierige Items hinsichtlich der jeweiligen kognitiven Anforderungen beurteilt und zu Kompetenzstufen oder -niveaus gruppiert (Luger-Bazinger, Freunberger & Itzlinger-Bruneforth, 2016).
- *Die Studien sind in ihrer Datengrundlage vergleichbar:* Die Modelle zu Bildungsstandards in den jeweiligen Ländern beruhen auf empirischen Large-Scale-Assessments und Skalierungen aus repräsentativen Stichproben. Es wird angenommen, dass mit steigender Schülerfähigkeit die Wahrscheinlichkeit steigt, schwierige Items zu lösen. Damit geht einher, dass Schüler/innen mit höherer Wahrscheinlichkeit leichtere Items und mit niedrigerer Wahrscheinlichkeit schwierigere Items lösen können.

Bei der Setzung und Bezeichnung von Stufen gibt es allerdings einen erheblichen Unterschied zwischen den Modellen der jeweiligen Länder, obwohl sich auf vergleichbare Theorien gestützt wird. Es geht um die Frage, ob kurze Zusammenfassungen der Kompetenzstufenbeschreibungen (im Sinne von „Kurzlabels") vorliegen oder nicht. Das Formulieren von solchen kurzen Beschreibungen hat sowohl Vor- als auch Nachteile: Um die „Abnehmerzugänglichkeit" (z. B. aufseiten der Lehrer/innen) zu erhöhen, helfen solche „Kurzlabels" oder kurze Zusammenfassungen der Kompetenzstufenbeschreibungen. Die Modelle werden so „greifbarer". Das Problem mit der Verkürzung ist aber, dass viele Aufgaben und Items dann nicht mehr zu den Bezeichnungen der Stufen passen. Insofern könnten solche Verkürzungen auch zu größerer Unzufriedenheit im Umgang mit den Modellen führen, etwa wenn Rückmeldungen an Lehrkräfte in den Schulen gegeben werden. Vor diesem Hintergrund wurde bei der Entwicklung der Kompetenzstufenmodelle in Österreich folgendermaßen vorgegangen: Das Kompetenzstufenmodell weist für jede Stufe die in Abbildung 2 genannten Beschreibungen aus, über die Schüler/innen typischerweise verfügen. Es wird dann zwischen vier Stufen unterschieden: „Stufe 3: Bildungsstandards übertroffen", „Stufe 2: Bildungsstandards erreicht", „Stufe 1: Bildungsstandards teilweise erreicht" und „unter Stufe 1: Bildungsstandards nicht erreicht" (BIFIE, 2016).

Hören: inhaltliche Beschreibung der einzelnen Kompetenzstufen

3	**Bildungsstandards übertroffen**	**ab 675 Punkten**

Die Schüler/innen verfügen über alle Kompetenzen der Stufe 2, darüber hinaus verstehen sie auch Hörtexte, deren Themenfelder über den vertrauten Kontext hinausgehen. Sie können auch weniger zentrale Detailinformationen entnehmen, nicht nur das Thema, sondern auch den Aufbau des Hörtexts erkennen und die Sprecherintention deuten. Die Schüler/innen können komplexere Schlüsse ziehen und ihre Reflexionen mit ihrem Vorwissen in Beziehung setzen.

2	**Bildungsstandards erreicht**	**470–674 Punkte**

Die Schüler/innen sind in der Lage, altersangemessene monologische und dialogische Hörtexte (fiktionale und nichtfiktionale Textsorten) zu verstehen, wenn die Themen aus vertrautem Kontext stammen. Sie können wichtige Detailinformationen entnehmen, textnahe Schlüsse ziehen und das Thema des Texts erkennen. Die Schüler/innen können die aufgenommenen Informationen eigenständig vergleichen und beurteilen, auch im Hinblick auf die Textfunktion und die Handlungsmotive. Sie können in Dialogen die Angemessenheit von Sprechhandlungen erkennen und einschätzen und greifen dazu auch auf parasprachliche Hinweise zurück.

1	**Bildungsstandards teilweise erreicht**	**378–469 Punkte**

Die Schüler/innen sind in der Lage, altersangemessene monologische und dialogische Hörtexte (fiktionale und nichtfiktionale Textsorten) zu verstehen, wenn die Themen aus vertrautem Kontext stammen. Die Schüler/innen können die Bedeutung ähnlich klingender Wörter im Satzkontext unterscheiden. Sie können wichtige und prominent präsentierte Detailinformationen entnehmen und textnahe Schlüsse ziehen. Sie können das Thema des Texts nicht immer sicher erkennen. Die Schüler/innen können die aufgenommenen Informationen beurteilen.

unter 1	**Bildungsstandards nicht erreicht**	**bis 377 Punkte**

Abbildung 2: Kompetenzstufen (BIFIE, 2016)

Kompetenzstufenmodelle zum Zuhören haben innerhalb der Fachdidaktik zu einer erheblichen Validitätsdiskussion geführt (Bredel, 2014): Alle Modelle zum Hörverstehen im deutschsprachigen Raum haben nämlich eine große Ähnlichkeit mit Modellen zum Leseverstehen (Knoepke et al., 2013). Vor diesem Hintergrund hat Böhme (2012, S. 92) mit Daten aus Deutschland eine vergleichende Analyse aus der Normierungsstudie für die Bildungsstandards in der Grundschule durchgeführt und gezeigt,

> dass sowohl für das Hör- als auch für das Leseverstehen allgemeine Aufgabenmerkmale, die sich auf das Textverstehen beziehen, bei der Erklärung der Itemschwierigkeit einen wichtigen Stellenwert einnehmen. Um größere Teile der Varianz in der Itemschwierigkeit erklären zu können, sind jedoch Merkmale nötig, die spezifische Eigenschaften des Hör- beziehungsweise Leseverstehens beschreiben.

Neumann (2012) hat zudem Aspekte von Zuhöraufgaben aus der Normierung der KMK-Bildungsstandards für den Mittleren Schulabschluss in Deutschland auf ihre Spezifik und Schwierigkeit im Vergleich zum Lesen untersucht. Aufseiten der Hörtexte haben sich die „Sprechgeschwindigkeit" und die „Worthäufigkeit" (der gleichen Wörter im Hörtext) als bedeutsames Kriterium für die Schwierigkeit herausgestellt. Aufseiten der Items ist es insbesondere das Aufgabenformat (offen vs. geschlossen), das sich auf die Schwierigkeit auswirkt.

Zuletzt haben Behrens et al. (2021) Validitätsfragen in den Mittelpunkt gestellt. In einer groß angelegten Studie wurde u. a. die Präsentation von Items und Stimuli manipuliert, sodass die Bedeutung der prosodischen Dimension gesprochener Texte differenzierter als bislang in die Modellierung von Zuhörfähigkeiten zu Bildungsstandards einbezogen wurde. Zusätzlich wurden die Daten hinsichtlich der Effekte der Itemdarbietung (gelesen vs. gehört), auch unterschieden nach der Art der Hörtexte und Aufgabenstellungen, ausgewertet. Eines der zentralen Ergebnisse ist, dass sich der Zusammenhang zwischen Leseverstehen und Hörverstehen verringert, wenn Items nur auf Basis prosodischer Merkmale wie Intonation oder Satzakzente gelöst werden können. Gleiches gilt, wenn die Items (wie die Hörtexte) nur noch als Audios angeboten werden. Vor diesem Hintergrund werden dort zwar Vorschläge für Kompetenzabstufungen formuliert, jedoch weitere Grundlagenforschung angemahnt.

Vor dem Hintergrund solcher Vorarbeiten werden im Folgenden Schlussfolgerungen für die Entwicklung der Tests gezogen.

2.4.2 Schlussfolgerungen für die Aufgabenentwicklung

Bei der Konstruktion der Tests wird unterschieden zwischen den Quellen der Information (den Hörtexten bzw. Stimuli) und den Fragestellungen und Arbeitsaufträgen („Items") an die Schüler/innen, mit denen dann die jeweilige Kompetenz gemessen wird (siehe oben). Mehrere Items und der Stimulus bilden zusammen eine Testaufgabe. Die Items sind durchwegs auf die Stimuli bezogen bzw. sollen nicht ohne den jeweiligen Stimulus zu lösen sein. Wesentlich ist, dass die Bearbeitung der Items auch selbst eine Schwierigkeit darstellt, die nicht zu unterschätzen ist. Die Schüler/innen müssen sich in der Testsituation sowohl (1) an Informationen aus dem Hörtext erinnern bzw. ein mentales Modell des Textes aufbauen als auch (2) Information im Rahmen eines Items verarbeiten. Beides spielt eine Rolle, wenn es um die Leistungen der Schüler/innen geht.

Zu (1): Mit Blick auf die Kompetenz, die diagnostiziert wird, ist also relevant, ob die Informationen im Hörtext für das Gesamtverständnis wesentlich sind oder nicht; wie

prominent die Information im Text ist. Zudem gibt es eine Reihe schwierigkeitsbestimmender Merkmale von Hörtexten, u. a. die Anzahl und Länge der Pausen, das Sprechtempo sowie prosodische Merkmale als Träger der Information u. v. m. (Buck, 2001). Ein Überblick über gängige Parameter ist in der folgenden Tabelle zu finden. Diese wurden als Orientierung auch für die Auswahl eher leichter und schwieriger Stimuli berücksichtigt.

Tabelle 3: Schwierigkeitssteuernde Aufgabenmerkmale nach Buck (2001, zitiert nach Behrens, Böhme und Krelle, 2009)

Schwierigkeitssteuernde Aufgabenmerkmale nach Buck (2001)		
	leicht	**schwierig**
Sprechtempo	langsam	schnell
Pausen	viele, längere	wenige, kurze
Aussprache, Dialekt	vertraut	unvertraut
Intonation	natürlich	unnatürlich, ungewöhnlich
Wörter	häufig verwendete	selten verwendete
Grammatik	schlicht	komplex
Sätze	linear verbunden	„verschachtelt“
Inhalte, Gedanken	ausdrücklich benannt	implizit, zu erschließen
Wiederholungen	hohe Redundanz	wenig oder keine Redundanz
Struktur	deutlich lineare Anordnung	nichtlineare Anordnung
Position von Beispielen	nach einer klaren Aussage	vor der eigentlichen Aussage
Thema	vertraut	unvertraut
verschiedene Dinge/Personen im Text	wenige, klar zu unterscheiden	viele, schlecht zu unterscheiden
Beziehungen von Elementen	festgelegt	veränderlich
Inhalt	konkret	abstrakt

Zu (2): Mit Blick auf das Aufgabenformat geht es sowohl um gegebene als auch erfragte Informationen. Wenn man beispielsweise erfragt, ob die eine oder andere Figur etwas im Text gesagt hat, wird die Existenz mehrerer Figuren präsupponiert (Bremerich-Vos & Böhme, 2009). Wie sich gesuchte und gegebene Informationen im Rahmen von Lesetestitems klassifizieren lassen, haben Kirsch, Jungeblut und Mosenthal bereits 1998 beschrieben. Für Hörverstehenstests lassen sich die Ergebnisse übertragen, u. a. mit Blick darauf, wie plausibel die Distraktoren, also die falschen Antwortoptionen sind oder wie konkret oder abstrakt die gesuchten und gegebenen Informationen sind etc. Schließlich spielt auch noch eine Rolle, wie anspruchsvoll das Auffinden einer oder mehrerer Informationen ist, welche Arten von Schlussfolgerungen im Spiel sind und welche Rolle das jeweilige Vor- und Sprachwissen spielt.

Ein wesentlicher Aspekt der Konstruktion von Items ist, auf welche kognitiven Aspekte des Hörverstehens sie sich beziehen lassen. Aus den oben formulierten theoretischen Grundlagen (siehe Abschnitt 2.2) leiten sich vier zentrale Anforderungen ab, die als „Subkompetenzen" für die Aufgabenentwicklung leitend sind und die sich auch als Fragen formulieren lassen.

- Können die Schüler/innen eine Information aus dem Hörtext im Rahmen eines Items wiedererkennen oder wiedergeben?
- Können sie darüber hinaus auf Basis von Welt- und Sprachwissen Schlussfolgerungen ziehen, die sich auf „benachbarte" Informationen im Hörtext beziehen?
- Können sie außerdem ihre Interpretationen nutzen, um den Text als Ganzen zu erfassen?
- Und: Können sie auf Aspekte des Textes bezogene Aussagen beurteilen, begründen und/oder reflektieren?

Es wird deutlich, dass sich solche Fähigkeiten der Unterscheidung von „hierarchieniedrigen" und „hierarchiehohen" Prozessen zuordnen lassen, wie sie oben diskutiert wurden. Als Subkompetenzen werden diese in der folgenden Abbildung 3 aufgelistet:

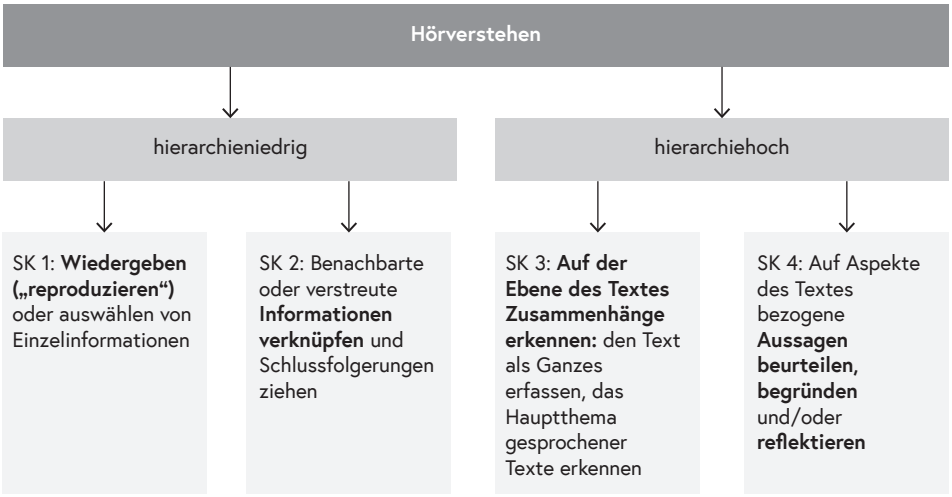


Abbildung 3: Subkompetenzen (SK) zum Testkonstrukt Hörverstehen

Eine solche Auflistung ist in Ergänzung zu den Formulierungen in den Bildungsstandards zu sehen (3.1), die mitunter auch als Könnensbeschreibungen bezeichnet werden (Längauer-Hohengaßner & Schaffer, 2013). Mit den aufgeführten Subkompetenzen sind auch keine Stufen- bzw. Niveauzuschreibungen gemeint. Es kann z. B. äußerst anspruchsvoll sein, eine Einzelinformation zu reproduzieren, wenn der Text lang und/oder die erfragte Information für das Gesamtverständnis nicht zentral ist. Andererseits können manche begründete Urteile sehr leicht zu fällen sein, sofern die Struktur des Textes solche Urteile erlaubt.

Die Subkompetenzen sind so gesehen Ausführungen zu den Bildungsstandards und dienen der Erleichterung der Item- und Aufgabenentwicklung, um die es in den folgenden Kapiteln gehen wird.

Literatur

Ackermann, M. (2006). Hörwörter – etymologisch. In V. Bernius, P. Kemper, R. Oehler & K.-H. Wellmann (Hrsg.), *Der Aufstand des Ohrs – die neue Lust am Hören* (S. 59–75). Göttingen: Vandenhoeck & Ruprecht.

Becker-Mrotzek, M. (2008). Gesprächskompetenz vermitteln und ermitteln. Gute Aufgaben im Bereich „Sprechen und Zuhören". In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Lernstandsbestimmung im Fach Deutsch. Gute Aufgaben für den Unterricht* (S. 52–77). Weinheim: Beltz.

Becker-Mrotzek, M. (2009). Mündliche Kommunikationskompetenz. In M. Becker-Mrotzek (Hrsg.), *Didaktik der mündlichen Kommunikation* (S. 66–83). Baltmannsweiler: Schneider Verlag Hohengehren.

Behrens, U. (2010). Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis. Edition Zuhören* (S. 31–50). Göttingen: Vandenhoeck & Ruprecht.

Behrens, U., Böhme, K. & Krelle, M. (2009). Zuhören – Operationalisierung und fachdidaktische Implikationen. In D. Granzer, O. Köller & A. Bremerich-Vos (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 357–376). Weinheim: Beltz.

Behrens, U., Käser-Leisibach, U., Krelle, M., Weirich, S. & Zingg Stamm, C. (2021). *stim·mig. Ergebnisse einer Studie zum Zuhören in der dritten Jahrgangsstufe.* Münster: Waxmann.

Behrens, U. & Krelle, M. (2014). Hörverstehen – ein Forschungsüberblick. *Didaktik Deutsch, 36,* 86–107. doi:10.25656/01:17208

BIFIE (Hrsg.). (2011). *Praxishandbuch für Deutsch, Lesen, Schreiben. 4. Schulstufe* (2., aktualisierte Fassung). Graz: Leykam. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/872cf2c46b548d7321a1174640f69dbe3a3bc4ab/bist_d_vs_praxishandbuch_deutsch_4_2011-08-22.pdf

BIFIE (2016). *Konstrukt- und Kompetenzstufenbeschreibung. Deutsch/Lesen/Schreiben 4. Schulstufe. Die Kompetenzstufen für die Überprüfung der Bildungsstandards* (2., aktualisierte Fassung). Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/fec772b59ab074e5cbe78f6278d1bcb2a8e6cad8/BIST-UE_D4_Konstruktbeschreibung.pdf

Böhme, K. (2012). *Methodische und didaktische Überlegungen sowie empirische Befunde zur Erfassung sprachlicher Kompetenzen im Deutschen. Analysen zu den Bildungsstandards im Fach Deutsch für den Primarbereich* (Dissertation). Berlin: Humboldt-

Universität zu Berlin. Verfügbar unter http://edoc.hu-berlin.de/dissertationen/boehme-katrin-2011-12-16/PDF/ boehme.pdf

Bredel, U. (2014). Anspruch und Wirklichkeit – Debattenbeitrag zu den Bildungsstandards. *Didaktik Deutsch, 36,* 5–8.

Breit, S., Bruneforth, M. & Schreiner, C. (Hrsg.). (2016). *Standardüberprüfung 2015. Deutsch, 4. Schulstufe. Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/0d879bba1b6fb81469e09bc99aab7e4916f88bef/BiSt_UE_D4_2015_Bundesergebnisbericht.pdf

Breit, S., Bruneforth, M. & Schreiner, C. (Hrsg.). (2017). *Standardüberprüfung 2016. Deutsch, 8. Schulstufe. Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/365d58afa26186150a317debd7f8f19c3d267ba2/BiSt_UE_D8_2016_Bundesergebnisbericht.pdf

Bremerich-Vos, A. (2009). Die Bildungsstandards Deutsch. In A. Bremerich-Vos, D. Granzer, U. Behrens & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret* (S. 14–42). Berlin: Cornelsen.

Bremerich-Vos, A., Behrens, U., Böhme, K., Krelle, M., Neumann, D., Robitzsch, A. et. al. (2010). Kompetenzstufenmodelle für das Fach Deutsch. In O. Köller, M. Knigge & B. Tesch (Hrsg.), *Sprachliche Kompetenzen im Ländervergleich* (S. 37–50). Münster: Waxmann.

Bremerich-Vos, A. & Böhme, K. (2009). Lesekompetenzdiagnostik – die Entwicklung eines Kompetenzstufenmodells für den Bereich Lesen. In D. Granzer, O. Köller & A. Bremerich-Vos (Hrsg.), *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule* (S. 219–249). Weinheim: Beltz.

Bremerich-Vos, A., Böhme, K., Krelle, M., Weirich, S. & Köller, O. (2012). Kompetenzstufenmodelle im Fach Deutsch. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011* (S. 56–71). Münster: Waxmann.

Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Sprachliche Kompetenzen im Fach Deutsch – Strukturanalysen und Validierungsbefunde. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 198–218). Weinheim: Beltz.

Brinker, K. & Sager, S. F. (2001) [1989]. *Linguistische Gesprächsanalyse* (3. Aufl.). Berlin: Schmidt.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Coninx, F. (2009). Hörüberprüfungen. Relevanz und Methoden. In M. Becker-Mrotzek (Hrsg.), *Mündliche Kommunikation und Gesprächsdidaktik. Handbuch Deutschunterricht in Theorie und Praxis* (DTP, Bd. 3, S. 458–468). Baltmannsweiler: Schneider Verlag Hohengehren.

EDK Schweizerische Konferenz der kantonalen Erziehungsdirektoren (Hrsg.). (2011). *Grundkompetenzen für die Schulsprache. Nationale Bildungsstandards.* Verfügbar unter https://edudoc.ch/record/96791?ln=de

Eriksson, B. (2009). Leistungsmessung. In M. Becker-Mrotzek (Hrsg.), *Mündliche Kommunikation und Gesprächsdidaktik. Handbuch Deutschunterricht in Theorie und Praxis* (DTP, Bd. 3., S. 445–457). Baltmannsweiler: Schneider Verlag Hohengehren.

Eriksson, B., Lindauer, T. & Sieber, P. (2008). HarmoS Schulsprache – Kompetenzbeschreibungen und Basisstandards. B*eiträge zur Lehrerbildung BzL, 3,* 338–350. doi:10.25656/01:13684

Fiehler, R. (Hrsg.). (1998). *Verständigungsprobleme und gestörte Kommunikation.* Opladen: Westdeutscher Verlag. Verfügbar unter http://www.verlag-gespraechsforschung.de/2002/probleme/probleme.pdf

Grotjahn, R. (2000). *Testen der Fertigkeit Hörverstehen. Leistungsmessung und Leistungsbeurteilung.* Studieneinheit. Hellenic Open University. Verfügbar unter https://docplayer.org/18084955-Leistungsmessung-und-leistungsbeurteilung.html

Grotjahn, R. (2005). Testen und Bewerten des Hörverstehens. In M. ó Dúill, R. Zahn & K. D. C. Höppner (Hrsg.), *Zusammenarbeiten. Eine Festschrift für Bernd Voss* (S. 115–144). Bochum: AKS-Verlag.

Gschwend, R. (2014). Zuhören und Hörverstehen – Aspekte, Ziele und Kompetenzen. In E. Grundler & C. Spiegel (Hrsg.), K*onzeptionen des Mündlichen* (S. 141–158). Bern: hep-Verlag.

Hagen, M. (2006). *Förderung des Hörens und Zuhörens in der Schule.* Göttingen: Vandenhoeck & Ruprecht.

Hagtvet, B. E. (2003). Listening comprehension and reading comprehension in poor decoders. Evidence for the importance of syntactic and semantic skills as well as phonological skills. *Reading and Writing, 16,* 505–539. doi:10.1023/A:1025521722900

Hartung. W. (1998). Perspektiven-Divergenzen als Verständigungsproblem. In R. Fiehler (Hrsg.), *Verständigungsprobleme und gestörte Kommunikation* (S. 63–79). Opladen: Westdeutscher Verlag. Verfügbar unter http://www.verlag-gespraechsforschung.de/2002/probleme/063-079.pdf

Hartung, M. (2004). Wie lässt sich Gesprächskompetenz wirksam und nachhaltig vermitteln? Ein Erfahrungsbericht aus der Praxis. In M. Becker-Mrotzek & G. Brünner (Hrsg.), *Analyse und Vermittlung von Gesprächskompetenz* (S. 47–66). Frankfurt am Main: Lang. Verfügbar unter http://www.verlag-gespraechsforschung.de/2004/becker1.htm

Hellbrück, J. & Ellermeier, W. (2004). *Hören. Physiologie, Psychologie und Pathologie* (2., aktualisierte und erweiterte Aufl.). Göttingen: Hogrefe.

Hennon, E., Hirsh-Pasek, K. & Golinkoff, R. (2000). Die besondere Reise vom Fötus zum spracherwerbenden Kind. In H. Grimm (Hrsg.), *Sprachentwicklung* (S. 41–103). Göttingen: Hogrefe.

Imhof, M. (2003). *Zuhören. Psychologische Aspekte auditiver Informationsverarbeitung.* Göttingen: Vandenhoeck & Ruprecht.

Imhof, M. (2010). Zuhören lernen und lehren. Psychologische Grundlagen zur Beschreibung und Förderung von Zuhörkompetenzen in Schule und Unterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis. Edition Zuhören.* Bd. 6 (S. 15–30). Göttingen: Vandenhoeck & Ruprecht.

Institut zur Qualitätsentwicklung im Bildungswesen (Hrsg.). (2013). *Kompetenzstufenmodell zu den Bildungsstandards im Kompetenzbereich Sprechen und Zuhören – Primarbereich. Auf Grundlage des Ländervergleichs 2011 überarbeiteter Entwurf in der Version vom 13. Februar 2013.* Verfügbar unter www.iqb.hu-berlin.de/bista/ksm

Kahlert, J. (2000). Der gute Ton in der Schule. Überlegungen zum pädagogischen Stellenwert des Zuhörens in der akustisch gestalteten Schule. In L. Huber & E. Odersky (Hrsg.), *Zuhören – Lernen – Verstehen* (S. 7–25). Braunschweig: Westermann.

Kintsch, W. (1998). *Comprehension. A paradigm for cognition.* Cambridge: Cambridge University Press.

Kirsch, I. S., Jungeblut, A. & Mosenthal, P. B. (1998). The measurement of adult literacy. In T. S. Murray, I. S. Kirsch & L. Jenkins (Hrsg.), *Adult literacy in OECD countries. Technical report on the first international adult literacy survey* (S. 105–134). Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel et al. (2003). *Zur Entwicklung nationaler Bildungsstandards. Eine Expertise.* Bonn: BMBF.

Knechtel, N. (2011). Mündliche und schriftliche Texte verstehen. Hör- und Leseverstehen in der Primarschule im Vergleich. In U. Behrens & B. Eriksson (Hrsg.), *Sprachliches Lernen zwischen Mündlichkeit und Schriftlichkeit* (S. 125–143). Bern: hep-Verlag.

Knoepke, J., Richter, T., Isberner, M.-B., Neeb, Y. & Naumann, J. (2013). Leseverstehen = Hörverstehen X Dekodieren? Ein stringenter Test der Simple View of Reading bei deutschsprachigen Grundschulkindern. In A. Redder & S. Weinert (Hrsg.), *Sprachförderung und Sprachdiagnostik. Interdisziplinäre Perspektiven* (S. 256–276). Münster: Waxmann.

Koch, P. & Österreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch, 36,* 15–43. doi:10.1515/9783110244922.15

Krelle, M. (2010). Zuhördidaktik – Anmerkungen zur Förderung rezeptiver Fähigkeiten des mündlichen Sprachgebrauchs im Deutschunterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis. Edition Zuhören.* Bd. 6 (S. 51–69). Göttingen: Vandenhoeck & Ruprecht.

Krelle, M. & Neumann, D. (2014). Sprechen und Zuhören. In U. Behrens, A. Bremerich-Vos., M. Krelle, K. Böhme & S. Hunger (Hrsg.), *Bildungsstandards für die Sekundarstufe 1. Deutsch konkret* (S. 14–45). Berlin: Cornelsen.

Krelle, M. & Prengel, J. (2014). Zur Konzeption von Zuhören im Rahmen der Vergleichsarbeiten für die dritte Klasse im Fach Deutsch. In E. Grundler & C. Spiegel (Hrsg.), *Konzeptionen des Mündlichen* (S. 210–228). Bern: hep-Verlag.

Kürschner, C. & Schnotz, W. (2008). Das Verhältnis gesprochener und geschriebener Sprache bei der Konstruktion mentaler Repräsentationen. *Psychologische Rundschau, 59* (3), 139–149. doi:10.1026/0033-3042.59.3.139

Kürschner, C., Schnotz, W. & Eid, M. (2006). Konstruktion mentaler Repräsentationen beim Hör- und Leseverstehen. *Zeitschrift für Medienpsychologie, 18* (2), 48–59. doi:10.1026/1617-6383.18.2.48

Längauer-Hohengaßner, H. & Schaffer, J. (2013). *Kompetenzbereich Hören. BIST-UE-D8H Begleittext Itemauswahl.* Unveröffentlichtes Manuskript, Salzburg.

Lipkowski, E. & Schüller, L. (2017). *Perspektiven für den inklusiven Deutschunterricht. Literatur- und sprachdidaktische Praxisbeispiele zum Thema Hören und Hören-schränkungen.* Münster: Waxmann.

Luger-Bazinger, C., Freunberger, R. & Itzlinger-Bruneforth, U. (2016). Standard-Setting. In C. Schreiner & S. Breit (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfungen* (S. 83–110). Wien: facultas.

Marx, H. (2000). Einfluss von Lehrmethoden und Hörverstehen auf das Lesenlernen im Grundschulalter. In R. Duit & C. von Rhöneck (Hrsg.), *Ergebnisse fachdidaktischer und psychologischer Lehr-Lern-Forschung. Beiträge zu einem Workshop an der Pädagogischen Hochschule Ludwigsburg* (S. 191–231). Kiel: IPN Institut für die Pädagogik der Naturwissenschaften, 2000, [Serie IPN, Nr. 169], 49.

Marx, H. & Jungmann, T. (2000). Abhängigkeit der Entwicklung des Leseverstehens von Hörverstehen und grundlegenden Lesefertigkeiten im Grundschulalter. Eine Prüfung des Simple View of Reading-Ansatzes. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 32* (2), 81–93. doi:10.1026//0049-8637.32.2.81

Neumann, D. (2012). *Schwierigkeitsbeeinflussende Merkmale bei Aufgaben zum Hörverständnis im Fach Deutsch in der Sekundarstufe I. KÖBES. Kölner Beiträge zur Sprachdidaktik.* Bd. 8. Duisburg: Gilles und Francke.

Porsch, R., Grotjahn, R. & Tesch, B. (2010). Hörverstehen und Hör-Sehverstehen in der Fremdsprache – unterschiedliche Konstrukte? *Zeitschrift für Fremdsprachenforschung, 21* (2), 143–189.

Pröscholdt, M. V., Michalik, A., Schneider, W., Duzy, D., Glück, D., Souvignier, E. et al. (2013). Effekte kombinierter Förderprogramme zur phonologischen Bewusstheit und zum Sprachverstehen auf die Entwicklung der phonologischen Bewusstheit von Kindergartenkindern mit und ohne Migrationshintergrund. *Frühe Bildung, 2* (3), 1–11. doi:10.1026/2191-9186/a000099

Rösch, H. (2011). *Deutsch als Fremd- und Zweitsprache.* Studienbuch. Berlin: Akademie Verlag.

Rogers, C. R. (1985). *Die nicht-direktive Beratung. Counseling and Psychotherapy.* Frankfurt am Main: Fischer.

Schwarz, M. (1992). *Einführung in die Kognitive Linguistik.* Tübingen: Francke.

Spiegel, C. (2006). Heißt Kommunizieren etwa auch Zuhören? In R. Wagner, A. Brunner & S. Voigt-Zimmermann (Hrsg.), *hören – lesen – sprechen* (S. 153–162). München: Reinhardt.

Spiegel, C. (2009). Zuhören im Gespräch. In M. Krelle & C. Spiegel (Hrsg.), *Sprechen und Kommunizieren. Entwicklungsperspektiven, Diagnosemöglichkeiten und Lernszenarien in Deutschdidaktik und Deutschunterricht* (S. 189–203). Baltmannsweiler: Schneider Verlag Hohengehren.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011.* Münster: Waxmann.

Thomas, A., Latschen D., Frank, E., Weishaupt, A. & Weishaupt, J. (2011a). *Bildungsstandards Hören. 4. Zwischenbericht Juni 2010.* Unveröffentlichtes Manuskript, Salzburg.

Thomas, A., Latschen, D., Weishaupt, A., Weishaupt, J., Frank, E. & Unterköfler-Klatzer, D. (2011b). Item-Generation für die Bildungsstandards D4 Hören. *Forschungszeitung, 4/5,* 12–15. Verfügbar unter https://www.ph-kaernten.ac.at/fileadmin/media/forschung/Forschungszeitung4.pdf

Waibel, S. & Eriksson, B. (2010). Bildungsstandards Zuhören – ein Bericht aus dem Schweizer Bildungsstandard-Projekt HarmoS. In V. Bernius & M. lmhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis. Edition Zuhören.* Bd. 6 (S. 69–80). Göttingen: Vandenhoeck & Ruprecht.

Wagner, Elvis (2007). Are they watching? Test-taker viewing behavior during an L2 video listening test. *Language Learning and Technology, 11* (1), 67–86. doi:10125/44089

Worthington, D. L. & Bodie, G. D. (2017). Defining Listening: A Historical, Theoretical, and Pragmatic Assessmen. In D. L. Worthington & G. D. Bodie (Hrsg.), *The Sourcebook of Listening Research. Methodology and Measures* (S. 3–17). Hoboken: Wiley.

Zingg Stamm, C., Käser-Leisibach, U., Behrens, U., Krelle, M. & Weirich, S. (2016). Neue Aufgabenformate für die Messung von Zuhörkompetenzen. In S. Keller & C. Reintjes (Hrsg.), *Aufgaben als Schlüssel zur Kompetenz* (S. 129–140). Münster: Waxmann.

Maria Engelbert-Kocher, Antonia Bachinger

Kapitel 3
Konstruktoperationalisierung: Testaufgabenentwicklung und -durchführung

Auf der Grundlage der österreichischen Bildungsstandards für das Fach *Deutsch, Lesen, Schreiben* in der 4. Schulstufe wurde unter Einbezug des aktuellen Forschungsstands ein Konstrukt dessen entwickelt, was theoretisch unter der Kompetenz des *Zuhörens* als Teil des Kompetenzbereichs *Hören, Sprechen und Miteinander-Reden* verstanden wird. Bei der Definition des Begriffs Zuhören müssen verschiedene Fähigkeiten unterschieden werden: zum einen der physische Vorgang des Hörens, der einen funktionierenden Hörapparat voraussetzt (Behrens, 2010), zum anderen der komplexe Vorgang der mehrstufigen Informationsverarbeitung, also des Zuhörens, der auf weitere Fähigkeiten zur Konzentration und Aufmerksamkeitssteuerung zugreift (Imhof, 2010) und ferner kommunikative Aspekte des Zuhörens. Häufig wird für die Informationsverarbeitung auch der Begriff des Hörverstehens verwendet. In Kapitel 2 – Konstrukt dieses Berichtsbands werden die Definition und die Weiterentwicklung des Konstrukts zum Hörverstehen sowie das zugrundeliegende Modell ausführlich beschrieben. In dem Konstrukt werden Aspekte von Zuhörkompetenz beschrieben, die im Rahmen von Leistungsstudien messbar gemacht werden sollen, indem die beschriebenen Kompetenzen mithilfe von Testaufgaben operationalisiert werden. Das zugrundeliegende Modell bietet gleichzeitig die Möglichkeit, Erkenntnisse darüber zu gewinnen, inwieweit die konstruierten Testinstrumente das messen, was sie messen sollen, sowie die Möglichkeit, die erhobenen Schülerleistungen auf Kompetenzstufen abzubilden.

In dem vorliegenden Kapitel werden zwei Phasen der Pilotierungsstudie näher vorgestellt. Zunächst wird der Prozess der Testaufgabenentwicklung, also der Operationalisierung des Konstrukts zu den Bildungsstandards in Testaufgaben, ausführlich beleuchtet. Im zweiten Teil werden Vorbereitung und Realisierung der Untersuchung beschrieben.

3.1 Anforderungen an Testaufgaben

Mit der regelmäßigen Überprüfung der österreichischen Bildungsstandards in verschiedenen Bereichen der Sprachkompetenz sollen die Qualitätssicherung an den Schulen und die Weiterentwicklung des Unterrichts evidenzbasiert unterstützt werden (BIFIE, 2016). Die im Jahr 2018 durchgeführte Untersuchung separierte die Messungen für die Kompetenzbereiche *Hören, Lesen, Rechtschreiben, Sprachbetrachtung* sowie *Verfassen von*

Texten (Textproduktion) in verschiedene Tests, die jeweils mit gesondertem Testmaterial erhoben wurden. Die Ausführungen in diesem Berichtsband fokussieren ausschließlich und eingehend auf die Kompetenz des *Zuhörens*. Die Ziele der D4-Studie 2018 wurden mit zweierlei Instrumenten verfolgt: Für die Überprüfung des Modells sowie für die Messung der Zuhörkompetenz, also die Erhebung von Leistungsdaten der Schüler/innen, wurde ein umfangreicher Pool von Testaufgaben entwickelt, von denen jede Schülerin/jeder Schüler je einen kleinen Teil in Einzelarbeit bearbeitete. Für die Erfassung der Hörgewohnheiten und den Einsatz von Zuhörstrategien im schulischen Kontext und in der Freizeit wurden eigens Schüler- und Lehrerfragebögen entwickelt. Näheres dazu findet sich im Abschnitt *3.3.1 Überblick über die Erhebungsinstrumente* in diesem Kapitel. Die Entwicklung der Testaufgaben und Fragebögen wurde vom Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS), ehemals BIFIE[1], koordiniert und in Zusammenarbeit mit Expertinnen und Experten der Fachdidaktik der Technischen Universität Chemnitz und mit österreichischen Lehrkräften realisiert. Die Ergebnisse aus den Fragebögen werden zukünftig an anderer Stelle veröffentlicht.

Die Testaufgaben dienen dazu, das Testkonstrukt zu den Bildungsstandards zu operationalisieren. Dementsprechend gelten für Testaufgaben besondere Anforderungen, die sich z. B. von Unterrichtsaufgaben unterscheiden. Testaufgaben sollen so konstruiert sein, dass die Schüler/innen ihre Fähigkeiten im Bereich *Hörverstehen* in der Testsituation zeigen können und bereit sind, dies zu tun. Im Unterschied zu Klassen- oder Prüfungsaufgaben im unterrichtlichen Kontext, die lerngruppenintern am Ende eines Lernprozesses stehen und ausschließlich für diese Zwecke und individualdiagnostisch erarbeitet werden, müssen Testaufgaben für Testzwecke geeignet sein und anderen Kriterien genügen. Einen Überblick über gängige Einteilungen von Aufgaben nach ihrer Funktion findet man z. B. bei Abraham und Müller (2009).

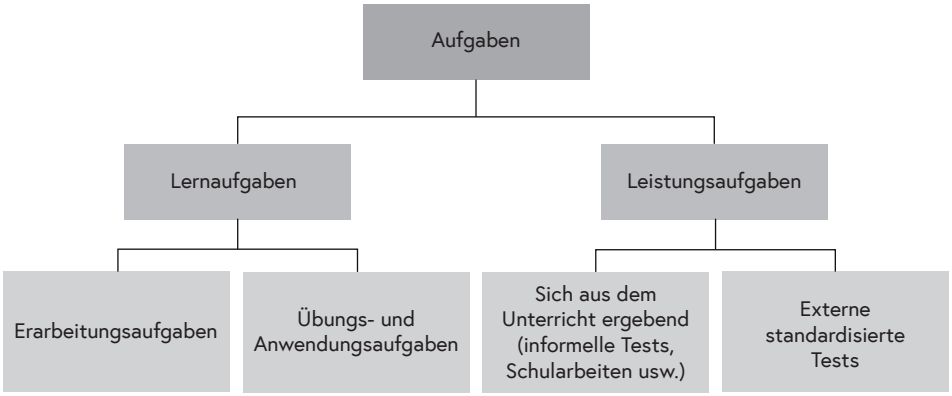


Abbildung 1: Zur Einteilung von Aufgaben nach ihrer Funktion in Lehr-Lern-Prozessen (gekürzt und angelehnt an Abraham & Müller, 2009, S. 6)

1 Aus dem Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) wurde am 1. Juli 2020 das Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS).

Bei den in diesem Bericht vorgestellten Aufgaben geht es um externe standardisierte Tests. Sie beziehen sich, anders als z. B. internationale Large-Scale-Studien wie PIRLS oder PISA, auf die in den Bildungsstandards beschriebenen Kompetenzen, sind wie typische Testaufgaben stark standardisiert und erfüllen inhaltliche und psychometrische Qualitätsmerkmale, um zuverlässige Messungen von Kompetenzen zu ermöglichen. Dabei wird üblicherweise auf Gütekriterien verwiesen, die solche Tests erfüllen müssen: eine hinreichende *Objektivität, Reliabilität* und *Validität*. Die *Objektivität* definiert sich dadurch, dass das Testergebnis unabhängig von der Person ist, die den Test durchführt, auswertet und interpretiert. Es darf also möglichst wenig Ermessensspielraum für den Testdurchführenden geben bzw. sind klare Vorgaben notwendig, die die gezeigten Leistungen als korrekt oder als falsch interpretieren (Moosbrugger & Kelava, 2012). Die *Reliabilität* bezeichnet, ob ein Test zuverlässig und mit möglichst geringen Messfehlern misst. Diese sind bei psychologischen Tests nicht ganz auszuschließen (Eid & Schmidt, 2014). Zu den wesentlichen Kriterien standardisierter Tests zählt zudem die *Validität,* deren Anspruch es ist, dass der Test auch wirklich das gewünschte Merkmal misst und nichts anderes, und dass die Schlussfolgerungen, die aus den Testergebnissen gezogen werden, zutreffend und somit gültig sind (Eid & Schmidt, 2014). Validität ist keine unveränderliche Eigenschaft des Testinstruments und an den Testzweck gebunden.

Solche Kriterien wurden in jedem Schritt des Testentwicklungsprozesses mitgedacht. Dies gilt auch für Nebentestgütekriterien wie Fairness, Zumutbarkeit etc. Am Ende des komplexen Aufgabenentwicklungsprozesses liegt nun ein geprüfter und konstruktabdeckender Aufgabenpool mit aktuellen Parametern vor, aus dem zukünftige Tests mit entsprechendem Sampling, Testdesign etc. gespeist werden können.

3.2 Testaufgabenentwicklung im Bereich Hörverstehen

3.2.1 Grundüberlegungen und Vorgehen

Für die Entwicklung von eigenen Testaufgaben im Bereich *Hörverstehen* liegen bereits einige Erfahrungen im deutschsprachigen Bereich der letzten Jahre vor, auf die sich die Expertise der einbezogenen Fachdidaktik stützt[2] (zu den Vorarbeiten des IQS, ehemals BIFIE, siehe Kapitel 2 – Konstrukt). Auf Basis dieser Erfahrungen und zur Qualitätssicherung wurde ein mehrschrittiger Prozess zur Aufgabenentwicklung festgelegt, der in Abbildung 2 dargestellt ist.

2 Die in diesem Kapitel beschriebene Testaufgabenentwicklung bezieht sich ausschließlich auf Testaufgaben, die zur Operationalisierung des weiterentwickelten Konstrukts erstellt wurden. Für nähere Informationen zum Konstrukt siehe Kapitel 2 – Konstrukt, für nähere Informationen zur Verlinkung von Item- und Personenparametern zwischen der BIST-Ü aus dem Jahr 2015 und der vorliegenden Pilotierungsstudie aus dem Jahr 2018 siehe Kapitel 6 – Verlinkung. Eine Übersicht über die referierten Studien bietet das Kapitel 1 – Überblick.

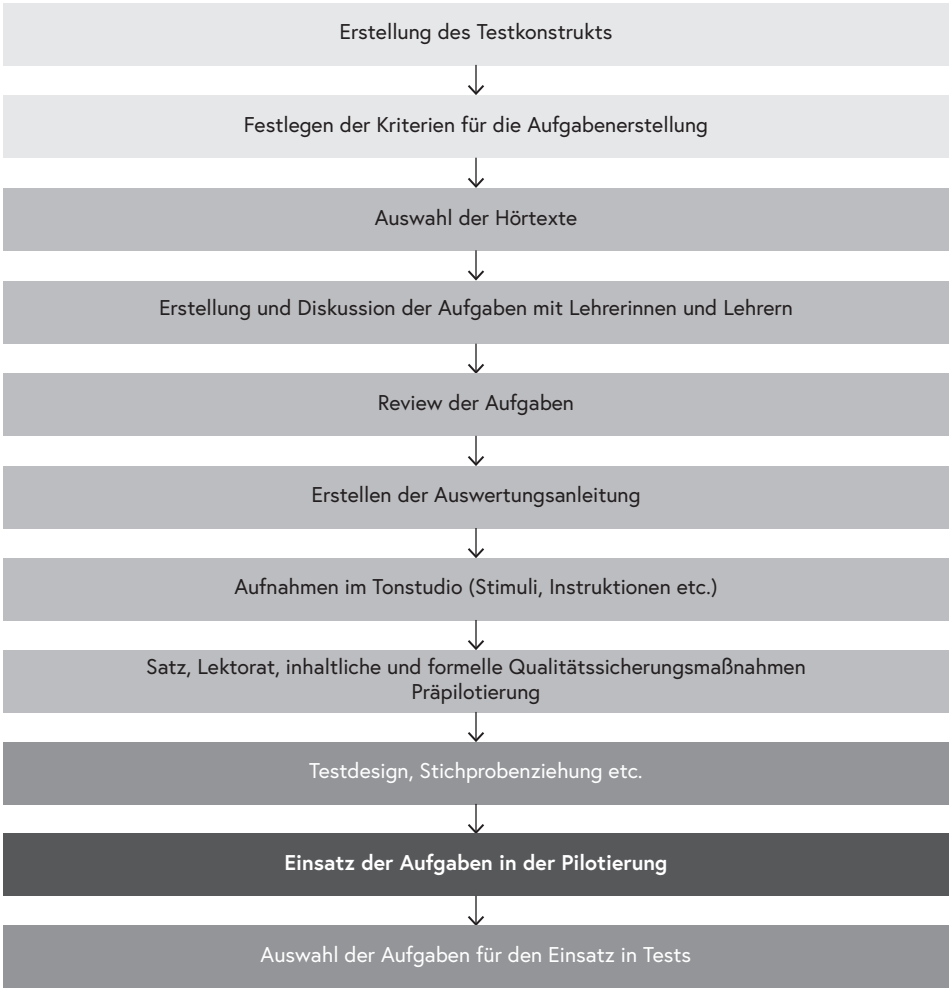


Abbildung 2: Testaufgabenentwicklungsprozess

In verschiedenen Projektphasen zur Erarbeitung des Testaufgabenpools wurden neben der Projektgruppe verschiedene Personen mit einschlägigen Erfahrungen eingebunden. Bei der Recherche nach geeigneten Hörtexten unterstützten Dr. Doris Rudlof-Garreis, Judith Neichl und Dr. Thomas Horwath. Mit der Entwicklung und Optimierung der Testaufgaben unter fachdidaktischer Leitung wurden Volkschullehrerinnen betraut, um das fachbezogene Professionswissen zu nutzen und die curriculare Validität der Testaufgaben zu sichern. Beteiligt waren dabei Andrea Greil, Johanna Leitner, Barbara Trinker, Gabriele Wagner und Martina Weiland, die neben Anna Windhofer auch die Prüfung der Fragebogeninstrumente übernahmen. Die Illustrationen der Testaufgaben erstellte Andreas Paar. Für die Tonaufnahmen und Montage der Hörtexte zeichnet Michele Gaggia von Digital Natural Sound verantwortlich.

Ausgehend vom Konstrukt und allgemeinen Richtlinien zur Testaufgabenentwicklung (z. B. Moosbrugger & Kelava, 2012; Pospeschill, 2010) wurden Leitfäden erarbeitet, die bei der Entwicklung von Testaufgaben dienlich sein und den umfassenden Prozess begleiten sollten. Dazu gehörten beispielsweise Kriterien für die Auswahl von Hörtexten, Richtlinien zur Erarbeitung von Testaufgaben oder Hinweise zu den Testgütekriterien usw. Für die Suche nach geeigneten Hörtexten wurden die Aufgabenentwickler/innen und Expertinnen und Experten für den österreichischen Kinderradio- und Hörbuchmarkt in mehreren Workshops unter Leitung von Fachdidaktikerinnen und Fachdidaktikern auf ihre Aufgabe vorbereitet und begleitet. Um das Feld potenziell geeigneter Hörtexte zu vergrößern, wurden auch schriftliche Texte gesucht und später im Tonstudio professionell vertont. In mehreren Durchgängen wurden die Testaufgaben anschließend nach psychometrischen und fachdidaktischen Kriterien überprüft und optimiert.

In anschließenden Workshops zur Testaufgabenentwicklung wurden zunächst Konstruktionsprinzipien vorgestellt, konkrete bewährte Formulierungen für Testfragen an die Hand gegeben und die ausgewählten Hörtexte zur Weiterarbeit an die Aufgabenentwickler/innen verteilt. Diese erarbeiteten in der nächsten Arbeitsphase für jeden Hörtext Testfragen und stellten sie auf einem gesicherten Portal zur Bewertung zur Verfügung. Die Expertinnen und Experten der Fachdidaktik prüften die Testaufgabenentwürfe und gaben sowohl zu einzelnen Testfragen als auch zu allgemeinen beobachteten Schwierigkeiten bei der Aufgabenentwicklung Rückmeldung an die Aufgabenentwickler/innen, auf deren Basis diese die Testaufgaben weiter optimieren konnten. In weiteren Workshops wurden die Aufgabenentwürfe intensiv diskutiert, überarbeitet und fertiggestellt. Details zur Konstruktion der Testfragen finden sich im Abschnitt 3.2.5 dieses Kapitels.

Im Mai 2018 konnten in der anschließenden Erprobung der Testaufgaben, der Pilotierung, an ca. 2.800 Schülerinnen und Schülern psychometrische Informationen über die Testaufgaben und Fähigkeitsverteilungen der Schüler/innen gewonnen werden. Die Auswertung der Pilotierung zeigte auch Fehlkonstruktionen einzelner Testaufgaben auf, die folglich für die weitere Verwendung des Tests überarbeitet oder aus dem Aufgabenpool gänzlich entfernt wurden.

Aus der Konzeption der Bildungsstandards als Beschreibung von gewünschten Lernergebnissen am Ende einer Schulart ergibt sich als geeigneter Messzeitpunkt das Ende der 4. Schulstufe. Eine weitere Grundüberlegung betrifft den Modus des Tests. Aus testorganisatorischen Überlegungen, die unter anderem die Ausstattung der Volksschulen betreffen, wurden ausschließlich Papier-Stift-Tests verwendet. Aktuell werden Testverfahren ausgelotet, die digital durchgeführt werden und damit die Möglichkeit eröffnen, adaptiv zu testen, das heißt, auf die Testteilnehmerin/den Testteilnehmer während der Untersuchung angepasste Testaufgaben zu präsentieren. Ermüdungserscheinungen durch zu schwierige oder zu leichte Testaufgaben bei der individuellen Testteilnehmerin/dem individuellen Testteilnehmer ließen sich auf diese Weise besser kontrollieren. Darüber hinaus erhöhen

computerisierte Testverfahren die Effizienz in Bezug auf den administrativen, personellen und finanziellen Aufwand (Asseburg, 2011).

Es gehört zu den Einschränkungen eines solchen Testverfahrens, dass zur Erfassung der Zuhörfähigkeiten weitere Kompetenzen erforderlich sind. Neben den oben genannten Fähigkeiten zur Konzentration und Aufmerksamkeitssteuerung müssen die Schüler/innen die Testaufgaben *lesen* und *schriftlich* bearbeiten können (ausführlicher zur Konstruktkonfundierung bei Böhme, Robitzsch & Busè, 2010). In Abbildung 3 ist verdeutlicht, dass sich die Tests hinsichtlich der Durchführungen zwischen Hörverstehen und Leseverstehen lediglich durch den Modus des Inputs unterscheiden (vgl. Kapitel 2 – Konstrukt).

	rezeptiv (Input): Hörtext	rezeptiv (Input): Testfragen	produktiv (Output): Testbearbeitung
Hörverstehen			
Leseverstehen			

Abbildung 3: Gegenüberstellung der Testsituationen im Bereich Hörverstehen und Leseverstehen

Um diese ungewollte Verschränkung gering zu halten, wurde bei der Konstruktion der Testaufgaben darauf geachtet, die Anteile dieser Anforderungen auf ein erforderliches Höchstmaß zu beschränken, ohne Schülerinnen und Schülern die Chance zu nehmen, ihre Fähigkeiten zeigen zu können. Zurzeit werden Möglichkeiten von Testverfahren erforscht und zunehmend genutzt, die diese ungewollten Überschneidungen weiter minimieren. Es wird beispielsweise angedacht, dass bei computerbasierten Testungen auch die Antwortoptionen von den Schülerinnen und Schülern individualisiert anhörbar sein sollen (vgl. Behrens, Käser-Leisibach, Krelle, Weirich & Zingg Stamm, 2020).

3.2.2 Aufbau der Testaufgaben

Der Aufbau einer Testaufgabe folgt immer dem gleichen Muster. Konkrete Beispiele sind in Kapitel 4 – Beispielaufgaben zu finden. Zunächst sieht die Testteilnehmerin/der Testteilnehmer ein Deckblatt, das den Titel der Testaufgabe, eine zum Hörtext passende Illustration und einen kurzen Text zeigt, der die Testteilnehmer/innen inhaltlich auf den darauffolgenden Hörtext vorbereitet (siehe Abbildung 4).

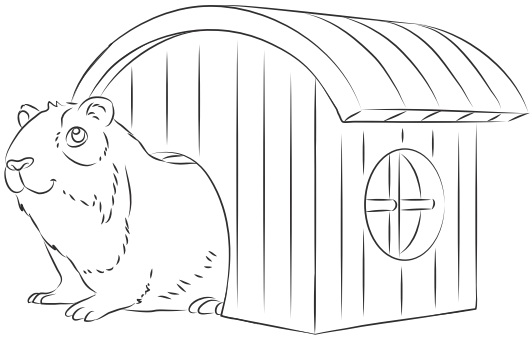
Zugleich werden die schriftlichen Informationen auch wortgleich von CD abgespielt. Zu hören sind also der Titel, der vorbereitende Text und ein Hörtext von ca. eineinhalb bis fünf Minuten Länge. Eine kurze Melodie signalisiert den Beginn des Hörtexts. Die Schüler/innen sehen während des Hörens die darauf bezogenen Testfragen (auch: Items) noch nicht und sollen sich beim Hören keine Notizen machen. Diese Bedingungen zielen darauf ab, dass die

D4H023600

Der alte Mann und das Meerschweinchen

Du hörst gleich einen Ausschnitt aus einer Geschichte. Blättere noch nicht um, sondern warte ab, bis wir die Geschichte zu Ende gehört haben.

Höre genau zu und pass gut auf.





Blättere bitte noch nicht um, sondern warte ab, bis du dazu aufgefordert wirst.

© Audio: Sparschuh, J. (2017). Der alte Mann und das Meerschweinchen. Auf *Der alte Mann und das Meerschweinchen* [CD]. Hamburg: Hörcompany.

Abbildung 4: Deckblatt einer Testaufgabe

Testteilnehmer/innen ihre Aufmerksamkeit ausschließlich auf das Zuhören lenken. Dabei evozieren der Titel, die Illustration und der kurze vorbereitende Text bei der Hörerin/dem Hörer bereits Hypothesen über Inhalt und Zweck des Texts. Für jede Testaufgabe wurde eine Illustration gefertigt, die im Testheft abgebildet und neben dem Titel der Testaufgabe und dem kurzen vorbereitenden Text das Einzige ist, das während des Hörens des Beitrags im Testheft zu sehen ist. Die Illustrationen sind so ausgewählt, dass sie nicht zur Bearbeitung

der Items beitragen, sie erhöhen vielmehr das situationale Interesse und helfen, den Fokus und den Blick zu konzentrieren, um ein mentales Modell des Gehörten zu entwickeln. Über das gesamte Testmaterial hinweg wurde auf ein stark vereinheitlichtes und übersichtliches Layout geachtet, um der/dem Testteilnehmenden die Orientierung zu erleichtern (vgl. Moosbrugger & Kelava, 2012).

Am Ende des Hörtexts erklingt die Melodie abermals und eine Instruktion zum Beginn der Aufgabenbearbeitung ist zu hören. Auf den folgenden Seiten sind auf den Hörtext bezogene Items zu finden, die die Schüler/innen im Testheft lesen und in einer vorgegebenen Zeit schriftlich bearbeiten. In Abbildung 5 wird die Abfolge einer typischen Testaufgabe im Hörverstehen deutlich.

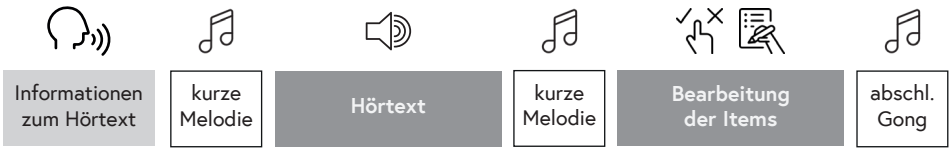


Abbildung 5: Ablauf der Bearbeitung einer Testaufgabe

Jede Testteilnehmerin/jeder Testteilnehmer bearbeitete vier solcher Testaufgaben hintereinander, die sich hinsichtlich verschiedener Kriterien unterscheiden, auf die später näher eingegangen wird. Im Rahmen der Testzeit von ca. 25 Minuten ist auch der Einsatz von mehr als vier Testaufgaben denkbar. Als Folge dessen wäre dann die inhaltliche Auseinandersetzung der Schüler/innen mit einem Hörtext kürzer und auf weniger Items beschränkt. Anzunehmen ist, dass die Begrenzung auf vier Hörtexte die Motivation und Aufmerksamkeitssteuerung der Schüler/innen hinreichend aufrechterhält. Dem steht gegenüber, dass die gezeigten Schülerleistungen umso mehr von der Qualität der vier Testaufgaben abhängen. Um dem entgegenzuwirken, lag der Fokus der Testentwicklung auf der Auswahl von geeigneten Hörbeiträgen, die der Lebenswirklichkeit von Kindern entsprechen, und auf einer hohen Qualität der Items.

Der Hörtext wird jeweils nur einmal gehört. Dieses Vorgehen wurde in der Vergangenheit ausführlich diskutiert (Buck, 2011; Grotjahn, 2012; Neumann, 2012; Ruhm et al., 2016). Letztlich spricht für die einmalige Präsentation, dass sie realistischen Zuhörsituationen entspricht, in denen in der Regel – in Abhängigkeit von der Quelle – Beiträge nur einmal gehört werden können. Das Verständnis eines Beitrags nach dem einmaligen Hören ist das, was mit dem vorliegenden Konstrukt gemessen werden soll. Was beim mehrmaligen Hören bei der Zuhörerin/beim Zuhörer passiert und welche anderen Fähigkeiten bei der Verarbeitung genutzt werden, ist bislang nicht bekannt (Buck, 2011). In dieser Weise wurde auch in vorangegangenen Studien der Bildungsstandardüberprüfung sowie den regelmäßigen Überprüfungen des Erreichens der Bildungsstandards durch das IQB in Deutschland verfahren.

Als eine entscheidende Frage bei der Konstruktion von Testaufgaben gilt die Position der Items in Bezug auf den Hörtext: Da sich die Items stets ausschließlich auf den Hörtext beziehen, können sie vor oder nach dem Hören präsentiert werden. Sind die Items vor dem Hören für die Schüler/innen zu sehen, ergibt sich die Frage, ob sie mit der Bearbeitung der Items bereits während des Hörens beginnen können oder erst im Anschluss (vgl. zusammenfassend bei Grotjahn, 2012; Neumann, 2012). Das gleichzeitige Hören des Texts und die Bearbeitung der Fragestellung, also das Lesen der Aufgabenstellung und das Verschriftlichen der Antwort, stellt für Schüler/innen der Primarstufe eine große Herausforderung dar. Es kann dazu führen, dass sich Kinder auf einzelne Fragen fokussieren und andere Fragen auslassen. Werden Items vor dem Hören präsentiert, sind nur kurze Hörtexte sinnvoll, aus denen die erfragten Informationen zielgerichtet herauszuhören sind. Im Sinne des Ziels, das Konstrukt möglichst breit zu operationalisieren, ist es wichtig, den jeweiligen Hörtext als Ganzes zu verarbeiten, um Informationen auf allen Ebenen zu erinnern bzw. zu rekonstruieren (vgl. Behrens et al., 2020). In der vorliegenden Pilotierungsstudie werden die Items aus diesem Grund nach dem Hörtext präsentiert.

Die Reihenfolge der Items innerhalb einer Testaufgabe richtet sich nach der inhaltlichen Logik des Textes. Die Wahl des Formats und die Schwierigkeit der ersten Items beeinflussen mutmaßlich die Motivation der Schülerin/des Schülers: Leichte Items, in denen kaum ausführliche Antworten verlangt werden, gehen schwierigeren und offenen Items voraus (Moosbrugger & Kelava, 2012). Um Schüler/innen, die mit der Bearbeitung der schwierigen Items überfordert sind, den Wiedereinstieg zu ermöglichen, flacht die Schwierigkeitskurve der Testaufgabe mit mittelschwierigen und leichten Items wieder ab.

In der vorliegenden Studie variiert die Anzahl der Items pro Testaufgabe zwischen drei und acht Items pro Testaufgabe. Die Menge ist nicht zufällig, sondern resultiert aus einer Reihe von Überlegungen: Denkbar sind auch Testaufgaben mit durchweg weniger Items. Eine geringere Anzahl an Items pro Testaufgabe erfüllt die Bedingungen der lokalen stochastischen Unabhängigkeit eher. Darunter versteht man, „dass die Antwort eines Probanden auf ein Item nicht davon abhängt, wie der gleiche Proband andere Items des Tests beantwortet hat“ (Moosbrugger & Kelava, 2012, S. 278). Dem stehen Aspekte der Testökonomie und Zumutbarkeit entgegen: Bearbeitet jede Schülerin/jeder Schüler möglichst viele Items pro Hörtext (auch: Stimulus), ist der zeitliche Testaufwand gering und die Anzahl der Testaufgaben, die jedes Kind bearbeitet, reduziert.

Ein weiteres zentrales Prinzip ist die Unabhängigkeit der Items untereinander, d. h., die Bearbeitung eines Items darf die Bearbeitung eines anderen Items nicht begünstigen oder erschweren, sofern sie gemeinsam in einem Testheft sind. Beispielsweise sollte sich die gesuchte Information eines Items nicht in der Aufgabenstellung eines anderen (benachbarten) Items finden lassen oder die Lösung eines Items nicht davon abhängen, dass ein vorhergehendes Item erfolgreich bearbeitet wurde. Auch innerhalb eines einzelnen geschlossenen Items sollen sich die Antwortoptionen nicht gegenseitig ausschließen, da das die Wahl der

Antwortoption erleichtert, und sie sollten sich alle in gleichem Maß voneinander unterscheiden, sodass nur genau eine Option zutreffend ist, beispielsweise im Wortschatz oder etwa den Grad der Abstraktionsebene betreffend. Ein probates Prüfverfahren hinsichtlich der Itemabhängigkeit und der Lösung von Items auf Basis von Weltwissen oder logischem Denken ist die probeweise Bearbeitung der Items ohne Vorliegen des Hörtextes. Auch statistische Kennwerte können nach der Pilotierung dabei helfen, lokale Abhängigkeiten zwischen Itempaaren herauszufinden. Bei der Q3-Statistik von Yen (1984) handelt es sich um eine bedingte Korrelation zwischen allen möglichen Itempaaren, bei der die Schülerfähigkeit herausgerechnet wird. Unter Kontrolle der Schülerfähigkeit sollten die Itempaare keinen bedeutsamen Zusammenhang mehr aufweisen. Falls doch, kann dies ein Hinweis auf den Einfluss einer weiteren Fähigkeit neben der eigentlichen Schülerfähigkeit sein, die auf die Beantwortung beider Items wirkt (Details zur Berechnung können in Freunberger et al., 2016, eingesehen werden). Beide Maßnahmen der Qualitätssicherung wurden hier angewendet.

Sowohl in der Auswahl der Hörtexte als auch in der Entwicklung der dazugehörigen Items wurde berücksichtigt, dass verschiedene Aspekte des Hörtexts abgefragt werden können. So reduziert sich einerseits das Risiko der Abhängigkeit zweier Items untereinander. Andererseits wird vermieden, dass Schüler/innen bei Nichtverstehen eines Textaspekts gleich mehrere Items nicht beantworten können. Die Verschiedenartigkeit der Anforderungen und die Länge des Hörtexts zwingt die Schüler/innen dazu, zwischen wichtigen und unwichtigen Informationen zu unterscheiden, wie dies auch in den Kompetenzbeschreibungen der Bildungsstandards definiert ist (BIFIE, 2016). Je mehr Items auf verschiedenen – auch höheren – Anforderungsniveaus eingesetzt werden, umso mehr tragen sie zur Auseinandersetzung mit dem Gehörten und dem mentalen Modell bei. Folglich ist der diagnostische Informationsgewinn über die Kompetenz der Schüler/innen hoch, ohne diese allzu stark zu beanspruchen.

Häufig hängt die Anzahl der Items pro Hörtext auch mit dessen Länge zusammen. Dabei spielen die Komplexität des Hörtexts und die Fülle sowie Prominenz von Detailinformationen eine Rolle, um möglichst viele zu bearbeitende Anforderungen auf ein breites Fähigkeitsspektrum zu stellen.

3.2.3 Die Hörtexte

Die Leistung der Schülerin/des Schülers wird in der Bearbeitung der Items sichtbar, hängt darüber hinaus aber auch wesentlich vom Input, also dem Stimulus, auf den sich die Items beziehen, ab. Die Auswahl und Präsentation der Hörtexte sind demnach mitentscheidend für die Untersuchung. Welche Merkmale die Verarbeitung eines Hörtexts erleichtern oder erschweren, wird hier erläutert. Für die vorliegende Untersuchung wurden 50 Hörtexte ausgewählt, die insgesamt einen variantenreichen Stimuluspool bilden. Für die Auswahl der Hörtexte und Texte galt eine Reihe von Kriterien und Hinweisen, von denen auf einige wesentliche nachfolgend genauer eingegangen wird.

Für die Textauswahl kamen größtenteils Hörtexte in Betracht, die auch Gegenstand von Unterricht sein könnten, z. B. szenische Lesungen zu Texten der Kinder- und Jugendliteratur, aber auch informatorische Texte, z. B. Features, Nachrichten u. v. m. Die eingesetzten Hörtexte sind als solche also weitgehend unterrichtsvalide. Für die Hörsituation gilt das allerdings nur bedingt: Normalerweise werden Hörtexte im Alltag „um ihrer selbst willen" gehört und nicht, um anschließend Testaufgaben dazu zu bearbeiten. Im Unterricht werden Hörtexte u. a. gehört, um allein oder mit anderen Unterrichtsaufgaben zu bearbeiten. Hier stehen allerdings Items mit den entsprechenden Anforderungen im Mittelpunkt (siehe oben).

Die vorliegenden Testaufgaben sind dementsprechend einer großen Bandbreite an Textsorten und Textfunktionen zuzuordnen, die an curricularen Inhalten orientiert sind und auch in der Lebenswelt der Kinder eine Rolle spielen. Es finden sich zu gleichen Teilen literarische Texte, wie etwa (Ausschnitte von) Erzählungen, Fabeln, Kurzgeschichten und Märchen, und expositorische Texte, darunter (Radio-)Nachrichten, Kinder- und Experteninterviews sowie Reportagen. Ein Überblick über die eingesetzten Aufgaben nach Textsorten ist in Tabelle 1 zu finden.

Tabelle 1: Überblick über eingesetzte Hörtexte

	Testaufgaben	Items
Literarische Hörtexte	25	138
Expositorische Hörtexte	25	116
	50	**254**

Der Pool weist sowohl konzeptionell schriftliche als auch montierte Hörtexte mit konzeptionell mündlichen Passagen auf, d. h., die Hörer/innen folgen beispielsweise Gesprächen, die typischerweise Merkmale der Mündlichkeit wie Interjektionen oder Emotionalität aufweisen (vgl. Ulrich & Michalak, 2019). Bei der Wahl von expositorischen Texten interessieren solche, die ihren Gegenstand vertiefend und möglicherweise von unterschiedlichen Positionen aus entfalten. Es bieten sich speziellere Themen an, um auszuschließen, dass die Kinder die präsentierten Informationen bereits kennen. In den Beispielaufgaben (siehe Kapitel 4 – Beispielaufgaben) etwa im *Kinderinterview mit einer Kinderbuchforscherin* oder bei einer Reportage zur Kunstsprache *Esperanto* ist diese Forderung eingelöst. Inhaltlich sind auf literarischer Seite Texte interessant, in denen die Figurenkonstellation Interesse weckt oder in deren Verlauf unerwartete Wendungen auftreten. In solchen Texten können Handlungsmotive hinterfragt werden und Emotionen, vielleicht auch Konflikte der Figuren, aufgegriffen werden, etwa wie im Text „Der alte Mann und das Meerschweinchen".

Allgemein lässt sich sagen, dass die Themenwahl des Stimulus besonderes Gewicht für die Validität des Tests hat, die den Höranlässen im schulischen und außerschulischen

Kontext möglichst entsprechen sollte. Um allen Testteilnehmenden möglichst vergleichbare Startbedingungen zu sichern, sollten die Texte nicht Bestandteil von österreichischem Lehrmaterial sein und bei der Zielgruppe möglichst nicht bekannt sein. In die Recherche wurden daher alle deutschsprachigen Märkte einbezogen. Bei der Übersetzbarkeit von schriftlichen Texten in auditives Material muss bedacht werden, ob sich das Thema für einen Hörtext eignet. So kommen z. B. Texte, die mit zahlreichen Bildern ergänzt werden, kaum in Frage. Die Texte sollten thematisch für möglichst viele Schüler/innen der 4. Schulstufe interessant und an ihrer Lebenswelt orientiert sein (siehe oben Zumutbarkeit). Themen, die emotional belastend oder verstörend für die Schüler/innen sein könnten, wie Scheidung, Gewalt oder schwere Krankheiten, behindern die zweckmäßige Verarbeitung des Gehörten und folglich auch die Bearbeitung der Items. Ähnlich verhält es sich mit allzu lustigen Hörtexten. Themen, die mutmaßlich nur bestimmte Gruppen (Geschlecht, Herkunft, Religion etc.) interessieren oder bestimmte Gruppen möglicherweise benachteiligen, wurden ausgeklammert (siehe oben Fairness). Die Themen sollten für Schüler/innen neuartig sein, um auszuschließen, dass die Items mit Weltwissen bearbeitet werden.

Die Texte weisen verschieden hohe Anforderungen auf, um das ganze Fähigkeitsspektrum von Schülerinnen und Schülern zu berücksichtigen. Der Einfluss von schwierigkeitsbeeinflussenden Merkmalen des Texts auf die Testleistung wurde bereits vielfach diskutiert und untersucht (Buck, 2011; Grotjahn, 2012; Neumann, 2012). Einige der erwähnten Studien beziehen sich dabei auf einen nachweisbaren Zusammenhang zwischen Hör- und Leseverstehensprozessen, wenn es darum geht, schwierigkeitsbeeinflussende Merkmale von Lesetexten auf die Informationsverarbeitung von auditivem Material zu übertragen (Bremerich-Vos, Böhme & Robitzsch, 2009). In anderen Studien konnte kein so eindeutiger Zusammenhang gezeigt werden (für einen Überblick siehe Grotjahn, 2012). Dennoch geben Kriterien für schwierigkeitsbeeinflussende Merkmale bezogen auf Lesetexte, wie die mittlere Satzlänge oder die Überlappung mit dem Grundwortschatz (Böhme et al., 2010), Hinweise, die auch für die Auswahl von Hörtexten dienlich sein können.

Die Merkmale, die die Schwierigkeit von Texten beeinflussen, liegen auf verschiedenen Ebenen. Zu den maßgeblichen sprachlichen Merkmalen gehören die Verständlichkeit in Bezug auf Wortwahl und Syntax[3], der Wortschatz (abstrakt vs. konkret und fachspezifisch vs. alltagssprachlich) bzw. die Frage, ob und mit wie viel Aufwand unbekannte Wörter oder Verhältnisse aus dem Zusammenhang erschlossen werden können. Auf der Ebene des Textes sind die inhaltliche Dichte, der Grad der Kohärenz, die Erzählstruktur in Bezug auf die Zeitgestaltung im Text und die Abstraktheit des Themas, die Größe des Figureninventars usw. von Bedeutung.

3 Wenngleich es hier um Hörtexte geht, können auch Parameter, die die Schwierigkeit von schriftlichen Texten beeinflussen, bedingt berücksichtigt werden. Ein an der Universität Regensburg von Wild und Pissarek entwickeltes Analysetool für Texte ermittelt die Schwierigkeit auf Basis von zählbaren Faktoren wie Wortlänge, Substantivierungen, dem Verhältnis von Inhalts- und Funktionswörtern, syntaktischen Maßen und Passivkonstruktionen. Das Ergebnis gibt näherungsweise das Lesealter in Schulstufen an. Die transkribierten Texte zeigen dabei eine Bandbreite für das Lesealter von 3,7 bis 8,8 (Wild und Pissarek, o. J.); Link siehe Literaturverzeichnis).

Unterschiede zeigen die Texte auch hinsichtlich der Audiolänge. Die Beiträge sind zwischen rund einer halben Minute und fünf Minuten lang. Aus noch längeren Texten konnten Ausschnitte verwendet werden, die eine Sinneinheit darstellen. Abhängig von der Textqualität bieten längere Beiträge eine größere inhaltliche Tiefe und mehr Gelegenheiten, sich vertiefend mit einem Thema auseinanderzusetzen. Dass gerade bei längeren Beiträgen das Arbeitsgedächtnis stärker beansprucht wird, deckt sich mit den lebensweltlichen Anforderungen. Aus testökonomischen Gründen und um den ungewollten Einfluss der Aufmerksamkeits- und Konzentrationsfähigkeit auf die Testleistung gering zu halten, wurde die maximale Länge der Hörtexte auf ca. fünf Minuten festgelegt.

Bezogen auf auditive Merkmale waren das Vorkommen und die Stärke von Akzenten oder Dialekten der Sprecher/innen, die Aussprache, die Sprechgeschwindigkeit sowie das Ausmaß an Hintergrundgeräuschen und -musik entscheidend bei der Auswahl der Stimuli. Zudem wurde darauf geachtet, dass in überwiegendem Ausmaß österreichische Sprecher/innen ausgewählt wurden. Die Anzahl und Unterscheidbarkeit der Sprecher/innen kann ebenfalls die Schwierigkeit beeinflussen: Sowohl monologische als auch dialogische Texte bergen Schwierigkeiten hinsichtlich Wiederholungen einerseits und Sprecherwechsel andererseits. Die Tonqualität der Aufnahme galt als ein wesentliches Auswahlkriterium: Inwieweit erhebliche Störgeräusche oder Lücken in der akustischen Wahrnehmbarkeit bei der Testteilnehmerin/beim Testteilnehmer kompensiert werden können, interessiert bei der Messung der Hörverstehenskompetenz nicht bzw. ist nicht konstruktrelevant.

Im Rahmen mehrerer Rückmeldeschleifen wurden über mehrere Monate zahlreiche Audiobeiträge und schriftliche Texte für die Testentwicklung zusammengetragen, erfasst und von den Lehrkräften und fachdidaktischen Expertinnen und Experten systematisch hinsichtlich ihrer Eignung nach den oben aufgeführten Kriterien beurteilt. Für die 50 verbleibenden Stimuli wurde bei den Rechteinhaberinnen/Rechteinhabern der produzierten Hörtexte bzw. der schriftlichen Texte das Nutzungsrecht erfragt.

3.2.4 Tonaufnahmen

Ein Großteil des Audiomaterials lag als Radiobeitrag, Podcast, Hörspiel oder -buch vor. In einigen Fällen wurden lange Beiträge auf eine in sich geschlossene Sinneinheit gekürzt, ohne den Charakter des Textes zu verändern. Ein anderer Teil der Stimuli wurde professionell in einem Tonstudio vertont. Dabei kamen ausgebildete Sprecher/innen zum Einsatz, die sich in Alter, Geschlecht und Stimmcharakter unterscheiden, um den Textinhalt und die -situation angemessen und realistisch darstellen zu können und um für die Zuhörer/innen gut differenzierbar zu sein. Letzteres ist insbesondere in dialogischen Beiträgen von Bedeutung für das Textverständnis und die Bearbeitung der Items. Anhand von Hörproben wurden die zu vertonenden Texte den Sprecher/innen entsprechend ihrer Stimmmerkmale zugeteilt.

Der Einsatz mehrerer Sprecher/innen erlaubte zudem, dass die Hörbeiträge, die zu einem gemeinsamen Testheft gehörten, für die Zuhörer/innen klar unterscheidbar waren. Die

Vertonung der Texte erfolgte erst nach Abschluss der Itementwicklung. Die Betonung von bestimmten Textstellen, die für die Itembearbeitung entscheidend sind, wurde gezielt durch Sprecher/innen-Manuskripte gesteuert.

Für Texte, die vertont werden mussten, wurden Manuskripte der Originalquellen für die Sprecher/innen gefertigt und ihre adäquate auditive Umsetzung im Tonstudio wurde kontrolliert. Für einige literarische Texte wurde auf die Bereitstellung von Manuskripten verzichtet, die Sprecherin/der Sprecher erzählte vielmehr die ihr/ihm skizzierte Geschichte frei nach. Das freie Erzählen wird stärker begleitet von Sprechpausen, Wiederholungen und variierender Lautstärke und ist stärker konzeptionell mündlich. Bei allen Eigenproduktionen wurde weitgehend auf die Einbindung von Hintergrundgeräuschen und Musik verzichtet.

Jedem Stimulus wurde ein kurzer auditiver und schriftlicher Hinweis zum Mitlesen vorangestellt, der die Zuhörer/innen auf die folgende Aufgabe vorbereitet, indem auf die zu erwartende Textsorte, das Thema, bei literarischen Texten auf die Autorin/den Autor und die einmalige Präsentation des Hörtexts hingewiesen wird, z. B.:

> „Du hörst gleich eine kurze Radiosendung. Blättere noch nicht um, sondern warte ab, bis wir die Sendung zu Ende gehört haben. Höre genau zu und pass gut auf."

Diese Hinweise dienen als Einleitung und entlasten die Zuhörer/innen in ihrer Erwartung auf den darauffolgenden Beitrag. Eine kurze Melodie kündigt den eigentlichen Stimulus an. Um den Stimulus von der Instruktion abzugrenzen, unterscheiden sich die Sprecher/innen beider Hörtexte. Anschließend wurden alle Teile des Stimulus montiert: der eingesprochene Titel des Hörtexts, die Situierung, kurze (stets gleiche) Melodien, die den Hörtext rahmen, der Hörtext und ein Gong, der den Bearbeitungsschluss markiert (siehe auch Abbildung 5).

3.2.5 Die Items

Für die sich hieran anschließende Entwicklung der Items wurden theoriegestützt Konstruktionsprinzipien diskutiert und erarbeitet, die sicherstellen, dass die im Konstrukt definierten Aspekte von Zuhörkompetenz bei der Operationalisierung angemessen und vollständig abgedeckt und die Qualitätsanforderungen für Testverfahren der pädagogischen Diagnostik erfüllt werden. Diese Konstruktionsprinzipien dienen den in die Testentwicklung involvierten Personen als Arbeitsgrundlage. Im Folgenden werden die wesentlichen Prinzipien vorgestellt.

Auf Basis von Richtlinien wurden die Aufgabenentwickler/innen auf ihre Aufgabe im Rahmen von Schulungen und mithilfe von Schulungsmaterial vorbereitet. Hier lag der Fokus insbesondere auf der Herangehensweise, bei der Entwicklung eines Items jeweils nur einen ausgewählten Aspekt von Hörverstehen zu operationalisieren. Die im Konstrukt ausdefinierten Subkompetenzen gaben einen klaren, gut verständlichen Rahmen für die Itementwicklung vor (siehe Kapitel 2 – Konstrukt).

Sofern die Texte nur schriftlich vorlagen, wurden sie zunächst behelfsweise eingesprochen, um den Eindruck der Testsituation zu simulieren. Es erwies sich für die Aufgabenentwicklung als hilfreich, den Hörtext zu Beginn einmal aufmerksam zu hören und dabei die Hauptaussagen und wesentlichen Motive bzw. angelegten Konflikte des Textes zu identifizieren. Die Einmaligkeit der Präsentation stellt eine nützliche Analogie zur Testsituation her, durch die die Unterscheidung wesentlicher von unwesentlichen Informationen gelingt. Nach der Itementwicklung durch die Aufgabenentwickler/innen wurden die Testaufgaben mehrfach diskutiert und mit Rückgriff auf Erkenntnisse aus der Psychometrie und Fachdidaktik überarbeitet. Mittels der Merkmale auf der Ebene der gesamten Testaufgabe sowie auf Ebene der einzelnen Items konnte die gewünschte Abdeckung des Konstrukts in seiner Breite stetig überwacht und Testaufgaben gezielt nachentwickelt werden.

Um das Konstrukt erschöpfend zu erfassen, wurde eine ausgeglichene Verteilung der im Kapitel 2 – Konstrukt erörterten Subkompetenzen über den gesamten Aufgabenpool hinweg angestrebt, die eine Differenzierung über Aussagen der Zuhörkompetenzen erlaubt. Die zu entwickelnden Items beziehen sich also auf verschiedene ‚Themen', die der Stimulus bereithält, und die in verschiedenen Anforderungen und Itemformaten (siehe weiter unten) erfasst werden können. Sowohl die sprachliche Anforderung, das Thema des Hörtexts als auch das Itemformat eignen sich dazu, die Schwierigkeit eines Items zu beeinflussen, was nicht bedeutet, dass sich jede Itemidee in ein beliebiges Itemformat umsetzen lässt.

Ein wesentliches Konstruktionsprinzip von Items ist, dass diese nur auf Basis des Hörstimulus korrekt gelöst werden können. Das bedeutet, dass der im Konstrukt definierte Aspekt der Zuhörkompetenz notwendig ist, um das Item korrekt zu bearbeiten. Im Rahmen der Kompetenzmessung geht es nicht darum, das Weltwissen der Schüler/innen abzufragen, wenngleich Vorwissen stets aktiviert wird. Ein differenziertes Bild über die gezeigte Leistung kann sich nur zeigen, wenn jeweils nur eine Frage bzw. nur ein Problem in einem Item behandelt wird. Kombinationen von mehreren Fragen in einem Item können den Testteilnehmer überfordern, zudem ist die Zuordnung der Beantwortung zur konkreten Frage erschwert – aus testtheoretischer Sicht wird davon unbedingt Abstand genommen (Moosbrugger & Kelava, 2012).

Bei der Itemkonstruktion wurden entsprechend dem oben genannten Prinzip Items gezielt auf isolierte Aspekte des Konstrukts hin entwickelt, etwa darauf, zentrale Einzelinformationen wiederzugeben, textnahe Schlussfolgerungen zu ziehen, auf der Ebene des gesamten Texts Zusammenhänge zu erkennen bzw. das zentrale Thema eines Hörtexts zu erfassen oder über den Hörtext zu reflektieren und diesen zu interpretieren. Dabei ist für die Schwierigkeit des Items maßgeblich, wie explizit und wie häufig die Informationen im Hörtext präsentiert sind und wie prominent die interessierende Stelle ist (Behrens, Böhme & Krelle, 2009). Auch bei hierarchieniedrigen Verstehensprozessen ist von Bedeutung, dass beim einmaligen Hören „das Arbeitsgedächtnis begrenzte Kapazität hat und dass man die Information relativ schnell und abschließend bearbeiten muss, weil man nicht die Möglichkeit hat, auf Information wiederholt zuzugreifen" (Imhof, 2010, S. 21). Es werden also nur

relevante Informationen des Hörstimulus erfragt, von denen man annehmen kann, dass sie zur Konstruktion des mentalen Modells etwas beitragen. So werden etwa Formulierungen aus dem Hörtext mutmaßlich nicht erinnert, wenn sie einmalig und irrelevant sind.

Sobald eine Itemidee entstanden ist, folgt die Wahl eines geeigneten Itemformats. Itemformate sind konkrete Bearbeitungsvorgaben für die Schülerlösungen. Zu unterscheiden sind geschlossene und offene Itemformate. Die in der vorliegenden Pilotierungsstudie verwendeten Itemformate haben sich in den vergangenen Studien bewährt und beschränken sich auf drei prominente Formate: das Multiple-Choice-Item und das Richtig-falsch-Item als geschlossene Itemformate sowie das offene Item. Alle sind zu vergleichbaren Anteilen im Aufgabenpool vertreten, wobei Multiple-Choice-Items etwas überwiegen; insgesamt lässt sich sagen, dass die gleichmäßige Verteilung der verschiedenartigen Itemformate zur Testfairness beiträgt. Für die geschlossenen Itemformate gelten formale Regeln für den Aufbau: In den Multiple-Choice-Items werden stets vier Antwortoptionen zur Auswahl gestellt, von denen jeweils eine korrekt ist. In den Richtig-falsch-Items werden ebenso vier Aussagen präsentiert, für die jeweils die Entscheidung getroffen werden muss, ob die Aussage zutrifft oder nicht. Die Anzahl richtiger bzw. falscher Aussagen ist nicht vorgegeben, daher muss der Testteilnehmende tatsächlich jede Aussage prüfen. Richtig-falsch-Items werden in einem Partial-Credit-System ausgewertet, haben aber für das Gesamtergebnis das gleiche Gewicht wie Multiple-Choice-Items. Dabei werden die Richtig-falsch-Items mit einer niedrigeren Wahrscheinlichkeit zufällig richtig gelöst, als es bei Multiple-Choice-Items der Fall ist. Wie die Itemformate zu bearbeiten sind, erfahren und üben die Schüler/innen vor Beginn des Tests. Alle Itemformate folgen der gleichen Struktur: ggf. Aussage bzw. Frage, Aufgabenstellung, Feld zur Bearbeitung durch die Schüler/innen, z. B. anzukreuzende Antwortoptionen oder Schreiblinien für freie Antworten (vgl. Abbildung 6). Die Aufgabenstellung enthält einen Operator (z. B. „Kreuze an“) und ist für vergleichbare Fragen gleichlautend.

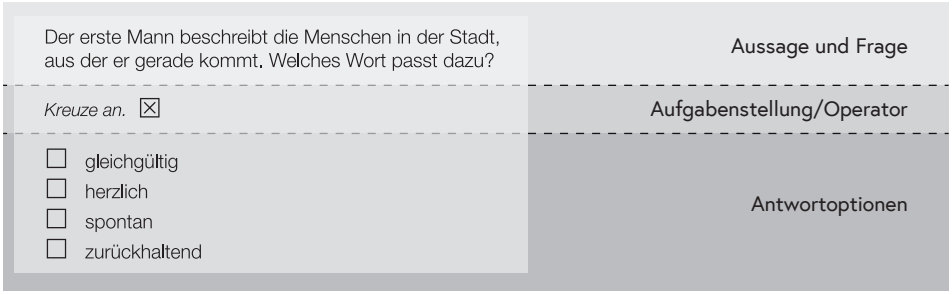


Abbildung 6: Aufbau eines geschlossenen Items

Verschiedene Merkmale erleichtern bzw. erschweren die Lösung geschlossener Items: Über das Ausmaß der sprachlichen Übereinstimmung zwischen der gesuchten Information in der Antwortoption, in der Fragestellung und im Text, über die Disjunktheit der Antwortoptionen bis zur Plausibilität der falschen Antwortoptionen (Moosbrugger & Kelava, 2012). Offene Itemformate laden zu Sprachproduktionen ein, indem sie offen gehaltene Fragen

oder Aufforderungen stellen, die die Schüler/innen eigenständig schriftlich formulieren. Ihr Einsatz ist insbesondere bei komplexeren Anforderungen sinnvoll: Offene Items sind nicht nur ein wichtiges Instrument zur Erfassung von höheren kognitiven Leistungen. Das Format eignet sich auch gut, anspruchsvollere Items zu stellen, die auf hierarchiehohe Anforderungen abzielen (siehe Kapitel 2 – Konstrukt), beispielweise hinsichtlich der Textstruktur oder Fragen, die über den Text durch Interpretationen oder Meinungen hinausweisen.

Zudem kann in diesem Format ausgeschlossen werden, dass sie wie Multiple-Choice-Items durch Ausschlussverfahren bearbeitet oder gar erraten werden. Eine zufällig richtige Antwort ist im offenen Itemformat kaum möglich. Das bedeutet andererseits nicht, dass das Itemformat allein die Schwierigkeit des Items bestimmt. Mit offenen Items können basale Informationen reproduziert werden und in geschlossenen Items können abstrakte Themen behandelt werden. Letzteres bietet sich an, wenn die gewünschte Information noch nicht von vielen Schülerinnen und Schülern in der 4. Jahrgangsstufe erwartet werden kann, z. B. wenn es um das Erkennen von Erzählperspektiven oder um Textsortenwissen geht oder wenn Aussagen wie Sprichwörter mit Textaussagen in Verbindung gebracht werden sollen.

Für die Operationalisierung einer Itemidee, also eines bestimmten zu messenden Aspekts, sind verschiedene Itemformate denkbar, die mit verschiedenen Anforderungen einhergehen. Lagen für das Thema eines Items mehrere Varianten vor, ist der Einsatz aller Varianten in einer Testaufgabe ausgeschlossen. In diesen Fällen wurden die Formate ausgewählt, die zu einer Ausgewogenheit hinsichtlich der Formate in einer Testaufgabe beitragen. Für alle Itemformate gilt, dass sie entsprechend der kognitiven Anforderung zu verwenden sind, beispielsweise bietet sich ein geschlossenes Itemformat an, wenn Informationen aus dem Hörtext wiedererkannt werden sollen.

Aus der Wahl des Itemformats ergeben sich auch unterschiedliche zu erwartende Bearbeitungszeiten für die Testteilnehmerin/den Testteilnehmer. Der Anteil nicht bearbeiteter Items in der Pilotierung, in denen also kein Lösungsversuch zu erkennen ist, entfällt zu knapp einer Hälfte auf offene Frageformate, unabhängig von der Position des Items im jeweiligen Testheft.

Im Bereich der Itemformulierung sind ebenso eine Reihe von Prinzipien zu beachten. Die Aufgabenstellungen sollen eindeutig, kurz und nur so komplex wie nötig formuliert sein, um Lesezeit zu sparen. Die Itemformulierung soll möglichst wenig Leseverstehenskompetenz verlangen. Bildungssprachliche Ausdrücke wie „begründe“ oder „kreuze an“ wurden in den Aufgabenstellungen zum einen vereinheitlicht, um die Testteilnehmer/innen durch Variationen nicht kognitiv zu belasten. Zum anderen wurden die Formulierungen mit Blick auf die Lexik einfach gehalten, um den Einfluss eines ausgeprägten Wortschatzes auf die gezeigte Leistung gering zu halten (vgl. dazu Ergebnisse von Heppt, Stanat, Dragon, Berendes & Weinert, 2014). Darüber hinaus kann die (sprachliche) Übereinstimmung von Informationen im Hörtext und der gegebenen Aufgabenstellung bewusst variiert werden.

Um eine hinreichende Differenzierung zwischen Kompetenzausprägungen der Schüler/innen vornehmen zu können, wurde eine möglichst hohe Bandbreite von Itemschwierigkeiten angestrebt. Mit den oben genannten Merkmalen kann die Schwierigkeit von Items gezielt beeinflusst werden. Im Rahmen der Aufgabenkonstruktion unterstützt die Verwaltung diverser Merkmale auf der Ebene der Stimuli und der Items die Steuerung der weiteren Itementwicklung und -überarbeitung. So lassen sich Informationen bezüglich der kognitiven Anforderungen (z. B. Subkompetenzen) als auch bezüglich äußerlicher Merkmale (z. B. Itemformate) beschreiben. Die Merkmale wurden zunächst als Zielmarken definiert und fortlaufend im Blick behalten. Die Subkompetenzen wurden zum Abschluss der Itementwicklung zunächst durch Personen unabhängig voneinander eingeschätzt und in einem ausführlichem Diskussionsprozess zugeordnet. Mithilfe dieser abstrahierten Ebene wurde eine ausgewogene Auswahl von mehreren Testaufgaben, die in einem Testheft zusammentreten, erstellt (siehe unten Testdesign).

3.2.6 Qualitätssicherung und Präpilotierung

Im Rahmen der Entwicklung der Testaufgaben kommen zahlreiche Qualitätssicherungsmaßnahmen zum Einsatz. Dazu zählen etwa der Review der Stimuli und der Testitems durch Lehrer/innen, Expertinnen und Experten, Fachexpertinnen und -experten für Testmethodik, Kontrollen der Höraufnahmen, der Grafiken, Lektorat der Items etc. Für die Kontrolle der Testhefte und weitere Qualitätssicherungsmaßnahmen siehe Kapitel 5 – Studiendesign.

In der Projektphase der Testaufgabenentwicklung wurde die Möglichkeit genutzt, einen Teil der Testaufgaben anhand einer kleinen, nicht repräsentativen Stichprobe zu erproben *(präpilotieren).* Dieses Verfahren bietet sich vor allem an, um die Bearbeitungsdauer, die Schwierigkeit von Items oder Themen der Stimuli abschätzen zu können und um die Auswertungshinweise der Items auf Basis authentischer Schülerlösungen zu präzisieren. Die Ergebnisse von präpilotierten Testaufgaben ermöglichen die Weiterentwicklung aufgrund fachdidaktischer und psychometrischer Kriterien. Insgesamt wurden acht Testaufgaben im Juni 2017 in drei Klassen präpilotiert.

Nach der Optimierung der Testaufgaben wurden sie mittels eines Testdesigns zum Einsatz in der D4-Studie 2018 auf Testhefte verteilt.

3.3 Durchführung der D4-Studie 2018

3.3.1 Überblick über die Erhebungsinstrumente

In jedem Jahr fanden Bildungsstandardüberprüfungen in Österreich in einem anderen Fach und einer anderen Schulstufe statt. Im Jahr 2018 fand die Bildungsstandardüberprüfung im Fach Mathematik in der 4. Schulstufe als Vollerhebung statt. Parallel wurde die Pilotierung für die hier vorgestellte Kompetenzmessung im Fach Deutsch in der 4. Schulstufe anhand

einer repräsentativen Stichprobe durchgeführt. Durch die Verknüpfung der Pilotierung im Fach Deutsch und der Vollerhebung im Fach Mathematik vergrößert sich die Datenbasis für Kontextinformationen mittels der Erhebungsinstrumente, auf die im Folgenden kurz eingegangen wird.

Wie bereits beschrieben wurde ein Großteil der Testaufgaben eigens für die Studie entwickelt. Dabei wurde jede Testaufgabe an jeweils mindestens 150 Kindern erprobt (siehe Kapitel 5 – Studiendesign). Anhand der Auswertung der Ergebnisse können sowohl Informationen über die Testaufgaben und Items bezüglich ihrer Eignung, das zugrundeliegende Konstrukt zu operationalisieren, festgestellt, als auch Kompetenzwerte für jede Schülerin/jeden Schüler ermittelt werden. Ein Einblick in den Testaufgabenpool wird in Kapitel 4 – Beispielaufgaben dieses Berichtsbands gegeben.

Über zwei aufeinanderfolgende Tage verteilt kamen Tests für die fünf Kompetenzbereiche aus dem Fach Deutsch zum Einsatz, wobei jeweils ein Kompetenzbereich in der Testgruppe gleichzeitig bearbeitet wurde (siehe Tabelle 6 in Kapitel 5 – Studiendesign). Für jeden Kompetenzbereich wurden gesonderte Testhefte eingesetzt. Während bei den meisten Testteilen verschiedene Testhefte gleichzeitig in einer Lerngruppe auf die Schüler/innen verteilt werden, können aufgrund der technischen Testbedingungen beim Zuhören alle Testteilnehmenden nur dieselbe Testheft-Variante in Bezug auf die Aufgaben erhalten. Das heißt, es handelt sich um eine klassenweise Administration.

Um für die große Zahl der Hörverstehensaufgaben hinreichend viele Leistungsdaten erheben zu können und zugleich die Testzeit für die einzelne Schülerin/den einzelnen Schüler in einem vertretbaren Rahmen zu halten, wurden die Testaufgaben auf verschiedene Testhefte verteilt. Mittels dieses Multi-Matrix-Designs wurde jeder/jedem Testteilnehmenden nur ein kleiner Teil der Testaufgaben vorgelegt. Jeweils vier Testaufgaben wurden zu einem Testheft zusammengestellt. Dabei wurde auf eine ausbalancierte Verteilung der Testaufgaben auf jede der vier, mindestens aber drei möglichen Positionen geachtet. Dadurch wird der Zusammenhang zwischen der Itemposition und der Lösungswahrscheinlichkeit verringert, wenngleich dieser damit nicht kontrolliert werden kann (Nagy, Frey, Nagengast, Becker & Rose, 2016). Dabei dienten zwei grobe Faktoren der Verteilung der Testaufgaben auf die Testheft-Varianten: die Zuteilung zu einer tendenziell kurzen oder tendenziell langen Testaufgabe und ihre Textsorte. Insgesamt kamen 50 dieser Testheft-Varianten in der vorliegenden Studie zum Einsatz.[4] Bei den 50 Testheft-Varianten handelt es sich um echte Parallelformen, die durch ein Multi-Matrix-Design Überschneidungen in den Items aufweisen und dadurch miteinander verlinkt werden können (siehe Kapitel 5 – Studiendesign). Schülerleistungen können somit pro Kompetenzbereich auf einer gemeinsamen Skala abgebildet werden. Zudem wurden alle geschlossenen Formate in einer zweiten Variante angeboten. Dafür wurden die Positionen der Antwortoptionen getauscht und es

4 Die 50 Testheft-Varianten dienten vornehmlich der Erprobung neuer Testaufgaben, aber auch für eine Brücken-, eine Administrationseffekt- und eine Validierungsstudie. Näheres zum Studiendesign siehe Kapitel 1 – Überblick und Kapitel 5 – Studiendesign.

entstanden sogenannte Jumble-Versionen jedes Testhefts (siehe Kapitel 5 – Studiendesign). Diese Jumble-Versionen können auch als Pseudo-Parallelversionen bezeichnet werden: Sie enthalten die gleichen Items in der gleichen Reihenfolge und unterscheiden sich nur in der Positionierung der Antwortoptionen, um Abschreiben zu verhindern. Dies ist durch die klassenweise Administration, in der die ganze Klasse die gleichen Stimuli und Items zur gleichen Zeit bearbeitet, notwendig.

Neben den Testinstrumenten zur Ermittlung von Leistungsständen wurden ergänzend Fragebögen für die Schüler/innen, deren Eltern, die Lehrkräfte der Testgruppen und die Schulleiter/innen eingesetzt, um Hintergrundinformationen zu erheben. Die Fragebögen und Kompetenztests wurden anonymisiert, es können keine Rückschlüsse auf einzelne Personen gezogen werden. Die Beantwortung der Fragebögen war für die Schüler/innen und Schulleiter/innen verpflichtend, hingegen für die Eltern und Lehrkräfte freiwillig. Die Teilnahme an der Befragung war für die Schulleiter/innen nur in der Vollerhebung in Mathematik verpflichtend. Die so erhobenen Daten können auch für die Auswertung in Deutsch herangezogen werden.

Die Schülerfragebögen konzentrieren sich neben Fragen zum Leistungsstand und zur Motivation im Fach Deutsch auf die Wahrnehmung des Lesens, Schreibens, Zuhörens, pragmatischer und kommunikativer Aspekte des Sprechens sowie der Sprachbetrachtung und Rechtschreibung im Unterricht und in der Freizeit. Die Befragung der Lehrkräfte, die in den getesteten Klassen Deutsch unterrichteten, zielt auf die Bedingungen schulischer Lernprozesse ab. Es konnten außerdem vergleichbare Fragen zum Deutschunterricht in den Lehrer- und Schülerfragebögen eingesetzt werden, sodass vergleichende Analysen möglich waren. Dabei interessieren allgemeine Fragen zum Fachunterricht sowie Fragen zur Unterrichtsgestaltung der Kompetenzbereiche, z. B. dem Stellenwert des Zuhörens im Unterricht.

3.3.2 Testablauf und -administration

Der Ablauf der Testungen folgte – wie bei empirischen Tests üblich – streng standardisierten Verfahren. Die Bedingungen in den Testsituationen möglichst ideal und vor allem standardisiert zu gestalten, ist Sache der Testorganisation und schließt Fragen der Raum- und Akustikgegebenheiten, der technischen Voraussetzungen zur Audiowiedergabe sowie einer störungsfreien Umgebung ein (Behrens, 2010). Die Testteilnehmenden wurden von ausgebildeten Testleiterinnen und Testleitern von der Vorbereitung bis zum Ende durch die zweitägige Erhebung geführt. Zur Vorbereitung erhielten die Testleiter/innen vorab eine Schulung und Material mit Hintergrundinformationen zur Studie, Hinweisen zur Durchführung und zum Vorgehen bei Problemfällen (siehe Kapitel 5 – Studiendesign).

Die Schüler/innen wurden im Klassenverband in ihrer Schule jeweils zu Beginn eines Schultags getestet. An dem ersten der beiden Testtage erhielten die Schüler/innen eine detaillierte

Einweisung in die Bearbeitung der verschiedenen Itemformate: In einem gesonderten Heft mit Beantwortungshinweisen folgten die Testteilnehmenden den Instruktionen der Testleiterin/des Testleiters, die durch die beispielhafte Bearbeitung und ggf. Korrektur von Items das Verständnis für die Formate bei allen Kindern sicherstellte.

Darüber hinaus verantwortete die Testleiterin/der Testleiter die Vergleichbarkeit der Testbedingungen, achtete auf die Einhaltung der vorgeschriebenen Bearbeitungszeiten, protokollierte die Teilnahmen und den Ablauf der Testsitzung und wahrte den Schutz der Daten und Testmaterialien sowie die sichere Aufbewahrung der Materialien in der Schule bis zur Abholung. Die Testleiterin/der Testleiter konnten jederzeit zur Bearbeitung der Formate befragt werden, durften jedoch keine inhaltlichen Fragen beantworten.

Der Hörverstehenstest erfolgte jeweils zu Beginn des zweiten Testtags und dauerte je nach eingesetzter Testheft-Variante ca. 25 Minuten (siehe Abbildung 7).

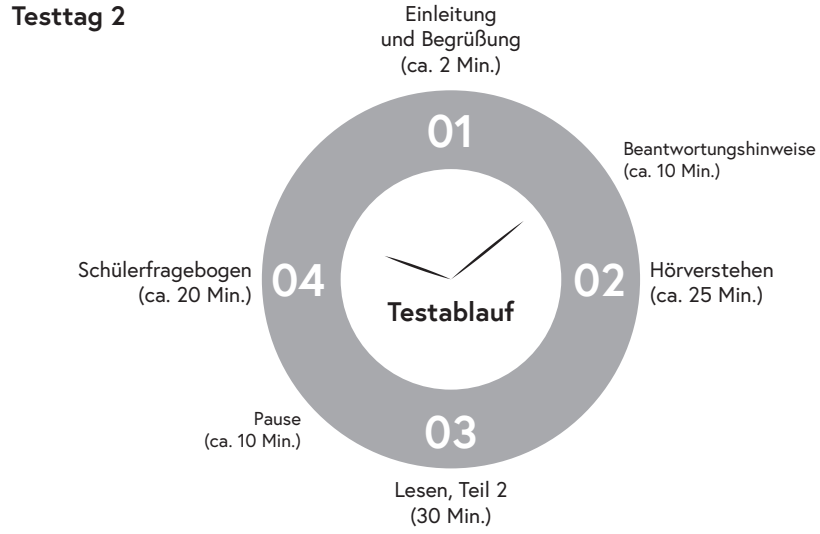


Abbildung 7: Ablauf am 2. Testtag

Wie bereits erwähnt bearbeiteten die Schüler/innen in dieser Zeit vier Testaufgaben zum Hörverstehen hintereinander, wobei für jede Testaufgabe die Bearbeitungszeit auf Basis der zu bearbeitenden Itemformate und der kognitiven Anforderungen festgelegt war. Ein Stopp-Zeichen am Ende jeder Testaufgabe signalisierte der Schülerin/dem Schüler, dass sie/er nicht weiterblättern durfte, auch wenn das Ende der Testzeit noch nicht erreicht war. Die Schüler/innen konnten ihre Bearbeitungen noch einmal überprüfen und diese bei Bedarf überarbeiten. Die Auswertung der Protokolle, in denen die Testleiter/innen unter anderem Beobachtungen während der Testbearbeitung festhalten konnten, zeigten, dass die vorgesehene Bearbeitungszeit in einigen Fällen als zu lang wahrgenommen wurde.

3.3.3 Auswertung der Schülerbearbeitungen in den Kompetenztests

Nach dem Rücklauf des bearbeiteten und pseudonymisierten Testmaterials wurde es zur Weiterverarbeitung digitalisiert (siehe auch Kapitel 5 – Studiendesign). Bei der Erfassung der Schülerantworten ist von Bedeutung, dass sich alle Testitems wie beschrieben in drei verschiedene Itemformate kategorisieren lassen: Multiple-Choice-Items und Richtig-falsch-Items als geschlossene Itemformate sowie offene Items, bei denen die Antwort durch die Schülerin bzw. den Schüler selbst formuliert und verschriftlicht werden muss. Geschlossene Itemformate können ökonomisch durch eine Software erfasst werden. Für die Erfassung und Interpretation der Items im freien Aufgabenformat hingegen ist der Einsatz von Kodiererinnen und Kodierern nötig, die alle handschriftlichen Schülerantworten als eingescannte Bilddateien prüfen und ausschließlich dichotom als richtig oder falsch bewerten *(Coding)*. Die Ergebnisse der Bewertung können für die statistischen Analysen verwendet werden.

Dazu wurden im Vorfeld für die ca. 70 Items mit offenem Itemformat auf Basis der Schülerantworten in der Präpilotierung und a priori die Kategorien für richtige und falsche Antworten genau beschrieben. Die Auswertungsanleitungen müssen dabei klar voneinander abgegrenzt sein und dürfen nur wenig Spielraum bei der Interpretation der konkreten Schülerantwort lassen, um eine objektive Testauswertung zu sichern. Je erschöpfender die Auswertungsanleitungen beschrieben sind, umso besser können sie variablen Schülerleistungen – auch in folgenden Studien – standhalten. Die Auswertungsanleitung ist somit wesentlicher Teil der Konstruktion des Items. Machten die Schüler/innen Fehler in der Rechtschreibung oder Grammatik, wirkten sich diese nicht auf die Bewertung ihrer Hörverstehensleistung aus.

Dem Coding der digitalisierten Schülerantworten geht eine gemeinsame Trainings- und Diskussionsphase mit den Kodiererinnen und Kodierern voraus. Durch die Einbindung mehrerer Kodierer/innen verringert sich zum einen der Aufwand des Codings, zum anderen erhöht sich die Objektivität des Verfahrens: Lässt sich das Item durch unabhängige Kodierer/innen konform bewerten, resultiert daraus eine hohe Übereinstimmung zwischen den Kodiererinnen und Kodierern, was wiederum ausschlaggebend dafür ist, einen korrekten Kompetenzwert zu ermitteln. Um die Übereinstimmung zwischen den unabhängigen Kodiererinnen und Kodierern festzustellen, wird nur ein zufällig ausgewählter Teil aller digitalisierten Schülerantworten mehrfach bewertet. Ein Teil der Schülerantworten wird von allen Kodiererinnen und Kodierern bewertet, ein weiterer Teil von jeweils zwei Kodiererinnen/Kodierern. Dass die Schülerfähigkeiten bei der Formulierung von Antworten in Sätzen oder Teilsätzen in der 4. Jahrgangsstufe mitunter noch etwas begrenzt sind bzw. das Schriftbild nicht vollständig ausgereift ist, erfordert zusätzlich ein gewisses Maß an Interpretation. Die Übereinstimmung zwischen den Kodiererinnen und Kodierern lag für die vorliegende Erhebung im Hörverstehen bei 81,53 %.

3.3.4 Itemanalyse und Itemselektion

Bevor die Daten der Pilotierung skaliert werden konnten, wurden die Items einer Evaluation unterzogen, um zu ermitteln, welche Items den empirischen Kriterien der Leistungsmessung entsprechen, etwa „dass die Schwierigkeit der Aufgaben hinreichend breit streut, sodass das gesamte Leistungsspektrum der in die Studie einbezogenen Schülerinnen und Schüler abgedeckt wird" (Baumert, Stanat & Demmrich, 2001, S. 40). Anhand der oben geschilderten Gütekriterien und der ermittelten Itemschwierigkeit sowie weiterer Maße der Itemanalyse wurden die Testaufgaben hinsichtlich ihrer Qualität beurteilt und solche aus dem Messinstrument ausgeschlossen, die sich als ungeeignet erwiesen. Dabei wurden Erkenntnisse zur Konstruktion von Items und Testaufgaben gewonnen, die für die weitere Erstellung von Testaufgabenpools nutzbar gemacht wurden.

Bei der Analyse der Items wurde ein Konstruktionsproblem bzgl. eines bestimmten Itemtyps offenen Formats sichtbar. Um dem Testteilnehmer hierarchiehohe kognitive Leistungen abzuverlangen, sind Items günstig, in denen Begründungen auf Aspekte des Texts bezogen produziert werden müssen, z. B. „Warum liegt Herr Möhring so oft faul auf dem Sofa? Begründe mit Hilfe von dem, was du gehört hast." Diese Items folgen dem Muster *Fragestellung* (Warum liegt Herr M…) – *Operator* (Begründe mit Hilfe …). In einigen Items wurde die Fragestellung mit Entscheidungsfragen oder „W-Fragen", etwa Interrogativpronomen, gestellt. Der darauffolgende Operator bedeutet eine zweite, zumal nachgestellte Aufforderung an die Testteilnehmerin/den Testteilnehmer. Für diese Art der Begründungsitems zeigte sich, dass ein Teil der Schüler/innen nur den ersten Teil der Frage bearbeitete, nicht jedoch die gewünschte Begründung. Alle Begründungsitems wurden daraufhin mit einem separaten Beurteilungsraster differenziert ausgewertet und es wurde geprüft, ob die korrekten Teillösungen ausreichen oder ob die betreffenden Items aus dem Testaufgabenpool ausgeschlossen werden mussten. Als ursächlich kommen die grafische Darstellung der Operatoren (nämlich abgesetzt von der Fragestellung) und die Redundanz dieses Textteils infrage: Der Operator bei Begründungsaufgaben ist stets gleichlautend und wird gegebenenfalls leicht überlesen. Diese Feststellung wurde bei der weiteren Arbeit an Testaufgaben bereits berücksichtigt. Nach Abschluss der Itemanalyse und Itemselektion umfasst der Test nur noch Items, die geeignet sind, Zuhörkompetenzen zu messen.

Im folgenden Kapitel – Beispielaufgaben werden beispielhaft vier Testaufgaben inkl. aller Items dargestellt, die Textauswahl genau beschrieben und fachdidaktisch kommentiert. Zudem werden verschiedene Kontextdaten (Lösungshäufigkeiten, Zuordnung zum Kompetenzstufenmodell etc.) angegeben. Es wird außerdem angeführt, welche Subkompetenzen des Konstrukts mit den verschiedenen Items operationalisiert werden.

Literatur

Abraham, U. & Müller, A. (2009). Aus Leistungsaufgaben lernen. *Praxis Deutsch, 36* (214), 4–12.

Asseburg, R. (2011). *Leistungsbereitschaft in Testsituationen. Motivation zur Bearbeitung adaptiver und nicht-adaptiver Leistungstests* (Wissenschaftliche Beiträge aus dem Tectum-Verlag, Bd. 24). Zugl.: Kiel, Univ., Diss., 2011. Marburg: Tectum-Verl.

Baumert, J., Stanat, P. & Demmrich, A. (2001). PISA 2000. Untersuchungsgegenstand, theoretische Grundlagen und Durchführung der Studie. In J. Baumert (Hrsg.), *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske + Budrich.

Behrens, U. (2010). Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (Edition Zuhören, Bd. 8, S. 31–50). Göttingen: Vandenhoeck & Ruprecht.

Behrens, U., Böhme, K. & Krelle, M. (2009). Zuhören – Operationalisierung und fachdidaktische Implikationen. In D. Granzer (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (Pädagogik, S. 357–375). Weinheim: Beltz.

Behrens, U., Käser-Leisibach, U., Krelle, M., Weirich, S. & Zingg Stamm, C. (2020). *stimmig. Ergebnisse einer Studie zum Zuhören in der dritten Jahrgangsstufe.* Münster: Waxmann.

BIFIE. (2016). *Konstrukt- und Kompetenzstufenbeschreibung in Deutsch/Lesen/Schreiben 4. Schulstufe* (2., aktualisierte Fassung). Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/fec772b59ab074e5cbe78f6278d1bcb2a8e6cad8/BIST-UE_D4_Konstruktbeschreibung.pdf

Böhme, K., Robitzsch, A. & Busè, A.-K. (2010). Zur Abgrenzung des Hörverstehens gegenüber dem Leseverstehen mit Hilfe schwierigkeitsbestimmender Merkmale bei der Entwicklung von Testaufgaben. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (Edition Zuhören, Bd. 8, S. 81–104). Göttingen: Vandenhoeck & Ruprecht.

Bremerich-Vos, A., Böhme, K. & Robitzsch, A. (2009). Sprachliche Kompetenzen im Fach Deutsch – Strukturanalysen und Validierungsbefunde. In D. Granzer (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (Pädagogik, S. 198–218). Weinheim: Beltz.

Buck, G. (2011). *Assessing listening* (The Cambridge language assessment series, 10. print). Cambridge: Cambridge Univ. Press.

Eid, M. & Schmidt, K. (2014). *Testtheorie und Testkonstruktion* (Bachelorstudium Psychologie). Göttingen: Hogrefe.

Freunberger, R., Robitzsch, A. & Luger-Bazinger, C. (2016). Statistische Analysen produktiver Kompetenzen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 225–258). Wien: facultas.

Grotjahn, R. (2012). Hörverstehen: Konstrukt und Messung. *FLuL – Fremdsprachen Lehren und Lernen, 41* (1), 72–86.

Heppt, B., Stanat, P., Dragon, N., Berendes, K. & Weinert, S. (2014). Bildungssprachliche Anforderungen und Hörverstehen bei Kindern mit deutscher und nicht-deutscher Familiensprache. *Zeitschrift für Pädagogische Psychologie, 28* (3), 139–149. doi:10.1024/1010-0652/a000130

Imhof, M. (2010). Zuhören lernen und lehren. Psychologische Grundlagen zur Beschreibung und Förderung von Zuhörkompetenzen in Schule und Unterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (Edition Zuhören, Bd. 8, S. 15–30). Göttingen: Vandenhoeck & Ruprecht.

Moosbrugger, H. & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage). Berlin: Springer.

Nagy, G., Frey, A., Nagengast, B., Becker, M. & Rose, N. (2016). Itempositionseffekte in Large-Scale-Assessments. In *Forschungsvorhaben in Ankopplung an Large-Scale-Assessments. Berlin: Bundesministerium für Bildung und Forschung* (S. 121–139). Berlin.

Neumann, D. (2012). *Schwierigkeitsbeeinflussende Merkmale bei Aufgaben zum Hörverstehen im Fach Deutsch in der Sekundarstufe I* (Kölner Beiträge zur Sprachdidaktik, Bd. 8). Duisburg: Gilles & Francke.

Pospeschill, M. (2010). *Testtheorie, Testkonstruktion, Testevaluation. Mit 77 Fragen zur Wiederholung* (utb.de-Bachelor-Bibliothek, Bd. 3431). München: Ernst Reinhardt Verlag.

Ruhm, R., Leitner-Jones, C., Kulmhofer, A., Kiefer, T., Mlakar, H. & Itzlinger-Bruneforth, U. (2016). Playing the Recording Once or Twice: Effects on Listening Test Performances. *International Journal of Listening, 30* (1–2), 67–83. doi:10.1080/10904018.2015.1104252

Ulrich, K. & Michalak, M. (2019). Konzeptionelle Mündlichkeit – Konzeptionelle Schriftlichkeit. In *Sprache im Fach.* München, Eichstätt: Universitätsbibliothek der Ludwig-Maximilians-Universität München. Verfügbar unter: https://epub.ub.uni-muenchen.de/61755/

Wild, J. & Pissarek, M. (o. J.). *Ratte. Regensburger Analysetool für Texte.* Verfügbar unter: https://www.uni-regensburg.de/sprache-literatur-kultur/germanistik-did/downloads/ratte/index.html

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8* (2), 125–145. doi:10.1177/014662168400800201

Antonia Bachinger, Michael Krelle, Maria Engelbert-Kocher

Kapitel 4
Beispieltestaufgaben mit fachdidaktischer Kommentierung

4.1 Einleitung

Im folgenden Kapitel werden Beispiele für verschiedene Testaufgaben dargestellt, die in der Pilotierung zu den Bildungsstandardüberprüfungen im Fach Deutsch im Jahr 2018 eingesetzt wurden (zum Aufbau der Studie siehe Kapitel 5 – Studiendesign). Die vorgestellten Beispielaufgaben dienen dazu, das Konstrukt (siehe Kapitel 2 – Konstrukt) und die Ergebnisse der Operationalisierung bzw. der Aufgabenentwicklung (siehe Kapitel 3 – Operationalisierung) zu illustrieren, mithilfe derer Aspekte des Zuhörens gemessen werden. In der Art der Darstellung und Formulierung orientieren wir uns an vergleichbaren Dokumenten, die z. B. im Rahmen der Vergleichsarbeiten in Deutschland (Krelle et al., 2016) bzw. im Rahmen der Bildungsstandardüberprüfung BIST-Ü[1] veröffentlicht sind. Schließlich ist auch die Konstruktion der Testaufgaben und die Operationalisierung des Konstrukts an gängigen internationalen Standards orientiert (Behrens, Böhme & Krelle, 2009).

Zudem ist die Darstellung an typische Beispiele von Testaufgaben angepasst: Zunächst ist die Testseite mit dem Titel, einem knappen Hinweis zu dem, was die Schüler/innen gleich hören werden, und einer Illustration abgedruckt, die während des Zuhörens zu sehen ist (siehe dazu Kapitel 3 – Operationalisierung). Anschließend werden ein Transkript bzw. die Textgrundlage der Hörtexte abgedruckt (beides lag den Schülerinnen und Schülern nicht vor). Dann werden in Kürze Informationen zur Art des Textes (z. B. Parabel ...) und dessen Entstehung (z. B. Radiosendung, Interview …) geboten, anschließend wird die Auswahl des Stimulus begründet.

Einen Überblick über alle Items der dargestellten Beispielaufgabe bietet jeweils eine Grafik. Die Abbildung zeigt, wie viele Items die Testaufgabe enthält (z. B. [5]), auf welcher Kompetenzstufe welches Item liegt (z. B. „2 Bildungsstandard erreicht"), in welchem Itemformat das Item konstruiert wurde (z. B. Multiple-Choice-Item ☒) und welche Sub-

1 Freigegebene Items der Standardüberprüfungen sind verfügbar unter https://www.iqs.gv.at/downloads/archiv-des-bifie/bildungsstandardueberpruefungen/freigegebene-items.

kompetenzen für die Itementwicklung unterschieden wurden (siehe Abbildungen 1 bis 4). Die in Kapitel 2 – Konstrukt erörterten Subkompetenzen sollen an dieser Stelle noch einmal genannt werden:

- *Können die Schüler/innen eine Information aus dem Hörtext im Rahmen eines Items wiedererkennen oder wiedergeben? (Subkompetenz 1)*
- *Können die Schüler/innen darüber hinaus auf Basis von Welt- und Sprachwissen Schlussfolgerungen ziehen, die sich auf „benachbarte" Informationen im Hörtext beziehen? (Subkompetenz 2)*
- *Können die Schüler/innen außerdem ihre Interpretationen nutzen, um den Text als Ganzes zu erfassen? (Subkompetenz 3)*
- *Können die Schüler/innen auf Aspekte des Textes bezogene Aussagen beurteilen, begründen und/oder reflektieren? (Subkompetenz 4)*

Die Items jeder Beispielaufgabe werden mit den jeweils richtigen Lösungen der geschlossenen Items dargestellt und durch Kurzkommentare illustriert. Darüber hinaus finden sich in der Kommentierung die Lösungshäufigkeit und an dieser Stelle erneut die Zuordnung der Items zur jeweiligen Kompetenzstufe. Bei offenen Items werden zudem prototypische richtige und falsche Schülerantworten aus der Pilotierung wiedergegeben.

Des Weiteren sei noch auf verschiedene weitere Aspekte hingewiesen:

(1) Bei den ersten drei Testaufgaben wird die Textvorlage abgedruckt, die u. a. Vorlage für die Aufnahmen im Tonstudio war. Bei der vierten Testaufgabe handelt es sich um ein konzeptionell mündliches Interview, welches keiner Textvorlage folgt. Deshalb wird hier ein Transkript der Aufnahme nach dem gesprächsanalytischen Transkriptionssystem GAT 2 (Selting et al., 2009) angeführt. Die Audiodateien für alle vier Testaufgaben können auf folgender Webseite des IQS oder über den QR-Code bei jeder Textvorlage eingesehen werden: www.iqs.gv.at/zuhoerbericht-beispielaufgaben.

(2) Es werden in diesem Kapitel Beispiele für geschlossene und offene Itemformate präsentiert (vgl. dazu Kapitel 3 – Operationalisierung). Bei dem geschlossenen Itemformat des Multiple Choice ist stets nur eine Antwortoption korrekt. Für das geschlossene Itemformat des Richtig-falsch-Items gilt, dass pro Zeile ein Kreuz korrekt ist. Kinder, die drei von vier Zeilen (auch: Subitems) richtig lösen, erhalten Teilpunkte; jene Kinder, die alle vier Zeilen richtig bearbeiten, erhalten die volle Punktzahl. Somit werden nur Punkte vergeben, wenn die Zahl der richtig gelösten Zeilen über der Ratewahrscheinlichkeit liegt. Bei der Auswertung der offenen Antwortformate ging es ausschließlich darum, von den Schülerantworten auf die zugrunde liegende Zuhörkompetenz zu schließen (vgl. dazu Kapitel 2 – Konstrukt). Da schriftliches Argumentieren und Erklären in der Primarstufe noch nicht zu den gängigen Unterrichtsinhalten gehört und für viele Schüler/innen herausfordernd ist, wurde bei der Bewertung der Argumentationslogik ein breites Spektrum an Antworten erlaubt. In der Folge fanden z. B. Fehler in der Rechtschreibung und der Grammatik bei der Einschätzung

der Leistungen keine Berücksichtigung. Die Schülerantworten sind in der hier wiedergegebenen Version orthografisch korrigiert.

(3) Schließlich sei noch erwähnt, dass die folgenden vier Testaufgaben nicht aus dem gleichen Testheft entstammen, aber theoretisch in dieser Form kombiniert sein könnten.

4.2 Beispiele für literarische Hörtexte

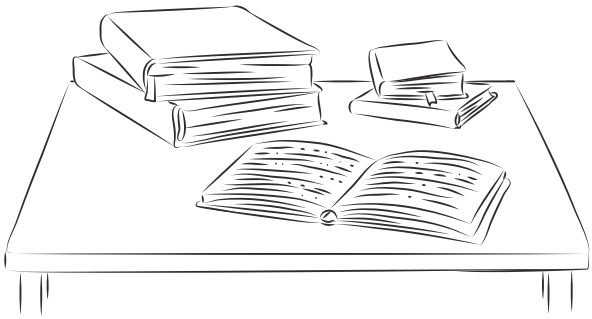
4.2.1 Wie man in den Wald hineinruft

D4H026300

Wie man in den Wald hineinruft

Du hörst gleich eine kurze Geschichte. Blättere noch nicht um, sondern warte ab, bis wir die Geschichte zu Ende gehört haben.

Höre genau zu und pass gut auf.





Blättere bitte noch nicht um, sondern warte ab, bis du dazu aufgefordert wirst.

Textvorlage (www.iqs.gv.at/zuhoerbericht-beispielaufgaben):

Vor den Toren der Stadt saß einmal ein alter Mann. Jeder, der in die Stadt wollte, kam an ihm vorbei. Ein Fremder hielt an und fragte den Alten: „Sag, wie sind die Menschen hier in der Stadt?"

„Wie waren sie denn dort, wo du zuletzt gewesen bist?", fragte der alte Mann zurück.

„Schrecklich. Sie waren gemein, unfreundlich, keiner half dem anderen."

„So, fürchte ich, werden sie auch hier sein", erwiderte der alte Mann.

Dann kam ein anderer Fremder zu dem alten Mann. Auch er fragte: „Sag mir doch alter Mann, wie sind die Menschen hier in der Stadt?"

„Wie waren sie denn dort, wo du zuletzt gewesen bist?", lautete die Gegenfrage.

„Wunderbar. Ich habe mich dort sehr wohl gefühlt. Sie waren freundlich, großzügig und stets hilfsbereit."

Darauf antwortete der alte Mann: „So etwa werden sie auch hier sein."

„Wie man in den Wald hineinruft" ist ein literarischer Text, in dem wesentliche Aspekte von Parabeln angelegt sind: Das Geschehen ist zeitlich und örtlich nicht genau eingeordnet („Vor den Toren der Stadt saß einmal ..."). Die Ereignisse werden verdichtet und chronologisch dargestellt. Die Figuren sind typisiert („ein alter Mann", „ein Fremder" etc.). Für das Textverständnis ist es erleichternd, dass sowohl die Figurenkonstellation als auch die Schauplätze leicht erschlossen werden können: Der alte Mann sitzt vor der Stadt, zwei fremde Männer kommen an ihm vorbei. Die Ereignisse auf der Bildebene lassen sich auf die Sach- oder Bedeutungsebene (Nickel-Bacon, 2014) aus dem Text erschließen, indem die Aussagen des alten Mannes vergegenwärtigt werden: „‚So, fürchte ich, werden sie auch hier sein', erwiderte der alte Mann." bzw. „So etwa werden sie auch hier sein.". Hier zeigt sich die zentrale Textaussage, dass die Sichtweise auf die Welt und das eigene Verhalten auch das Handeln anderer beeinflussen kann.

Damit zielt die Textauswahl auf einen zentralen Gegenstand des Deutschunterrichts in der Primarstufe ab: Schüler/innen sollen sich beim Zuhören literarischer Texte u. a. mit der metaphorischen und symbolischen Ebene von Texten auseinandersetzen können (Spinner, 2006). Einen wichtigen Hinweis auf diese metaphorische und symbolische Ebene kann hier der Titel liefern. Die Entschlüsselung der eigentlichen Textaussage gelingt den Kindern allerdings auch, wenn die nur angedeutete Redewendung „*Wie man in den Wald hineinruft, so schallt es heraus*" noch nicht bekannt ist.

Sprachlich stellt der Text keine größere Herausforderung für die Schüler/innen dar. Der Hörtext wird als szenische Lesung angeboten. Ein Sprecher liest den gesamten Text; es muss nicht zwischen unterschiedlichen Stimmeigenschaften unterschieden werden. Allerdings verändert der Sprecher mehrmals seine Stimme, um Figurenmerkmale herauszustellen oder zu verdeutlichen. Für das Textverstehen ist es günstig, dass die wörtliche Rede der Figuren in unterschiedlicher Klangfarbe und Lautstärke intoniert wird. So wird etwa der erste Fremde betont überheblich inszeniert, der zweite Fremde hingegen als zugewandt und freundlich dargestellt. Der Wechsel zwischen den Figuren wird durch eine angemessene Pausensetzung signalisiert, insbesondere an den Stellen, in denen es keine explizite Markierung der Figurenrede gibt.

Der Hörtext hat einen zeitlichen Umfang von etwas mehr als einer Minute (01:14). Diese Länge erlaubt es, den Text in seinen Details zu verstehen, auch wenn man die Aufmerksamkeit noch nicht sehr lange auf gesprochene Sprache richten kann. Die im Testheft dargestellte Illustration gibt keinen Hinweis auf die Lösung der Items.

Die Schüler/innen haben fünf Minuten Zeit, um fünf Items zu diesem Hörtext zu bearbeiten. Es handelt sich dabei um drei Multiple-Choice-Items, ein Richtig-falsch-Item und ein offenes Item, die im Folgenden abgedruckt sind. Die vergleichsweise lange Bearbeitungszeit ergibt sich aus der zu lesenden Textmenge in den Items und daraus, dass im Rahmen des offenen Items argumentiert werden muss (vgl. dazu die Kapitel 2 – Konstrukt und Kapitel 3 – Operationalisierung).

Items mit fachdidaktischer Kommentierung

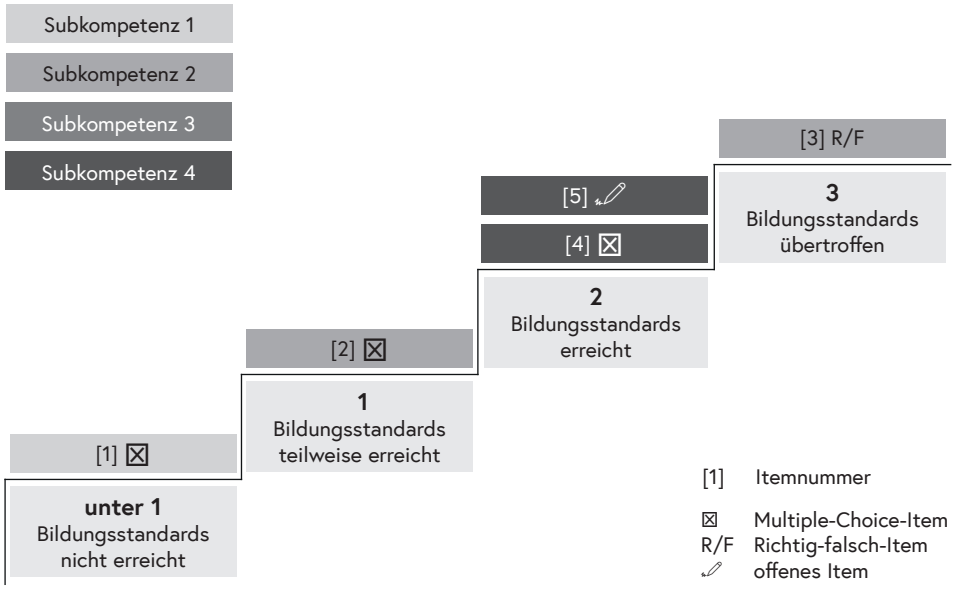


Abbildung 1: Einordnung der Items mit Kompetenzstufen für „Wie man in den Wald hineinruft“

1 Wo sitzt der alte Mann?

Kreuze an. ☒

- ☒ vor der Stadt
- ☐ in der Gasse
- ☐ auf der Bank
- ☐ vor dem Haus

D4H026311

Bei diesem Multiple-Choice-Item geht es um eine wesentliche Information, die für das Verständnis der Hauptfigur und des Schauplatzes der Handlung wichtig ist. Dass der alte Mann vor der Stadt sitzt, können 82 % der Schüler/innen hier richtig beantworten (Kompetenzstufe unter 1). Erleichternd wirkt sich dabei aus, dass die Antwortoptionen alle sprachlich einfach zu entschlüsseln sind. Schwierig ist für manche Schüler/innen allerdings, dass die falschen Antwortoptionen (Distraktoren) plausible Lösungen sein könnten, wenn man den Text nicht hinreichend verstanden hat.

2 Der zweite Fremde beschreibt die Menschen in der Stadt, aus der er gerade kommt. Welches Wort passt dazu?

Kreuze an. ☒

- ☐ gleichgültig
- ☒ herzlich
- ☐ spontan
- ☐ zurückhaltend

D4H026321

Auch diese Frage bezieht sich auf wesentliche Informationen im Text: die Aussagen des zweiten Fremden. Im Rahmen eines Multiple-Choice-Items muss hier erkannt werden, dass der zweite Fremde positiv über die Menschen in der vorherigen Stadt spricht. Insofern ist hier auch elementares Sprachwissen im Spiel, wenn von der Aussage „Sie waren freundlich, großzügig und stets hilfsbereit" im Text auf die Formulierung „herzlich" im Item gefolgert werden muss. Wie bei dem ersten Item wirkt sich dabei erleichternd aus, dass die kurzen Antwortoptionen sprachlich einfach zu entschlüsseln sind. Allerdings sind auch hier die falschen Antwortoptionen für manche Schüler/innen plausible Lösungen: Insgesamt haben 72 % der Schüler/innen dieses Item richtig gelöst (Kompetenzstufe 1).

3 Welche Aussage passt zur Geschichte?

Kreuze für jede Zeile an. ☒

	ja	nein
Der alte Mann belügt die beiden Menschen.	☐	☒
Der alte Mann möchte die beiden Fremden vor den Menschen warnen.	☐	☒
Der alte Mann spricht mit Fremden, die in die Stadt kommen.	☒	☐
Der alte Mann will die Fremden zum Nachdenken anregen.	☒	☐

D4H026331

Dieses Item erfordert die Fähigkeit, die Hauptfigur des alten Mannes ansatzweise einzuschätzen und im Textkontext zu verstehen. Die Schüler/innen müssen dazu Aussagen über die Handlungen des alten Mannes auf Passung zum Text hin untersuchen. 47 % der Kinder haben dieses Item zumindest teilweise gelöst (3 von 4 richtig gesetzten Kreuzen), 31 % haben dieses Item vollständig gelöst (Kompetenzstufe 3).

4 Welche Aussage stimmt?

Kreuze an. ☒

In der Geschichte geht es darum, ...

- ☐ dass man immer vor Reisen fragen soll, wie die Menschen dort sind.
- ☒ dass sich die Menschen so verhalten, wie man sich ihnen gegenüber verhält.
- ☐ dass man dem alten Mann nicht trauen darf, weil er immer etwas anderes sagt.
- ☐ dass sich die Menschen in verschiedenen Städten unterschiedlich verhalten.

D4H026341

Hier müssen die Schüler/innen eine zentrale Aussage des Textes im Rahmen eines Multiple-Choice-Items identifizieren. Dazu muss zunächst der Text als Ganzes verstanden und dann müssen die Aussagen im Item geprüft werden. Erleichternd wirkt sich aus, dass manche Antwortoptionen vergleichsweise leicht ausgeschlossen werden können. Dass beispielsweise der alte Mann „immer etwas anderes sagt", spielt im Text keine Rolle. Erwartungsgemäß erfüllen noch nicht alle der getesteten Schüler/innen solche Anforderungen: 42 % der Kinder haben dieses Item richtig gelöst (Kompetenzstufe 2).

5 Warum antwortet der alte Mann nicht gleich, sondern stellt den beiden Fremden zuerst eine Frage?

Begründe mithilfe von dem, was du gehört hast.

D4H026351

Die Schüler/innen sollen hier eine begründete Antwort auf Basis des Hörtextes formulieren. Dabei ist es erforderlich, auf die Motive des alten Mannes zu schließen. Im Rahmen eines offenen Itemformats geht es hier ansatzweise auch um schriftliches Argumentieren. Richtig gewertet wurden z. B. begründete Antworten, die auf einen Zusammenhang von eigenem und fremdem Verhalten abzielen, z. B.: „Der alte Mann möchte herausfinden, was die Fremden über andere Menschen denken." oder auch „Weil er sie nachdenken lassen wollte.". Typische als falsch bewertete Antworten waren etwa: „Weil er wissen möchte, ob ihm die Menschen in der anderen Stadt gut gefallen haben." oder „Weil er so sein möchte wie in der anderen Stadt.". Erwartungsgemäß handelt es sich hier um eine Anforderung für leistungsstarke Kinder im Zuhören: 32 % der Schüler/innen haben dieses Item richtig gelöst, 11 % haben es nicht bearbeitet. 57 % der Schüler/innen haben dieses Item falsch gelöst (Kompetenzstufe 2).

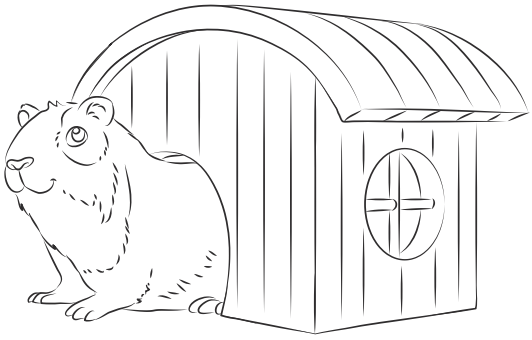
4.2.2 Der alte Mann und das Meerschweinchen

D4H023600

Der alte Mann und das Meerschweinchen

Du hörst gleich einen Ausschnitt aus einer Geschichte. Blättere noch nicht um, sondern warte ab, bis wir die Geschichte zu Ende gehört haben.

Höre genau zu und pass gut auf.





Blättere bitte noch nicht um, sondern warte ab, bis du dazu aufgefordert wirst.

© Audio: Sparschuh, J. (2017). Der alte Mann und das Meerschweinchen. Auf *Der alte Mann und das Meerschweinchen* [CD]. Hamburg: Hörcompany.

Textvorlage (www.iqs.gv.at/zuhoerbericht-beispielaufgaben):

Über den Brillenrand hinweg schielte er immer wieder in die Küchenecke.

Dort stand ein Käfig. Und in dem Käfig, auf wackeligen Hinterbeinchen, stand ein Meerschweinchen. Mit beiden Vorderpfoten klammerte es sich

an den Gitterstäben fest. Nicht, dass es daran rüttelte, nein, aber es schaute Herrn Möhring doch ziemlich vorwurfsvoll an.

„Ich rühre mich hier nicht vom Fleck. Wann legt der mal die Zeitung weg?“

Herr Möhring legte die Zeitung weg.

Er schaute ins Meerschweinchen-Ratgeberbuch, das Angelina ihm für alle Fälle dagelassen hatte, falls er mal etwas nicht wusste.

>Denke stets daran: Dein Meerschweinchen braucht liebevolle Ansprache!<

Nachdem er lange hin und her überlegt hatte, was er sagen sollte, sagte Herr Möhring schließlich: „Hallo.“ Aufmerksam schaute er über den Brillenrand. Doch das Meerschweinchen antwortete ihm nicht. Es drehte sich weg. Na gut, da sagte er erstmal auch nichts mehr.

„Aber hallo?“, dachte Ottilie. Er musste sie ja nicht gleich knuddeln wie Angelina, wenn die aus der Schule kam. Aber eine etwas nettere Begrüßung hätte sie sich schon gewünscht.

Herr Möhring war jetzt im Ratgeberbuch beim hochinteressanten Kapitel „Wissenswertes“ angelangt:

>Meerschweinchen leben am liebsten in Gemeinschaft. Einzelhaltung kann sie krank machen. Stirbt von zwei Meerschweinchen eines, sitzt das Zurückgebliebene in einer Käfigecke und frisst nicht mehr.<

Stimmt, dachte Herr Möhring: Oder es frisst trotzdem, so wie ich. Und liegt den ganzen Tag nur faul auf dem Sofa herum.

Gerade wollte er aufstehen und ins Wohnzimmer zu seinem geliebten Sofa schlurfen, da hörte er aus der Ecke einen leisen, heiseren Fieps.

Also, das war jetzt Ottilies vorläufig letzter Versuch mit diesem Herrn endlich mal Kontakt aufzunehmen: „Fiep!“

„Wer schweigt da so betreten? UAWG! (Um Antwort wird gebeten!)“

Herr Möhring antwortete nicht. Er steckte die Hände in die Hosentaschen.

„Ja, ja, piep-piep“, sagte er nach einer Weile. Worauf hatte er sich da nur wieder eingelassen? Doch Angelina hatte ihm erklärt, dass Fiepen und Quieken typisch

für Meerschweinchen ist. So stand es ja auch in dem dicken Meerschweinchen-Ratgeberbuch:

>Fiept oder quiekt ein Meerschweinchen, so will es dich begrüßen.<

Hm, was sollte man nun dazu sagen?

Quelle: Jens Sparschuh: Der alte Mann und das Meerschweinchen
Copyright © 2017 Gerstenberg Verlag, Hildesheim

„Der alte Mann und das Meerschweinchen" ist ein Auszug aus dem gleichnamigen literarischen Text von Jens Sparschuh, der hier als szenische Lesung angeboten wird (Sparschuh, 2017, S. 44–47). Der Hörtext weist typische Merkmale von Kurzgeschichten für Kinder auf: Eine kurze Exposition dient dazu, den Zuhörerinnen und Zuhörern den Einstieg in die Handlung zu ermöglichen. Die Geschehnisse werden weitgehend verdichtet und chronologisch dargestellt: Die erzählte Zeit umfasst nur einen kurzen Zeitraum. Der Schauplatz der Handlung ist auf die Wohnung von Herrn Möhring begrenzt. Das erzählwürdige Ereignis ist die schwierige Kontaktaufnahme zwischen Herrn Möhring und dem Meerschweinchen. Dynamik bekommt die Erzählung durch den auktorialen Erzähler, der die Perspektiven, Gedanken und Gefühle beider Figuren wechselseitig darstellt. So erhalten die Zuhörer/innen einen polyperspektivischen Blick auf die Kommunikationsversuche: Das Meerschweinchen versucht, die Aufmerksamkeit und Zuneigung von Herrn Möhring zu erregen, was zunächst misslingt („Er musste sie ja nicht gleich knuddeln wie Angelina, wenn die aus der Schule kam. Aber eine etwas nettere Begrüßung hätte sie sich schon gewünscht."). Herr Möhring versucht hingegen mithilfe eines „Meerschweinchen-Ratgeberbuchs", das Verhalten des Meerschweinchens zu deuten.

Dass Tiere sprechen können und Gefühle zeigen, die wiederum (erwachsene) menschliche Figuren nicht verstehen, ist in Erzählungen für Kinder nicht ungewöhnlich. Dieses Motiv zeigt sich in zahlreichen Werken der Kinder- und Jugendliteratur. Hier ist es allerdings auch für Herrn Möhring schwierig, sich auszudrücken und Kontakt aufzunehmen (z. B.: „Nachdem er lange hin und her überlegt hatte, was er sagen sollte, sagte Herr Möhring schließlich: ‚Hallo.'").

Damit zielt auch dieser Hörtext auf einen zentralen Gegenstand des Deutsch- bzw. des Literaturunterrichts in der Primarstufe ab: Kinder sollen sich identifizierend und abgrenzend mit literarischen Figuren auseinandersetzen und Verständnis für mögliche Gedanken und Gefühle von Figuren entwickeln (Spinner, 2006). In dieser Perspektive kann man den Text als Plädoyer für Verständnis sowie für das Überbrücken von Fremdheit und Barrieren ansehen.

Sprachlich stellt auch dieser Hörtext keine größere Herausforderung für die Schüler/innen dar und es gibt nur wenige anspruchsvolle Wortgruppen (z. B. „[Herr Möhring war] beim hochinteressanten Kapitel ‚Wissenswertes' angelangt"). Solche Stellen können im Textkontext entschlüsselt werden oder sind für das Gesamtverständnis nicht wesentlich. In Bezug auf die sprachliche Gestaltung sind allerdings besonders die direkten Reden im Text

beachtenswert. So spricht das Meerschweinchen Ottilie (meist) in Reimen (z. B. „Ich rühre mich hier nicht vom Fleck. Wann legt der mal die Zeitung weg?“). In der gedruckten Form des Textes sind solche Textstellen als Comic mit Sprechblasen dargestellt. Insofern bestehen gewisse Unterschiede zwischen der Druck- und der Hörbuchversion.

Wie beim ersten Aufgabenbeispiel („Wie man in den Wald hineinruft“) wird auch dieser Hörtext als szenische Lesung angeboten. Das bedeutet, dass ein Sprecher den gesamten Text liest; es muss nicht zwischen unterschiedlichen Stimmeigenschaften unterschieden werden. Dabei verändert der Sprecher mehrmals seine Stimme, um Figurenmerkmale herauszustellen oder zu verdeutlichen. Für das Textverstehen ist es günstig, dass die wörtliche Rede der Figuren in unterschiedlicher Klangfarbe und Lautstärke intoniert wird.

Der Hörtext hat einen zeitlichen Umfang von ca. drei Minuten (3:06 Min.). Diese Länge erlaubt es noch, den Text in seinen Details zu verstehen, auch wenn die Schüler/innen die Aufmerksamkeit noch nicht sehr lange auf gesprochene Sprache richten können. Die im Testheft dargebotene Illustration stellt einen direkten Bezug zur Geschichte her, gibt aber dennoch keinen Hinweis auf die Bearbeitung der Items. Dass der Text ein Auszug ist, wird den Schülerinnen und Schülern zwar vorab gesagt, für das Gesamtverständnis ist diese Information aber weniger relevant. So wird in diesem Ausschnitt nicht erwähnt, dass Angelina die eigentliche Besitzerin des Meerschweinchens ist. Diese Information kann erschlossen werden bzw. wird in den Items nicht thematisiert.

Die Schüler/innen haben insgesamt sechs Minuten Zeit, um sieben Items zu diesem Text zu bearbeiten. Davon sind vier Items Multiple-Choice-Items, zwei Richtig-falsch-Items und eines ein offenes Item, die im Folgenden dargestellt werden.

Items mit fachdidaktischer Kommentierung

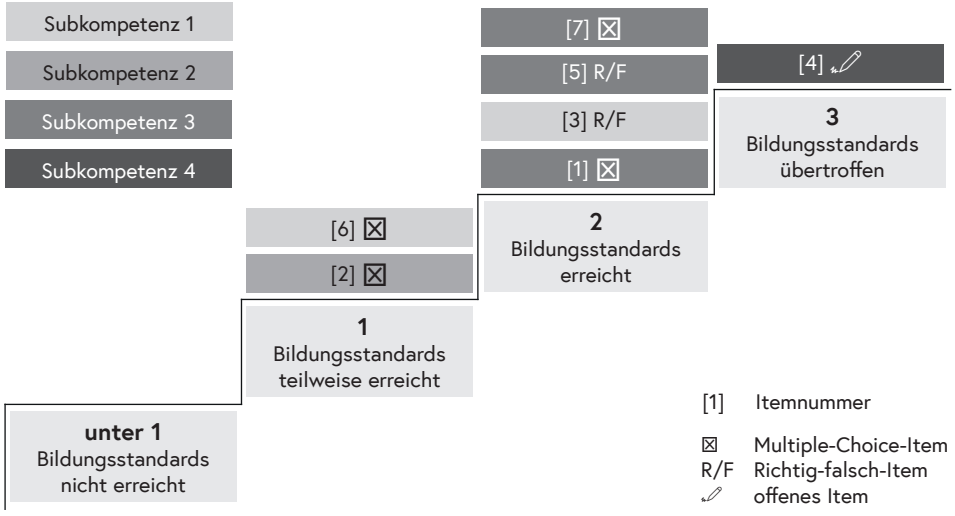


Abbildung 2: Einordnung der Items mit Kompetenzstufen für „Der alte Mann und das Meerschweinchen“

1 Was passt am besten zu Herrn Möhring?

Kreuze an. ☒

- ☐ Er ist ungeduldig.
- ☒ Er ist schweigsam.
- ☐ Er ist aufgeregt.
- ☐ Er ist ängstlich.

D4H023611

Um dieses Item bearbeiten zu können, muss der Text als Ganzes verstanden werden. Dabei ist von den Schülerinnen und Schülern nachzuvollziehen, dass das Verhalten der zentralen Figur (Herr Möhring) über den ganzen Text hinweg eher zurückhaltend und schweigsam ist. Vor diesem Hintergrund können die anderen Antwortoptionen leicht ausgeschlossen werden. Angst, Ungeduld und Aufregung spielen keine bzw. eine untergeordnete Rolle im Text. Bei diesem Item geht es erwartungsgemäß um eine Anforderung, die nur etwas mehr als die Hälfte (58 %) der Schüler/innen in der Testung bewältigen konnte (Kompetenzstufe 2).

2 Warum schaut das Meerschweinchen Herrn Möhring vorwurfsvoll an?

Kreuze an. ☒

- ☒ Es bekommt zu wenig Aufmerksamkeit.
- ☐ Es ist in einem kleinen Käfig eingesperrt.
- ☐ Es friert sehr in der kalten Küche.
- ☐ Es möchte etwas zu essen bekommen.

D4H023621

Hier steht die Reaktion des Meerschweinchens auf das Verhalten von Herrn Möhring zu Beginn des Textes im Mittelpunkt. Dabei geht es um ein zentrales Motiv des Textes, das in der Perspektive des Meerschweinchens nachvollzogen werden muss. Die Antwort kann aus dem ersten Drittel des Textes geschlossen werden, wenn es heißt: „Ich rühre mich hier nicht vom Fleck. Wann legt der mal die Zeitung weg?“ Dieses Item wurde von 71 % der Schüler/innen richtig gelöst (Kompetenzstufe 1).

3 Was liest Herr Möhring im Ratgeberbuch?

Kreuze für jede Zeile an. ☒

	richtig	falsch
Meerschweinchen leben gerne in Gemeinschaft.	☒	☐
Meerschweinchen fressen sehr viel, wenn sie fröhlich sind.	☐	☒
Meerschweinchen brauchen liebevolle Ansprache.	☒	☐
Meerschweinchen quieken laut, wenn sie Angst haben.	☐	☒

D4H023641

Zum Lösen dieses Items müssen gezielt mehrere Informationen identifiziert werden, die Herr Möhring im Ratgeberbuch liest. Dabei muss zwischen falschen und richtigen Antworten unterschieden werden, wobei alle Informationen aus dem Item im Text eine Rolle spielen. Das Format des Richtig-falsch-Items erschwert das Lösen des Items. Allerdings wirkt es sich erleichternd aus, dass das Item als teilweise richtig bearbeitet gilt, wenn in mindestens drei der vier Zeilen das richtige Kästchen angekreuzt wurde. Diese Anforderung haben 72 % der Schüler/innen bewältigt, 56 % haben dieses Item vollständig korrekt gelöst (volle Punktzahl, Kompetenzstufe 2).

4 Warum liegt Herr Möhring so oft faul auf dem Sofa?

Schreib auf.

D4H023651

Um dieses Item zu lösen, müssen die Schüler/innen eine eigenständige Begründung formulieren, die sich auf die Handlungsmotive von Herrn Möhring bezieht. Die Schwierigkeit besteht darin, dass die geforderten Gründe nicht direkt aus dem Hörtext hergeleitet werden können. Insofern ist es erforderlich, ein Verständnis für mögliche Gedanken und Gefühle von Herrn Möhring zu entwickeln. Erschwerend wirkt sich zudem aus, dass die Schüler/innen die richtige Lösung eigenständig produzieren müssen. Allerdings ist die richtige Antwort sprachlich einfach darzustellen, z. B.: „Weil er ganz alleine ist." oder „Seine Frau ist gestorben und darum ist er sehr einsam.". Falsche Antworten von Schülerinnen und Schülern in der Pilotierung waren etwa: „Weil er nicht weiß, was er sonst tun soll.", „Er frisst sehr viel.", „Weil er fernschauen will." bzw. „Weil er das Meerschweinchen nicht mag.". Rund 26 % der Schüler/innen haben dieses Item richtig gelöst, 14 % haben dieses Item nicht bearbeitet. 61 % der Schüler/innen haben dieses Item falsch gelöst (Kompetenzstufe 3).

5 Welches Problem gibt es zwischen dem Meerschweinchen und Herrn Möhring?

Kreuze für jede Zeile an. ☒

	richtig	falsch
Sie wissen noch nicht, wie sie miteinander umgehen sollen.	☒	☐
Sie sind wütend aufeinander, weil sie sich nicht mögen.	☐	☒
Sie wollen lieber allein sein, weil der andere nur stört.	☐	☒
Sie sind eifersüchtig aufeinander, weil beide Angelina mögen.	☐	☒

D4H023661

Um dieses Item lösen zu können, ist es erforderlich, den Text als Ganzes zu verstehen, indem die Figurenkonstellation in den Blick genommen wird. Die Schüler/innen müssen dazu Aussagen über die Beziehung der Figuren auf Passung zum Text hin untersuchen. Dabei sind auch die Begründungen innerhalb der Aussagen zu vergegenwärtigen. 65 % der Kinder haben dieses Item zumindest teilweise gelöst (3 von 4 richtig gesetzten Kreuzen), 50 % haben dieses Item vollständig gelöst (Kompetenzstufe 2).

6 Das Meerschweinchen sagt „UAWG". Was bedeutet die Abkürzung?

Kreuze an. ☒

- ☐ Um Aufschub wird gebeten.
- ☐ Um Anfrage wird gebeten.
- ☐ Um Auskunft wird gebeten.
- ☒ Um Antwort wird gebeten.

D4H023671

Hier muss eine prominente Information wiedergegeben werden, die explizit im Hörtext genannt wird. Um das Item zu lösen, kann aus einer Reihe von sprachlich ähnlichen Vorgaben ausgewählt werden: Dabei können die erste und die zweite Antwortoption („Um Aufschub ...", „Um Anfrage ...") sehr leicht ausgeschlossen werden. Schwieriger ist es, die dritte Antwortoption als falsch zu identifizieren. Das Item wurde von 75 % der Kinder korrekt gelöst (Kompetenzstufe 1).

7 Woran erkennt man, dass Herr Möhring und das Meerschweinchen sprechen?

Kreuze an. ☒

Wenn sie sprechen, ...

- ☐ dann wird der Sprecher besonders laut.
- ☐ dann hörst du einen zweiten Sprecher.
- ☒ dann verstellt der Sprecher seine Stimme.
- ☐ dann wird der Sprecher von Musik begleitet.

D4H023671

Zur Lösung dieses Multiple-Choice-Items ist es erforderlich, eine Aussage auszuwählen, die sich auf die stimmlichen Merkmale des Sprechers bezieht. Das richtige Beantworten der Frage wird dadurch erleichtert, dass aus einer Reihe von einfachen Vorgaben ausgewählt werden kann: Musik kommt im Hörtext nicht vor. Dass nur ein Sprecher zu hören ist, kann leicht erschlossen werden. 49 % der Schüler/innen haben dieses Item folglich richtig gelöst (Kompetenzstufe 2).

4.3 Beispiele für expositorische Hörtexte

4.3.1 Mi ne komprenis tion

D4H027500

Mi ne komprenis tion

Du hörst gleich eine kurze Radiosendung. Blättere noch nicht um, sondern warte ab, bis wir die Sendung zu Ende gehört haben.

Höre genau zu und pass gut auf.





Blättere bitte noch nicht um, sondern warte ab, bis du dazu aufgefordert wirst.

Textvorlage (www.iqs.gv.at/zuhoerbericht-beispielaufgaben):

Mi ne komprenis tion.

Oje, diese Überschrift ist aber merkwürdig. Die kann man ja gar nicht verstehen. Aber weißt du was: Genau das ist mit dieser Überschrift auch gemeint. „Das habe ich nicht verstanden", würde man statt „Mi ne komprenis tion"

auf Deutsch sagen. Trotzdem: eine merkwürdige Sache! Was könnte denn das für eine Sprache sein?

Das will ich dir verraten. Es ist eine Sprache, die man in keinem Land der Welt spricht. Und doch spricht man sie in so gut wie allen Ländern. Diese Sprache hat keine lange Geschichte, wie etwa das Deutsche, das Russische oder das Chinesische. Sie hat sich auch nicht entwickelt, wie es Sprachen normalerweise machen. Aber trotzdem ist sie da. Und sie wird – auch wenn die genaue Zahl nicht bekannt ist – von mehreren Millionen Menschen gesprochen. Du wunderst dich?

Diese Sprache ist da, weil sie jemand erfunden hat. Und zwar der polnische Augenarzt Ludwik Lejzer Zamenhof. Er hat alle ihre Wörter erfunden, also zum Beispiel, dass „Blume" „floro" heißt und „Füllfeder" „fontplumo". Außerdem hat er all die Regeln erfunden, wie man die Wörter richtig verwendet, also zum Beispiel, dass man „Mi estas" sagen muss, wenn man „Ich bin" meint. Und er hat der Sprache auch ihren Namen gegeben: Esperanto.

Die Sprache Esperanto gibt es nun bereits seit weit über hundert Jahren. Für eine Sprache ist das aber kein Alter. Esperanto ist im Vergleich zum Griechischen oder zum Hebräischen geradezu ein kleines Pflänzchen. Aber ein Pflänzchen mit einem genauen Plan für sein Leben. Esperanto will nämlich eine leichte Sprache sein, die jeder lernen kann. Esperanto will dazu beitragen, dass sich Menschen aus allen Ländern miteinander unterhalten können und keiner einen „Startvorteil" durch seine zuerst erlernte Sprache hat. Die Idee von Esperanto ist: Diese Sprache gehört allen und niemandem. Denn jeder, egal wie alt er ist, kann sie zusätzlich zu seiner Sprache erlernen. Esperanto will, dass Menschen aufeinander zugehen, das Gemeinsame entdecken, sich als eine Einheit fühlen.

Stell dir vor, du möchtest mit Kindern aus der ganzen Welt sprechen, heute mit einem Jungen aus China, morgen mit einem Mädchen aus Spanien … Weil du aber weder Chinesisch noch Spanisch sprichst und die beiden kein Deutsch verstehen, müsstest du zaubern können, um mit ihnen zu sprechen. Oder ihr lernt alle Esperanto. Sie ist wie eine Geheimsprache, die jeder lernen kann, der es will.

Freilich: Heute gibt es auch andere Möglichkeiten, sich mit den meisten Menschen auszutauschen, zum Beispiel das Englische, das man fast schon auf der ganzen Welt versteht und spricht. Aber es gibt Englischsprechende mit einem „Startvorteil" – und deshalb ist die Grundidee der einen Sprache, die jeder gleich lernen muss, um sich dann mit allen unterhalten zu können, noch immer spannend. Das Pflänzchen Esperanto könnte noch weiterwachsen und eine richtig große Pflanze werden. Eine Sprache, die in Zukunft noch viel mehr Menschen sprechen werden als heute. Wer weiß?

Eble. Das ist Esperanto und heißt: „Vielleicht."

„Mi ne komprenis tion" ist eine Radioproduktion für Kinder und ein informierender Text. Am Beispiel von Esperanto wird das Thema Kunstsprache entfaltet. Strukturell besteht der Hörtext durchweg aus Passagen, in denen eine professionelle Sprecherin einen vorbereiteten Text wiedergibt. Das Sprechtempo ist moderat. Typische Montagen bzw. Zusammenstellungen mit Tönen, Geräuschen, Interviewteilen oder Musik gibt es – anders als beim unten präsentierten Hörtext „Kinderbuchforscherin" – nicht, sodass die Schüler/innen die Aufmerksamkeit nur auf eine Sprecherin richten müssen. Für das Textverständnis erleichternd ist zudem, dass die erwachsene Sprecherin deutlich betont spricht; Atemgeräusche, Füll- und Pausenlaute sind kaum zu hören oder wurden bei der Produktion im Tonstudio gelöscht. Die Textteile der erwachsenen Sprecherin sind an nur wenigen Stellen sprachlich anspruchsvoll. Vor allem am Anfang ist der Text auch syntaktisch einfach. Des Weiteren enthält der Text Passagen, die für „geskriptete" gesprochene Texte typisch sind, wie etwa Ellipsen („Trotzdem: eine merkwürdige Sache!") und rhetorische Fragen („Was könnte denn das für eine Sprache sein?"). Auch die längeren Sätze in der Mitte des Textes sind strukturell einfach. Darüber hinaus werden zentrale Informationen mehrfach angegeben. Die Schüler/innen müssen allerdings schon beim Zuhören zwischen zentralen und weniger zentralen Informationen unterscheiden können.

Die im Testheft dargebotene Illustration gibt keinen Hinweis auf die Bearbeitung der Items. Nachdem die Schüler/innen den Text gehört haben (ca. 3:50 Minuten), sind sechs Items zu bearbeiten, die im Folgenden abgebildet und kommentiert sind. Für die Bearbeitung haben die Schüler/innen etwa fünf Minuten Zeit. Insgesamt handelt es sich um zwei Multiple-Choice-Items, zwei Richtig-falsch-Items und zwei offene Items.

Items mit fachdidaktischer Kommentierung

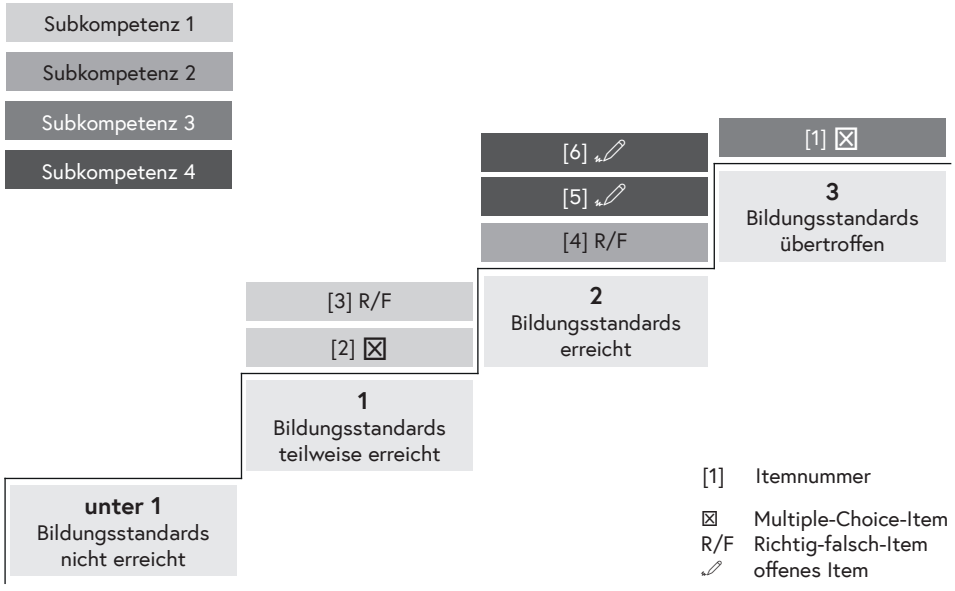


Abbildung 3: Einordnung der Items mit Kompetenzstufen für „Mi ne komprenis tion"

1 Was gibt es in diesem Text?

Kreuze an. ☒

- ☐ Es gibt es einen spannenden Höhepunkt.
- ☐ Es werden Personen genau beschrieben.
- ☐ Es wird erklärt, wie man Sprachen lernt.
- ☒ Es kommen viele sachliche Informationen vor.

D4H027511

Das erste Item zu diesem Stimulus bezieht sich auf den Radiobeitrag als Ganzes. Die Anforderung, den Text als Ganzes zu erfassen, wird noch nicht von allen Schülerinnen und Schülern bewältigt: Das Multiple-Choice-Item wurde von 30 % der Schüler/innen gelöst und gilt als schwierig (Kompetenzstufe 3). Erschwerend wirkt sich hier auch aus, dass die falschen Antwortoptionen durchaus plausibel sind, wenn man die Funktion des Textes (noch) nicht erkennen kann.

2 Woher kommen die wichtigsten Regeln für Esperanto?

Kreuze an. ☒

- ☒ Sie wurden von einem Menschen erfunden.
- ☐ Sie haben sich im Lauf der Zeit entwickelt.
- ☐ Sie wurden in einem alten Buch entdeckt.
- ☐ Sie gehen auf eine Geheimsprache zurück.

D4H027521

Zur Lösung dieses Multiple-Choice-Items ist es erforderlich, aus einer Reihe von möglichen Ursprüngen die richtige Antwortoption zu identifizieren, die im Text diskutiert wird. Dabei ist elementares Weltwissen im Spiel: Wer z. B. weiß, dass Esperanto eine Kunstsprache ist, kann die zweite Antwortoption vergleichsweise leicht ausschließen. Dieses Wissen wird außerdem im Text vermittelt, weshalb Schüler/innen, die dieses Weltwissen nicht haben, diese Antwortoption gleichfalls ausschließen können. Auch ist es weniger plausibel, dass Regeln für Esperanto „in einem alten Buch entdeckt" wurden. Im Text wird jedoch eingehend über den Erfinder der Kunstsprache gesprochen. Dementsprechend können 73 % der Schüler/innen dieses Item richtig lösen (Kompetenzstufe 1).

3 Stimmt diese Aussage über Ludwik Lejzer Zamenhof?

Kreuze für jede Zeile an. ☒

	ja	nein
Er war ein polnischer Arzt.	☒	☐
Er wurde vor 50 Jahren geboren.	☐	☒
Er hat der Sprache den Namen gegeben.	☒	☐
Er hat viele Wörter aus dem Chinesischen genommen.	☐	☒

D4H027531

Hier sind Entscheidungen unter Rückbezug auf im Text getroffene Aussagen über Ludwik Lejzer Zamenhof gefragt. Es müssen vier Antwortoptionen geprüft werden, wobei nur zwei der genannten Aussagen explizit im Text vorkommen: „Er war ein polnischer Arzt" und „Er hat der Sprache den Namen gegeben". Das Richtig-falsch-Format erschwert die Lösung. Das Item gilt als richtig bearbeitet, wenn alle Kästchen richtig angekreuzt sind. 65 % der Schüler/innen haben dieses Item vollständig gelöst, 74 % haben dieses Item zumindest teilweise gelöst, das heißt, sie haben 3 von 4 richtigen Antwortoptionen gewählt (Kompetenzstufe 1).

4 Stimmt diese Aussage über Esperanto?

Kreuze für jede Zeile an. ☒

	ja	nein
Esperanto ist älter als Hebräisch und Griechisch.	☐	☒
Esperanto soll möglichst leicht zu lernen sein.	☒	☐
Esperanto wird nur in wenigen Ländern gesprochen.	☐	☒
Esperanto soll die Menschen näher zusammenbringen.	☒	☐

D4H027541

Auch hier sind Entscheidungen unter Rückbezug auf im Text getroffene Aussagen zu treffen. Dabei geht es aber – anders als in dem Item oben – um die Kunstsprache Esperanto. Auch hier wird das Lösen durch das Richtig-falsch-Format erschwert. Zudem sind einige Informationen über den Text verstreut zu finden. 65 % der Schüler/innen haben dieses Item zumindest teilweise richtig gelöst, indem sie drei von vier Aussagen richtig angekreuzt haben. 50 % der Schüler/innen haben das Item vollständig richtig gelöst (Kompetenzstufe 2).

5 Warum finden Engländer wohl Esperanto <u>nicht</u> so wichtig?

Begründe mithilfe von dem, was du gehört hast.

D4H027551

Zur richtigen Lösung müssen die Schüler/innen eine These – „Engländer finden Esperanto nicht so wichtig" –, die mit der Fragestellung verbunden ist, auf den Text beziehen und eine Begründung finden. Dabei ist es erforderlich, textnahe Schlüsse zu ziehen. Erschwerend wirkt sich aus, dass die Schüler/innen die richtige Lösung eigenständig produzieren müssen. Als richtig gelten Antworten, die zeigen, dass Engländer die am weitesten verbreitete Sprache – Englisch – sprechen und wahrscheinlich keine zweite Weltsprache benötigen, z. B.: „Weil fast überall schon Englisch gesprochen wird" oder „Weil sie Englisch sprechen und deshalb überall verstanden werden". 49 % der Schüler/innen haben dieses Item richtig gelöst (Kompetenzstufe 2).

6 Die Sprecherin vergleicht Esperanto mit einer Pflanze.

Erkläre, was hier mit „Pflanze“ gemeint ist.

D4H027561

Auch das letzte Item zu diesem Stimulus ist ein offenes Item. Darin wird auf die im Text angelegte übertragende Bedeutung von „Pflanze“ referiert. Die besondere Herausforderung dieses Items liegt dementsprechend in einem textnahen Schluss, z. B., dass die Sprache wie die Pflanze wachsen muss. Typische Antworten der Schüler/innen sind etwa „Die Sprache ist auch klein, aber sie wächst noch.“ bzw. „Dass sie wie eine Pflanze auch noch wächst.“. Zu den falschen Antworten gehört zum Beispiel: „Es gibt eine Pflanze, die so heißt.“ Auch hier gilt allerdings, was bereits in der Einleitung zu diesem Kapitel thematisiert wurde: Es wird eine größere Bandbreite an Formulierungen akzeptiert, da das schriftliche Erklären in der Primarstufe nur bedingt behandelt wird. Nach dieser Auswertung haben 51 % der Kinder dieses Item richtig beantwortet (Kompetenzstufe 2).

4.3.2 Die Kinderbuchforscherin Ylva Schwinghammer

D4H021200

Die Kinderbuchforscherin Ylva Schwinghammer

Du hörst gleich einen Ausschnitt aus einem kurzen Interview von Schülerinnen und Schülern mit der Kinderbuchforscherin Ylva Schwinghammer. Blättere noch nicht um, sondern warte ab, bis wir das Interview zu Ende gehört haben.

Höre genau zu und pass gut auf.





Blättere bitte noch nicht um, sondern warte ab, bis du dazu aufgefordert wirst.

©Audio: Mit freundlicher Genehmigung von Doris Rudlof-Garreis und www.ohrenklick.at

Textvorlage (www.iqs.gv.at/zuhoerbericht-beispielaufgaben)

Lara: HALlo– (.)
ich bin die Lara;
Stefan: hallo (.) ich heiße STEfan–

Nikolaus: ich bin der nikolaus;
Kian: hallo ich bin der <<schmunzelnd> KIan>–
Ylva: ICH bin die ylva schWINGhammer¬– (-)
und ich bin an der universität grAz WISSENschaftlerin;
Stefan: ylva du hast ZWEIhundert KINDerbücher gelesen (.) über_s mittelALTer, sind in den hh° mittelalterbüchern AUCH: (.) GAUKler vorkomMEN?
Ylva: JA:: die kommen auch in GANZ GANZ vielen büchern vor,
sie waren KASCHperl,
aber auch arTISten,
manchmal waren sie auch SÄNGer,
manchmal haben sie auch (.) THEater gespielt;
ja=
=also sehr VIELseitig und quasi wie das FERNsehen des mittelalters;
Lara: wieso waren die RITter für dich interessant,
und NICHT die KAIser?
oder die RÖmer?
Ylva: naja= über das mittelalter gibt_es eben GANZ ganz viele geschichten und GANZ viele vo:rSTELlungen–=
=es wird immer als so eine DÜ:Stere zeit (.) beschrieben;=
=und wir DENKen heute immer es war dort alles SO schlecht=
=aber EIgentlich wIssen wir gar nicht so: VIEL (.) darüber,=
=und DAS hoab i irgendwie sehr spannend gefunden;=
=und auch weil es eine so: LANGe zeit ist–=
=also das mittelalter °hh hat ja TAUsend jahre gedauert;
Nikolaus: welche: fe:hler machen die: °hh kinderbuchautoren?
Ylva: also zuerst muss man natürlich SAGen–=
=dass h° (.) die meisten autoren und auTORinnen natürlich (.) h° sich GANZ GUT vorher inforMIEren,
und KAUM oder keine fehler machen;
die fehler, die pasSIEREN, sind dann (-)°hh recht KLEIne fehler,
und ganz VIELE davon haben mit der entdeckung (.) AMERikas zu tun; (.)
indirekt; (.)
man sagt ja auch-=
=also (.) zu dem zeitpunkt, wo AMERika entdeckt worden ist-=
=das ist das ENDE des mittelALTers, (.)
und danach fangt (.) die neuzeit an;
°h das heißt aber AUCH, beVOR amerika (.) entdeckt worden ist–
°h gab_s ganz viele dinge- (.)
die WIR HEUte ganz selbstverständlich haben, (.) noch NI:CHT.
und das wären zum beispiel (.) °h karTOFfeln;
Kian: die karTOFfel hat der CHRISToph koLUMBus zu uns gebracht;

Ylva:	GEnau (.) ganz richtig;
	°h jetzt wÄ:r-=
	=das wäre so ein FEHler, den ein kinderbuchautor (.) MACHen KÖNn:te,=
	=und den auch ein paar mAchen, °h
Stefan:	YLva?
	wie LIE:st du (.) als FO:Rscherin so: °h ein kinderbuch?
Ylva:	also als ALLERerstes fange ich immer GAR nicht an zu lesen, sondern schaue mir das buch NUR ganz genau an;
	°h also-
	WIE schaut_s AUSSen aus;
	wer hat_s geSCHRIEben?
	wie VIEle SEITen hat_s?
	°h für WE:N ist es geschrieben,
	das heißt, ist_es eher für (.) °h KINDer, die gerade erst zu LEsen ANfangen?
	oder (-) für kinder, die schon ein bisschen älter sind, und BESser lesen können,
	°hh DANN schaue ich mir an-
	sind BILDer drinnen,
	wie GRO:SS ist es geschrieben,=
	=das heißt, wie groß sind die BUCHstaben?
	wie viele ZEIlen sind auf einer SEIte-
	also (.) das alles, noch BEVOR ich angefangen hab, die geschichte (.) oder den text der drinnen steht, <überlappend; unverständlich> überhaupt zu le:sen;
Stefan:	was machst du DANN als NÄ:CHStes?
Ylva:	als NÄCHstes fange ich dann WIRKlich an, das Buch (.) °h zu LEsen;
	aber das mache ich NICHT (.) GEMÜTlich (.) im BE:TT (.) oder auf¬_m SOfa; °h
	sOndern meistens an meinem SCHREIBtisch,
	u:nd ich habe dann immer den LAPtop-=
	=oder den computer, (.)
	neben DEM BUCH
	°hh damit ich, wenn ich irgendwas interesSANTes seh im buch,
	°h das sofort AUFschreiben kann,
	°h A:BER AUCH, dass ich das NACHschlagen kann-
	in anderen bÜCHern- (-),
	°h die ich in meiner bibliothek hab;
Kian:	DANKke für_s INTerview;
Ylva:	bitte=GERN geschehen;

Dieser expositorische Hörtext ist ein Interview mit der Kinderbuchforscherin Ylva Schwinghammer. Das Interview wurde von Kindern für Kinder aufgenommen und von Doris Rudolf-Garreis für www.ohrenklick.at produziert. Vor diesem Hintergrund sind die Fragen im Interview an der Lebenswelt der kindlichen Zuhörer/innen ausgerichtet. Insofern geht es neben Fragen zum Berufsbild auch um konkrete Arbeitsabläufe sowie um biografische Aspekte der Kinderbuchforscherin Ylva Schwinghammer.

Der Hörtext weist typische Merkmale mündlicher Sprache auf, die häufig in Radiointerviews, Features oder auch in Podcasts zum Tragen kommen: Im Interview wechseln sich Fragen und Antworten jeweils ab. Der Text ist aber montiert. Das heißt, dass zwar Teile des Interviews im (autorisierten) „O-Ton" vorgespielt werden, jedoch manche Teile durch die Schüler/innen im Nachhinein im Tonstudio geschnitten wurden. Bei manchen montierten Hörtexten können solche Schnitte zur Schwierigkeit des Hörtextes beitragen. Die Schwierigkeit ist insbesondere dann erhöht, wenn zusätzlich noch Geräusche und Musik unterlegt werden, wenn die Dichte an Informationen erhöht oder Wiederholungen bzw. *advanced organizers* entfernt werden. Der vorliegende Hörtext stellt diesbezüglich keine besonderen Herausforderungen an die Zuhörer/innen. Die wenigen Hintergrundgeräusche sind durchwegs leise; es kommt kaum zu Überlappungen. Die Sprecher/innen lassen sich deutlich unterscheiden, wenn diese Unterscheidung für das Verstehen von Bedeutung ist. Sie sprechen weitgehend deutlich. Die wenigen Atemgeräusche sowie Füll- und Pauselaute stören das Textverstehen nicht; teils sind sie in der Produktion mithilfe von Filtern unterdrückt oder herausgeschnitten worden. Insofern ist das Interview im Hinblick auf die Klangqualität im Nachhinein bearbeitet worden.

Für die Antworten der Kinderbuchforscherin Ylva Schwinghammer gelten die Regeln mündlicher Sprache: Großteils ist das, was die Forscherin sagt, unvorbereitet und in der Situation der Befragung entstanden. Es finden sich dementsprechend typische Merkmale gesprochener Sprache, z. B. Verschleifungen, Auslassungen, leichte dialektale Färbung u. v. m. Es handelt sich hier im Wesentlichen um authentisches Material, wie es für autorisierte Interviews typisch ist. Thematisch ist das Interview unterrichtsvalide, da hier mit dem Thema Kinderliteratur ein wesentlicher Aspekt des Deutschunterrichts behandelt wird.

Insgesamt dauert es knapp drei Minuten (2:53 Min), das Interview anzuhören. Die im Testheft dargestellte Illustration gibt auch hier keinen Hinweis auf die Bearbeitung der fünf Items.

Items mit fachdidaktischer Kommentierung

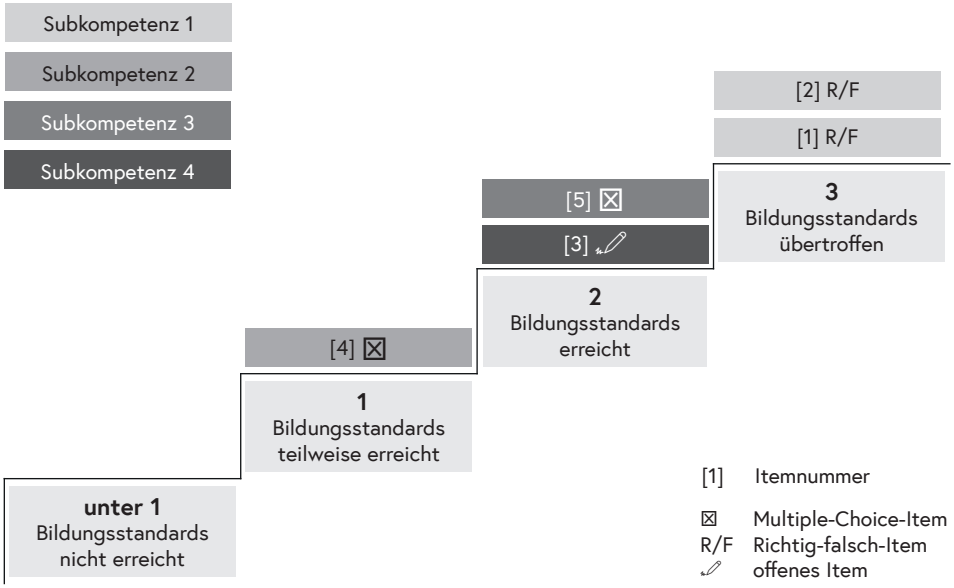


Abbildung 4: Einordnung der Items mit Kompetenzstufen für „Die Kinderbuchforscherin Ylva Schwinghammer“

1 Was wird im Interview behauptet?

Kreuze für jede Zeile an. ☒

	richtig	falsch
Bücher waren wie das Internet des Mittelalters.	☐	☒
Gaukler waren wie die Fernseher des Mittelalters.	☒	☐
Schriftsteller waren wie die Reporter des Mittelalters.	☐	☒
Geschichten waren wie das Theater des Mittelalters.	☐	☒

D4H021211

Hier müssen Behauptungen auf ihre Passung zum Text geprüft werden. Dazu ist es erforderlich, sich eine Reihe von Aussagen zu vergegenwärtigen. Allerdings ist nur eine Behauptung richtig. Im Text heißt es über Gaukler: „Sie waren Kasperl, aber auch Artisten, manchmal waren sie auch Sänger, manchmal haben sie auch Theater gespielt. Ja, also sehr vielseitig und quasi wie das Fernsehen des Mittelalters.“ 24 % der Kinder haben dieses Item zumindest teilweise gelöst (3 von 4 richtig gesetzten Kreuzen), 10 % haben dieses Item vollständig gelöst (Kompetenzstufe 3).

2 Warum findet Ylva das Mittelalter so interessant?

Kreuze für jede Zeile an. ☒

	ja	nein
Weil es eine sehr düstere Zeit war.	☐	☒
Weil es Ritter und Könige gab.	☒	☐
Weil man nicht so viel darüber weiß.	☒	☐
Weil es sehr lange gedauert hat.	☒	☐

D4H021221

Auch hier müssen mehrere Antwortoptionen auf Passung zum Text geprüft werden. Dazu ist es erforderlich, die Interessen der Kinderbuchforscherin wiederzuerkennen. Das Lösen des Items wird dadurch erschwert, dass alle Begründungen durchaus plausible Lösungen sein könnten. Daher handelt es sich hier um eine anspruchsvolle Leistung, die nur die stärksten Schüler/innen bewältigen können: 4 % der Schüler/innen haben dieses Item vollständig gelöst, 22 % der Kinder haben dieses Item zumindest teilweise gelöst (3 von 4 richtig gesetzten Kreuzen, Kompetenzstufe 3).

3 In einer Geschichte über das Mittelalter gibt es Schweinsbraten und Kartoffeln zu essen. Was stimmt hier nicht?

Erkläre mithilfe von dem, was du gehört hast.

D4H0211231

Hier ist eine Aussage auf einen Fehler zu prüfen. Dazu muss die Information aus dem Hörtext reproduziert werden, dass es viele Dinge, die heute selbstverständlich seien, noch nicht gab, bevor Amerika entdeckt wurde. Als Beispiel wird die Kartoffel genannt. Im Gespräch antwortet darauf ein Kind, dass Christoph Kolumbus die Kartoffel gebracht habe. Die Kinderbuchforscherin sagt darauf, dass das richtig sei und dass das ein Fehler sei, den Kinderbuchautoren machen könnten und den manche auch machen würden. Eine typische richtige Antwort war etwa „Dass es Kartoffeln noch gar nicht gab". Zu den falschen Antowrten zählen etwa jene, in denen argumentiert wird, dass es noch keinen Schweinsbraten gab. Da das schriftliche Erklären, das hier ansatzweise gefordert ist, in der Primarstufe noch nicht zu den gängigen Unterrichtsinhalten zählt, wird hier eine große Bandbreite an Formulierungen als richtig gewertet. 35 % der Schüler/innen lösen dieses Item richtig (Kompetenzstufe 2).

4 Womit beginnt Ylva, wenn sie ein Buch erforscht?

Kreuze an. ☒

- ☐ Sie sucht im Internet nach Informationen.
- ☒ Sie sieht sich das Buch genau an.
- ☐ Sie schreibt etwas über das Buch auf.
- ☐ Sie liest das Buch aufmerksam durch.

D4H021241

Zur Lösung dieses Multiple-Choice-Items ist es erforderlich, eine zentrale Information aus einer Antwort der Kinderbuchforscherin wiederzugeben. Erschwerend wirkt sich aus, dass mehrere der in den Antwortoptionen angebotenen Informationen richtig sein könnten, insbesondere, dass sie sich das Buch aufmerksam durchliest. Vor diesem Hintergrund verwundert es nicht, dass dieses Item lediglich von 73 % der Schüler/innen richtig gelöst wurde, obwohl die Wiedergabe einer prominenten Information gefragt ist (Kompetenzstufe 1).

5 Worum geht es hier hauptsächlich?

Kreuze an. ☒

- ☐ Wie man das Mittelalter am besten erforscht.
- ☐ Wo Ylva am liebsten arbeitet.
- ☐ Was Kinder über das Mittelalter wissen wollen.
- ☒ Womit sich Ylva beschäftigt.

D4H021251

Das letzte Item dieser Testaufgabe zielt auf das zentrale Thema des Interviews ab. Dafür müssen auf der Ebene des Textes mehrere Informationen herangezogen werden. Das Lösen des Items wird dadurch erleichtert, dass aus einer Reihe von Vorgaben ausgewählt werden kann (Multiple-Choice-Format). Erschwerend wirkt sich aus, dass alle Informationen, die im Item angeboten werden, im Text mehr oder weniger vorkommen. Fast die Hälfte der Schüler/innen (47 %) hat die richtige Antwortoption ausgewählt: „Womit sich Ylva beschäftigt." (Kompetenzstufe 2).

4.4 Zur fachdidaktischen Einordnung der Aufgabenentwicklung

Die Testaufgaben, die in diesem Kapitel vorgestellt wurden, sind nur eine kleine Auswahl dessen, was die Grundlage der Pilotierungen und Erhebungen ist (zu den Erhebungen BIST-Ü und iKM^{PLUS} siehe Kapitel 1 – Überblick). Insofern stehen sie in mancher Hinsicht prototypisch für die Art der Aufgabenentwicklung und illustrieren die Prinzipien der Testentwicklung. Nicht alle Prinzipien der Aufgabenentwicklung, die in Kapitel 3 – Operationalisierung thematisiert werden, konnten in diesem Kapitel illustriert werden.

Zudem wird in der Forschung häufig thematisiert, dass es einen erheblichen Unterschied zwischen Lern- und Leistungssituationen gibt (Abraham & Müller, 2009; Wiesner et al., 2017). Wir haben dementsprechend zwar Hörtexte ausgewählt, die auch Gegenstand des Unterrichtsgeschehens sein könnten; die hier präsentierten Texte und Items folgen aber dennoch der Logik von Testaufgaben: „Insofern liegt es auf der Hand, dass dringend davon abzuraten ist, solche Aufgaben (ohne Anpassung) für Lernzwecke einzusetzen. Mit ihnen lässt sich in der vorliegenden Form weder Zuhören erwerben noch lässt sich mit ihnen Zuhören üben!“ (Krelle & Prengel, 2014, S. 225). Hierzu braucht es andere Aufgabentypen und -formen, die in der umfangreichen fachdidaktischen Literatur sowie diversen Fachzeitschriften beschrieben sind. Eine Auswahl an Ideen für den Unterricht, die sich mit Zuhören im weiteren Sinn und somit auch mit digitalen Medien auseinandersetzen, kann der folgenden Liste entnommen werden.

Empfehlungen für den Unterricht

„Deutsch differenziert“:

Kalkavan, Z., Osburg, C., Sasse, A., Sengelhoff, B. & Valtin, R. (Hrsg.). (2014). Digitale Medien. Neue Lern- und Arbeitsformen [Themenheft]. *Deutsch differenziert,* (1).

Kalkavan-Aydin, Z., Osburg, C., Sasse, A., Sengelhoff, B. & Valtin, R. (Hrsg.). (2015). Zuhören und Sprechen. Hörfähigkeit schulen. Verstehen fördern [Themenheft]. *Deutsch differenziert,* (1).

„Grundschule“:

Dühlmeier, B., Greiling, A., Knörzer, M., Lüpkes, J., Meyer, K., Oldenburg, I. et al. (Hrsg.). (2018). Keine Angst vor Tablet und Co. Wie Kinder mit digitalen Medien – und gleichzeitig auch etwas über sie – lernen können [Themenheft]. *Grundschule,* (7).

Dühlmeier, B., Greiling, A., Knörzer, M., Lüpkes, J., Meyer, K., Oldenburg, I. et al. (Hrsg.). (2018). Stille Kinder im Blick. Wie Sie zurückhaltenden und schüchternen Schülern verhelfen können, ihr volles Potential zu entfalten [Themenheft]. *Grundschule,* (10).

„Grundschule Deutsch“:

Becker, S. H., Geister, A., Merklinger, D., Niemann, H., Rathmann, C., Vach, K. et. al. (Hrsg.). (2016). Hören und Zuhören [Themenheft]. *Grundschule Deutsch,* (52).

Friedrich Verlag (Hrsg.). (2013). Sachtexte multimedial [Themenheft]. *Grundschule Deutsch,* (39).

Kallmeyer bei Friedrich in Velber Verlag (Hrsg.). (2007). Geschichten zum Hören [Themenheft]. *Grundschule Deutsch,* (15).

Wildemann, A. & Becker, S. H. (Hrsg.). (2010). Hören-Sehen-Klicken [Themenheft]. *Grundschule Deutsch,* (26).

„Grundschulunterricht Deutsch“:

Blaseio, B., Kleinschmidt-Bräutigam, M., Kruse, N., Kurtzmann, G., Neubert, B. & Wittowske, S. (Hrsg.). (2018). Digitale Medien [Themenheft]. *Grundschulunterricht Deutsch, 65* (4).

Blaseio, B., Kleinschmidt-Bräutigam, M., Kruse, N., Ladel, S., Neubert, B. & Wittkowske, S. (Hrsg.). (2017). Sprache trifft Musik [Themenheft]. *Grundschulunterricht Deutsch, 64* (2).

Giest, H., Kleinschmidt-Bräutigam, M., Kruse, N., Neubert, B., Ladel, S. & Wittkowske, S. (Hrsg.). (2014). Hörmedien [Themenheft]. *Grundschulunterricht Deutsch, 61* (3).

Oldenbourg Verlag (Hrsg.). (2008). Ganz Ohr: Hören (lernen) [Themenheft]. *Grundschulunterricht Deutsch,* (1).

Oldenbourg Verlag (Hrsg.). (2011). Medien und Mediengeschichten [Themenheft]. *Grundschulunterricht Deutsch,* (1).

„Praxis Grundschule“:

Westermann Verlag (Hrsg.). (2011). Ein Musical für Kinder. Hören – Erleben – Aufführen [Themenheft]. *Praxis Grundschule,* (3).

„Praxis Deutsch“:

Abraham, U., Baurmann, J., Eisenberg, P., Feilke, H., Hurrelmann, B., Kammler, C. et al. (Hrsg.). (2004). Literatur hören und hörbar machen [Themenheft]. *Praxis Deutsch, 31* (185).

Baurmann, J., Gerth, K., Haas, G., Kügler, H., Ludwig, O., Menzel, W. et al. (Hrsg.). (1988). Hören und Zuhören [Themenheft]. *Praxis Deutsch, 15* (88).

„Themenheft“:

BIFIE (2011). Themenheft für den Kompetenzbereich „Hören, Sprechen und Miteinander-Reden“. Graz: Leykam. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/98b5afbb1c15f83927932f0e5f7c94e9d645cd5b/bist_de_vs_themenheft_hsm_2011-11-14.pdf

Literatur

Abraham, U. & Müller, A. (2009). Aus Leistungsaufgaben lernen. *Praxis Deutsch, 36* (214), 4–12.

Behrens, U. (2013). Zuhörkompetenzen und ihre Förderung in Primar- und Sekundarstufe. In S. Gailberger & F. Wietzke (Hrsg.), *Handbuch kompetenzorientierter Deutschunterricht* (S. 384–399). Weinheim: Beltz.

Behrens, U., Böhme, K. & Krelle, M. (2009). Zuhören – Operationalisierung und fachdidaktische Implikationen. In A. Bremerich-Vos, D. Granzer & O. Köller (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule* (S. 357–375). Weinheim: Beltz.

Krelle, M., Dünschede, S., Bremerich-Vos, A., Bittins, P., Dörnhaus, S., Engelbert, M. et al. (2016). *Vergleichsarbeiten 2015. 3. Jahrgangsstufe (VERA-3) Deutsch – Didaktischer Aufgabenkommentar „Zuhören"*. Berlin: IQB.

Krelle, M. & Prengel, J. (2014). Zur Konzeption von Zuhören im Rahmen der Vergleichsarbeiten für die dritte Klasse im Fach Deutsch. In E. Grundler & C. Spiegel (Hrsg.), *Konzeptionen des Mündlichen* (S. 208–226). Bern: hep-Verlag.

Nickel-Bacon, I. (2014). Fabeln und Parabeln in Literaturwissenschaft und -didaktik. In K. Erlemann, I. Nickel-Bacon & A. Loose (Hrsg.), *Gleichnisse – Fabeln – Parabeln.* (S. 70–113). Tübingen: A. Francke (UTB).

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K. et al. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung: Online-Zeitschrift zur verbalen Interaktion* (10), 353–402.

Sparschuh, J. (2017), *Der alte Mann und das Meerschweinchen.* Hildesheim: Gerstenberg.

Spinner, K. H. (2006). Literarisches Lernen. *Praxis Deutsch* (200), 6–16.

Wiesner, C., Illetschko, M., George, A. C., Breit, S., Süss-Stepancik, E. & Schreiner, C. (2017). Aufgabenkultur. *BIFIE-Journal, 1,* 6–10. doi: 10.17888/bifiejournal-1.2017-1-2.

Benedikt O. J. Winter, Manuel Wick, Stefan Gugerell, Gloria Schneeweiß

Kapitel 5
Studiendesign und Durchführung der Pilotierung

5.1 Einleitung

In diesem Kapitel wird auf das Studiendesign und auf die Grundlagen zur Studiendurchführung der Pilotierung in Deutsch auf der 4. Schulstufe 2018 (kurz: D4-Studie 2018[1]) eingegangen (siehe dazu auch Kapitel 1 – Übersicht). Zuerst werden in Abschnitt 5.2 die notwendigen *Field Operations,* also die Kommunikation zu wichtigen Stakeholderinnen und Stakeholdern, behandelt. Anschließend gehen wir in Abschnitt 5.3 näher auf die Datengrundlage der D4-Studie 2018 und auf die Stichprobenziehung auf Schul- und Klassenebene ein. Die Abschnitte 5.4 bzw. 5.5 behandeln die Durchführung und den Ablauf der D4-Studie 2018 bzw. die anschließende Aufbereitung und Analyse der Daten. Abschließend wird in Abschnitt 5.6 die Teilnahmequote für die Erhebung von *Zuhören* sowie die Repräsentativität der realisierten Stichprobe dargestellt.

5.2 Field Operations

Für die Organisation von Bildungsstandardüberprüfungen und Pilotierungen ist ein enger Kontakt des BIFIE, jetzt IQS[2], zu wichtigen Stakeholdern im österreichischen Schulsystem notwendig. Im Folgenden wird auf die Kommunikation mit dem damaligen Bundesministerium für Bildung (BMB), aktuell Bundesministerium für Bildung, Wissenschaft und Forschung (BMBWF), mit den Schulen sowie mit den Testleiterinnen und Testleitern im Zuge der D4-Studie 2018 eingegangen.

5.2.1 Übermittlung der Datengrundlage

An den von 2012 bis 2019 durchgeführten Bildungsstandardüberprüfungen nahmen im Primarbereich laut Bildungsstandardverordnung alle öffentlichen und privaten Volksschulen

1 Der Begriff D4-Studie wird hier als Überbegriff für die 2018 durchgeführten Studien, also die Administrationseffekt-, Pilotierungs-, Validierungs- und die Brückenstudie verwendet.

2 Aus dem Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des österreichischen Schulwesens (BIFIE) wurde am 1. Juli 2020 das Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS).

mit gesetzlich geregelter Schulartbezeichnung und auf Dauer verliehenem Öffentlichkeitsrecht teil (§ 1 Abs. 1 Verordnung über Bildungsstandards im Schulwesen, BGBl. II Nr. 1/2009 i. d. F. BGBl. Nr. 185/2012). Die im Vorfeld an einer Stichprobe durchgeführten Pilotierungen waren für die gezogenen Schulen und Schüler/innen verpflichtend, so auch die D4-Studie 2018 (§ 6 Abs. 2 BIFIE-Gesetz, BGBl I 25/2008 i. d. F. I 50/2019). Von der D4-Studie 2018 ausgenommen waren lediglich außerordentliche Schüler/innen sowie Schüler/innen mit sonderpädagogischem Förderbedarf, die in Deutsch nach dem Lehrplan der Allgemeinen Sonderschule (ASO) oder nach dem Lehrplan einer niedrigeren Schulstufe unterrichtet wurden. Schüler/innen mit Körper- oder Sinnesbehinderung nahmen dann nicht teil, wenn sie selbst mit allenfalls im Unterricht zur Verfügung stehenden Hilfsmitteln unter den standardisierten Testbedingungen die gestellten Aufgaben voraussichtlich nicht lösen konnten (§ 1 Abs. 3 Verordnung über Bildungsstandards im Schulwesen, BGBl. II Nr. 1/2009 i. d. F. BGBl. Nr. 185/2012).

Im Sommer 2017 stand das BIFIE bzgl. der Datenbasis (Zielgruppenschulen) für die 2018 durchzuführende Bildungsstandardüberprüfung in Mathematik in Kommunikation mit dem BMB. Diese Datenbasis, die dem BIFIE vom BMB übermittelt wurde, stellte auch die Datengrundlage für die Stichprobenziehung auf Schulebene (siehe Abschnitt 5.3) für die D4-Pilotierung dar.

5.2.2 Kontaktaufnahme mit den Schulen

Im Herbst 2017 informierte das BIFIE die teilnehmenden Schulen über die bevorstehende Bildungsstandardüberprüfung in Mathematik sowie die Pilotierung in Deutsch auf der vierten Schulstufe. Die Schulen bekamen dabei Informationen zu den Testterminen, der Testadministration etc. Über die online zur Verfügung gestellte Plattform zur Organisation der Standardüberprüfung (PLATO)[3] übermittelten die teilnehmenden Schulen u. a. Kontaktdaten, Testtermine, Anzahl der Klassen, Unterrichtsgruppen und Anzahl sowie ggf. Ausschlussgründe der Schüler/innen der 4. Schulstufe. Bei Fragen zur Datenübermittlung stand den Schulen ein Hotline-Support zur Verfügung.

Nach abgeschlossener Datenübermittlung wurden Kontroll- und Plausibilitätschecks der übermittelten Daten durchgeführt und ggf. Korrekturen vorgenommen. Diese Checks betreffen etwa die gemeldeten Klassen und Schüler/innen, den Anteil ausgenommener Schüler/innen je Klasse und die Bestätigung eines Testtermins für die Pilotierung. Die korrigierte Datenbasis wurde im Anschluss für die Stichprobenziehung auf Klassenebene (siehe Abschnitt 5.3) verwendet. Pro Schule wurde eine Klasse für die Teilnahme an der D4-Studie 2018 ausgewählt. Ab November 2017 konnten die Schulleiter/innen die ausgewählte Klasse über die Onlineplattform PLATO einsehen.

In der nächsten Schulinformation im März 2018 erhielten die teilnehmenden Schulen detaillierte Informationen zur Standardüberprüfung in Mathematik sowie zur D4-Studie

3 https://plato.bifie.at

2018. Im Rahmen dieser Schulaussendung wurden den Schulleiterinnen und Schulleitern sowie den Lehrerinnen und Lehrern Zugangslinks zu Fragebögen übermittelt, mit denen Kontextinformationen erhoben wurden. Außerdem wurden postalische Informationsschreiben an Klassenlehrer/innen und Eltern/Erziehungsberechtigte zur Verteilung an die Schulen übermittelt.

5.2.3 Testleiter/innen

Die D4-Studie 2018 in Deutsch wurde von eigens geschulten externen Testleiterinnen und Testleitern, also von einer sogenannten schulfremden Lehrperson durchgeführt. Der Einsatz von externen Testleiterinnen und Testleitern dient der Qualitätssicherung. Die Testleiter/innen wurden in Zusammenarbeit mit den Bundeslandkoordinatorinnen und -koordinatoren an den Pädagogischen Hochschulen rekrutiert und im März 2018 geschult. Testleiter/innen, die seit 2012 im Rahmen der Bildungsstandardüberprüfungen (inkl. Pilotierungen) bereits eine TL-Schulung besucht hatten, konnten statt der Präsenzschulung ein Onlineupdate absolvieren. Dies traf auf 98 % der Testleiter/innen zu. Alle Schulungsmaßnahmen und Schulungsunterlagen wurden vom BIFIE konzipiert und auch die Präsenzschulung (ca. 1 Halbtag) wurde von BIFIE-Mitarbeiterinnen und -Mitarbeitern abgehalten. Für die Durchführung der Studie standen die externen Testleiter/innen in einem vertraglichen Verhältnis mit dem BIFIE.

5.3 Datengrundlage und Stichprobenziehung

Als Datengrundlage für die Stichprobenziehung der D4-Studie 2018 auf Schulebene dienten die vom BMB bereitgestellten Schul- und Klassendaten aus dem Schuljahr 2016/17 (siehe Tabelle 1). Nachgelagerte Ziehungen (Klassen- oder Schülerebene) basierten auf Informationen, die direkt von den Schulen erhoben wurden.

Tabelle 1: Vom BMB gemeldete Datengrundlage

Bundesland	Anzahl Schulen	Anzahl Schüler/innen	Schüler/innen in Prozent
Burgenland	168	2.481	3,28
Kärnten	227	4.736	6,26
Niederösterreich	599	14.435	19,09
Oberösterreich	538	13.408	17,73
Salzburg	177	4.783	6,32
Steiermark	450	10.080	13,33
Tirol	364	6.547	8,66
Vorarlberg	162	3.828	5,06
Wien	258	15.329	20,27
Gesamtergebnis	**2.943**	**75.627**	**100,00**

5.3.1 Ausschluss von Klein- und Kleinstschulen

Bei Pilotierungen im Rahmen der Bildungsstandardüberprüfungen nehmen nur Schulen mit einer bestimmten Mindestgröße teil. Der Ausschluss von Klein- und Kleinstschulen hat vor allem organisatorische und finanzielle Gründe. Beispielsweise liegen Klein- und Kleinstschulen oft sehr dezentral, und außerdem könnten in diesen Schulen pro Testsitzung nur sehr wenige Schüler/innen getestet werden. Die definierten Größen der ausgenommenen Schulen sind datengetrieben, das heißt, die Ausschlüsse sollen einen möglichst kleinen Anteil an der Schülerpopulation darstellen. Die Kriterien überschneiden sich, so sind z. B. alle Kleinstschulen auch Kleinschulen. Es wird davon ausgegangen, dass die vorgenommenen Ausschlüsse die Testergebnisse nur unwesentlich beeinflussen.

Kleinstschulen

In die erste Kategorie von von der D4-Studie 2018 ausgenommenen Schulen fallen die kleinsten Schulen im österreichischen Schulsystem. Als Kleinstschulen werden Schulen bezeichnet, die in Summe nur eine Klasse über alle Schulstufen hinweg führen. Dies traf bundesweit auf 197 Schulen (6,69 %) mit 849 Schülerinnen und Schülern (1,12 %) der Datenbasis vom BMB zu.

Kleinschulen

Neben Kleinstschulen wurden Kleinschulen ebenfalls von der D4-Studie 2018 ausgenommen. Kleinschulen sind dabei Schulen, die entweder zwei bis drei Klassen im Bereich 1. bis 4. Schulstufe oder maximal 12 Schüler/innen auf der 4. Schulstufe aufweisen. Diese Kriterien trafen auf 1.026 Schulen (34,86 %) mit 8.504 Schülerinnen und Schülern (11,24 %) der Datengrundlage zu.

Schulen mit zu geringer Klassengröße

In die dritte Kategorie von ausgenommenen Schulen fallen schließlich auch noch Schulen mit einer zu geringen Klassengröße. Konkret wurden jene Schulen nicht für die D4-Studie 2018 herangezogen, deren Klassen auf der 4. Schulstufe durchschnittlich weniger als 6 Schüler/innen aufwiesen. Das bedeutete einen Ausschluss von 1.175 Schülerinnen und Schülern (1,55 %) an 278 Schulen (9,44 %).

In Summe wurden 1.033 aller vom BMB gemeldeten Schulen (35,1 %) und 8.689 Schüler/innen (11,49 %) von der D4-Studie 2018 ausgenommen, da mindestens ein Ausschlusskriterium bzgl. der Schulgröße auf sie zutraf. Nach dem Ausschluss aller Kleinstschulen, Kleinschulen sowie Schulen mit zu geringer Klassengröße verblieben 66.938 Schüler/innen (88,51 %) an 1.910 Schulen (64,89 %) als Grundlage für die Stichprobenziehung. Obwohl relativ viele Schulen aufgrund ihrer Größe von der Datenbasis ausgenommen werden, betrifft dies nur einen Bruchteil der Schülerpopulation.

5.3.2 Stichprobenziehung I

Um die Repräsentativität der für Pilotierungen verwendeten Stichproben zu gewährleisten, wird ein zweistufiges Verfahren nach der *Probability-Proportional-to-Size-Sampling*-Methode mit Stratifizierung nach expliziten und impliziten Strata vorgenommen (George, Itzlinger-Bruneforth & Trendtel, 2015). Die Stratifizierungsvariablen sind Gruppierungsmerkmale, innerhalb derer die Variabilität der Zielvariable (z. B. Schülerleistung) geringer ist als zwischen den Strata. Zunächst wird die Population anhand der expliziten Strata, in der Volksschule sind dies die Bundesländer Österreichs, in Teilgruppen untergliedert, aus denen dann *Proportional-to-Size*-Stichproben gezogen werden. Anschließend wird innerhalb der expliziten Strata anhand weiterer Merkmale impliziter Stratifizierungsmerkmale sortiert, um die Realibilität, also die Genauigkeit der Stichrprobenschätzer zu verbessern (OECD, 2012). Die verwendeten impliziten Strata sind der Schulbezirk und die Schulgröße. Dadurch wird eine gleichverteilte Auswahl der Schulen über diese Variablen hinweg bewerkstelligt. Sofern die verwendeten Merkmale mit der Zielvariable korreliert sind, verbessert dies infolge die Stichprobenschätzer. Aus der stratifizierten Datengrundlage werden entsprechend dem Anteil des Stratums an der Population Schulen für die Stichprobe gezogen.

Im zweiten Schritt wurde je Schule genau eine Klasse per Zufallsauswahl ermittelt, die an der D4-Studie 2018 teilnahm. Insgesamt wurden 278 Schulen gezogen. Zum Zeitpunkt des Samplings I entsprach dies einer erwarteten Stichprobengröße von n = 4.990 Schülerinnen und Schülern. Eine Übersicht über das Ergebnis der Stichprobenziehung findet sich in Tabelle 2. Vertiefende technische Details zur Stichprobenziehung lassen sich dem Technischen Bericht (George, Itzlinger-Bruneforth & Trendtel, 2015) oder dem Kapitel „Stichprobenziehung“ im Buch „Large-Scale Assessment mit R“ (George, Oberwimmer & Itzlinger-Bruneforth, 2016, S. 51–78) entnehmen.

Tabelle 2: Übersicht über das Ergebnis der ersten Stichprobenziehung

Bundesland	Stichprobe *n*			Population *N*		
	Anzahl Schulen	Anzahl erwarteter Schüler/innen	erwartete Schüler/innen in Prozent	Anzahl Schulen	Anzahl Schüler/innen	Schüler/innen in Prozent
Burgenland	8	133	2,67	62	1.620	2,42
Kärnten	18	288	5,77	141	3.946	5,90
Niederösterreich	52	927	18,58	401	12.546	18,74
Oberösterreich	50	870	17,43	387	11.993	17,92
Salzburg	18	343	6,87	122	4.335	6,48
Steiermark	36	642	12,87	277	8.673	12,96
Tirol	23	419	8,40	178	5.299	7,92
Vorarlberg	16	257	5,15	90	3.303	4,93
Wien	57	1.111	22,26	252	15.223	22,74
Gesamtergebnis	278	4.990	100,00	1.910	66.938	100,00

5.3.3 Stichprobenziehung II und Testheftrotation

Die Stichprobenziehung II bestimmte jene Klassen, die an der D4-Studie 2018 teilnahmen. Pro ausgewählter Schule wurde dabei genau eine Klasse gezogen. Zunächst wurde je Schule eine teilnehmende Klasse per Zufallsauswahl bestimmt und anschließend jeder Schülerin/jedem Schüler ein bestimmtes Testheft zugewiesen. Die Zuteilung erfolgte dabei im Regelfall rotiert, sodass keine zwei aufeinanderfolgenden Schüler/innen gleiche Testhefte erhielten. *Zuhören* stellt dabei allerdings einen Sonderfall dar, da eine teilnehmende Klasse jeweils die gleichen Audiodateien hörte. Insofern beinhalteten die Testhefte auch die gleichen Items zu den vorgegebenen Stimuli. Die Antwortoptionen der Items im geschlossenen Format waren allerdings in gejumbelter Reihenfolge, also durcheinandergewürfelt angeordnet, um ein Abschreiben zu erschweren (siehe dazu Kapitel 3 – Operationalisierung).

Die Ergebnisse der Stichprobenziehung II sind in Tabelle 3 dargestellt. Aufgrund der aktualisierten Datengrundlage weichen sie leicht von den zunächst erwarteten Schülerzahlen (vgl. Tabelle 2) ab.

Tabelle 3: Ergebnis der Stichprobenziehung II

Bundesland	Stichprobe *n*		Population *N*	
	Anzahl Schüler/innen	Schüler/innen in Prozent	Vergleich Anzahl Schüler/innen	Schüler/innen in Prozent
Burgenland	131	2,39	1.620	2,42
Kärnten	322	5,87	3.946	5,90
Niederösterreich	1.020	18,60	12.546	18,74
Oberösterreich	898	16,37	11.993	17,92
Salzburg	369	6,73	4.335	6,48
Steiermark	718	13,09	8.673	12,96
Tirol	432	7,88	5.299	7,92
Vorarlberg	321	5,85	3.303	4,93
Wien	1.273	23,21	15.223	22,74
Gesamtergebnis	**5.484**	**100,00**	**66.938**	**100,00**

5.3.4 Teilstichproben und (Sub-)Studiendesign

Aufgrund einer Konstrukt- sowie Administrationsänderung beim *Zuhören* waren mehrere Teilstichproben innerhalb der D4-Studie 2018 notwendig, um verschiedene Effekte berechnen und berücksichtigen zu können (siehe Kapitel 6 – Verlinkung und Kapitel 8 – Konstruktäquivalenz). An dieser Stelle soll nur auf die Rahmenbedingungen des Studiendesigns eingegangen werden, für detailliertere Einblicke in den (allgemeinen) Prozess des

Testdesigns auf Testheftebene sei auf das Kapitel „Testdesign" von Kiefer, Kuhn und Fellinger (2016, S. 149–181) verwiesen.

Die D4-Pilotierung im *Zuhören* setzt sich aus vier Substudien und folglich vier Teilstichproben zusammen (siehe Abbildung 1). Um trotz der Änderungen im Vergleich zur Erhebung von 2015 verlinken zu können, wurde eine Brückenstudie durchgeführt. Dabei wurden zwei Testhefte mit Items der Bildungsstandardüberprüfungen D4 2015 mit alter Administrationsform eingesetzt. Über die Annahme äquivalenter Gruppen lässt sich die Brückenstudie mit der Substudie verbinden (Näheres dazu siehe Kapitel 6 – Verlinkung). Für die Berechnung der Auswirkungen der Administrationsänderungen wurde eines der beiden Testhefte von 2015 in einer eigenen Substichprobe mit neuer Administrationsform eingesetzt. Durch den Vergleich zum Einsatz mit der alten Administrationsform in der Brückenstudie lässt sich so der Unterschied ermitteln, der auf die Administration zurückzuführen ist. Eine weitere Stichprobe wurde zur Validierung verwendet. Dabei wurden hybride Testhefte eingesetzt, welche sowohl aus alten als auch aus neuen Items bestehen und mit der neuen Administrationsform vorgegeben wurden. Zuletzt verbleibt die eigentliche Pilotierungsstudie, in der anhand des neuen Konstrukts neu entwickelte Items eingesetzt wurden, inklusive neuer Administrationsform (siehe dazu Kapitel 2 – Konstrukt und Kapitel 3 – Operationalisierung). Eine kompakte Übersicht über die Substudien lässt sich Tabelle 4 entnehmen.

Vorab wurden Überlegungen hinsichtlich notwendiger Stichprobengrößen zur ausreichend genauen Berechnung der jeweiligen Effekte vorgenommen, wonach in der Brücken-, der Administrations- und der Validierungsstudie zumindest 250 Schüler/innen ein Testheft bearbeiten sollten. Für die Pilotierungsstudie galt ein Richtwert von 150 Schülerinnen und Schülern je Item.

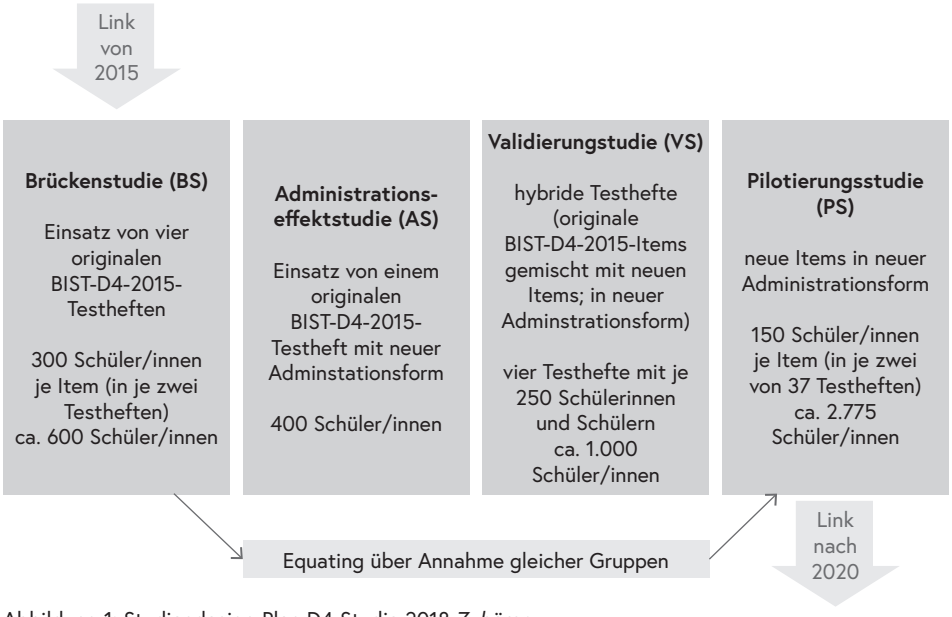


Abbildung 1: Studiendesign-Plan D4-Studie 2018 *Zuhören*

Für die einzelnen Teilbereiche (Brückenstudie [BS], Administrationseffektstudie [AS], Validierungsstudie [VS] und die verbleibende Pilotierung [PS]) wurde die Stichprobe auf ebenfalls repräsentative Substichproben aufgeteilt. Auch hier wurde stratifiziert und *Proportional-to-Size* gesampelt. Vorab wurde mit ca. 4.800 Schülerinnen und Schülern gerechnet, nach der Stichprobenziehung I wurde von 4.990 Schülerinnen und Schülern ausgegangen. In der nachgelagerten Stichprobenziehung II, welche mit aktuelleren Daten durchgeführt wurde, die von den Schulen selbst stammen, fanden sich insgesamt 5.484 Schüler/innen. Diese Größenunterschiede ergeben sich einerseits durch die Aktualisierung der Datengrundlage und andererseits daraus, dass vorab mit der durchschnittlichen Klassengröße gerechnet wurde, bei der tatsächlichen Ziehung dann aber mit höherer Wahrscheinlichkeit größere Klassen gezogen werden. Dadurch resultiert schlussendlich eine größere Stichprobe. Diese wurde auf die vier Teilstudien aufgeteilt (siehe Tabelle 4 und Tabelle 5). Aus Erfahrung von früheren Samplings und Pilotierungen gleicht sich dieser Überschuss in der Stichprobe durch Ausfälle (krank, nicht bearbeitet, sonst. Verhinderung) zumeist wieder aus. Die Anzahl bearbeiteter Testhefte betrug insgesamt 5.044, was nahe an den ursprünglich erwarteten 4.990 Schülerinnen und Schülern liegt (siehe Tabelle 4).

Tabelle 4: Verteilung der Stichprobe *Zuhören* auf Teilstudien

Teilstudie	Abkürzung	Items	Administration	Anzahl Testhefte	N Schüler/innen Sampling	N Testhefte bearbeitet
Brückenstudie	BS	alt	alt	4	878	806
Administrationseffektstudie	AS	alt	neu	1	540	490
Validierungsstudie	VS	alt & neu	neu	8	1.028	950
Pilotierungsstudie	PS	neu	neu	37	3.038	2.798
Gesamt				**50**	**5.484**	**5.044**

5.4 Durchführung der Pilotierung

Um die Schüler/innen der 4. Schulstufe nicht zu überfordern, fand die Studie in Deutsch 2018 – wie auch die Bildungsstandardüberprüfung in Deutsch 2015 – an zwei aufeinanderfolgenden Tagen statt (regulärer Testtermin: 18. und 19. April 2018; Ersatztermin: 24. und 25. April 2018). Die Testmaterialien wurden in 278 Paketen am 5. April 2018 an den jeweiligen Schulstandorten zugestellt.

5.4.1 Ablauf der D4-Studie 2018

Schulleitung und externe Testleitung trafen sich am Testtag rund 30 Minuten vor Beginn der D4-Studie 2018, um die notwendigen Vorbereitungsarbeiten zu treffen (Übergabe

Tabelle 5: Verteilung Schüleranzahl auf Substudien im Vergleich zur Population im Bereich *Zuhören*

Bundesland	Brückenstudie		Administrationseffekt-studie		Validierungsstudie		Pilotierungsstudie		Population *N*	
	Anzahl Schüler/innen	Schüler/innen in Prozent	Anzahl Schüler/innen	Schüler/innen in Prozent	Anzahl Schüler/innen	Schüler/innen in Prozent	Anzahl Schüler/innen	Schüler/innen in Prozent	Anzahl Schüler/innen	Schüler/innen in Prozent
Burgenland	21	2,39	15	2,78	31	3,02	64	2,11	1.620	2,42
Kärnten	42	4,78	41	7,59	66	6,42	173	5,69	3.946	5,90
Niederösterreich	166	18,91	97	17,96	193	18,77	564	18,56	12.546	18,74
Oberösterreich	141	16,06	81	15,00	164	15,95	512	16,85	11.993	17,92
Salzburg	60	6,83	40	7,41	79	7,68	190	6,25	4.335	6,48
Steiermark	136	15,49	57	10,56	127	12,35	398	13,10	8.673	12,96
Tirol	67	7,63	38	7,04	72	7,00	255	8,39	5.299	7,92
Vorarlberg	61	6,95	23	4,26	60	5,84	177	5,83	3.303	4,93
Wien	184	20,96	148	27,41	236	22,96	705	23,21	15.223	22,74
Gesamtergebnis	**878**	**100,00**	**540**	**100,00**	**1.028**	**100,00**	**3.038**	**100,00**	**66.938**	**100,00**

und Entsiegelung des Testmaterials, Vorbereitung der Testsitzung etc.). Die Studie begann jeweils mit der ersten Unterrichtseinheit und dauerte inklusive Begrüßung und Pausen je Testtag rund 2 Stunden. In jeder Klasse wurde eine von zwei Durchführungsvarianten angewandt.

Tabelle 6: Ablauf der Studie

Testtag	Testteil	Bearbeitungsdauer	
		Variante A	Variante B
Testtag 1	Lückendiktat	ca. 5 min	ca. 5 min
	Rechtschreiben und Sprachbetrachtung	33 min	33 min
	Verfassen von Texten (Textproduktion)	20 min	20 min
	Lesen: Teil 1	1 min	1 min
	Schülerfragebogen	ca. 20 min	-
Testtag 2	**Zuhören**	**ca. 20 min**	**ca. 25 min**
	Verfassen von Texten (Prozessorientierte Teilkompetenzen)	20 min	-
	Lesen: Teil 2	30 min	30 min
	Schülerfragebogen	-	ca. 20 min

Wie aus Tabelle 6 ersichtlich, wurde der Testteil *Zuhören* jeweils zu Beginn des zweiten Testtags administriert. Beide Varianten beinhalten zwar gleich viele Items, sind aber aufgrund der Audiolängen der verschiedenen Items nicht exakt gleich lang (siehe Kapitel 3 – Operationalisierung). Es wurden vier unterschiedliche Versionen des Schülerfragebogens eingesetzt, um eine größere Anzahl verschiedener Fragen abdecken zu können und gleichzeitig die Belastung der Schüler/innen so gering wie möglich zu halten. Zusätzlich zu den Schülerfragebögen kamen auch Elternfragebögen zum Einsatz, die im Rahmen der Standardüberprüfung Mathematik an die Schüler/innen ausgegeben wurden. Eltern, deren Kinder an der D4-Studie 2018 teilnahmen, erhielten eine eigene Version des Fragebogens.

Nach der jeweiligen Testsitzung wurde das Testmaterial von Schulleitung und Testleitung versiegelt und ein Versiegelungsprotokoll unterzeichnet. Die Schulleitung übernahm eine Teilnahmeliste, auf der von der Testleitung der Anwesenheitsstatus der Schüler/innen protokolliert wurde, und übermittelte dem BIFIE den Teilnahmestatus der Schüler/innen über die Onlineplattform PLATO. Die versiegelten 278 Testpakete wurden am 2. Mai 2018 von den Schulen abgeholt und dem BIFIE zur weiteren Bearbeitung zugestellt.

5.4.2 Aufgaben der Testleiter/innen

Die Aufgaben der externen Testleiter/innen am Testtag umfassten Vorbereitung, Durchführung und Nachbereitung der Testsitzung (siehe Abschnitt 5.2.3). Im Rahmen der Vorbereitung trafen sich die Testleiter/innen ca. 30 Minuten vor Testbeginn mit den Schulleitungen und übernahmen die Testmaterialien. Ab diesem Zeitpunkt verantworteten die Testleiter/innen die Gewährleistung der Sicherheit der Testmaterialien. Nach der gemeinsamen Entsiegelung des Testmaterials kontrollierte die Testleitung die Testmaterialien hinsichtlich Vollständigkeit und Korrektheit. Unmittelbar vor der Durchführung der Studie bereitete die Testleitung den Testraum für die Überprüfung vor (Arbeitsplätze vorbereiten, Funktionstüchtigkeit des CD-Players überprüfen, Fineliner verteilen etc.).

Um ein hohes Maß an Standardisierung zu gewährleisten, wurden bei der Studiendurchführung alle Instruktionen wortwörtlich vorgelesen und Ablauf sowie Testzeiten exakt eingehalten. Die Testleiter/innen durften inhaltliche Fragen der Schüler/innen zu den Testitems nicht beantworten und mussten sichergehen, dass während der Testsitzung keine elektronischen Geräte (z. B. Handy, Laptops, Kameras) verwendet wurden. Während der Testsitzung wurden außerdem verschiedene Formulare (Teilnahmeliste, Testsitzungsprotokoll) von der Testleitung ausgefüllt, um die genauen Testzeiten und die Anwesenheit der Schüler/innen zu protokollieren.

Im Anschluss an die Testsitzung wurden die Testmaterialien wieder gemeinsam mit der Schulleitung versiegelt und das Versiegelungsprotokoll unterzeichnet. Während der Testtage stand den Testleitungen und Schulleitungen ein Hotline-Support bei Fragen und Problemen zur Verfügung.

5.5 Aufbereitung und Analyse der Daten

In diesem Abschnitt wird auf die *Datenerfassung* und die *Datenaufbereitung* im Anschluss an die Durchführung der D4-Studie 2018 eingegangen.

5.5.1 Datenerfassung

Der erste Schritt der Datenerfassung bestand in der Rücklaufkontrolle der gesamten Testmaterialien. Dabei wurde erfasst, ob die Testmaterialien vollständig zurückgeschickt wurden. Von den 41.255 eingesetzten Testheften und Fragebögen wurden 41.252 retourniert. Das entspricht einer Rücklaufquote von 99,99 %. Im Testteil *Zuhören* wurden 5.761 von 5.762 Testheften an das BIFE retourniert. Die Anzahl der Testhefte liegt über der Zahl der Schüler/innen, da jede Klasse zusätzlich ein Reservetestheft erhält. Die Gesamtzahl der

versendeten Testhefte setzt sich folglich aus 5.484 regulären Testheften plus 278 Reservetestheften zusammen, insgesamt 5.762.

Tabelle 7: Retournierte und bearbeitete Testhefte

Testteil	Gesamtzahl retournierter Testhefte	retournierte Testhefte Prozent	Gesamtzahl bearbeiteter Testhefte	bearbeitete Testhefte in Prozent
Lückendiktat	5.762	100,00	5.054	87,71
Zuhören	**5.761**	**99,98**	**5.045**	**87,57**
Lesen: Teil 2	5.762	100 ,00	5.039	87,45
Rechtschreiben und Sprachbetrachtung	5.761	99,98	5.054	87,73
Schülerfragebogen	5.761	99,98	5.037	87,43
Lesen: Teil 1	5.762	100,00	5.022	97,16
Verfassen von Texten (Textproduktion)	5.762	100,00	4.813	83,53
Verfassen von Texten (Prozessorientierte Teilkompetenzen)	921[a]	100,00	806	87,51

[a] Verfassen von Texten (Prozessorientierte Teilkompetenzen) wurde nur in einer Teilstichprobe eingesetzt.

Die kontrollierten Testhefte und Fragebögen wurden mithilfe von Hochleistungsscannern digitalisiert. Eine Datenerfassungssoftware wertete dabei die von den Schülerinnen und Schülern geschlossenen Items wie etwa Multiple-Choice-Items elektronisch aus (zu den Itemformaten, die im Testteil *Zuhören* zum Einsatz kamen, siehe Kapitel 3 – Operationalisierung). Antworten, die nicht eindeutig identifiziert werden konnten (durch Mehrfachantworten, Scan- oder Druckfehler), wurden von geschultem Personal gesichtet und anhand genau definierter Richtlinien nachbearbeitet. 28.405 von 2.152.250 Schülerantworten mussten auf diese Weise nachbearbeitet werden. Dies entspricht einer Nachbearbeitungsquote von 1,32 %. Im Testteil *Zuhören* zeigte sich ein ähnliches Bild: Von insgesamt 390.372 elektronisch ausgelesenen Schülerantworten mussten 4.198 nachbearbeitet werden. Das entspricht einer Nachbearbeitungsquote von 1,08 %.

Neben den geschlossenen Itemformaten, die von der Software interpretiert und ggf. verifiziert wurden, gab es ebenfalls offene bzw. halboffene Formate. Da bei diesen Formaten die Antworten frei formuliert wurden, war eine Bewertung der Schülerantworten durch entsprechend geschultes Personal sowie Expertinnen und Experten erforderlich. Im Rahmen der Standardüberprüfung erfolgte dies in der Plattform für Coding & Rating (CORA)[4], einer Webanwendung zur Bewertung von Schülerantworten mithilfe gescannter Bilddateien.

4 https://cora-login.bifie.at/

Die gescannten 16.963 Schülerantworten der 69 Kodieritems des Testteils *Zuhören* wurden den Expertinnen und Experten auf der Plattform angezeigt und von ihnen als „richtig" oder „falsch" kodiert. Um sichergehen zu können, dass diese Formate von den zahlreichen verschiedenen Koderinnen und Kodern gleich bewertet wurden und nicht eine Person bei einem Format strenger als eine andere bewertete, wurden *Single-* und *Multiple-Coding* eingesetzt. Beim *Single-Coding* wurde eine Schülerantwort nur von einer Koderin/einem Koder bewertet. Beim *Multiple-Coding* bewerteten hingegen mehrere Koder/innen dieselbe Schüleraufgabe unabhängig voneinander. Ziel war die Feststellung der Qualität der Kodierung, gemessen an der Übereinstimmung zwischen den Koderinnen und Kodern. Im *Zuhören* konnten im Durchschnitt 81,53 %[5] Übereinstimmung erreicht werden.

5.5.2 Datenaufbereitung

In der Datenaufbereitung wurden die gewonnenen Daten der unterschiedlichen Quellen zu einem Gesamtdatensatz zusammengeführt. Außerdem wurden Qualitäts- und Vollständigkeitskontrollen durchgeführt. So wurde beispielsweise überprüft, ob für alle eingesetzten Testitems nur gültige und plausible Werte in den Daten vorkamen. Bei auffälligen Werten musste ausgeschlossen werden, da diese Werte auf einen technischen Fehler in der Datenerhebung zurückzuführen waren. Weiters wurde eine Datenbereinigung der Rohdatensätze vorgenommen. Zum Bereich der Datenaufbereitung gehörte ebenfalls, Informationen aus diversen Quellen zu berücksichtigen und ggf. notwendige Korrekturen an den Daten vorzunehmen. Zu diesem Zweck wurden die Informationen aus 546 retournierten Testsitzungsprotokollen und Teilnahmelisten ausgewertet und etwaige Auffälligkeiten an den Daten mithilfe dieser Informationen behoben. In 16 Fällen wurden Testheftvertauschungen festgestellt und entsprechend korrigiert.

Tabelle 8: Testheftvertauschungen

Testteil	Anzahl vertauschter und korrigierter Testhefte	Häufigkeit in Prozent
Lückendiktat	6	0,1
Zuhören	**2**	**0,03**
Lesen: Teil 2	2	0,03
Rechtschreiben und Sprachbetrachtung	2	0,03
Schülerfragebogen	0	0
Lesen: Teil 1	4	0,07
Verfassen von Texten (Textproduktion)	0	0
Verfassen von Texten (Prozessorientierte Teilkompetenzen)	0	0

5 Übereinstimmung um zufällige Übereinstimmung bereinigt. Als Maßzahl wurde Cohens Kappa verwendet, einem statistischen Maß für die Übereinstimmung von Beurteilerinnen und Beurteilern (auch: Interrater-Reliabilität).

Beim Testteil *Zuhören* wurde zudem überprüft, ob durch technische Defekte der gelieferten CDs Schüler/innen Testaufgaben nicht bearbeiten konnten. Dazu wurde in den Testsitzungsprotokollen folgende Frage gestellt: „Führten technische Defekte der CDs dazu, dass einzelne Hörbeiträge des Testteils *Zuhören* nicht bearbeitet werden konnten, weil sie von keiner der beiden CDs abspielbar waren?" Diese Frage wurde nur in einer einzigen Testsitzung bejaht. Eine Kontrolle der Schülerantworten dieser Klasse ergab jedoch keinerlei Auffälligkeiten und der beanstandete Hörbeitrag wurde von allen Schülerinnen und Schülern bearbeitet. Es wurde in diesem Fall deshalb von einer minderen technischen Beeinträchtigung ausgegangen.

5.6 Realisierte Stichprobe und Teilnahmequote im Bereich *Zuhören*

Die Teilnahmequote bei der D4-Studie 2018 war generell hoch. Von den 5.484 regulären Testheften im *Zuhören* wurden 5.013 bearbeitet, wobei ein Heft von einer Schülerin/einem Schüler mit einem Ausschlussgrund bearbeitet wurde.[6] Von den 278 Reservetestheften wurden 37 verwendet, wovon 32 auch bearbeitet wurden. Die restlichen 241 Reservetesthefte wurden nicht benötigt. Insgesamt befanden sich in der Stichprobe somit 5.044 bearbeitete Testhefte, was einer Teilnahmequote von knapp 92 % (5.044 realisierte Stichprobe/5.484 erwartete Stichprobe) entspricht. Die Geschlechterverteilung war mit 2.505 Mädchen (49,7 %) und 2.539 Burschen (50,3 %) ausgeglichen und repräsentativ für die 4. Schulstufe (vgl. z. B. BIFIE, 2019). Das durchschnittliche Alter betrug 10,4 Jahre (SD = 0,48 Jahre) mit einer Spanne von 9,08 bis 14,08 Jahren. Der Migrationsstatus konnte aufgrund fehlender Elternangaben in den Kontextfragebögen nicht für alle Schüler/innen ermittelt werden (296 fehlende Werte). Von den verbleibenden 4.748 Schülerinnen und Schülern wiesen 1.070 (22,5%) einen Migrationshintergrund[7] auf.

Um etwaige Verzerrungen durch Antwortausfälle[8] (Non-Responses) zu vermeiden, wurden für die statistischen Analysen Schülergewichte verwendet. Zusätzlich wurden durch die Verwendung von Gewichten die unterschiedlichen Ziehungswahrscheinlichkeiten der Schüler/innen ausgeglichen. Für eine detaillierte Beschreibung des Vorgehens sei auf den Technischen Bericht zur Stichprobenziehung in Bildungsstandardüberprüfungen und Pilotierungen verwiesen (George, Itzlinger-Bruneforth & Trendtel, 2015). Alle Zahlen und Tabellen im vorliegenden Kapitel stellen deskriptive Statistiken dar und basieren daher auf ungewichteten Daten.

6 Schüler/innen mit einem Ausschlussgrund sollten nicht an der Pilotierung der D4-Studie teilnehmen. Das entsprechende Testheft wurde von allen Auswertungen ausgenommen.

7 BIFIE-Definition von Migrationshintergrund in Anlehnung an die Definition der OECD: Einem Kind wird Migrationshintergrund zugeordnet, wenn beide Elternteile im nichtdeutschen Ausland geboren wurden. Aufgrund der sprachlichen Nähe wird Kindern, deren Eltern aus Deutschland stammen, kein Migrationshintergrund zugeordnet.

8 Damit ist der Ausfall sämtlicher Antworten gemeint (z. B. wegen Krankheit), nicht aber fehlende Angaben zu einzelnen Items (Item Non-Response).

Tabelle 9: Ausschlussgründe und Bearbeitet-Status nach Testhefttyp

Testhefttyp	Bearbeitet-Status	Ausschlusscode[a]						Gesamtergebnis
		0	1	2	4	7	NA	
Normales Testheft	nicht bearbeitet	225	87	24	6	129	0	471
	bearbeitet	5.012	1	0	0	0	0	5.013
Reservetestheft	nicht bearbeitet	3	2	0	0	0	241	246
	bearbeitet	32	0	0	0	0	0	32
Gesamtergebnis		**5.272**	**90**	**24**	**6**	**129**	**241**	**5.762**

[a]Anmerkung: 0 = kein Ausschluss; 1 = Außerordentliche/r Schüler/in; 2 = Körper- oder Sinnesbeeinträchtigung; 4 = ASO-Lehrplan Deutsch; 7 = ASO-Lehrplan Deutsch & Mathematik; NA = nicht angegeben

Tabelle 10: Bearbeitet-Status je Bundesland und Vergleich der Prozentanteile der Bundesländer in der realisierten Stichprobe gegenüber der Population

	Bearbeitet-Status			Realisierte Stichprobe	Population
Bundesland	nicht bearbeitet	bearbeitet	Gesamtergebnis	Schüler/innen in Prozent (bearbeitet)	Schüler/innen in Prozent
Burgenland	18	121	139	2,40	2,42
Kärnten	44	296	340	5,87	5,90
Niederösterreich	107	965	1.072	19,13	18,74
Oberösterreich	101	846	947	16,77	17,92
Salzburg	47	341	388	6,76	6,48
Steiermark	105	649	754	12,87	12,96
Tirol	49	406	455	8,05	7,92
Vorarlberg	53	284	337	5,63	4,93
Wien	194	1.136	1.330	22,52	22,74
Gesamtergebnis	**718**	**5.044**	**5.762**	**100,00**	**100,00**

Zu den 5.044 bearbeiteten Testheften gab es 4.826 ausgefüllte Schülerfragebögen, sprich 95,6 % der Schüler/innen, die den Test bearbeiteten, füllten auch den Schülerfragebogen aus. Weitere 171 Fragebögen wurden von Schülerinnen und Schülern ausgefüllt, die den Test nicht bearbeiteten. Dies betrifft ausschließlich Schüler/innen, die den Fragebogen am ersten Testtag ausfüllten und am zweiten Testtag nicht anwesend waren (Variante A, siehe Tabelle 6).

Bei den Elternfragebögen gab es zwei verschiedene Versionen, wobei eine Version speziell für die D4-Studie 2018 entwickelt und verwendet wurde. Zu den 5.044 von Schülerinnen und Schülern bearbeiteten Heften gab es 4.237 bearbeitete Elternfragebögen und weitere 138 zu Schülerinnen und Schülern, die den Test nicht bearbeiteten (analog zum Schülerfragebogen). Insgesamt gab es Fragebogendaten von 4.375 Eltern bzw. Erziehungsberechtigten, was unter Berücksichtigung der freiwilligen Teilnahme sehr hoch ist (4.375/5.044 = 86,7 %).

Für die Stichprobe der D4-Studie 2018 wurde der Lehrerfragebogen um einen zusätzlichen deutschspezifischen Teil erweitert. Für 564 Unterrichtsgruppen bzw. 538 Klassen an 250 Schulen liegen Lehrerantworten vor. Die Schulleiterfragebögen wurden an allen 278 Schulen der Stichprobe ausgefüllt.

5.7 Schlusswort

Abschließend soll nochmals die Wichtigkeit der qualitätssichernden Maßnahmen herausgestrichen werden, die im vorliegenden Kapitel angeführt wurden. Diese kommen in der Vorbereitung etwa in der Kommunikation zu wichtigen Stakeholderinnen und Stakeholdern zum Einsatz. Auch in der Durchführung, etwa durch die externen Testleiter/innen, spielt die Qualitätssicherung eine wichtige Rolle. In der Nachbereitung wird die Qualität etwa durch die Kontrollen bei der Datenerfassung und -aufbereitung sichergestellt.

Letztlich ist nochmals auf die grundlegende Bedeutsamkeit einer repräsentativen Stichprobenziehung hinzuweisen. Um aussagekräftige Rückschlüsse auf die interessierende Grundgesamtheit treffen zu können, muss die verwendete Stichprobe in den relevanten Merkmalen die gleiche Zusammensetzung aufweisen. Das vorliegende Kapitel zeigt auf, dass dies für die Stichprobe der D4-Studie 2018 sowie für alle Substudien gewährleistet ist. Zum einen für die Verteilung über die Bundesländer als explizites und zum anderen für verschiedene Schulgrößen als implizites Stratifizierungsmerkmal. Dadurch ist die Stichprobe infolge auch für diverse Schülermerkmale repräsentativ (Geschlecht, Migrationshintergrund etc.).

Die Basis, um die Ergebnisse der D4-Studie 2018 auf die Population verallgemeinern zu können, ist damit gegeben. Die Ergebnisse der Pilotierungsstudie werden in Kapitel 7 – Ergebnisse dieses Berichtsbands auch unter Berücksichtigung der hier beschriebenen Merkmale dargestellt.

Literatur

BIFIE (2019). *Standardüberprüfung 2018. Mathematik, 4. Schulstufe. Bundesergebnisbericht.* Salzburg. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/884d9cba36794bb78355f4df430488117b0740a1/BiSt_UE_M4_2018_Bundesergebnisbericht.pdf

George, A. C., Itzlinger-Bruneforth, U. & Trendtel, M. (2015). *Technischer Bericht zur Stichprobenziehung in Bildungsstandardüberprüfungen und Pilotierungen. Technische Dokumentation – BIST-Ü.* Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/c3da2b6bfba9bc128524409c000030729b1d2b8a/TD_Stichprobenziehung.pdf

George, A. C., Oberwimmer, K. & Itzlinger-Bruneforth, U. (2016). Stichprobenziehung. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 51–78). Wien: Facultas.

Kiefer, T., Kuhn, J.-T. & Fellinger, R. (2016). Testdesign. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 149–181). Wien: Facultas.

OECD (2012). *PISA 2009 Technical Report.* OECD Publishing. doi:10.1787/9789264167872-en

Michael Ober, Matthias Trendtel

Kapitel 6
Verlinkung zur ersten Bildungsstandardüberprüfung und Bestimmung des Administrationseffekts

6.1 Einleitung

Mit den Bildungsstandardüberprüfungen (BIST-Ü), angelegt als Large-Scale Assessments (LSAs), werden in Österreich Kompetenzstände in den Schulfächern Deutsch, Mathematik und Englisch von Schülerinnen und Schülern der 4. und der 8. Schulstufe in regelmäßigen Abständen gemessen. Durch die Periodizität dieser zentral vorgegebenen Leistungsmessungen von Teilen des österreichischen Schulwesens wird das Ziel verfolgt, Schul- und Unterrichtsentwicklung zu unterstützen, indem Entwicklungen mittlerer Kompetenzstände über längere Zeiträume dargestellt werden (Schreiner & Breit, 2016). Hierfür wird, wie es in den meisten LSAs üblich ist, nicht eine Kohorte (z. B. Schüler/innen eines Jahrgangs) in einem längsschnittlichen Design wiederholt gemessen, sondern in einem querschnittlichen Design werden mittlere Kompetenzstände vergleichbarer Kohorten (z. B. Schüler/innen einer bestimmten Schulstufe) über verschiedene Erhebungswellen hinweg untersucht (Lietz, Cresswell, Rust & Adams, 2017). So wurde im Jahr 2015 das Fach Deutsch, einschließlich des Kompetenzbereichs Zuhören, an Schülerinnen und Schülern der 4. Klasse (BIST-Ü D4 2015) getestet, die zweite für das Jahr 2020 geplante Erhebung ist jedoch, bedingt durch die Umstellung der Bildungsstandardüberprüfungen auf die iKM^{PLUS}, entfallen (siehe Kapitel 1 – Übersicht). Für eine Beschreibung des Kompetenzbereichs Zuhören sowie dessen Operationalisierung siehe die Kapitel 1–4.

Kompetenzstände werden bei den BIST-Ü mit Kompetenztests gemessen, die aus einer bestimmten Menge an Items bestehen. Aus dem Antwortverhalten der getesteten Schüler/innen kann dann auf Ausprägungen der zu messenden Kompetenz rückgeschlossen werden. Da das Antwortverhalten sowohl von Personen- als auch von Itemeigenschaften beeinflusst wird, ist es notwendig, gewisse Anforderungen an die eingesetzten Kompetenztests zu stellen, damit unterschiedliches Antwortverhalten in klar zuordenbarer Weise auf die personen- bzw. itemspezifischen Unterschiede zurückgeführt werden kann (González & Wiberg, 2017). Konkret werden bei den BIST-Ü mithilfe einer Menge an gemeinsamen Items – sogenannten Ankeritems – Ergebnisse aus Kompetenztests verschiedener Erhebungswellen in

einem Common-Items-Nonequivalent-Groups-Design (Kolen & Brennan, 2014) mit einem Linking-Verfahren vergleichbar gemacht (siehe Abbildung 1).

Voraussetzung für ein solches Linking ist unter anderem, dass (1) die Ankeritems repräsentativ für das zu messende Konstrukt sind, (2) die Linking-Stichprobe repräsentativ für die zu untersuchende Population ist und (3) die Testadministration sowie (4) die Testinstrumente hinsichtlich Inhalt und Schwierigkeit möglichst konstant gehalten werden (Trendtel, Pham & Yanagida, 2016). Zwischen den Erhebungswellen 2015 und 2018 kam es für den Kompetenzbereich Zuhören allerdings zu Änderungen hinsichtlich der Testadministration und der Operationalisierung, sodass ein Linking mit Standardverfahren unter Umständen keine validen Vergleiche zwischen den beiden Erhebungswellen (siehe Kapitel 1 – Übersicht) zulässt.

6.2 Änderungen in der Administration und Operationalisierung

Im Wesentlichen gab es zwei Änderungen bezüglich der Testadministration: Zum einen wurden allgemeine Bearbeitungshinweise bzw. Instruktionen am Anfang des Testhefts und die Beispielitems am Anfang der Testhefte gestrichen. Stattdessen gab es kurze Instruktionen vor jedem Stimulus. Zum anderen wurde die Möglichkeit der Voransicht einzelner Items, in dem Sinne, dass Schüler/innen bereits während des Hörens eines Stimulus die zugehörigen Items mitlesen können (Vorentlastung oder *Preview*), verhindert.

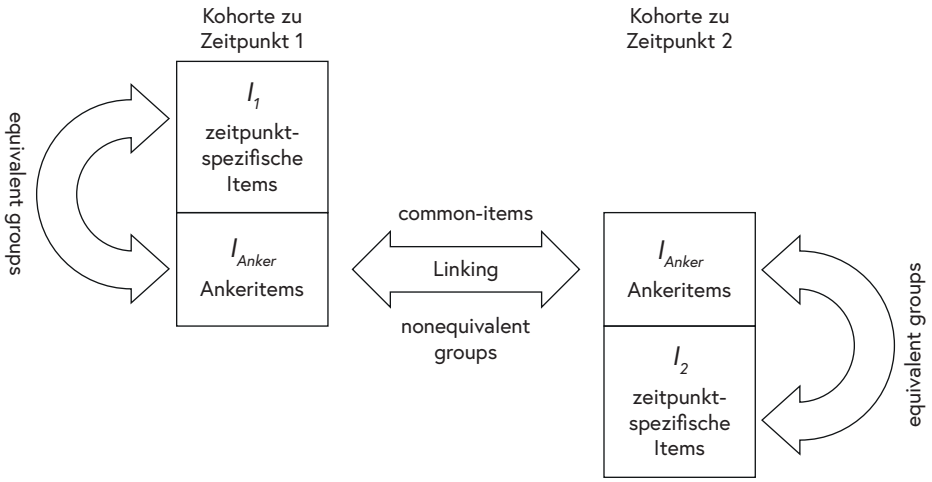


Abbildung 1: Schematische Darstellung eines Common-Items-Nonequivalent-Groups-Design für das Verlinken zweier Studien

Hinsichtlich der Instruktionen zur Bearbeitung einzelner Items wird in der Literatur gefordert, dass Items möglichst eindeutig formuliert sind und den Testpersonen klar ist, wie die einzelnen Items zu bearbeiten sind (Baranowski, 2006). Letzteres wird häufig über klar

formulierte Instruktionen sichergestellt. Wird gegen dieses Prinzip der Itemkonstruktion verstoßen, kann dies das Antwortverhalten negativ beeinflussen (Downing, 2005). Für einzelne Itemformate wiederum konnten teilweise keine Effekte von (kleineren) Änderungen in Iteminstruktionen auf das Antwortverhalten gefunden werden (Violata, 1991). Insgesamt kann allerdings nicht ausgeschlossen werden, dass Änderungen in den Iteminstruktionen das Antwortverhalten beeinflussen, selbst wenn sie – wie im vorliegenden Fall – nicht zu Verstößen gegen Prinzipien der Itemkonstruktion führen.

Über Vorentlastungen im Bereich Textverarbeitung gibt es einen Konsens, dass Vorabinformationen den Aufbau eines zur Verarbeitung des Textes notwendigen situativen Modells beeinflussen (z. B. Schnotz, 1988; siehe dazu auch Kapitel 2 – Konstrukt). So ist insbesondere bei Kompetenzmessungen im Bereich Zuhören anzunehmen, dass bereits mit Kenntnis der Fragestellung vor dem eigentlich Gehörten ein situatives Modell zumindest angelegt wird (z. B. Behrens, 2010). Somit muss bei der Verlinkung der Studien die Möglichkeit berücksichtigt werden, dass diese Änderung bezüglich der Vorentlastung bei den getesteten Kindern einen Einfluss auf das Antwortverhalten hat. Darüber hinaus gibt es empirische Hinweise darauf, dass sich Mädchen und Burschen hinsichtlich des Aufbaus von situativen Modellen bei der Textverarbeitung unterscheiden (z. B. Wassenburg, de Koning, de Vries, Boonstra & van de Schoot, 2017, für den Bereich Lesen), sodass ein differenzieller Effekt nach Geschlecht durch diese Veränderung in der Testadministration nicht ausgeschlossen werden kann. Eine ähnliche Argumentation trifft auf Kinder mit und ohne Migrationshintergrund zu (z. B. McElvany, Ohle, El-Khechen, Hardy & Cinar, 2017).

Neben den Änderungen in der Administrationsform kam es hinsichtlich der Operationalisierung der Zuhörkompetenz bei den österreichischen Bildungsstandardüberprüfungen zu Änderungen. So wurden im Vergleich zur Überprüfung im Jahr 2015 deutlich längere Hörstimuli mit insgesamt mehr Items pro Stimulus eingesetzt. Darüber hinaus wurde erstmals in diesem Kompetenzbereich mit offenen Items getestet und es wurde nur authentisches Hörmaterial als Stimulus eingesetzt. „Authentisch" heißt hier, dass ausschließlich unterrichtsvalide Stimuli verwendet werden, denen Kinder auch im Deutschunterricht begegnen könnten. Bei der inhaltlichen Definition des Konstrukts wurde dem literarischen Bereich ein höheres Gewicht beigemessen, als dies 2015 noch der Fall war. Die Ergebnisse empirischer Untersuchungen hinsichtlich der (Nicht-)Äquivalenz dieser beiden Operationalisierungen werden in Kapitel 8 – Konstruktäquivalenz in diesem Berichtsband dargestellt.

6.3 Untersuchungsanliegen

Um – auch für weitere Studien – abschätzen zu können, welche Auswirkungen Änderungen in der Testadministration auf gezeigte Testleistungen haben, wird untersucht, welcher Unterschied sich im Antwortverhalten zwischen den beiden Administrationsformen zeigt. Konkret wird der Frage nachgegangen:

1. Beeinflussen die Änderung in den Iteminstruktionen und der Wegfall der Möglichkeit zur Vorentlastung das Antwortverhalten von Schülerinnen und Schülern bei der Kompetenzmessung im Bereich Zuhören? Gibt es dahingehend differenzielle Effekte nach Geschlecht und Migrationshintergrund?

Um die Verlinkung der Pilotierung 2018 mit der BIST-Ü 2015 über eine Brückenstudie – und somit eine stabile Trendfortschreibung bezüglich des Kompetenzbereichs Zuhören – zu stützen, wird untersucht, ob sich die notwendige Annahme äquivalenter Gruppen (vgl. Abbildung 1 in Kapitel 5 – Studiendesign dieses Berichtsbands) halten lässt:

2. Unterscheiden sich die Zusammensetzungen von Brückenstudienstichprobe und Pilotierungsstichprobe untereinander bzw. von jener der Gesamtpopulation hinsichtlich demographischer Merkmale und/oder anderer Kompetenzen im Fach Deutsch?

6.4 Methode

6.4.1 Stichprobe

Die Stichprobenziehung für die Pilotierung der D4-Studie 2018 unterteilte sich wie üblich in zwei Schritte: die Schulenziehung und die nachfolgende Klassenziehung. Die Vorgehensweise bei der Schulenziehung entsprach einem stratifizierten Sampling mit dem Bundesland als explizite und dem Bezirk und der Schulgröße als implizite Stratifizierungsvariable (für eine genauere Beschreibung siehe Abschnitt 5.3.1 in diesem Berichtsband). Insgesamt wurden 278 Schulen gezogen, wobei die Ziehung in jedem expliziten Stratum proportional zu den Anteilen der Schüler/innen in dem entsprechenden Bundesland erfolgte. In einem weiteren Schritt wurde diese Stichprobe auf die vier Teilstudien der Pilotierung – d. h. Brücken-, Administrationseffekt-, Validierungs- und eigentliche Pilotierungsstudie – derart aufgeteilt, dass auch für jede Teilstudie das erwähnte Proportionalitätskriterium bezüglich Bundesländern erfüllt war (vgl. Abschnitt 5.3.4). In der Klassenziehung wurde schließlich, basierend auf den aktualisierten Daten der gezogenen und in Folge kontaktierten Schulen, genau eine Klasse pro Schule gesampelt und damit die Anzahl an Schülerinnen und Schülern in den Schulen konkretisiert.

Wie sich die 278 Schulen aus der Schulenziehung und die daraus resultierenden Anzahlen an gesampelten Schülerinnen und Schülern auf die für die Pilotierung relevanten Substudien verteilen, ist in Tabelle 1 ersichtlich. Als Ergänzung ist des Weiteren die entsprechende Anzahl an tatsächlich getesteten Schülerinnen und Schülern (für die es analysierbare Daten gibt) angegeben; diese Werte sind ausfallsbedingt etwas niedriger, erfüllen aber noch die notwendigen Vorgaben an die Stichprobengrößen (vgl. mit Abbildung 1 in Abschnitt 5.3.4). Die prozentuelle Verteilung der gesampelten Schüler/innen in Bezug auf die Bundesländer repräsentiert für alle Teilstudien jene der Gesamtpopulation hinreichend gut (siehe Tabelle 5 in Abschnitt 5.3.4).

Tabelle 1: Verteilung von Schulen und Schülerinnen und Schülern auf die Teilstudien

Teilstudie	Anzahl Schulen	Anzahl gesampelter Schüler/innen	Anzahl getesteter Schüler/innen
Brückenstudie (BS)	43	878	806
Administrationseffektstudie (AS)	26	540	490
Validierungsstudie (VS)	55	1.028	950
Pilotierungsstudie (PS)	154	3.038	2.798
Gesamtstudie	278	5.484	5.044

Im Rahmen von Analysen nach bestimmten demografischen Merkmalen ist auch die Zusammensetzung der Population in den Stichproben der Teilstudien bezüglich Alter, Geschlecht, Migrationshintergrund oder sozioökonomischem Status relevant. Tabelle 2 gibt einen entsprechenden Überblick, die angegebenen Werte berechnen sich aus den getesteten Schülerinnen und Schülern, die auf die zu testende Schülerpopulation hochgewichtet werden. Für Variablen, die in nicht vollständig beobachteter Form vorliegen, da sie indirekt durch Fragebögen erhoben wurden (der Migrationshintergrund über das Geburtsland der Eltern und der sozioökonomische Status, HISEI, über die berufliche Tätigkeit der Eltern), sind zusätzlich die Missing-Raten angegeben. Der HISEI ergibt sich als der höhere *International Socio-Economic Index of Occupational Status* (ISEI) von Vater und Mutter (Ganzeboom, De Graaf & Treiman, 1992).

Tabelle 2: Zusammensetzung der Teilstudien-Populationen nach demografischen Merkmalen

Teilstudie	Gewichtete Stichprobengröße	Alter		Anteil Mädchen (in %)	Anteil Schüler/innen mit Migrationshintergrund (in %)	Missing-Rate (in %)	HISEI		Missing-Rate (in %)
		M	SD				M	SD	
BS	11.298	10,40	0,50	50,50	26,93	6,59	50,68	22,11	9,16
AS	6.767	10,40	0,51	51,03	19,14	7,92	51,62	21,90	10,85
VS	13.983	10,40	0,49	51,40	20,42	5,01	49,35	20,90	7,74
PS	40.122	10,40	0,48	48,40	20,96	5,69	50,23	21,26	9,35
Gesamtstudie	72.169	10,40	0,49	49,55	21,62	5,91	50,25	21,40	9,15
Population	76.447	10,39	0,48	49,13	22,97	–	48,78	21,04	–

Als Quelle für die Werte der Gesamtpopulation in Tabelle 2 dienen die Daten der Vollerhebung, die ebenfalls im Frühjahr 2018 in Mathematik in der 4. Schulstufe (BIST-Ü M4 2018), mit einer Schüler/innen-Anzahl von 73.780 stattgefundenen hat. Das entspricht dem getesteten Anteil von insgesamt 76.447 zu testenden Schülerinnen und Schülern. Fehlende

Werte beim Migrationsstatus und beim HISEI wurden im Rahmen der Analysen für diese Vollerhebung mit dem R-Paket mice (van Buuren et al., 2021) multipel imputiert (Enders, 2010; Lüdtke, Robitzsch, Trautwein & Köller, 2007; Robitzsch, Pham & Yanagida, 2016), sodass keine Missings mehr vorliegen.

Ein Vergleich der einzelnen Teilstudien untereinander und mit der Population anhand der Werte in Tabelle 2 zeigt, dass die entsprechenden Stichproben hinsichtlich der diskutierten demografischen Merkmale eine relativ ähnliche Zusammensetzung besitzen. Aufgrund der Annahme äquivalenter Gruppen als Basis für die in Abschnitt 6.3 erwähnte stabile Trendberichterstattung ist dabei die Gegenüberstellung von Brücken- und Pilotierungsstudie von besonderer Bedeutung. Auch in diesem konkreten Fall gibt es recht gute Übereinstimmungen, mit Ausnahme eines auffällig hohen Anteils an Schülerinnen und Schülern mit Migrationshintergrund in der Brückenstudie bei vergleichbarer Missing-Rate.

Für die Einschränkung der Stichprobe einer Teilstudie auf einzelne Testhefte muss die Repräsentativität hinsichtlich Bundesländern nicht mehr gegeben sein. Beispielsweise ergeben sich konkret für einzelne Testhefte der Brückenstudie aufgrund von kleinen Populationszahlen Verteilungen, bei denen einige Bundesländer gar nicht abgedeckt sind (vgl. mit Tabelle 4 im nächsten Abschnitt 6.4.2). Ferner können für einzelne Testhefte auch die Kennwerte der betrachteten demografischen Merkmale stärker vom Populationswert abweichen (siehe Tabelle 5 ebenfalls in Abschnitt 6.4.2).

6.4.2 Messinstrument

Als Ergebnis des Testdesigns für die Pilotierung mit insgesamt 50 Testheften, wobei jumblingbedingte Testheft-Varianten (vgl. Abschnitt 5.3.3) nicht unterschieden werden, gibt es insbesondere vier Testhefte für die Brückenstudie und ein Testheft für die Administrationseffektstudie (neben acht bzw. 37 Testheften für die Validierungs- bzw. Pilotierungsstudie); Jumbling bedeutet, dass die Antwortoptionen bei geschlossenen Itemformaten vertauscht sind. Das Testheft der Administrationseffektstudie stimmt inhaltlich, d. h. im Hinblick auf die darin enthaltenen Stimuli und Items, mit einem der Testhefte aus der Brückenstudie überein und unterscheidet sich lediglich durch die neue, bereits beschriebene Administrationsform. Der Fokus liegt in diesem Kapitel – gemäß dem in Abschnitt 6.3 dargelegten Untersuchungsanliegen – auf der Brücken- und der Administrationseffektstudie; die Validierungsstudie wird hingegen in Kapitel 8 – Konstruktäquivalenz – und die eigentliche Pilotierungsstudie in Kapitel 7 – Ergebnisse – genauer betrachtet.

Bei den vier Testheften der Brückenstudie handelt es sich um originale Testhefte aus den BIST-Ü in der ursprünglichen Administrationsform, die der Verlinkung zur BIST-Ü D4 2015 dienen und damit eine fortlaufende Trendberichterstattung garantieren sollen. Die

nachstehende Tabelle zeigt, wie sich die Schulen und Schüler/innen aus der Stichprobe für die Brückenstudie auf die zugehörigen Testhefte verteilen. TH39 ist hier jenes Testheft, welches in modifizierter Form dem Testheft der Administrationseffektstudie entspricht.

Tabelle 3: Aufteilung der Brückenstudien-Stichprobe auf die entsprechenden Testhefte

Brückenstudien-Testheft	Anzahl Items	Anzahl Schulen	Anzahl gesampelter Schüler/innen	Anzahl getesteter Schüler/innen
TH38	25	11	233	205
TH39	25	10	200	185
TH40	26	11	227	216
TH41	26	11	218	200
BS-gesamt	34	43	878	806

Die bereits angesprochene eingeschränkte Repräsentativität einzelner Testheft-Populationen äußert sich insbesondere bei dem sehr wichtigen Testheft TH39. Die mangelnde Repräsentativität wird anhand der Verteilungen bezüglich Bundesländern und weiterer demografischer Merkmale, einschließlich einer Gegenüberstellung mit der Administrationseffektstudie, in den beiden nachfolgenden Tabellen 4 und 5 festgehalten.

Tabelle 4: Vergleich von Brückenstudie (Testheft TH39) und Administrationseffektstudie

Bundesland	Anzahl Schulen		Anteil (getesteter) Schüler/innen (in %)		
	BS-TH39	AS	BS-TH39	AS	Population
Burgenland	0	1	0,00	2,86	2,42
Kärnten	0	2	0,00	6,53	5,90
Niederösterreich	2	5	20,00	18,98	18,74
Oberösterreich	2	4	16,22	15,10	17,92
Salzburg	0	2	0,00	7,76	6,48
Steiermark	2	3	20,00	10,82	12,96
Tirol	1	2	10,27	7,35	7,92
Vorarlberg	1	1	7,57	3,88	4,93
Wien	2	6	25,95	26,73	22,74
Gesamt	10	26	100,00	100,00	100,00

Tabelle 5: Vergleich von Brückenstudie (Testheft TH39) und Administrationseffektstudie

Teilstudie	Gewichtete Stichprobengröße	Alter		Anteil Mädchen (in %)	Anteil Schüler/innen mit Migrationshintergrund (in %)	Missing-Rate (in %)	HISEI		Missing-Rate (in %)
		M	SD				M	SD	
BS-TH39	2.539	10,43	0,54	52,33	27,99	7,36	49,16	23,22	9,57
AS	6.767	10,40	0,51	51,03	19,14	7,92	51,62	21,90	10,85
Gesamtstudie	72.169	10,40	0,49	49,55	21,62	5,91	50,25	21,40	9,15
Population	76.447	10,39	0,48	49,13	22,97	–	48,78	21,04	–

Im Hinblick auf die Konstruktabdeckung in den Teilstudien ist des Weiteren die Anzahl der Items in den Testheften speziell von Brücken- und Administrationseffektstudie interessant. Die vier nichtdisjunkten Hefte der Brückenstudie beinhalten – als Resultat des Testdesigns zur BIST-Ü D4 2015 – insgesamt 34 Items. Eine genauere Aufschlüsselung nach Testheften ist in Tabelle 3 gegeben; die Administrationseffektstudie mit Testheft TH39 besteht demnach aus nur 25 Items. Während bei der Brückenstudie davon ausgegangen werden kann, dass das zu messende Konstrukt zumindest auf Gesamtebene hinreichend gut abgedeckt ist, könnte dies bei der Administrationseffektstudie eventuell nicht mehr der Fall sein. Als Vergleichsbasis dient hier die Anzahl verwendeter Link-Items für den PISA-Trend in Lesen von 2000 bis 2006 (Mazzeo & von Davier, 2009).

6.4.3 Fehlende Werte

Für die Untersuchung eines möglichen differenziellen Effekts nach Migrationshintergrund wurde aufgrund von nicht vernachlässigbaren Missing-Raten bei der damit assoziierten Variable, die im konkreten Fall für die Administrationsstudie 7,92 %, für die Brückenstudie 6,59 % und speziell bei TH39 7,36 % betrugen (siehe auch Tabelle 2 bzw. 5), eine multiple Imputation der fehlenden Werte durchgeführt (Enders, 2010; Lüdtke et al., 2007; Robitzsch, Pham & Yanagida, 2016). Für das Studium eines differenziellen Effekts hinsichtlich Geschlechts war dieser Schritt nicht notwendig, da die zugehörige Variable vollständig beobachtet vorlag. Im Rahmen des Imputationsprozesses wurden weitere Variablen, die Missings aufwiesen, insbesondere die Variable HISEI, aber auch Variablen mit Informationen etwa über den sprachlichen Hintergrund oder die Lernmotivation der Schüler/innen imputiert. Die Umsetzung in R erfolgte mit dem Paket mice (van Buuren et al., 2021).

Zusätzlich zur Imputation wurden noch Plausible Values (Mislevy, 1991; Robitzsch et al., 2016; von Davier, Gonzalez & Mislevy, 2009; Wu, 2005) gezogen, wofür das R-Paket TAM (Robitzsch, Kiefer & Wu, 2021; Funktion tam.pv) zur Anwendung kam. Als Ergebnis lagen schließlich 10 sogenannte PV-Datensätze vor, ohne Missings bei den ursprünglich betroffenen Variablen und mit Plausible Values als Fähigkeitswerten. Diese Datensätze

konnten dann für Analysen auf (Teil-)Populationsebene herangezogen werden. Eine analoge Vorgehensweise kam auch bei der Auswertung der Pilotierungsstudie zum Tragen (siehe Kapitel 7 – Ergebnisse). Tabelle 6 fasst die Imputationsergebnisse für die Variablen zum Migrations- und zum sozioökonomischen Status zusammen; ein Vergleich mit den Tabellen 2 und 5 zeigt, dass es lediglich moderate Unterschiede gibt.

Tabelle 6: Migrationsstatus und HISEI vor und nach der Imputation

Teilstudie	Gewichtete Stichprobengröße	Vor Imputation			Nach Imputation		
		Anteil Schüler/innen mit Migrationshintergrund (in %)	HISEI		Anteil Schüler/innen mit Migrationshintergrund (in %)	HISEI	
			M	SD		M	SD
BS-TH39	2.539	27,99	49,16	23,22	29,89	49,67	22,20
BS	11.298	26,93	50,68	22,11	28,79	50,65	21,43
AS	6.767	19,14	51,62	21,90	20,81	52,09	21,14
PS	40.122	20,96	50,23	21,26	22,32	50,44	20,65
Population	76.447	–	–	–	22,97	48,78	21,04

6.4.4 Post-hoc-Gewichtung

Die in Tabelle 6 ersichtlichen Abweichungen zwischen einer einzelnen Teilstudien-Stichprobe und der Gesamtpopulation beim Anteil der Schüler/innen mit Migrationshintergrund konnten durch Adjustierung der Schülergewichte ausgegeglichen werden. Diese wurde, im Sinne der weiteren Analysen in diesem Kapitel, speziell für die Brückenstudie (und separat für TH39) sowie die Administrationseffektstudie durchgeführt. Bei der Regewichtung wurden individuelle Korrekturfaktoren für die vier Subgruppen, die sich als Kombinationen aus Migrationsstatus und Geschlecht ergeben, berücksichtigt. Die modifizierten Gewichte flossen in weiterer Folge sowohl in die im nachfolgenden Abschnitt 6.4.5 beschriebenen Skalierungs- und Linking-Prozesse als auch in den PV-Imputationsprozess ein.

6.4.5 Analysestrategie

In diesem Abschnitt werden die mit den Teilstudien der D4H-Pilotierung durchgeführten Skalierungs- und Verlinkungs- bzw. Equatingschritte beschrieben. In einem vorbereitenden Schritt wurde die Population der BIST-Ü D4 2015, Bereich Zuhören, auf die Linking-Testhefte der Pilotierung (d. h. auf die Testhefte der Brückenstudie) eingeschränkt, mit anschließender Neuskalierung der Daten. Das Ergebnis wurde so transformiert, dass für die Populationsparameter mit Mittelwert $M = 0$ und Standardabweichung $SD = 1$ neu definierte Referenzwerte auf der Logit-Metrik vorlagen.

Die (getestete) Gesamtpopulation für Zuhören in der BIST-Ü D4 2015 bestand aus 75.297 Schülerinnen und Schülern, an denen 23 Testhefte mit insgesamt 135 Items administriert wurden. Im Mittel wurden rund 27 Items pro Testheft eingesetzt. Auf die getestete Linking-Teilpopulation entfielen entsprechend 1.383 Schüler/innen, welche die vier Link-Testhefte der späteren Brückenstudie mit insgesamt 34 Items bearbeiteten (siehe auch Tabelle 3 in Abschnitt 6.4.2). Die Linking-Anforderungen (1) und (2) gemäß Abschnitt 6.1 bezüglich Repräsentativität der Item- und Individuen-Auswahl werden damit für diese Erhebung (basierend auf einem geeigneten Testdesign 2015) als erfüllt angesehen.

In einem ersten Schritt erfolgte die Skalierung der einzelnen Teilstudien innerhalb der D4H-Pilotierung, d. h. der Brückenstudie, der Administrationseffektstudie, der Validierungsstudie und der Pilotierungsstudie. Das bereits hervorgehobene Testheft TH39 der Brückenstudie wurde noch einmal separat skaliert. Bei allen Skalierungsprozessen kam das Rasch-Modell zur Anwendung (Rasch, 1960). Als Ergebnis lagen jeweils Werte für die Itemparameter und Personenfähigkeiten (WLEs; Warm, 1989) vor. Die praktische Umsetzung in R wurde mit dem Paket TAM realisiert (Robitzsch, Kiefer & Wu, 2021; Funktionen tam.mml und tam.wle).

Der zweite Schritt beinhaltete die Verlinkung der Teilstudien der D4H-Pilotierung untereinander bzw. mit der BIST-Ü D4 2015. Die Brückenstudie wurde gemäß dem in Abschnitt 6.1 beschriebenen Common-Items-Nonequivalent-Groups-Design (vgl. Abbildung 1) mit der auf die Linking-Testhefte eingeschränkten Erhebung von 2015 verlinkt. Ankeritems sind die 34 Linkitems. Als konkrete Verlinkungsmethode kam das Linking nach Haberman (Haberman, 2009) zum Einsatz; die Berechnungen wurden mit dem R-Paket sirt (Robitzsch, 2021; Funktion linking.haberman) durchgeführt. Die daraus resultierenden Linking-Parameter (M = 0 und SD = 1 für die BIST-Ü 2015 und davon abweichende Werte für die Brückenstudie) entsprechen den Populationsparametern der beiden Studien auf der Logit-Metrik. Mit ihnen konnten in weiterer Folge noch die WLEs und die PVs mittels linearer Vorschrift passend transformiert werden.

Die Pilotierungsstudie konnte unter Annahme äquivalenter Gruppen über ein lineares Equating (Kolen & Brennan, 2014) mit der Brückenstudie verknüpft werden (vgl. Abbildung 1 in Abschnitt 5.3.4 in Kapitel 5 – Studiendesign), d. h., die Metrik der Pilotierungsstudie wurde so transformiert, dass die PV-basierten Fähigkeitsverteilungen der beiden Studien den gleichen Mittelwert und die gleiche Standardabweichung aufweisen. Für die Analyse der Validierungsstudie fanden zwar Skalierungsprozesse statt (siehe Kapitel 8 – Konstruktäquivalenz), es war jedoch keine Verlinkung zu anderen Teilstudien erforderlich.

Zur Messung des Administrationseffekts wurde die entsprechende Studie mit der Brückenstudie, genauer mit der Einschränkung der Brückenstudie auf das separat skalierte Testheft TH39, über ein Haberman-Linking verknüpft (mit normierten Linking- bzw. Populationsparametern M = 0 und SD = 1 auf der Logit-Metrik für die TH39-Brückenstudie). Unter der

Annahme äquivalenter Gruppen, die ein inhaltlich identes Testheft bearbeitet haben, sollten mögliche Unterschiede in den Itemantworten rein auf die veränderte Administrationsform zurückzuführen sein.

Zur Untersuchung auf Vorhandensein eines differenziellen Effekts nach Geschlecht oder Migrationshintergrund wurden die Differenzen von Subgruppen-Mittelwerten beider Studien auf der Logit-Metrik betrachtet. Die Berechnungen basieren auf den imputierten Werten für die Migrationsvariable, den adjustierten Schülergewichten und den PVs als Fähigkeitsschätzer. Die praktische Durchführung erfolgte unter Verwendung des R-Pakets BIFIEsurvey (Robitzsch & Oberwimmer, 2019).

6.5 Ergebnisse

In diesem Abschnitt sollen die Erkenntnisse zu den Untersuchungsanliegen aus 6.3 diskutiert werden, im Vordergrund stehen also die Brückenstudie und die Administrationseffektstudie. Die Populationskennwerte für die BIST-Ü D4 2015 wurden, wie bereits dargelegt, mit Mittelwert M = 0 und Standardabweichung SD = 1 als neue Baseline-Werte festgelegt. Durch die beschriebenen Linking- und Equating-Schritte konnten die Ergebnisse der Teilstudien, sofern erforderlich, mit diesen Baseline-Werten vergleichbar gemacht werden. Von Interesse sind in diesem Kapitel einerseits der Kohortentrend in Deutsch Zuhören von 2015 auf 2018 und andererseits der Effekt einer abgeänderten Administrationsform. Für den zweiten Punkt war keine Verlinkung zur BIST-Ü notwendig, vielmehr wurde die TH39-Brückenstudien-Stichprobe wiederum auf M = 0 und SD = 1 renormiert.

Die beiden nächsten Tabellen geben eine Übersicht über die Ergebnisse zum Kohortentrend und zum Administrationseffekt. Die linkingspezifischen Standardfehler des Mittelwerts in Tabelle 7 wurden mittels Jackknife-Resampling-Methode berechnet (Robitzsch, Dörfler, Pfost & Artelt, 2011). Die Standardfehler in Tabelle 8 beinhalten die Varianzkomponenten für Imputation und Resampling, wobei mit Letzterem der Stichprobenfehler erfasst wird.

Tabelle 7: Linking-Ergebnisse auf der Logit-Metrik

	M	SD	SE_{Link}
Kohortentrend			
BIST-Ü 2015	0,00	1,00	–
BS 2018	–0,20	1,14	0,04
Administrationseffekt			
BS-TH39 2018	0,00	1,00	–
AS 2018	0,44	1,01	0,11

Tabelle 8: Mittelwerte und zugehörige Standardfehler auf der Logit-Metrik für Subgruppen

	Geschlecht				Migrationsstatus			
	M_{Bursch}	M_{Maed}	SE_{Bursch}	SE_{Maed}	M_{Mig}	$M_{keinMig}$	SE_{Mig}	$SE_{keinMig}$
BIST-Ü 2015	–0,09	0,09	0,01	0,01	–0,63	0,14	0,02	0,01
BS 2018	–0,23	–0,18	0,01	0,04	–0,98	0,03	0,04	0,02
Kohortentrend	–0,14	–0,27	–	–	–0,35	–0,11	–	–
BS-TH39 2018	0,04	–0,04	0,03	0,03	–0,63	0,19	0,06	0,03
AS 2018	0,40	0,48	0,02	0,03	–0,45	0,70	0,06	0,02
Administrationseffekt	0,36	0,52	–	–	0,18	0,51	–	–

Die Auswertung der Brückenstudie und deren Verlinkung mit der BIST-Ü 2015 zeigt, dass es eine deutliche Verschlechterung der Schüler/innen im Kompetenzbereich Zuhören von rund 0,20 Punkten auf der Logit-Metrik gibt. Der negative Kohortentrend kann nicht nur auf Gesamtebene, sondern in unterschiedlichem Ausmaß auch auf Ebene von Subgruppen beobachtet werden, wobei er bei den Mädchen (–0,27 Logit-Punkte) deutlich stärker als bei den Burschen (–0,14 Logit-Punkte) ausgeprägt ist, und sich als noch extremerer Unterschied zwischen Schüler/innen mit und ohne Migrationshintergrund (–0,35 bzw. –0,11 Logit-Punkte) bemerkbar macht.

Die Verlinkung der Administrationseffektstudie mit der auf das Testheft TH39 eingeschränkten Brückenstudie zeigt, dass es – entgegen den Erwartungen (vgl. Abschnitt 6.2) – einen positiven Administrationseffekt von rund 0,44 Punkten auf der Logit-Metrik gibt. Eine Untersuchung nach den Populationsmerkmalen Geschlecht und Migrationsstatus lässt vermuten, dass dieser Effekt differenzieller Natur ist: Er ist einerseits bei den Mädchen (0,52 Logit-Punkte) größer als bei den Burschen (0,36 Logit-Punkte) und andererseits bei Schüler/innen ohne Migrationshintergrund (0,51 Logit-Punkte) größer als bei Schülerinnen und Schülern mit Migrationshintergrund (0,18 Logit-Punkte).

6.6 Diskussion

Tendenziell kann bei Trendmessungen im Rahmen der Bildungsstandardüberprüfungen in Österreich eine Verbesserung der Schülerleistungen beobachtet werden. Dies trifft sowohl auf der Primarstufe für die Mathematik-Erhebungen 2013 und 2018 zu (BIFIE, 2019, S. 59, Abbildung 27) als auch auf der Sekundarstufe für die Überprüfungen in Mathematik 2012 und 2017 (BIFIE, 2018, S. 65, Abbildung 32) sowie für die Überprüfungen in Englisch 2013 und 2019 (BIFIE, 2020, S. 102, Abbildung 72 für die Gesamtkompetenz und speziell S. 93, Abbildung 61 für die Teilkompetenz Zuhören). Diesem Muster entsprechend hätte man bei Deutsch Zuhören ebenfalls einen positiven Trend erwarten dürfen – trotz einer geringeren Zeitspanne von nur drei Jahren zwischen den beiden Messungen 2015 und 2018. Demgegenüber wurde jedoch bereits auch in anderen Studien ein negativer Trend in Deutsch

Zuhören in der Primarstufe festgestellt, etwa bei den IQB-Bildungstrends in Deutschland (IQB, 2017, S. 159, Abbildung 6.4). Aus zwei Messungen schon eine allgemeinere Entwicklung abzuleiten, wäre allerdings zu verfrüht, wie wiederum die PIRLS-Studie mit drei Erhebungen in den Jahren 2006, 2011 und 2016 zeigt (BIFIE, 2017); hier ergibt sich bei der ländervergleichenden Lesekompetenz am Ende der Primarstufe zumindest für Österreich ein V-förmiger Verlauf (siehe S. 42, Abbildung 2.3.1).

Der positive Trend bedingt durch die neue Administrationsform ist überraschend und kann nur schwer erklärt werden. Der in Abschnitt 6.2 beschriebene Wegfall der Item-Voransicht hätte erwarten lassen, dass ein erfolgreiches Bearbeiten der Items dadurch erschwert wird; dies scheint offensichtlich nicht der Fall zu sein. Ein Erklärungsansatz könnte sein, dass Schüler/innen sich jetzt rein auf das Zuhören konzentrieren und die Informationen dadurch besser aufnehmen, wohingegen das zum Zuhören zeitgleich erfolgende Mitlesen von Antwortmöglichkeiten der Items eher vom eigentlichen Zuhörprozess ablenkt. Zudem liegt die Vermutung nahe, dass sich Schüler/innen auf die Information für einzelne Items konzentrieren, die sie im *Preview* gelesen haben oder auf die sie während des Lesens ihren Blick richten. Andere Informationen – etwa für sonstige Items innerhalb eines Testlets – geraten aus dem Fokus, insbesondere wenn die Lesemenge groß ist. Somit kann die Lösung bei Testlets zusätzlich durch den *Preview* bei Items erschwert werden.

Die Brückenstudie ist zentral sowohl für die Verlinkung mit der BIST-Ü 2015 als auch für die Bestimmung des Administrationseffekts. Durch Post-hoc-Gewichtung konnte eine suboptimale Stichprobenzusammensetzung (deutlich höherer Migrationsanteil in Relation zur Gesamtpopulation, vgl. Tabelle 2 oder 6) korrigiert werden. Eine mangelnde Repräsentativität der Stichprobe für diese Teilstudie hätte in bedingter Form den vorhandenen negativen Kohortentrend noch weiter ins Negative verzerrt. Analog wurde mit Post-hoc-Gewichtung die Stichprobe für die Administrationseffektstudie angepasst. In Bezug auf den Administrationseffekt fällt aber noch als Argument ins Gewicht, dass dessen Messung auf lediglich einem einzigen Testheft mit nur 25 Items basiert, womit insgesamt die notwendigen Linking-Voraussetzungen in nicht ausreichendem Maß erfüllt sind.

Literatur

Baranowski, R. A. (2006). Item Editing and Editorial Review. In S. M. Downing & T. M. Haladyna (Hrsg.), *Handbook of Test Development* (S. 349–357). Mahwah, NJ: Lawrence Erlbaum Associates.

Behrens, U. (2010). Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 31–50). Göttingen: Vandenhoeck & Ruprecht.

BIFIE (2017). *PIRLS 2016. Die Lesekompetenz am Ende der Volksschule – Erste Ergebnisse.* Salzburg: BIFIE. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/72fbeba6e55bfb402a7ee76c77e712d0c58e36da/PIRLS_2016_Erste_Ergebnisse_final_web.pdf

BIFIE (2018). *Standardüberprüfung 2017, Mathematik, 8. Schulstufe – Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/4a28609fd6414dbc257274688ffa37e44e4a3cf7/BiSt_UE_M8_2017_Bundesergebnisbericht.pdf

BIFIE (2019). *Standardüberprüfung 2018, Mathematik, 4. Schulstufe – Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/884d9cba36794bb78355f4df430488117b0740a1/BiSt_UE_M4_2018_Bundesergebnisbericht.pdf

BIFIE (2020). *Standardüberprüfung 2019, Englisch, 8. Schulstufe – Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/b2f25d7876f79a9aae779089a6a2eda898fe41a2/BiSt_UE_E8_2019_Bundesergebnisbericht.pdf

Downing, S. M. (2005). The Effects of Violating Standard Item Writing Principles on Tests and Students: The Consequences of Using Flawed Test Items on Achievement Examinations in Medical Education. *Advances in Health Sciences Education, 10,* 133–143. doi:10.1007/s10459-004-4019-5

Enders, C. K. (2010). *Applied Missing Data Analysis.* New York, NY: Guilford Press.

Ganzeboom, H. B. G., De Graaf, P. M. & Treiman, D. J. (1992). A Standard International Socio-Economic Index of Occupational Status. *Social Science Research, 21* (1), 1–56. doi:10.1016/0049-089X(92)90017-B

González, J. & Wiberg, M. (2017). *Applying Test Equating Methods – Using R.* Cham: Springer.

Haberman, S. J. (2009). *Linking Parameter Estimates Derived from an Item Response Model Through Separate Calibrations* (ETS Research Report Nr. RR–09–40). Princeton, NJ: Educational Testing Service.

IQB (2017). *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich.* Berlin: IQB. Verfügbar unter: https://www.iqb.hu-berlin.de/bt/BT2016/Bericht/BT2016_Bericht.pdf

Kolen, M. J. & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking.* Heidelberg: Springer.

Lietz, P., Cresswell, J. C., Rust, K. F. & Adams, R. J. (2017). Implementation of Large-Scale Education Assessments. In P. Lietz, J. C. Cresswell, K. F. Rust & R. J. Adams (Hrsg.), *Implementation of Large-Scale Education Assessments* (S. 1–25). Hoboken, NJ: Wiley & Sons.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau, 58* (2), 103–117. doi:10.1026/0033-3042.58.2.103

Mazzeo, J. & von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results. *Education Working Papers EDU/PISA/GB (2008), 28,* 1–45.

McElvany, N., Ohle, A., El-Khechen, W., Hardy, I. & Cinar, M. (2017). Förderung sprachlicher Kompetenzen – Das Potenzial der Familiensprache für den Wortschatzerwerb aus Texten. *Zeitschrift für Pädagogische Psychologie, 31* (1), 13–25. doi:10.1024/1010-0652/a000189

Mislevy, R. J. (1991). Randomization-Based Inference about Latent Variables from Complex Samples. *Psychometrika, 56* (2), 177–196. doi:10.1007/BF02294457

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Nielsen & Lydiche.

Robitzsch, A. (2021). *sirt: Supplementary Item Response Theory Models. R package version 3.10-118.*

Robitzsch, A., Dörfler, T., Pfost, M. & Artelt, C. (2011). Die Bedeutung der Itemauswahl und der Modellwahl für die längsschnittliche Erfassung von Kompetenzen: Lesekompetenzentwicklung in der Primarstufe. *Zeitschrift für Entwicklungspsychologie und pädagogische Psychologie, 43,* 213–227. doi:10.1026/0049-8637/a000052

Robitzsch, A., Kiefer, T. & Wu, M. (2021). *TAM: Test Analysis Modules. R package version 3.7-16.*

Robitzsch, A. & Oberwimmer, K. (2019). *BIFIEsurvey: Tools for Survey Statistics in Educational Assessment. R package version 3.3-12.*

Robitzsch, A., Pham, G. & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessments mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 259–293). Wien: Facultas.

Schnotz, W. (1988). Instructional Implications of Text Processing Research. *European Journal of Psychology of Education, 3* (2), 111–121. doi:10.1007/BF03172649

Schreiner, C. & Breit, S. (2016). Konzeption der Überprüfung der Bildungsstandards in Österreich. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessments mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 1–19). Wien: Facultas.

Trendtel, M., Pham, G. & Yanagida, T. (2016). Skalierung und Linking. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der osterreichischen Bildungsstandardüberprüfung* (S. 185–224). Wien: Facultas.

van Buuren, S., Groothuis-Oudshoorn, K., Vink, G., Schouten, R., Robitzsch, A., Rockenschaub, P. et al. (2021). *mice: Multivariate Imputation by Chained Equations. R package version 3.13–0.*

Violata, C. (1991). Item Difficulty and Discrimination as a Function of Stem Completeness. *Psychological reports, 69* (3), 739–743. doi:10.2466/pr0.1991.69.3.739

von Davier, M., Gonzalez, E. & Mislevy, R. J. (2009). What are Plausible Values and Why are they Useful? *IERI monograph series, 2,* 9–36.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54,* 427–450. doi:10.1007/BF02294627

Wassenburg, S., de Koning, B., de Vries, M., Boonstra, A. & van de Schoot, M. (2017). Gender Differences in Mental Simulation During Sentence and Word Processing. *Journal of Research in Reading, 40* (3), 274–296. doi:10.1111/1467-9817.12066

Wu, M. (2005). The Role of Plausible Values in Large-Scale Surveys. *Studies in Educational Evaluation, 31* (2), 114–128. doi:10.1016/j.stueduc.2005.05.005

Antonia Bachinger, Roman Freunberger, Michael Krelle

Kapitel 7
Ergebnisse der Schülerinnen und Schüler im Zuhören

7.1 Einleitung

Im Frühjahr 2018 wurde eine Pilotierung für die Bildungsstandardüberprüfung in der 4. Schulstufe in Deutsch 2020 durchgeführt. Aufgrund der Neugestaltung der Erhebungen im Rahmen des Pädagogik-Pakets (BMBWF, 2021) wurde die geplante Bildungsstandardüberprüfung 2020 nicht durchgeführt (siehe dazu Kapitel 1 – Überblick). Abgesehen von den Neuerungen, die im Rahmen des Pädagogik-Pakets notwendig waren, waren die Ziele dieser Pilotierung ...

1. festzustellen, ob die neu entwickelten Testaufgaben für den Einsatz in der Bildungsstandardüberprüfung geeignet sind (siehe Kapitel 3 – Operationalisierung und Kapitel 4 – Beispielaufgaben) und wie schwierig die Testaufgaben sind (siehe Kapitel 6 – Verlinkung),
2. zu klären, ob die Testaufgaben das Konstrukt valide messen (siehe Kapitel 8 – Konstruktäquivalenz) und
3. die Ergebnisse mit der Bildungsstandardüberprüfung 2015 zu verknüpfen, weshalb eine Brückenstudie durchgeführt wurde (siehe Kapitel 5 – Studiendesign und Kapitel 6 – Verlinkung).

Mit den hier vorliegenden Daten legen die Autorinnen/Autoren dieses Kapitels zentrale Ergebnisse der Pilotierung in einer dem Bundesergebnisbericht ähnlichen Form vor. Somit dient der vorliegende Bericht auch als Überbrückung zwischen der BIST-Ü 2015 und der geplanten iKM^{PLUS} (siehe Kapitel 1 – Überblick). Die Ergebnisse sind hinsichtlich der Bundesländer und weiterer Merkmale (z. B. Schulgröße) repräsentativ für Österreich (siehe Kapitel 5 – Studiendesign).

Im folgenden Kapitel werden in einem ersten Schritt Zuhörkompetenzen der Schüler/innen mit Blick auf die Kompetenzstufen gezeigt. Anschließend werden die Mittelwerte unter Einbeziehung verschiedener Subgruppen diskutiert. Die Festlegung der Subgruppen orientiert sich an Merkmalen der Person, wie Geschlecht, Migrationshintergrund etc.[1] Bevor die von

1 Vergleichswerte zum Zuhören aus dem Jahr 2015 sind hier nicht abgedruckt. Sie können dem Bundesergebnisbericht der Bildungsstandardüberprüfung 2015 entnommen werden (Breit, Bruneforth & Schreiner, 2016, S. 84–90).

den Kindern erreichten Kompetenzstände berichtet werden, ist noch Folgendes mit Blick auf die Auswahl der Kontextmerkmale für die Einteilung in Subgruppen anzumerken: Aus Platzgründen und aufgrund der Tatsache, dass nicht zu allen möglichen Merkmalen Informationen über die Schüler/innen vorlagen, werden nur ausgewählte Merkmale berichtet. Das sind hier Geschlecht, Migrationshintergrund, Urbanisierungsgrad, Bildungsabschluss der Eltern sowie Leistungsheterogenität und Zufriedenheit in der Schule bzw. in der Klasse. Diese Merkmale orientieren sich an den nationalen Bildungsberichten (Oberwimmer, Vogtenhuber, Lassnigg & Schreiner, 2019), gehen aber in Teilen über diese hinaus (z. B. mit Blick auf die Besiedelungsdichte oder Aspekte der Zufriedenheit).

Bei der Auswahl der Merkmale spielen auch inhaltliche Gründe eine Rolle: Häufig wird geschlechtsbezogenen Disparitäten eine gewichtige Rolle zugeschrieben, wenn über Fragen der Bildungsgerechtigkeit diskutiert wird (Schipolowski, Wittig, Weirich & Böhme, 2017, S. 187), so auch im Grundsatzerlass „Reflexive Geschlechterpädagogik und Gleichstellung" (BMBWF, 2018). Vor diesem Hintergrund berichten wir hier über vergleichende Ergebnisse von Burschen und Mädchen beim Zuhören.

Infolge steigender Zahlen von Kindern mit Zuwanderungshintergrund im Bildungssystem (auch schon vor der BIST-Ü 2015) werden – wie international üblich – Analysen zu zuwanderungsbezogenen Disparitäten diskutiert (Neubacher, Freunberger et al., 2019a, S. 75). Dabei ist zu bedenken, was z. B. Rjosk, Haag, Heppt und Stanat (2017) sowie Haag, Kocaj, Jansen und Kuhl (2017) ausführlicher diskutieren. Demnach birgt eine solche Perspektive die Gefahr, dass eventuelle Kompetenznachteile auf eine den Gruppen inhärente Andersartigkeit zurückgeführt würden und so zu Ausgrenzung führen könnten: „Empirische Befunde zu zuwanderungsbedingten Disparitäten ermöglichen es jedoch, weiterhin bestehende Benachteiligungen sichtbar zu machen und eine Grundlage für Bemühungen um deren Verringerung zu schaffen" (Rjosk et al., 2017, S. 237).

Eine Reihe von Studien konnte zudem zeigen, dass die Kompetenzentwicklung der Schüler/innen mit dem sozialen Status bzw. dem kulturellen Kapital ihrer Eltern zusammenhängt (Haag et al., 2017; BMF, 2020, S. 147). Hier wird insbesondere das Ziel formuliert, dass die Chancen- und Geschlechtergerechtigkeit im Bildungswesen verbessert werden soll. Als einen Aspekt davon werden soziale Disparitäten vor dem Hintergrund des Bildungsabschlusses der Eltern dargestellt. Bezüglich des elterlichen Bildungshintergrunds kann etwa die Frage gestellt werden, ob die Eltern die Chance, ihre Kinder beim Lernen zu unterstützen bzw. ihnen selbst etwas zu vermitteln, nützen.

Regionale Unterschiede kommen ins Spiel, wenn man den Urbanisierungsgrad des Schulstandorts miteinbezieht. Vergleichbare Untersuchungen zeigen hier bedeutsame Unterschiede beim Zuhören (Behrens & Krelle, 2014, S. 97). Von besonderer Bedeutung ist hier die enge Verknüpfung verschiedener Kontextmerkmale. So leben etwa Kinder mit Migrationshintergrund vermehrt in dicht besiedelten Gemeinden und ihre Eltern haben tendenziell einen geringeren Bildungshintergrund (Oberwimmer, Baumegger & Vogtenhuber, 2019, S. 38–46).

Um Disparitäten innerhalb der Klasse darzustellen, wurden Daten der Leistungsheterogenität zugrunde gelegt. Unter Leistungsheterogenität versteht man Variabilität der Schülerleistung in Zuhören innerhalb einer Klasse, die über die Punktedifferenz (Range) ermittelt werden. Dieser Wert ist vor allem im Hinblick auf Differenzierung im Unterricht von Bedeutung.

Schließlich wurden noch Merkmale der Schulzufriedenheit der Schüler/innen aufgenommen, um auf die Daten auf der sogenannten „sozialen Ebene des Zuhörens" (Gschwend, 2014, siehe auch Kapitel 2 – Konstrukt) hinzuweisen. Die Zufriedenheit in Schule und Klasse wird sowohl in internationalen Studien wie PISA (siehe dazu etwa Peña-López, 2017) und PIRLS/IGLU (IEA & BIFIE, 2015), aber auch in den Bildungsstandardüberprüfungen üblicherweise erhoben. Neben dem Selbstkonzept und anderen Faktoren kann die Zufriedenheit als Prädiktor für die Leistung gesehen werden (Wiesner, Schreiner, Breit & Pacher, 2017). Die dabei verwendete Methodik wird weiter unten im Anschluss an das Testkonstrukt zum Zuhören dargestellt (Abschnitt 7.3).

7.2 Zuhören in den Bildungsstandardüberprüfungen

In der Bildungsstandardüberprüfung 2015 wurde im Fach Deutsch eine Reihe von Kompetenzen getestet: die Kompetenzbereiche *Lesen, Verfassen von Texten* (Schreiben), *Rechtschreiben, Sprachbewusstsein* sowie Aspekte des *Zuhörens* (aus dem Kompetenzbereich *Hören, Sprechen, Miteinander-Reden)* (BIFIE, 2016b, S. 4). Der letztgenannte Kompetenzbereich wurde in der Bildungsstandardüberprüfung in zwei getrennten Testteilen überprüft, da *Sprechen* nur anhand einer Stichprobe erhoben wurde (Breit et al., 2016, S. 91). Die Überprüfung des *Zuhörens* erfolgte in der Bildungsstandardüberprüfung flächendeckend gemeinsam mit anderen Testteilen wie *Lesen* und *Rechtschreiben* (zur Begriffsklärung Hören, Zuhören und Hörverstehen siehe Kapitel 2 – Konstrukt).

Welche Aspekte des Zuhörens Gegenstand von Bildungsstandardüberprüfungen sind, ist in der Verordnung „Bildungsstandards im Schulwesen" geregelt (BGB1. II Nr. 1/2009). Die *Kompetenzen*[2] in der Verordnung legen für das Ende der vierten Schulstufe u. a. fest, dass Schüler/innen „anderen verstehend zuhören" und „Informationen einholen und sie an andere weitergeben" können sollen. Damit geht es um die Fähigkeit, altersgemäße mündliche und im Fall der Überprüfung medial vermittelte Texte und Kommunikationsformen zu verstehen (Behrens 2010). Im Konstrukt zum Testen von Zuhörkompetenz (siehe Kapitel 2 – Konstrukt) werden auf theoretischer Basis die folgenden Subkompetenzen, auch Dimensionen festgelegt: die Wiedergabe bzw. Auswahl von Einzelinformationen (Subkompetenz 1), Schlussfolgerungen ziehen bzw. Informationen (lokal) verknüpfen

2 Unter Kompetenzen werden die mit den konkreten Standards verbundenen Leistungsanforderungen verstanden, die Verhalten und Handlungen beschreiben. Sie sind die unterste Gliederungsebene der Verordnung über Bildungsstandards im Schulwesen (BGB1. II Nr. 1/2009).

(Subkompetenz 2), auf der Ebene des Textes Zusammenhänge erkennen (Subkompetenz 3) und auf den Text bezogene Aussagen beurteilen, begründen oder darüber reflektieren (Subkompetenz 4). Zudem kann zwischen dem Hörverstehen von literarischen Texten und von Sachtexten unterschieden werden (siehe Kapitel 2 – Konstrukt und Imhof [2010]). Nicht erhoben werden weitere wichtige Aspekte des Zuhörens bzw. mündlicher Kommunikation wie Kooperation, Aufmerksamkeitssteuerung und interaktive Verfahren des Zuhörens im Gespräch. Es wird also bewusst auf ein eingeschränktes Konzept von Zuhören gesetzt, um Vergleichbarkeit zwischen großen Gruppen von Schülerinnen und Schülern herzustellen (Krelle & Prengel, 2014, S. 224). Für ein komplexeres Konstrukt würden deutlich umfangreichere Parameter benötigt, die z. B. bei Behrens (2010) beschrieben sind. Aspekte der Motivation werden anhand von Fragebögen erhoben, die auch Teil der Ergebnisdarstellung in diesem Kapitel sind.

Das Kompetenzstufenmodell wurde im Rahmen der Bildungsstandardüberprüfung 2015 entwickelt und unter anderem für die kriteriale Ergebnisrückmeldung an die Schüler/innen und Lehrer/innen verwendet (Eibl, Schwaiger & Breit, 2018, S. 30)[3]. Für Schüler/innen, die die Bildungsstandards nicht erreichen, liegen keine Beschreibungen vor (siehe Kapitel 2 – Konstrukt). Schüler/innen, die die Bildungsstandards in Hörverstehen teilweise erreichen, können Hörtexten wichtige und prominent präsentierte Detailinformationen entnehmen und einfache Schlüsse ziehen sowie Informationen (ansatzweise) beurteilen. Allerdings können sie das Thema des Textes nicht immer sicher erkennen.

Schüler/innen, die die Bildungsstandards erreichen, können darüber hinaus altersangemessene fiktionale und nichtfiktionale Hörtexte auch in ihrer Gänze verstehen, wenn die Themen aus einem für sie vertrauten Kontext stammen. Das zeigt sich, indem sie Gehörtem wichtige Detailinformationen abgewinnen, auch komplexere Schlüsse ziehen und das Thema des Hörtextes erkennen können. Diese Schüler/innen können die Informationen zudem eigenständig vergleichen und beurteilen, auch im Hinblick auf die Textfunktion und die Handlungsmotive. Sie können im Rahmen von gehörten Dialogen die Angemessenheit sprachlicher Handlungen erkennen und einschätzen und greifen dafür auch auf parasprachliche Hinweise zurück. Schüler/innen auf Kompetenzstufe 2 erreichen somit die Regelstandards, die am Ende der 4. Schulstufe erreicht werden sollen (Schreiner & Wiesner, 2019, S. 24–25).

Schüler/innen, die die Bildungsstandards übertreffen, verfügen zusätzlich über folgende Kompetenzen: Sie sind in der Lage, Hörtexte zu verstehen, die nicht aus einem vertrauten Kontext stammen und können unter anderem nicht nur das Thema, sondern auch den Aufbau erkennen. Diese Schüler/innen können komplexere Schlüsse ziehen und ihre Reflexionen mit ihrem Vorwissen in Beziehung setzen.

3 Zur Entwicklung des Kompetenzstufenmodells siehe Breit et al. (2016, S. 100) und Luger-Bazinger, Freunberger und Itzlinger-Bruneforth (2015).

Hören: inhaltliche Beschreibung der einzelnen Kompetenzstufen

3	**Bildungsstandards übertroffen**	**ab 675 Punkten**

Die Schüler/innen verfügen über alle Kompetenzen der Stufe 2, darüber hinaus verstehen sie auch Hörtexte, deren Themenfelder über den vertrauten Kontext hinausgehen. Sie können auch weniger zentrale Detailinformationen entnehmen, nicht nur das Thema, sondern auch den Aufbau des Hörtexts erkennen und die Sprecherintention deuten. Die Schüler/innen können komplexere Schlüsse ziehen und ihre Reflexionen mit ihrem Vorwissen in Beziehung setzen.

2	**Bildungsstandards erreicht**	**470–674 Punkte**

Die Schüler/innen sind in der Lage, altersangemessene monologische und dialogische Hörtexte (fiktionale und nichtfiktionale Textsorten) zu verstehen, wenn die Themen aus vertrautem Kontext stammen. Sie können wichtige Detailinformationen entnehmen, textnahe Schlüsse ziehen und das Thema des Texts erkennen. Die Schüler/innen können die aufgenommenen Informationen eigenständig vergleichen und beurteilen, auch im Hinblick auf die Textfunktion und die Handlungsmotive. Sie können in Dialogen die Angemessenheit von Sprechhandlungen erkennen und einschätzen und greifen dazu auch auf parasprachliche Hinweise zurück.

1	**Bildungsstandards teilweise erreicht**	**378–469 Punkte**

Die Schüler/innen sind in der Lage, altersangemessene monologische und dialogische Hörtexte (fiktionale und nichtfiktionale Textsorten) zu verstehen, wenn die Themen aus vertrautem Kontext stammen. Die Schüler/innen können die Bedeutung ähnlich klingender Wörter im Satzkontext unterscheiden. Sie können wichtige und prominent präsentierte Detailinformationen entnehmen und textnahe Schlüsse ziehen. Sie können das Thema des Texts nicht immer sicher erkennen. Die Schüler/innen können die aufgenommenen Informationen beurteilen.

unter 1	**Bildungsstandards nicht erreicht**	**bis 377 Punkte**

Abbildung 1: Kompetenzstufenmodell

7.3 Datenbasis und Methodik

Die Erhebung zum Zuhören, die hier die Datenbasis darstellt, wurde als Teilstudie der Pilotierung zur Bildungsstandardüberprüfung 2020 in der 4. Schulstufe durchgeführt. Die Pilotierung fand im April 2018 statt und umfasste für die Aufgabenpilotierung in Zuhören insgesamt 2.798 Schüler/innen (48,7 % weiblich). Im Gesamten wurden 154 Klassen mit 5 bis 25 Schülerinnen und Schülern pro Klasse (M = 18,2, SD = 3,8) getestet (siehe Kapitel 5 – Studiendesign). Im Folgenden werden die Datenbasis und die Methodik – auch im Hinblick auf die verschiedenen Kontextfaktoren – beschrieben.

Zuerst muss festgehalten werden, dass WLEs (*Weighted Likelihood Estimates;* Warm, 1989), welche den Schülerinnen und Schülern sowie Lehrpersonen als Personenschätzer rückgemeldet wurden, als Schätzer auf Populationsebene ungeeignet sind, da sie zu verzerrten Parameterschätzungen bei statistischen Analysen führen. In Large-Scale Assessments hat sich daher die Methode der *Plausible-Value-Ziehung* etabliert (Wu, 2005). Für die Analyse der Pilotierungsdaten wurden in diesem Sinne insgesamt 10 Plausible Values gezogen. Die Methodik wurde in Anlehnung an Verfahren aus der Bildungsstandardüberprüfung gewählt (Robitzsch, Pham & Yanagida, 2016). Hierbei wurden durch ein latentes Regressionsmodell (Hintergrundmodell) mit ausgewählten Kovariaten und Rohscores die latenten Kompetenzen durch ein multiples Imputationsverfahren gezogen. Die nachfolgenden statistischen Kenngrößen (z. B. Mittelwerte für Subgruppenvergleiche) werden durch eine Mittelung der Schätzungen über die 10 Imputationsdatensätze berechnet (Bruneforth, Oberwimmer & Robitzsch, 2016; Little & Rubin, 2002). Da die Hintergrundvariablen auch

fehlende Werte enthielten, wurden diese in einem gemeinsamen Schritt mit der Plausible-Value-Ziehung ebenfalls multipel imputiert.

Die Pilotierungsstichprobe wurde stratifiziert gezogen, dabei kam ein zweistufiges Verfahren nach der *Probability-Proportional-to-Size-Sampling*-Methode mit Stratifizierung nach expliziten und impliziten Strata zur Anwendung (für Details siehe Kapitel 5 – Studiendesign). Um das komplexe Stichprobendesign adäquat zu berücksichtigen, wurde für die Schätzung der Standardfehler ein Resampling-Verfahren mittels *Jackknife*-Zonen verwendet, welches zu erwartungstreuen Schätzern führt (George, Oberwimmer & Itzlinger-Bruneforth, 2016). Die Berechnungen wurden in der freien Software R (R Development Core Team, 2019) mit dem Paket BIFIEsurvey (BIFIE, 2016a) durchgeführt.

Mit Blick auf die oben dargestellten Kontextfaktoren gilt:

1. 78 % der getesteten Schüler/innen haben keinen Migrationshintergrund. Die Zuordnung des Migrationshintergrunds im Studiendesign erfolgt in Anlehnung an die Definition der OECD, in der das Geburtsland der Eltern und nicht die Sprache als Kriterium angelegt ist. Einem Kind wird demnach Migrationshintergrund nur dann zugeordnet, wenn beide Eltern im Ausland geboren wurden. Eine Ausnahme bilden hier allerdings Kinder aus Deutschland: Diese werden aufgrund der sprachlichen Nähe nicht zur Gruppe der Kinder mit Migrationshintergrund gezählt (Oberwimmer, Baumegger et al., 2019).
   Die Leistungsunterschiede nach Migrationshintergrund werden nachfolgend auch um den Sozialstatus korrigiert dargestellt.[4] Im Sinne des *Potential-Outcome*-Ansatzes wird hier der *Average Treatment Effect* berechnet (Gangl, 2010; Lüdtke, Robitzsch, Köller & Winkelmann, 2010). Für beide Subgruppen liegt separat jeweils ein Regressionsmodell mit der Zuhörkompetenz als abhängige Variable und dem Sozialstatus (mit quadratischen und kubischen Termen) als unabhängige Variable zugrunde. Die Modelle dienen dann zur Vorhersage der Werte der jeweils anderen Subgruppe. Dabei wird eine gleiche Verteilung des Sozialstatus in beiden Gruppen angenommen. Damit kann die Aussage getroffen werden, welcher Unterschied bestehen würde, hätten Schüler/innen aus beiden Gruppen den gleichen Sozialstatus (Freunberger et al., 2014).
2. Zusätzlich wurde der Urbanisierungsgrad (im Sinne des „Degree of Urbanisation" von Eurostat) als Kontextmerkmal für nachfolgende Analysen herangezogen: 28 % der Schüler/innen besuchen Schulen in dicht besiedelten Gebieten, 41 % in mittel besiedelten Gebieten und 31 % in dünn besiedelten Gebieten.
3. Berücksichtigt wurde zudem die elterliche Ausbildung: 7 % der Kinder haben Eltern, die höchstens einen Pflichtschulabschluss haben, 41 % haben einen Elternteil, der zumindest eine Berufsbildung abgeschlossen hat, 22 % haben mindestens einen Elternteil mit Matura und 30 % haben zumindest einen Elternteil, der einen tertiären Bildungsabschluss hat.

4 Bei der Berechnung des Sozialstatus werden, wie in Large-Scale-Studien üblich, die höchste Ausbildung der Eltern, die Anzahl der Bücher zu Hause und der HISEI *(Highest International Socio-Economic Index of occupational status)* berücksichtigt (Freunberger, Robitzsch & Pham, 2014; Ganzeboom, De Graaf & Treiman, 1992).

4. Für die Darstellung der Leistungsheterogenität wurde ähnlich zu Mayrhofer et al. (2019, S. 152) der Perzentilabstand des 10-%- und 90-%-Perzentils der Leistungen der Schüler/innen innerhalb der Klassen berechnet. Dabei ergeben sich Differenzen von 117 bis 363 Punkten für die Klassen. Diese Werte wurden in drei Kategorien[5] eingeteilt, wobei hierfür der Median (bei ca. 220 Punkten[6]) als Ankerpunkt verwendet wurde. Von diesem ausgehend wurden die Kategoriengrenzen wie folgt bestimmt: 0; Median – 25 Punkte; Median + 25 Punkte; 400. Die Kategorien führen die Bezeichnungen geringe, mittlere und hohe Heterogenität. 46 der insgesamt 154 Klassen der Pilotierung werden als Klassen mit niedriger Leistungsheterogenität eingeteilt, 62 mit mittlerer und 46 mit hoher Leistungsheterogenität.[7]
5. Neben den Leistungsdaten wurden für die vorliegenden Analysen auch Angaben der Schüler/innen aus den Fragebögen verwendet. Dieses Vorgehen ist in Übereinstimmung mit dem Bundesergebnisbericht von 2015 (Breit et al., 2016) und den Nationalen Bildungsberichten (Oberwimmer, Vogtenhuber et al., 2019). In der Pilotierung wurden vier Versionen der Fragebögen in einem rotierten Design eingesetzt, mit dem Ziel, eine erhöhte Anzahl an Fragebogenitems pilotieren zu können (siehe Kapitel 5 – Studiendesign). Die beiden hier aufgegriffenen Fragen zu *Zufriedenheit mit der Klasse* und *wie gerne die Schüler/innen in die Schule gehen* werden standardmäßig in allen Fragebögen zur BIST-Ü und zu deren Pilotierungen gestellt. Die Ergebnisse sind in Abbildung 2 und 3 dargestellt. Es zeigt sich, dass Schüler/innen mit Migrationshintergrund häufiger angeben, sehr gerne oder gerne in die Schule zu gehen als Schüler/innen ohne Migrationshintergrund. Die Zufriedenheit in der Klasse ist in beiden Gruppen sehr hoch. Diese Daten werden im Ergebnisteil als Grundlage verwendet, um die Leistungsdaten detaillierter einschätzen und kontextualisieren zu können (siehe Abschnitt 7.4.6 und 7.5).[8]

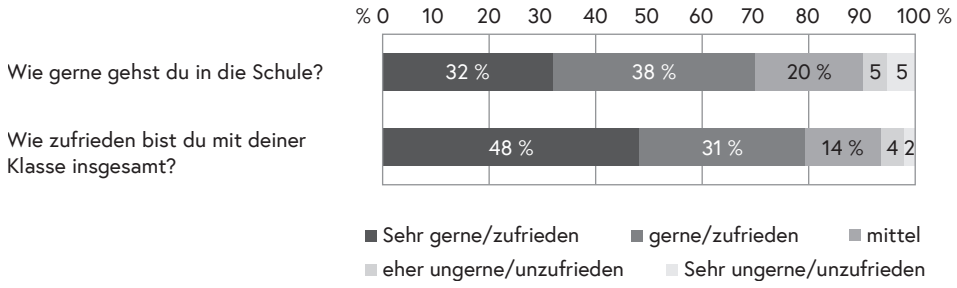


Abbildung 2: Zufriedenheit der Schüler/innen

5 Durch die kleinere Stichprobe würde eine Kategorisierung ähnlich wie im Nationalen Bildungsbericht 2018 (Mayrhofer et al., 2019) zu geringen Zellenbesetzungen führen. Daher verwenden wir hier nur drei Kategorien anstelle von sechs. Eine zu geringe Zellenbesetzung bedeutet, dass bei sechs Kategorien zu wenige Schulklassen pro Heterogenitätskategorie vorliegen würden.

6 Die Medianwerte schwanken über die 10 Plausible Values (PV) (s. u.) zwischen 207 und 228 Punkten.

7 Diese deskriptiven Ergebnisse basieren auf einem PV-Datensatz, leichte Schwankungen über die PV-Datensätze hinweg können vorhanden sein.

8 Angaben der Schüler/innen zum fachlichen Selbstkonzept wurden für die Analysen nicht herangezogen, weil diese sich auf das Fach Deutsch/Lesen/Schreiben bzw. den Deutschunterricht im Allgemeinen beziehen, z. B.: „Normalerweise bin ich gut in Deutsch" bzw. „Ich lerne schnell in Deutsch." (Breit et al., 2016, S. 31).

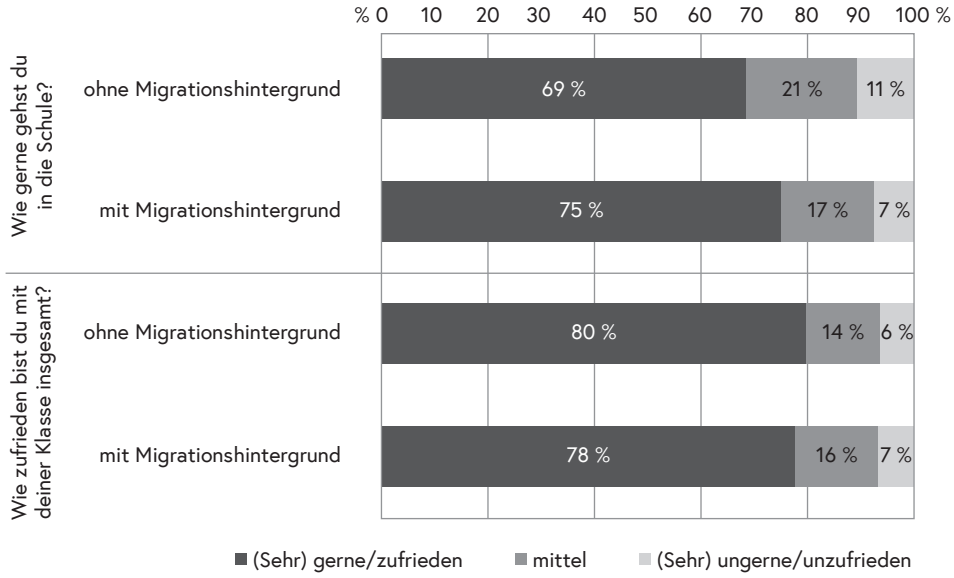


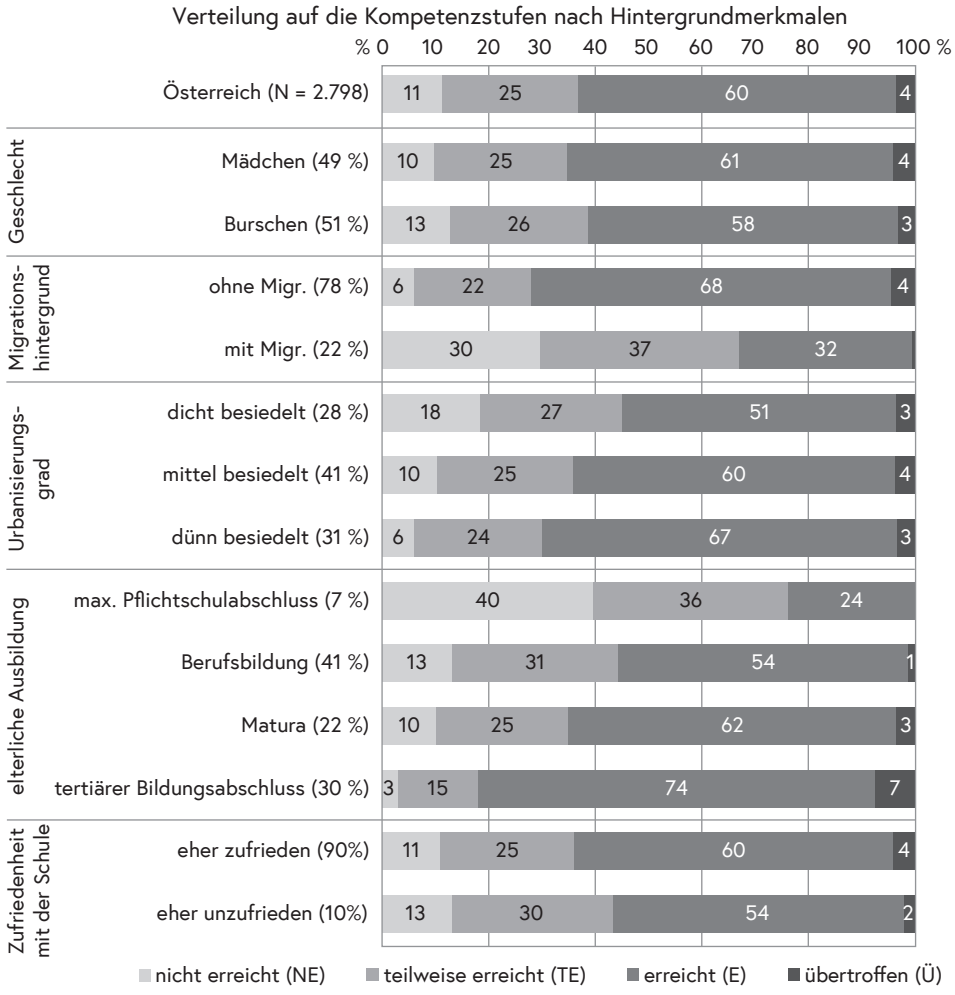
Abbildung 3: Zufriedenheit der Schüler/innen nach Migrationshintergrund

Anhand der hier vorliegenden Ergebnisse sind Rückschlüsse über die Veränderung der Kompetenzen (z. B. zur Bildungsstandardüberprüfung 2015 in der 4. Schulstufe im Fach Deutsch) nicht zulässig. Die Schülerdaten wurden nicht mit vorhergehenden Studien verlinkt und die Skalen (Gesamtskala und die Skalen der Subkompetenzen) wurden auf einen Mittelwert von 500 und eine Standardabweichung von 100 fixiert (für weitere Details und Analysen zum Linkingverfahren wird auf Kapitel 6 – Verlinkung in diesem Berichtsband verwiesen). Die Schwellenwerte zur Einteilung der Schülerkompetenzen nach Kompetenzstufen wurden aus dem Standard-Setting für die Bildungsstandardüberprüfung 2015 herangezogen.

Angaben zur statistischen Signifikanz von Kennwerten können den Tabellen entnommen werden und werden aus Lesbarkeitsgründen nicht im Text erwähnt.

7.4 Ergebnisse – Kompetenzen in Zuhören

In Abbildung 4 wird die Zuhörkompetenz einerseits für Österreich gesamt und andererseits getrennt nach unterschiedlichen Subgruppen dargestellt. Es ist ersichtlich, dass insgesamt 11 % der österreichischen Schüler/innen in der Pilotierungsstudie die Bildungsstandards nicht erreichen, 25 % die Bildungsstandards teilweise erreichen und 60 % der Schüler/innen die Bildungsstandards in Zuhören erreichen. Die höchste Kompetenzstufe „Bildungsstandards übertroffen“ erreichen 4 % der Schüler/innen in der 4. Schulstufe.



Standardfehler der Kompetenzstufenverteilungen		NE	TE	E	Ü
	Österreich (N = 2798)	1,1 %	1,3 %	1,6 %	0,5 %
Geschlecht	Mädchen (49 %)	1,3 %	1,8 %	1,9 %	0,7 %
	Burschen (51 %)	1,4 %	1,8 %	2,0 %	0,6 %
Migrationshintergrund	ohne Migr. (78 %)	0,9 %	1,4 %	1,5 %	0,6 %
	mit Migr. (22 %)	2,8 %	3,1 %	3,7 %	0,4 %
Urbanisierungsgrad	dicht besiedelt (28 %)	2,8 %	3,1 %	3,9 %	1,1 %
	mittel besiedelt (41 %)	1,5 %	2,1 %	2,5 %	0,8 %
	dünn besiedelt (31 %)	1,1 %	1,7 %	1,9 %	0,8 %
elterliche Ausbildung	max. Pflichtschulabschluss (7 %)	4,8 %	4,6 %	3,8 %	0,0 %
	Berufsbildung (41 %)	1,4 %	2,1 %	2,2 %	0,4 %
	Matura (22 %)	1,7 %	2,2 %	2,6 %	0,9 %
	tertiärer Bildungsabschluss (30 %)	0,8 %	1,6 %	1,9 %	1,2 %
Zufriedenheit mit der Schule	eher zufrieden (90 %)	1,1 %	1,4 %	1,7 %	0,5 %
	eher unzufrieden (10 %)	2,8 %	4,3 %	4,3 %	1,4 %

Abbildung 4: Verteilungen auf den Kompetenzstufen für Österreich gesamt und Standardfehler nach Kompetenzstufenverteilung

Im folgenden Abschnitt werden die Ergebnisse nach verschiedenen Subgruppen näher betrachtet. Dargestellt werden sowohl das Erreichen der verschiedenen Kompetenzstufen als auch die Mittelwerte. Die Mittelwerte werden für die vier Subkompetenzen gemeinsam (siehe Tabelle 1) und jeweils für die einzelnen Subkompetenzen dargestellt (siehe Tabelle 2).

Tabelle 1: Mittelwertvergleiche nach unterschiedlichen Kategorien

		M	SE	Diff.
Geschlecht	Mädchen	506	4,5	*Ref.*
	Burschen	494	4,3	12**
Migrationshintergrund	ohne Migr.	521	3,5	*Ref.*
	mit Migr.	428	7,4	93***
Urbanisierungsgrad	dicht besiedelt	480	9,8	34***
	mittel besiedelt	503	6,2	11
	dünn besiedelt	514	4,0	*Ref.*
Höchste elterliche Ausbildung	max. Pflichtschulabschluss	401	8,4	79***
	Berufsbildung	480	4,1	*Ref.*
	Matura	503	5,1	–23***
	tertiärer Bildungsabschluss	549	5,1	–69***
Zufriedenheit mit der Schule	(eher) zufrieden	502	4,0	*Ref.*
	(eher) unzufrieden	483	7,6	19*

Anmerkungen: M = Mittelwert; SE = Standardfehler des Mittelwerts; Diff. = Differenz der Mittelwerte; *Ref.* = Referenzkategorie, mit der die jeweiligen anderen Kategorien verglichen wurden; *** = p < .001; ** = p < .01; * = p < .05.

7.4.1 Geschlecht

Ein wichtiger Indikator für Bildungsgerechtigkeit ist neben anderen Merkmalen wie Migrationshintergrund und Bildungshintergrund der Unterschied zwischen den Geschlechtern. Insgesamt ist der Anteil an jenen Kindern, die die Bildungsstandards nicht erreichen, bei den Burschen mit 13 % etwas höher als bei den Mädchen (10 %). 25 % erreichen die Bildungsstandards nur teilweise, auch hier ist der Anteil der Burschen etwas höher (26 % im Vergleich zu 25 % bei den Mädchen). 60 % der Schüler/innen erreichen die Bildungsstandards in Zuhören, das sind 58 % der Burschen und 61 % der Mädchen. 4 % der Mädchen und 3 % der Burschen übertreffen die Bildungsstandards.

Der Unterschied des Mittelwerts über die Kompetenzstufen hinweg liegt bei 12 Punkten auf der 500er-Metrik und ist somit statistisch signifikant (siehe Tabelle 1). Vergleicht man die Mittelwerte auf der Ebene der Subkompetenzen, so zeigen sich unterschiedlich deutliche Differenzen zwischen Mädchen und Burschen. Keine bedeutenden Mittelwertunterschiede

Tabelle 2: Mittelwertvergleiche der Subkompetenzen

		Subkompetenz 1			Subkompetenz 2			Subkompetenz 3			Subkompetenz 4		
		M	SE	Diff.	M	SE	Diff.	M	SE	Diff.	M	SE	Diff.
Geschlecht	Mädchen	500	4,9	*Ref.*	508	5,0	*Ref.*	505	5,3	*Ref.*	513	4,9	*Ref.*
	Burschen	500	4,5	0	492	5,6	16**	495	4,7	10	487	5,2	26***
Migrationshintergrund	ohne Migr.	520	3,7	*Ref.*	521	4,0	*Ref.*	520	4,3	*Ref.*	517	3,8	*Ref.*
	mit Migr.	430	6,7	90***	428	7,8	93***	430	6,4	90***	440	8,6	77***
Urbanisierungsgrad	dicht besiedelt	480	9,3	37**	480	11,2	36**	483	10,7	28*	485	10,7	28*
	mittel besiedelt	502	6,6	15	502	7,2	14	504	6,3	7	501	7,5	12
	dünn besiedelt	517	5,4	*Ref.*	516	5,5	*Ref.*	511	5,6	*Ref.*	513	4,9	*Ref.*
Höchste elterliche Ausbildung	max. Pflichtschulabschluss	405	9,5	80***	394	9,9	85***	407	9,5	72***	412	9,7	71***
	Berufsbildung	485	4,7	*Ref.*	479	5,0	*Ref.*	479	4,9	*Ref.*	483	5,0	*Ref.*
	Matura	502	5,9	–17*	502	6,1	–23**	502	5,3	–23**	503	6,1	–20**
	tertiärer Bildungsabschluss	542	5,2	–57***	552	5,7	–73***	549	5,7	–70***	543	5,8	–60***
Zufriedenheit mit der Schule	(eher) zufrieden	501	4,2	*Ref.*	501	4,9	*Ref.*	503	4,4	*Ref.*	504	4,7	*Ref.*
	(eher) unzufrieden	489	8,4	12	490	7,6	11	471	7,8	32***	468	8,5	36***

Anmerkungen: M = Mittelwert; SE = Standardfehler des Mittelwerts; Diff. = Differenz der Mittelwerte; *Ref.* = Referenzkategorie, mit der die jeweiligen anderen Kategorien verglichen wurden; *** = p < .001; ** = p <. 01; * = p < .05.

gibt es bei Subkompetenz 1, also beim Wiedergeben und Auswählen von Informationen. Am deutlichsten fallen hingegen die Unterschiede bei Subkompetenz 4 aus, d. h. beim Bewerten von und Reflektieren über Textinhalte, bei der die Mädchen 513 Punkte erreichen, die Burschen hingegen nur 487 (siehe Tabelle 2).

7.4.2 Migrationshintergrund

Der Anteil der Kinder, der die Bildungsstandards nicht erreicht, ist unter den Kindern mit Migrationshintergrund mit 30 % deutlich höher als unter den Kindern ohne Migrationshintergrund mit 6 %. Die beiden höchsten Kompetenzstufen 2 und 3 erreichen 68 % respektive 4 % der Kinder ohne Migrationshintergrund, bei Kindern mit Migrationshintergrund liegt der Wert bei 32 % beziehungsweise unter 1 % (siehe Abbildung 4).

Ein ähnliches Bild zeigt sich auch, wenn man die Mittelwerte gegenüberstellt. Der Mittelwert liegt über alle Subkompetenzen hinweg bei Schülerinnen und Schülern ohne Migrationshintergrund bei 521, bei Schülerinnen und Schülern mit Migrationshintergrund jedoch bei 428 (siehe Tabelle 1). Auf Ebene der Subkompetenzen liegt die Differenz der Mittelwerte zwischen 77 (Subkompetenz 4, also dem Reflektieren) und 93 (Subkompetenz 2, der lokalen Schlussfolgerung) auf der 500er-Metrik. Diese Unterschiede sind sowohl für die einzelnen Subkompetenzen als auch für Zuhören im Gesamten signifikant (siehe Tabelle 2).

Die Mittelwertdifferenzen verringern sich deutlich, bleiben aber weiterhin bestehen, wenn sie um den Sozialstatus korrigiert werden. Beim Gesamtwert liegt die Reduktion der Differenz bei 21 Punkten, auf Ebene der Subkompetenzen liegt die Reduktion der Differenz zwischen 18 und 23 Punkten (siehe Abbildung 5). Kinder ohne Migrationshintergrund haben also unabhängig von ihrem Sozialstatus im Zuhören einen ausgeprägten Vorteil gegenüber Kindern mit Migrationshintergrund.

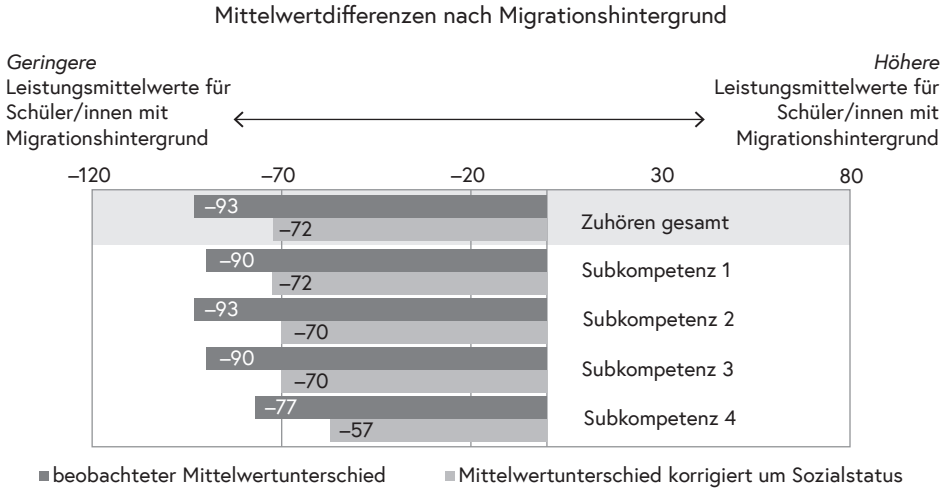


Abbildung 5: Mittelwertunterschiede zwischen Kindern mit und ohne Migrationshintergrund in Abhängigkeit des Sozialstatus

7.4.3 Urbanisierungsgrad

Betrachtet man die Unterschiede zwischen den verschiedenen Gruppen des Urbanisierungsgrads, zeigt sich, dass deutliche Disparitäten zugunsten der Schüler/innen in überwiegend ländlichen Gemeinden vorliegen. Nur 6 % erreichen die Bildungsstandards nicht, 24 % erreichen sie teilweise. Somit liegt der Anteil der Schüler/innen im ländlichen Raum, der die Bildungsstandards erreicht bzw. übertrifft, bei knapp 70 % im Vergleich zu 64 % in mittel besiedelten Gebieten und 54 % in dicht besiedelten Gebieten (siehe Abbildung 4). Über 18 % der Kinder aus überwiegend städtischen Gemeinden erreichen die Bildungsstandards nicht, 27 % erreichen sie teilweise. 51 % der Schüler/innen aus dicht besiedelten Gebieten erreichen die Bildungsstandards, nur knapp über 3 % erreichen die Stufe „Bildungsstandards übertroffen". In mittel besiedelten Gebieten erreichen 10 % der Schüler/innen die Bildungsstandards nicht, 25 % der Schüler/innen erreichen sie teilweise, 60 % erreichen die Bildungsstandards und 4 % übertreffen sie.

Über die Subkompetenzen hinweg ist der Mittelwert von Schülerinnen und Schülern, die Schulen in dünn besiedelten Gebieten besuchen, um 34 Punkte höher als bei Schülerinnen und Schülern aus überwiegend städtischen Gemeinden (siehe Tabelle 1; 514 im Gegensatz zu 480 Punkten). Schüler/innen, die Schulen in mittel besiedelten Gebieten besuchen, erreichen einen Mittelwert von 503.

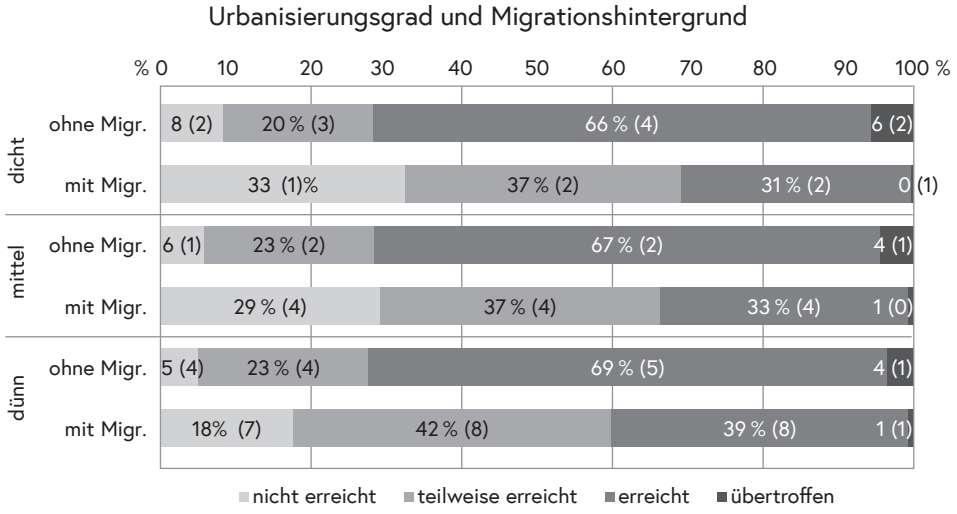


Abbildung 6: Kompetenzstufenverteilungen in Prozent (mit Werten der Standardfehler in Klammern) dargestellt nach Urbanisierungsgrad und Migrationshintergrund

Wie in Abbildung 6 zu sehen ist, zeigen sich interessante Ergebnisse, wenn zusätzlich nach Migrationshintergrund differenziert wird. Hier wird deutlich, dass sich die Verteilungen bei Schülerinnen und Schülern ohne Migrationshintergrund in den einzelnen Urbanisierungsgraden kaum unterscheiden, bei Schülerinnen und Schülern mit Migrationshintergrund allerdings deutliche Unterschiede erkennbar sind. Hier scheinen vor allem Schüler/innen

mit Migrationshintergrund in dicht besiedelten Gebieten stärkere Defizite aufzuweisen als in weniger dicht besiedelten Gemeinden. In dicht besiedelten Gemeinden liegen die Anteile der Schüler/innen auf der Stufe „Bildungsstandards nicht erreicht" mit 33 % noch etwas über jenem von Österreich gesamt (30 %, siehe Abbildung 4). 37 % der Kinder in dicht besiedelten Gebieten mit Migrationshintergrund erreichen die Bildungsstandards teilweise, 31 % erreichen sie. Von jenen Schülerinnen und Schülern mit Migrationshintergrund, die Schulen in dünn besiedelten Gebieten besuchen, erreichen 18 % die Bildungsstandards nicht, 42 % erreichen sie teilweise, 39 % erreichen die Bildungsstandards und 1 % der Schüler/innen übertreffen sie.

7.4.4 Leistungsheterogenität

Allgemein ist anzumerken, dass in den Volksschulen eine höhere Heterogenität der Leistungen als in der Sekundarstufe vorzufinden ist (Mayrhofer et al., 2019, S. 152). Dies zeigt sich auch in den Ergebnissen für Zuhören, wo die niedrigste Perzentildifferenz bei 117 Punkten liegt. Von den insgesamt 154 untersuchten Klassen aus der Pilotierung werden 46 als Klassen mit niedriger, 62 mit mittlerer und 46 mit hoher Leistungsheterogenität eingeteilt. Es kann nun betrachtet werden, wie viel Prozent der Schüler/innen Schulen mit unterschiedlicher Leistungsheterogenität besuchen und wie sich diese Anteile nach Urbanisierungsgrad verteilen. Mit 38 % besuchen die Schüler/innen in dünn besiedelten Gebieten eher Schulen mit geringer Leistungsheterogenität beim Zuhören als Schüler/innen in mittel bzw. dicht besiedelten Gebieten (27 % bzw. 29 %, siehe Abbildung 7). Ein Ergebnis, welches auch für Lesen in der Bildungsstandardüberprüfung 2015 in der 4. Schulstufe im Fach Deutsch gezeigt werden konnte (Neubacher, Freunberger et al., 2019, S. 223).

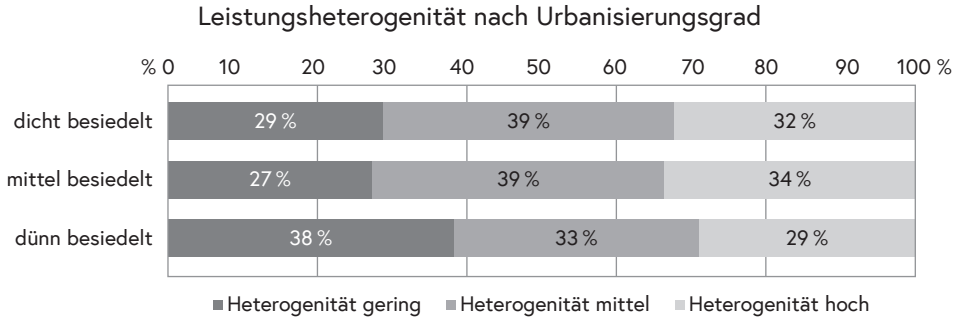


Abbildung 7: Die Leistungsheterogenität in den Klassen, dargestellt nach dem Urbanisierungsgrad

7.4.5 Bildungsabschluss der Eltern

Deutliche Unterschiede in den Schülerleistungen zeigen sich, wenn man die Ergebnisse der Schüler/innen unter Berücksichtigung des höchsten Bildungsabschlusses der Eltern betrachtet. Zur Gruppe der Kinder, deren Eltern maximal einen Pflichtschulabschluss haben, zählen immerhin noch 7 % der österreichischen Kinder, wie schon in Abschnitt 7.3 (3) berichtet wurde. Von jenen Kindern, deren Eltern maximal einen Pflichtschulabschluss

haben, erreichen 40 % die Bildungsstandards nicht. Nur 24 % der Schüler/innen aus dieser Gruppe erreichen die Bildungsstandards, es gibt in dieser Gruppe keine Kinder, die die Bildungsstandards übertreffen (siehe Abbildung 4).

Die größte Gruppe bilden mit 41 % jene Schüler/innen, deren Eltern höchstens eine Berufsausbildung abgeschlossen haben. 13 % dieser Schüler/innen erreichen die Bildungsstandards nicht, 31 % erreichen sie teilweise. Der Anteil der Schüler/innen, die die Bildungsstandards erreichen, ist mit 54 % deutlich höher als bei jenen Kindern, deren Eltern höchstens die Pflichtschule abgeschlossen haben. Der Anteil von Kindern, die die Bildungsstandards übertreffen, liegt bei knapp über 1 % (siehe Abbildung 4).

Von den Schülerinnen und Schülern, deren Eltern höchstens Matura haben (22 %), erreichen 10 % die Bildungsstandards nicht, 25 % erreichen die Bildungsstandards teilweise. Immerhin 61 % der Kinder erreichen die Bildungsstandards und 3 % übertreffen sie (siehe Abbildung 4).

30 % der Kinder haben Eltern, die eine universitäre oder gleichwertige Ausbildung abgeschlossen haben, d. h., die einen tertiären Bildungsabschluss (Universität, Fachhochschule, Pädagogische Hochschule, Akademie) haben. Von diesen Kindern übertreffen 7 % die Bildungsstandards und weitere 74 % erreichen sie. Nur 3 % respektive 15 % erreichen die Bildungsstandards nicht bzw. nur teilweise (siehe Abbildung 4).

Der Mittelwert liegt bei 401 Punkten bei Kindern, deren Eltern maximal einen Pflichtschulabschluss haben. Bei Kindern, deren Eltern höchstens eine Berufsausbildung abgeschlossen haben, liegt der Mittelwert bei 480 Punkten. Der Mittelwert von Kindern, deren Eltern höchstens Matura haben, liegt bei 503. Der Mittelwert ist bei Schüler/innen, deren Eltern einen tertiären Bildungsabschluss haben, mit 549 am höchsten im Vergleich zu den anderen Gruppen. Die Unterschiede zur Referenzkategorie (Berufsbildung als höchste elterliche Ausbildung) sind jeweils signifikant (siehe Tabelle 1).

7.4.6 Zufriedenheit

Aspekte der Zufriedenheit der Schüler/innen in der Schule wurden mittels Selbstauskunft im Schülerfragebogen erhoben (siehe Abbildung 2). Von jenen 90 % der Schüler/innen, die mit ihrer Schule sehr oder eher zufrieden sind, erreichen 11 % die Bildungsstandards nicht, 25 % erreichen die Bildungsstandards nur teilweise, 60 % erreichen die Bildungsstandards und 4 % übertreffen sie. 10 % der Schüler/innen gaben im Fragebogen an, sehr oder eher unzufrieden zu sein. Von ihnen erreichen 13 % die Bildungsstandards nicht, 30 % erreichen die Bildungsstandards nur teilweise, 54 % erreichen sie und nur 2 % übertreffen die Bildungsstandards (siehe Abbildung 4). Eine leichte Tendenz besteht dahingehend, dass mit ihrer Schule unzufriedene Schüler/innen vermehrt die Bildungsstandards nicht oder nur teilweise erreichen. Der Mittelwert von Kindern, die (eher) zufrieden sind, liegt bei 502, jener von Kindern, die (eher) unzufrieden sind, bei 483, die Unterschiede sind signifikant.

7.5 Zusammenfassung und Diskussion der Ergebnisse

Die vorliegenden Analysen zur Pilotierung zeigen insgesamt grundsätzlich positive Ergebnisse hinsichtlich der Kompetenzerreichung, insofern immerhin 64 % der Schüler/innen in Österreich die Bildungsstandards mindestens erreichen. Insgesamt 11 % der Schüler/innen erreichen in der Pilotierungsstudie die Bildungsstandards nicht, 25 % der Kinder erreichen die Bildungsstandards teilweise. Betrachtet man hier die Ergebnisse mit dem anderen, rezeptiven Kompetenzbereich, dem Lesen, so zeigen sich Unterschiede: In der Bildungsstandardüberprüfung 2015 erreichen 13 % der Schüler/innen die Bildungsstandards nicht, 25 % erreichen sie nur teilweise. 56 % der Kinder erreichen am Ende der 4. Schulstufe die Bildungsstandards in Lesen und 6 % übertreffen sie.

Beim Zuhören zeigen sich Unterschiede zwischen Mädchen und Burschen. Der Leistungsmittelwert der Burschen liegt beim Zuhören bei 494 (SE = 4,3), der Mittelwert der Mädchen bei 506 (SE = 4,5). Diese Disparitäten sind auch in anderen Kompetenzbereichen beobachtbar: Vorteile bei sprachlichen Kompetenzen zeigen sich bei Mädchen schon im Vorschulalter. Die Vorteile sind zudem – etwa in Lesen – am Ende der Volksschulzeit ausgeprägter als am Beginn, beim literarischen Lesen sind die Vorteile für Mädchen deutlicher erkennbar als beim Lesen von Sachtexten (Stanat & Bergann, S. 517–518). Die stärksten Unterschiede in der nationalen Bildungsstandardüberprüfung zeigen sich in Lesen in der 8. Schulstufe bei den Risikoleserinnen und -lesern (Neubacher, Ober, Wimmer & Hartl, 2019, S. 73).

Die Unterschiede zwischen Mädchen und Burschen beim Zuhören sind in den verschiedenen Subkompetenzen unterschiedlich ausgeprägt. Insbesondere bei Subkompetenz 4, also dem Beurteilen, Begründen und Reflektieren, treten deutliche Unterschiede zwischen den Geschlechtern auf. Ein praktischer Grund kann sein, dass Burschen mit der gleichen Kompetenz bei offenen Aufgaben schlechter abschneiden, was bisher anhand von Lesetests gezeigt werden konnte (Schwabe, McElvany & Trendtel, 2015; Trendtel, Schwabe & Fellinger, 2016, S. 225), weil sie diese häufiger nicht bearbeiten. Vermutet wird, dass das mit den höheren sprachproduktiven Kompetenzen oder mit der intrinsischen Motivation der Mädchen zusammenhängt. Vor allem Subkompetenz 4 wird häufig mit einem offenen Antwortformat erhoben.

Auch bei anderen Studien zum Leseverstehen gibt es deutliche Geschlechterdisparitäten. In der PIRLS-Studie zeigt sich beispielsweise, dass Mädchen die Burschen beim Teilbereich *literarische Erfahrungen machen* signifikant übertreffen. Der Unterschied ist hier also deutlicher als bei informierenden Texten. Die Mädchen schneiden außerdem bei hierarchiehöheren Prozessen wie dem Interpretieren und Bewerten besser ab als die Burschen (Salchegger et al., 2017, S. 51). Es wird angenommen, dass das Argumentieren über Texte bzw. die Anschlusskommunikation größere Gemeinsamkeiten mit dem Leseverstehen hat als hierarchieniedrige Hörverstehensprozesse (Grotjahn, 2005).

Betrachtet man die Ergebnisse aus der Bildungsstandardüberprüfung 2015, so zeigen sich beim Zuhören geringe Unterschiede zwischen Mädchen und Burschen im Vergleich

zu anderen Kompetenzbereichen. In Lesen lag der Unterschied bei der Erhebung 2015 bei 31 Punkten, bei Rechtschreiben bei 33 Punkten und bei Textproduktion und Sprachbetrachtung bei jeweils 25 Punkten, während auch 2015 mit 10 Punkten nur geringe geschlechterspezifische Disparitäten in Zuhören zu beobachten waren (Breit et al., 2016, S. 99–100). Vergleichbare Ergebnisse finden sich auch in Deutschland in der Erhebung über das Erreichen der Bildungsstandards (Bildungstrend) in der Primarstufe 2016, in der der Unterschied bei 12 Punkten lag und statistisch signifikant war (Schipolowski et al., 2017, S. 195).

Wie in Abschnitt 7.4.2. gezeigte wurde, gibt es deutliche Leistungsunterschiede beim Zuhören zwischen Kindern mit Migrationshintergrund und Kindern ohne Migrationshintergrund. Diese Unterschiede zeigten sich auch schon in der Bildungsstandardüberprüfung 2015. Hier wird argumentiert, dass dies für die Gruppe von Kindern mit Migrationshintergrund, die die Bildungsstandards nicht erreicht, „eine systematische Benachteiligung im Unterricht insgesamt [bedeutet], der primär sprachlich vermittelt wird" (Breit et al., 2016, S. 86).

Migrationsbedingte Disparitäten sind auch in anderen Kompetenzbereichen wie Rechtschreibung und Sprachbetrachtung feststellbar, allerdings fallen diese Unterschiede in anderen Kompetenzbereichen deutlich geringer aus, wenn man den Sozialstatus berücksichtigt (Breit et al., 2016, S. 101). In anderen (Teil-)Studien zu den Bildungsstandards zeigen sich also in den Kompetenzbereichen meist geringere Unterschiede in den mittleren Leistungen zwischen Schülerinnen und Schülern mit und Schülerinnen und Schülern ohne Migrationshintergrund, wenn der Sozialstatus einbezogen wird. Dieses Ergebnis zeigt sich allerdings wie in anderen Studien nicht beim Zuhören (Stanat, Pant, Richter & Böhme, 2012, S. 295). In der vorliegenden Studie bleiben bei gleichem Sozialstatus die Mittelwertdifferenzen beim Migrationshintergrund weitgehend erhalten. Die Leistungsunterschiede zwischen Kindern mit und ohne Migrationshintergrund können also nicht restlos durch den im Mittel verschiedenen Sozialstatus erklärt werden. Auf Ebene der Subkompetenzen gibt es keine besonderen Auffälligkeiten. Es kann angenommen werden, dass andere Faktoren wie etwa Aspekte der Erstsprache auf das Zuhören einen großen Einfluss haben (Behrens & Krelle, 2014, S. 99).

Auffällig ist in der vorliegenden Studie außerdem, dass der beobachtbare Mittelwertunterschied ebenso wie der um den Sozialstatus korrigierte Mittelwertunterschied bei Subkompetenz 4 (Reflektieren) geringer ist als bei den anderen Subkompetenzen. Die Werte zwischen den Subkompetenzen können allerdings nur bedingt verglichen werden, da diese jeweils separat skaliert wurden.

Deutliche Unterschiede zeigen sich auch zwischen den verschiedenen Gruppen nach Urbanisierungsgrad zugunsten der Schüler/innen, die Schulen in überwiegend ländlichen Gemeinden besuchen (siehe Abbildung 4). Die Unterschiede zwischen den Schülerinnen und Schülern aus überwiegend städtischem bzw. überwiegend ländlichem Gebiet hängen zum Teil auch damit zusammen, dass der Anteil an Kindern mit Migrationshintergrund unterschiedlich hoch ist. Im überwiegend urbanen Raum ist der Anteil an Kindern mit

Migrationshintergrund höher (siehe Abbildung 6): Während der Anteil an Schülerinnen und Schülern mit Migrationshintergrund in dünn besiedelten, überwiegend ländlichen Gemeinden mit 6 % sehr gering ist, liegt er in dicht besiedelten, überwiegend städtischen Gemeinden mit 38 % deutlich darüber. In Wien liegt der Anteil sogar bei 42 % (Oberwimmer, Baumegger et al., 2019, S. 43–44). Unterschiede bei Leistungsheterogenität nach Urbanisierungsgrad zeigen sich insbesondere zwischen dünn besiedelten Gebieten auf der einen Seite und mittel bzw. dicht besiedelten Gebieten auf der anderen Seite (siehe Abbildung 7).

Es zeigen sich deutliche Unterschiede in den Schülerleistungen, wenn die Ergebnisse der Schüler/innen unter Berücksichtigung des höchsten Bildungsabschlusses der Eltern betrachtet werden (siehe Abbildung 4). Es wird angenommen, dass dieses Merkmal Rückschlüsse auf die Bildungsnähe der Familie geben kann, was wiederum die kulturelle und kommunikative Praxis beeinflusst (Oberwimmer, Baumegger et al., 2019, S. 38). Besonders auffällig sind die Leistungen von Kindern, deren Eltern maximal einen Pflichtschulabschluss haben. Dazu zählen immerhin noch 7 % der österreichischen Kinder in der 4. Schulstufe. Die Gruppe wird allerdings kleiner (Oberwimmer, Baumegger et al., 2019, S. 38). In dieser Gruppe ist der Anteil an Kindern mit Migrationshintergrund überproportional groß, allerdings ist der Bildungsstand der Eltern mit Migrationshintergrund heterogen, tendenziell geringer und er ändert sich rasch. Generell lässt sich festhalten, dass das Bildungsniveau steigt und in der jüngeren Generation höher ist als in der gesamten Bevölkerung, diese Entwicklung muss allerdings langfristig betrachtet werden.

Es muss abschließend angemerkt werden, dass sich in Zuhören geringere Kompetenzen bei bildungsfernem Hintergrund, einem niedrigen sozioökonomischen Status der Familie sowie einer nichtdeutschen Alltagssprache zeigen. Dies ist auch in anderen Kompetenzbereichen ähnlich. Treten mehrere dieser Faktoren gleichzeitig auf, so spricht man von einer Kumulation sozialer Risikofaktoren. Diese können einem erfolgreichen Bildungserwerb umso hinderlicher sein (Oberwimmer, Baumegger et al., 2019, S. 27).

Vor allem zum Abbau dieser Nachteile braucht es noch weitere empirische Untersuchungen zur Wirksamkeit von Fördermethoden (Behrens & Krelle, 2014, S. 99–100; Imhof, 2010, S. 27). Zuhören kommt auch deshalb eine große Bedeutung zu, weil es sich um eine Kompetenz handelt, die in allen Fächern eine notwendige und zentrale Voraussetzung ist, um dem Unterricht folgen zu können und sich beteiligen zu können (Behrens & Eriksson, 2009, S. 10). Mündliche Kommunikation ist schließlich zentral für die aktive und passive Teilnahme in allen Fächern, weil Lerninhalte insbesondere in der Primarstufe häufig mündlich präsentiert werden und neue Kompetenzen über mündliche Kommunikation vermittelt werden (Alexander, 2012; BIFIE, 2011, S. 13–14; Kaldahl, Bachinger & Rijlaarsdam, 2019).

Literatur

Alexander, R. (2012). Improving oracy and classroom talk in English schools: Achievements and challenges. *White Rose Research – Online,* 1–15. Verfügbar unter https://eprints.whiterose.ac.uk/76321/1/R_alexander_improving_oracy.pdf

Behrens, U. (2010). Aspekte eines Kompetenzmodells zum Zuhören und Möglichkeiten ihrer Testung. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis* (S. 31–50). Göttingen: Vandenhoeck & Ruprecht.

Behrens, U. & Eriksson, B. (2009). Sprechen und Zuhören. In A. Bremerich-Vos, D. Granzer, U. Behrens & O. Köller (Hrsg.), *Bildungsstandards für die Grundschule: Deutsch konkret* (Bd. 3, S. 43). Berlin: Cornelsen.

Behrens, U. & Krelle, M. (2014). Hörverstehen – Ein Forschungsüberblick. *Didaktik Deutsch, 20* (36), 86–107. doi:10.25656/01:17208

BIFIE (2011). *Themenheft für den Kompetenzbereich „Hören, Sprechen und Miteinander-Reden". Deutsch, Lesen, Schreiben. Volksschule Grundstufe I + II.* Graz: Leykam.

BIFIE (2016a). *BIFIEsurvey: Tools for survey statistics in educational assessment* (Version R package version 1.8–0).

BIFIE (2016b). *Konstrukt- und Kompetenzstufenbeschreibung in Deutsch/Lesen/Schreiben 4. Schulstufe. Die Kompetenzstufen für die Überprüfung der Bildungsstandards* (2. Aufl.). Salzburg: BIFIE. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/fec772b59ab074e5cbe78f6278d1bcb2a8e6cad8/BIST-UE_D4_Konstruktbeschreibung.pdf

BMBWF (2018). *Grundsatzerlass „Reflexive Geschlechterpädagogik und Gleichstellung".* Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/schulrecht/rs/2018_21.html

BMBWF (2021). *Pädagogik-Paket.* Verfügbar unter: https://www.bmbwf.gv.at/Themen/schule/zrp/pp.html

BMF (2020). *Teilheft. Bundesvoranschlag 2020 Untergliederung 30. Bildung.* Verfügbar unter: https://service.bmf.gv.at/Budget/Budgets/2020/bfg/teilhefte/UG30/UG30_Teilheft_2020.pdf

BMUKK (2009). *Verordnung der Bundesministerin für Unterricht, Kunst und Kultur über Bildungsstandards im Schulwesen BGBl. II Nr. 1/2009 – Hauptdokument* [BGBl. II Nr 1/2009]. Verfügbar unter: http://www.ris.bka.gv.at/Dokument.wxe?Abfrage=BgblAuth&Dokumentnummer=BGBLA_2009_II_1

Breit, S., Bruneforth, M. & Schreiner, C. (2016). *Standardüberprüfung 2015 Deutsch, 4. Schulstufe. Bundesergebnisbericht, Salzburg.* Salzburg: BIFIE. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/0d879bba1b6fb81469e09bc99aab7e4916f88bef/BiSt_UE_D4_2015_Bundesergebnisbericht.pdf.

Bruneforth, M., Oberwimmer, K. & Robitzsch, A. (2016). Reporting und Analysen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 333–362). Wien: Facultas.

Eibl, S., Schwaiger, S. & Breit, S. (2018). Bildungsstandardüberprüfung Deutsch: Kompetenzmodelle als Grundlage für die Aufgabenentwicklung. In G. Sigott (Hrsg.), *Language Testing in Austria: Taking Stock / Sprachtesten in Österreich: Eine Bestandsaufnahme* (S. 23–42). Frankfurt am Main: Peter Lang.

Freunberger, R., Robitzsch, A. & Pham, G. (2014). *Hintergrundvariablen und spezielle Analysen. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013.* Salzburg: BIFIE. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/4e61c7bcf3f12cc9bc06c19ea372e5f721cedb40/TD_Hintergrundvariablen_E8.pdf.

Gangl, M. (2010). Causal Inference in Sociological Research. *Annual Review of Sociology, 36* (1), 21–47. doi:10.1146/annurev.soc.012809.102702

Ganzeboom, H. B., De Graaf, P. M. & Treiman, D. J. (1992). A Standard International Socio-Economic Index of Occupational Status. *Social Science Research, 21* (1), 1–56. doi:10.1016/0049-089X(92)90017-B

George, A. C., Oberwimmer, K. & Itzlinger-Bruneforth, U. (2016). Stichprobenziehung. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 51–81). Wien: Facultas.

Grotjahn, R. (2005). Testen und Bewerten des Hörverstehens. In M. Ó Dúill, R. Zahn & K. D. C. Höppner (Hrsg.), *Zusammenarbeiten: Eine Festschrift für Bernd Voss* (S. 115–144). Bochum: AKS-Verlag.

Gschwend, R. (2014). Zuhören und Hörverstehen-Aspekte, Ziele und Kompetenzen. In E. Grundler & C. Spiegel (Hrsg.), *Konzeptionen des Mündlichen. Wissenschaftliche Perspektiven und didaktische Konsequenzen* (Bd. 3, S. 143–160). Bern: hep.

Haag, N., Kocaj, A., Jansen, M. & Kuhl, P. (2017). Soziale Disparitäten. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich.* Münster: Waxmann.

IEA & BIFIE (2015). *PIRLS 2016. Schülerfragebogen.* Salzburg: IEA / BIFIE. Verfügbar unter https://www.iqs.gv.at/_Resources/Persistent/1d82b45c01b64feb64852aa7988c4baf30055d03/PIRLS_2016_Schuelerfragebogen.pdf.

Imhof, M. (2010). Zuhören lernen und lehren. Psychologische Grundlagen zur Beschreibung und Förderung von Zuhörkompetenzen in Schule und Unterricht. In V. Bernius & M. Imhof (Hrsg.), *Zuhörkompetenz in Unterricht und Schule. Beiträge aus Wissenschaft und Praxis.* Göttingen: Vandenhoeck & Ruprecht.

Kaldahl, A.-G., Bachinger, A. & Rijlaarsdam, G. (2019). Oracy Matters. Introduction to the special issue on oracy. Contribution to a special issue on Assessing Oracy, edited by Anne-Grete Kaldahl, Antonia Bachinger, and Gert Rijlaarsdam. *L1-Educational Studies in Language and Literature, 19,* 1–9. doi:10.17239/L1ESLL-2019.19.03.06

Krelle, M. & Prengel, J. (2014). Zur Konzeption von Zuhören im Rahmen der Vergleichsarbeiten für die dritte Klasse im Fach Deutsch. In E. Grundler & C. Spiegel (Hrsg.), *Konzeptionen des Mündlichen* (S. 208–226). Bern: hep-Verlag.

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data.* New Jersey: Wiley.

Lüdtke, O., Robitzsch, A., Köller, O. & Winkelmann, H. (2010). Kausale Effekte in der Empirischen Bildungsforschung. Ein Vergleich verschiedener Ansätze zur Schätzung des Effekts des Einschulungsalters. In W. Bos, E. Klieme & O. Köller (Hrsg.), *Schulische Lernangelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert* (S. 257–284). Münster: Waxmann.

Luger-Bazinger, C., Freunberger, R. & Itzlinger-Bruneforth, U. (2015). Das Standard-Setting als Beitrag zur validen Kompetenzdiagnostik. In M. Stock, P. Schlögl, K. Schmid &

D. Moser (Hrsg.), *Kompetent – wofür? Life Skills – Beruflichkeit – Persönlichkeitsentwicklung* (S. 191–205). Innsbruck: Studienverlag.

Mayrhofer, L., Oberwimmer, K., Toferer, B., Neubacher, M., Freunberger, R., Vogtenhuber, S. et al. (2019). Indikatoren C: Prozesse des Schulsystems. In K. Oberwimmer, S. Vogtenhuber, L. Lassnig & C. Schreiner (Hrsg.), *Nationaler Bildungsbericht Österreich, Band 1: Das Schulsystem im Spiegel von Daten und Indikatoren* (S. 123–196). Graz: Leykam. doi:10.17888/nbb2018-1-C.3

Neubacher, M., Freunberger, R., Schreiner, C., Vogtenhuber, S., Oberwimmer, K., Höller, I. et al. (2019). Indikatoren D: Output – Ergebnisse des Schulsystems. In K. Oberwimmer, S. Vogtenhuber, L. Lassnigg & C. Schreiner (Hrsg.), *Nationaler Bildungsbericht Österreich 2018* (S. 197–270). Graz: Leykam. doi:10.17888/nbb2018-1-D.2

Neubacher, M., Ober, M., Wimmer, C. & Hartl, M. (2019). Die Kompetenzen der Schüler/innen in der Zusammenschau aller Standardüberprüfungen des ersten Zyklus. In A. C. George, C. Schreiner, C. Wiesner, K. Pacher & M. Pointinger (Hrsg.), *Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven* (Bd. 1, S. 55–85). Münster: Waxmann. doi:10.31244/9783830991489

Oberwimmer, K., Baumegger, D. & Vogtenhuber, S. (2019). Indikatoren A: Kontext des Schul-und Bildungswesens. In K. Oberwimmer, S. Vogtenhuber, L. Lassnigg & C. Schreiner (Hrsg.), *Nationaler Bildungsbericht Österreich Band 1. Das Schulsystem im Spiegel von Daten und Indikatoren.* Graz: Leykam. doi:10.17888/nbb2018-1-A

Oberwimmer, K., Vogtenhuber, S., Lassnigg, L. & Schreiner, C. (2019). *Nationaler Bildungsbericht Österreich 2018, Band 1. Das Schulsystem im Spiegel von Daten und Indikatoren.* Graz: Leykam. doi:10.17888/nbb2018-1.4

Peña-López, I. (2017). PISA 2015 Results (Volume III). Students' Well-Being. Paris: OECD Publishing. doi:10.1787/9789264273856-en

R Development Core Team. (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rjosk, C., Haag, N., Heppt, B. & Stanat, P. (2017). Zuwanderungsbezogene Disparitäten. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 237–275). Münster: Waxmann.

Robitzsch, A., Pham, G. & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 259–294). Wien: Facultas.

Salchegger, S., Suchan, B., Widauer, K., Höller, I., Toferer, B. & Glaeser, A. (2017). Lesekompetenz im internationalen Vergleich. In C. Wallner-Paschon, U. Itzlinger-Bruneforth & C. Schreiner (Hrsg.), *PIRLS 2016. Die Lesekompetenz am Ende der Volksschule. Erste Ergebnisse.* Graz: Leykam.

Schipolowski, S., Wittig, J., Weirich, S. & Böhme, K. (2017). Geschlechtsbezogene Disparitäten. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 187–212). Münster: Waxmann.

Schreiner, C. & Wiesner, C. (2019). Die Überprüfung der Bildungsstandards in Österreich: der erste Zyklus als Meilenstein für die Schul- und Unterrichtsentwicklung – eine ge-

lungene Innovation im österreichischen Schulsystem. In A. C. George, C. Schreiner, C. Wiesner, K. Pacher & M. Pointinger (Hrsg.), *Kompetenzmessungen im österreichischen Schulsystem: Analysen, Methoden & Perspektiven* (S. 13–53). Münster: Waxmann. Verfügbar unter https://www.waxmann.com/index.php?eID=download&buchnr=3925

Schwabe, F., McElvany, N. & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly, 50* (2), 219–232. doi:10.1002/rrq.92

Stanat, P. & Bergann, S. (2009). Geschlechtsbezogene Disparitäten in der Bildung. In R. Tippelt & B. Schmidt (Hrsg.), *Handbuch Bildungsforschung* (S. 513–527): VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-91831-0_27

Stanat, P., Pant, H. A., Richter, D. & Böhme, K. (2012). IQB-Ländervergleich 2011: Zusammenfassung und Einordnung der Befunde. In P. Stanat, H. A. Pant, K. Böhme & D. Richter (Hrsg.), *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Deutsch und Mathematik. Ergebnisse des IQB-Ländervergleichs 2011.* Münster: Waxmann.

Trendtel, M., Schwabe, F. & Fellinger, R. (2016). Differenzielles Itemfunktionieren in Subgruppen. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (S. 111–147). Wien: Facultas.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54,* 427–450. doi:10.1007/BF02294627

Wiesner, C., Schreiner, C., Breit, S. & Pacher, K. (2017). Bildungsstandards und kompetenzorientierter Unterricht. *BIFIE-Journal, 1* (1), 1–5. doi:10.17888/bifiejournal-1.2017-1-1

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31* (2), 114–128. doi:10.1016/j.stueduc.2005.05.005

Gabriele von Eichhorn, Matthias Trendtel[1]

Kapitel 8
Aspekte psychometrischer Konstruktäquivalenz – Untersuchung zweier Operationalisierungen von Zuhörkompetenz

8.1 Einleitung

Ein zentrales Ziel der regelmäßigen Bildungsstandardüberprüfungen war, auf Systemebene Erkenntnisse darüber zu gewinnen, ob und wie sich die Leistungen der österreichischen Schüler/innen über die Zeit entwickeln. Um dies zu ermöglichen, wurden die zu verschiedenen Zeitpunkten gemessenen Kompetenzwerte auf einer gemeinsamen Metrik verlinkt (Trendtel, Pham & Yanagida, 2016). Das genaue Vorgehen bei der Verankerung des Testinstruments auf der so genannten „Bildungsstandard-Metrik" (BIST-Metrik) wird in Kapitel 6 – Verlinkung beschrieben. Eine Grundannahme bzw. Voraussetzung für eine aussagekräftige Verlinkung ist, dass die erhobenen Konstrukte hinreichend äquivalent über die Erhebungszeitpunkte sind, da sonst nicht gewährleistet werden kann, ob (Gruppen-) Unterschiede in der beobachteten Testleistung auf Fähigkeitsunterschiede in den erhobenen Konstrukten zurückzuführen sind, oder ob diese durch Eigenschaften der Messinstrumente entstehen (Holland, Dorans & Petersen, 2007).

Nach der Weiterentwicklung der Operationalisierung der Zuhörkompetenz (siehe Kapitel 2 – Konstrukt und Tabelle 1 in diesem Kapitel) stellt sich somit die Frage, ob mit den Testaufgaben – zum einen aus der Bildungsstandardüberprüfung 2015 und zum anderen aus der Pilotierungsstudie 2018 – die gleiche latente Fähigkeit gemessen wird und Konstruktäquivalenz vorliegt.

Einer solchen Fragestellung wurde in der Forschung bereits häufig nachgegangen: So wurden bereits verschiedene Studien mit unterschiedlichen Methoden verglichen, z. B. die deutschen Ländervergleiche mit dem *Programme for International Student Assessment* (PISA; Hartig & Frey, 2012; Jude, Hartig, Schipolowski, Böhme & Stanat, 2013), mit der *Progress in International Reading Literacy Study* (PIRLS; in Deutschland IGLU) und der *Trends in International Mathematics and Science Study* (TIMSS; Böhme et al., 2014). Auch

1 Die Autorin und der Autor haben zu gleichen Teilen zu dem Kapitel beigetragen.

das Nationale Bildungspanel (NEPS) und PISA wurden bereits vergleichend untersucht und mit den Ländervergleichen in Zusammenhang gebracht (Ehmke, van den Ham, Sälzer, Heine & Prenzel, 2020; van den Ham, Ehmke, Nissen & Roppelt, 2017; van den Ham, Nissen, Ehmke, Sälzer & Roppelt, 2014). Ziele solcher Untersuchungen sind beispielsweise, unterschiedliche Ergebnisse oder Kompetenzstufenverteilungen zu erklären oder die Metrik eines nationalen Tests auf der Metrik eines internationalen Tests zu verorten. Durch eine solche Vorgehensweise ist es häufig möglich, kriteriale Referenzwerte, die für die Metrik eines bestimmten Testinstruments abgeleitet wurden, auch auf ein anderes Testinstrument anzuwenden (Ehmke et al., 2020). Durch unterschiedliche Zielsetzungen und Fragestellungen wurde die Frage der Konstruktäquivalenz auch schon durch vielfältige Methoden untersucht. Den meisten Studien gemeinsam ist dabei ein *Single-Group*-Design (z. B. Dorans, Moses & Eignor, 2010), wodurch die Struktur der gemessenen Konstrukte innerhalb von Personen bei angenommenen konstanten Personeneigenschaften untersucht werden können.

Es gehört zur Natur von Kompetenzmessungen, die über einen längeren Zeitraum durchgeführt werden, Anpassungen der getesteten Konstrukte an die sich ändernden Bedingungen von Schule und Lernen vorzunehmen (z. B. Bruneforth & Höller, 2019; Stedman, 2008). Dennoch sind Studien zur Konstruktäquivalenz innerhalb einer Monitoring-Studie nicht sehr verbreitet. Das vorliegende Kapitel behandelt zwei Konstruktoperationalisierungen innerhalb eines Kompetenzmessungs-Kontexts: die Bildungsstandardüberprüfung der Kompetenz Zuhören auf der vierten Jahrgangsstufe 2015 und die der erneuten Überprüfung vorgelagerte Pilotierungsstudie im Jahr 2018, in der die weiterentwickelte Operationalisierung zum ersten Mal administriert wurde. Aufgrund der Weiterentwicklung der österreichischen Kompetenzmessungen zur *individuellen Kompetenzmessung PLUS* (iKM^{PLUS}), die ab dem Schuljahr 2021/22 stufenweise eingeführt wird, findet die nächste Überprüfung von *Zuhören* im Zyklusmodul ab 2024 statt (siehe Kapitel 1 – Überblick). Diese dient unter anderem dem Zweck, Ergebnisse zur Entwicklung der Zuhörkompetenzen der österreichischen Schüler/innen zu berichten.

Beiden Studien liegen die österreichischen Bildungsstandards zugrunde (BGBl. II Nr. 1/2009 i. d. F. BGBl. Nr. 548/2020). Die jeweilige Operationalisierung dieser Standards unterscheidet sich allerdings zwischen den Studien. In beiden Konstruktoperationalisierungen liegt der Fokus auf der Kompetenz der Schüler/innen, anderen aufmerksam zuzuhören, in der Weiterentwicklung wurden jedoch auch weitere relevante Bildungsstandards aus dem Bereich „Hören, Sprechen und Miteinander-Reden" hinzugezogen (siehe Kapitel 2 – Konstrukt).

Im Folgenden werden zunächst die wesentlichen Änderungen in der Konstruktoperationalisierung angeführt, die Fragestellungen der Studie vorgestellt und anschließend verschiedene Evidenzen abgeleitet, die Hinweise darauf geben können, inwieweit die weiterentwickelte Operationalisierung mit der ursprünglichen, 2015 eingesetzten Operationalisierung hinreichend ähnliche Eigenschaften aufweist. Betrachtet werden die latente Korrelation zwischen den aus den beiden Operationalisierungen abgeleiteten Testwerten, Korrelationen mit weiteren leistungsverwandten Variablen, Unterschiede in der Testletvarianz und schließlich der Varianzanteil auf Klassenebene.

Tabelle 1: Änderungen in der Konstruktoperationalisierung der Kompetenz Zuhören von 2015 zu 2018.

	Konstrukt 2015	Konstrukt 2018
Deskriptoren der Bildungsstandards	Fokus auf dem Deskriptor „Die Schüler/innen können anderen aufmerksam zuhören.“	Fokus auf dem Deskriptor „Die Schüler/innen können anderen aufmerksam zuhören.“ Hinzunahme von anderen Deskriptoren aus dem Bereich „Hören, Sprechen und Miteinander-Reden“ (siehe Kapitel 2 – Konstrukt)
Subkompetenzen	Neun Subkompetenzen (K 1–K 9), davon K1 bis K8 operationalisiert K 1: Die Bedeutung ähnlich klingender Wörter unterscheiden können K 2: Aus Texten Detailinformationen entnehmen K 3: Aus Texten die wesentlichen Informationen/Abläufe/Zusammenhänge entnehmen K 4: Aus Texten die Grundgedanken erfassen und weiterführen und/oder Schlussfolgerungen ziehen K 5: Die Bedeutung parasprachlicher Informationen erkennen und die Mitteilungsabsicht derselben einschätzen können K 6: Die Inhalte eines Textes auch bei Auftreten von Stör- bzw. Hintergrundgeräuschen erfassen können K 7: Die eigene Verstehensleistung kontrollieren (Comprehension Monitoring) K 8: Die Angemessenheit von Sprachhandlungen (Sprachkonventionen) erkennen und einschätzen können K 9: Einen altersadäquaten Wortschatz passiv verstehen und die Bedeutung der Wörter erfassen können	Vier Subkompetenzen (SK 1–SK 4): SK 1: Wiedergabe bzw. Auswahl von Einzelinformation SK 2: Schlussfolgerungen ziehen bzw. Informationen (lokal) verknüpfen SK 3: auf der Ebene des Textes Zusammenhänge erkennen SK 4: auf den Text bezogene Aussagen beurteilen, begründen oder darüber reflektieren
Textart	semiauthentischer Charakter	Texte entsprechen Hörerfahrungen der Kinder (z. B. Hörspiele, Radiosendungen)
Textsorten	fiktional und nichtfiktional	expositorisch und literarisch (zu gleichen Teilen)
Textlänge	zwischen 0:12 und 1:55 Min. (Mittelwert BIST 2015: 0:51 Min.)	zwischen 0:37 und 4:22 Min. lang (Mittelwert Pilot 2018: 1:45 Min.)
Testletlänge	1–4 Items pro Stimulus (Mittelwert BIST 2015: 2,3)	3–8 Items pro Stimulus (Mittelwert Pilot 2018: 5,2)
Itemformate	Multiple-Choice-Items mit einem Attraktor und drei Distraktoren, Richtig-falsch-Items mit jeweils vier Aussagen	Multiple-Choice-Items mit einem Attraktor und drei Distraktoren, Richtig-falsch-Items mit jeweils vier Aussagen, offene bzw. halboffene Items
Administration	Text von Audioquelle, einmaliges Abspielen, teilweise Itempreview	Text von Audioquelle, einmaliges Abspielen, kein Itempreview

8.2 Änderungen der Operationalisierung des Kompetenzkonstrukts

Im Anschluss an die Bildungsstandardüberprüfung in Deutsch 2015 wurde eine Weiterentwicklung des Tests beschlossen. Es sollten sowohl die zugrundeliegenden Konstrukte als auch die Testaufgaben adaptiert und verbessert werden. Die inhaltlichen Neuerungen umfassen Anpassungen an die aktuelle Forschung und Änderungen in der Auslegung der zugrundeliegenden Deskriptoren der Bildungsstandards (für Genaueres siehe Kapitel 2–4 in diesem Berichtsband).

Die Konstruktoperationalisierung von Zuhören in der 4. Schulstufe wurde in der Bildungsstandardüberprüfung (BIST-Ü) 2015 wie folgt zusammengefasst (Breit, Bruneforth & Schreiner, 2016, S. 84):

> Hörverstehen umfasst den Prozess und die Fähigkeit der komplexen, situationsbezogenen und unmittelbaren Verarbeitung von Gehörtem. Es werden die Fähigkeiten der Schüler/innen berichtet, monologische und dialogische Hörtexte semiauthentischen Charakters zu verstehen, den Hörtexten wichtige Detailinformationen zu entnehmen und textnahe Schlüsse zu ziehen. Auch die Fähigkeiten, Informationen zu vergleichen und zu beurteilen sowie die Angemessenheit von Sprechhandlungen zu erkennen und einzuschätzen, werden anhand von fiktionalen und nichtfiktionalen Hörtexten erhoben.

Folgende Änderungen ergeben sich nun mit Blick auf die aktuelle Operationalisierung (siehe Kapitel 2 – Konstrukt):

1. Während die Hörtexte in der Konstruktoperationalisierung der BIST-Ü (2015) *semiscripted,* d. h. auf der Basis von Stichworten, eingesprochen wurden, um authentische Situationen nachzuspielen, wurden in der aktuellen Weiterentwicklung nur solche Hörtexte eingesetzt, die den authentischen Zuhörerfahrungen der Kinder entsprechen. Dies sind überwiegend Hörtexte, in denen konzeptionell schriftliche Sprache mündlich vorgetragen wird (z. B. szenische Lesungen, Radiosendungen) sowie einige konzeptionell mündliche Hörtexte (z. B. Interviews). Näheres zur Unterscheidung konzeptioneller Schriftlichkeit und Mündlichkeit findet sich in Koch und Österreicher (1985).
2. Zudem wurde im Testdesign das Verhältnis von expositorischen und literarischen Texten in den Blick genommen. Der Anteil an literarischen Texten wurde erhöht. Weiterhin wurden die kognitiven Anforderungen der Testaufgaben in den Blick genommen. Die Testaufgaben können dementsprechend jeweils einer von vier Subkompetenzen zugeordnet werden, die zu gleichen Teilen im Test vorkommen:
   - die Wiedergabe bzw. Auswahl von Einzelinformation (Subkompetenz 1),
   - Schlussfolgerungen ziehen bzw. Informationen (lokal) verknüpfen (Subkompetenz 2),
   - auf der Ebene des Textes Zusammenhänge erkennen (Subkompetenz 3) und
   - auf den Text bezogene Aussagen beurteilen, begründen oder darüber reflektieren (Subkompetenz 4).

Die Definitionen der Subkompetenzen waren in Teilen bereits 2015 im Konstrukt angelegt, wurden nun jedoch stärker konturiert, um einen stärkeren Bezug auf die Bildungsstandards herstellen zu können.

3. Des Weiteren wurden – wie international üblich – längere Hörtexte eingesetzt, um die Bildung mentaler Modelle besser zu evozieren. Damit geht einher, dass nunmehr vier (statt zwölf) Hörtexte von einer Schülerin/einem Schüler zu bearbeiten sind. Um die Testzeit und die Hörtexte ökonomisch zu nutzen, sind mehr Items je Text zu bearbeiten. Dabei werden offene Itemformate administriert, um einen besseren Einblick in die sprachproduktive Auseinandersetzung der Schüler/innen mit den Hörtexten zu bekommen und das tiefere Verständnis der Texte besser messen zu können.

Alles zusammengenommen dienen die Änderungen dem Zweck, die Messung der Zuhörkompetenz besser in den aktuellen Forschungskontext einzubetten und sie überdies unterrichtsvalider zu gestalten. Dies ist hinsichtlich der Inhaltsvalidität eine positive Entwicklung, deren Einfluss auf die Äquivalenz oder Inäquivalenz der Konstrukte im Folgenden untersucht wird.

8.3 Herleitung der Fragestellungen

Die oben angeführten Änderungen betreffen teilweise die dem Konstrukt zugrunde liegenden Bildungsstandards sowie wesentliche Test- und Aufgabeneigenschaften. Besonders relevant für die Konstruktäquivalenz erscheinen hier die Auswahl der erhobenen Subkompetenzen und der Textarten, die Länge der Hörtexte und die Inkludierung offener Itemformate.

Die erhobenen Subkompetenzen bestimmen, welche Fähigkeiten die Schüler/innen bei der Bearbeitung des Tests abrufen und zeigen sollen. Einige im Konstrukt von 2015 noch erfasste Aspekte sind nun nicht mehr Teil der Operationalisierung. Dies stellt eine inhaltlich relevante Änderung dar. Weiterhin wurde der Anteil literarischer Texte erhöht. Best, Floyd und McNamara (2008) argumentieren, dass Drittklässler/innen ein besseres Leseverständnis von literarischen Texten als von expositorischen Texten aufwiesen. Expositorische Texte hingegen waren stärker vom Weltwissen der Schüler/innen abhängig. Das legt nahe, dass bestimmte Schülergruppen unterschiedlich gute Voraussetzungen für unterschiedliche Textarten mitbringen. Überträgt man diesen Genreeffekt auf das Hörverstehen, kann sich eine Änderung der Zusammensetzung der Textarten auch auf die Konstruktäquivalenz auswirken.

Ein weiterer Aspekt, der eine Auswirkung auf die Äquivalenz der Konstrukte haben kann, ist die Länge der Hörtexte. Durch die längeren Hörtexte in der Operationalisierung von 2018 wird das Arbeitsgedächtnis der Bearbeitenden stärker beansprucht als in der Operationalisierung von 2015 (siehe Kapitel 3 – Operationalisierung). Das kann sich auf das Bearbeitungs- und Antwortverhalten der Schüler/innen auswirken, was sich beispielsweise in unterschiedlichen Testwerten oder auch Reliabilitäten niederschlagen kann (Henning, 1990).

Zuletzt stellt die Auswahl passender Itemformate einen integralen Bestandteil in der Operationalisierung eines Konstrukts dar (Rodriguez, 2002). Jedes Itemformat bringt spezifische Eigenschaften mit sich, die seine Eignung für einen bestimmten Zweck mitbestimmen. Geschlossene bzw. Multiple-Choice-Items eignen sich für alle Arten des Messens von Fähigkeiten durch Abruf, Verständnis oder Anwendung (Haladyna & Rodriguez, 2013). Offene Itemformate hingegen scheinen eine weitere Bandbreite (komplexer) kognitiver Funktionen wie die Synthese und Evaluation von Information zu evozieren und werden daher häufiger für diese Zwecke eingesetzt (Martinez, 1999; Rauch & Hartig, 2010; Rodriguez, 2002). Verschiedene Itemformate können weiterhin in unterschiedlicher Weise mit Persönlichkeitseigenschaften der Testteilnehmer/innen interagieren (z. B. Solheim, 2011) oder einen Einfluss auf die empirische Schwierigkeit einer Testaufgabe haben (In'nami & Koizumi, 2009). Studien zur Äquivalenz der Kompetenzmessung durch verschiedene Itemformate sind trotz einer langen Forschungstradition inkonklusiv: Sowohl Traub (1993) als auch Rodriguez (2003) fanden in ihren respektiven Literaturreviews zu diesem Thema heterogene Ergebnisse, wobei die Itemformate mit steigender Ähnlichkeit in Inhalt und Itemstamm stärker zusammenhingen. Bennett (1991) gab allerdings zu bedenken, dass auch bei hoch korrelierten Leistungen unterschiedliche kognitive Prozesse hinter der Bearbeitung von offenen und geschlossenen Items liegen können. Die Reliabilität und Validität eines Tests scheinen aber oft nur bedingt vom eingesetzten Itemformat abhängig zu sein (Schult & Sparfeldt, 2016). Die Evidenzen zur Unterschiedlichkeit offener und geschlossener Formate legen nahe, dass die Entscheidung, offene Items in die Operationalisierung von 2018 zu integrieren, einen Einfluss auf die Äquivalenz der beiden Konstrukte ausüben kann.

Aus all diesen Gründen ist aus psychometrischer und fachdidaktischer Sicht durchaus ein gewisser Grad an „Inäquivalenz" möglich und erwartbar. Das 2018 eingesetzte Konstrukt wird allerdings in Zukunft in den Zyklusmodulen der Kompetenzmessung iKM^{PLUS} zum Einsatz kommen. Ein wesentlicher Testzweck dieser Module ist die Feststellung der Kompetenzentwicklung innerhalb der österreichischen Schülerpopulation, weswegen eine Verknüpfung zur vorangegangenen Erhebung 2015 erforderlich ist. Daher ist das Ziel dieser Studie nicht, die Frage nach Äquivalenz mit „ja" oder „nein" zu beantworten. Vielmehr ist das Ausmaß an (In-)Äquivalenz zwischen den beiden Konstrukten von Interesse, um Interpretationshilfen für den in Zukunft berichteten Kohortentrend ableiten zu können. Stellen sich die Konstrukte als stark äquivalent heraus, ist dies ein stärkendes Argument für die Verlinkung der beiden Studien und den Kohortentrend. Gibt es hingegen Evidenzen, die gegen eine vollständige Konstruktäquivalenz sprechen, schränkt dies die Interpretierbarkeit des Kohortentrends ein. Nachfolgend werden die vier Fragestellungen hergeleitet, die in dieser Studie als Evidenzen untersucht werden.

Modellvergleiche und Zusammenhang zwischen den Konstruktoperationalisierungen

Durch mehrdimensionale Item-Response-Theory-Modellierungen können Evidenzen für die empirische Abgrenzbarkeit der Konstruktoperationalisierungen abgeleitet werden. Zum einen können die Modellfits verschiedener Modellspezifikationen verglichen werden. Zum anderen erhält man durch eine explizite Modellierung den latenten Zusammenhang zwischen den Konstruktoperationalisierungen. Dieser sollte hoch ausgeprägt sein – dies

ist für die Annahme von Konstruktäquivalenz eine notwendige (aber nicht hinreichende) Bedingung. Ein hoher Zusammenhang der beiden Konstruktoperationalisierungen würde darauf hindeuten, dass Schüler/innen mit beiden Testinstrumenten zu ähnlichen Ergebnissen kommen würden. Dies führt zu Fragestellung 1:

1. Lassen sich die beiden Konstrukte besser getrennt oder gemeinsam modellieren? Wie hoch ist der Zusammenhang zwischen den 2015 und 2018 operationalisierten Konstrukten?

Nomologisches Netzwerk: Zusammenhänge mit weiteren Variablen

Hartig und Frey (2012, S. 13) weisen darauf hin, dass die „bloße Betrachtung korrelativer Zusammenhänge zur Beurteilung der Abgrenzbarkeit von Messverfahren kein vollständiges Bild zeichnet". Daher sollten auch Zusammenhänge mit relevanten weiteren Variablen herangezogen werden, um einen umfassenderen Eindruck über Unterschiede und Gemeinsamkeiten der Konstrukte und deren Operationalisierung zu bekommen. Als Evidenz für konvergente und divergente Validität erwartet man hier, dass der Zusammenhang zwischen den beiden Zuhören-Konstruktoperationalisierungen höher ist als mit Konstrukten, die etwas anderes zu messen versuchen, beispielsweise Aspekte mathematischer Kompetenz.

Die vorherige Fragestellung bezieht sich auf den Zusammenhang zwischen Konstruktoperationalisierungen der Kompetenz Zuhören. Um die beiden Konstruktoperationalisierungen vergleichen zu können, ist es allerdings auch notwendig, Zusammenhänge mit anderen leistungs- und konstruktrelevanten Variablen wie dem fachbezogenen Selbstkonzept, der Lernfreude und der Schulfreude zu explorieren. Zusammenhänge zwischen dem fachbezogenen Selbstkonzept und der gemessenen Kompetenz sind gut belegt, mit der Lernfreude bestehen eher kleine Zusammenhänge (Lohbeck, Hagenauer & Moschner, 2016; Schreiner et al., 2019). Dennoch sind diese motivationalen Faktoren wesentlich für das Lernen und stehen mit dem Kompetenzerwerb in Verbindung (Pekrun, 2011). Bei Konstruktäquivalenz würde man ähnlich hohe Zusammenhänge zwischen den Konstrukten von 2015 und 2018 mit den Fragebogenvariablen erwarten. Im Sinne eines solchen nomologischen Netzwerks lassen sich ebenfalls Evidenzen für die Konstruktvalidität ableiten. Somit stellt sich Fragestellung 2 folgendermaßen dar:

2. Wie hoch sind die Zusammenhänge mit den Konstrukten Lesekompetenz, Mathematikkompetenz, Zugehörigkeitsgefühl, Lernfreude und Selbstkonzept im Fach Deutsch und unterscheiden sich dabei die beiden Konstruktoperationalisierungen?

Testletvarianz

Eine wesentliche Änderung im Vergleich zum Zuhören-Konstrukt von 2015 war die Entwicklung und der Einsatz längerer Stimuli mit einer größeren Anzahl an Items pro Stimulus. Längere Hörtexte sind häufig komplexer und bieten eine größere inhaltliche Tiefe als kürzere Texte. Dadurch sind sie zum einen interessanter für die testteilnehmenden Kinder, zum anderen können schwierigere Testitems dazu entwickelt werden (Weirich, Bachinger, Trendtel & Krelle, 2019). Weiterhin evozieren sie komplexere mentale Modelle, wodurch die Zuhörkompetenz der Schüler/innen stärker gefordert ist. Um die begrenzte Testzeit ökonomisch zu nutzen, ist es notwendig, zu längeren Hörtexten auch mehr Testitems in

einem sogenannten *Testlet*[2] zu administrieren (Wainer, Bradlow & Wang, 2007). Betrachtet man die Forschung zum Einsatz mehrerer Testitems zu einem Stimulus, findet man unterschiedliche Positionen. Als wesentlicher Vorteil wird eine erhöhte Validität angeführt, als Nachteil dagegen ein größeres Risiko für lokale Abhängigkeiten zwischen Testitems, die einem Stimulus zugehören.

Robitzsch (2015, S. 35) schreibt zu den Vorteilen, dass möglicherweise „erst durch die Bearbeitung mehrerer Items zu einem Testlet alle Facetten des zu messenden Konstrukts aufgedeckt werden" (siehe auch Ferrara, Huynh & Michaels, 1999). Auf diese Weise drückt sich das Reliabilitäts-Validitäts-Dilemma (Brennan, 2001) aus: In Testlets geschachtelte Items haben unter Umständen eine verminderte Reliabilität (zählt man die durch Testletfaktoren erzeugte Varianz zur Fehlervarianz), dafür aber unter Umständen eine höhere Validität. Die Validität eines Tests kann allerdings nicht unter nur rein empirisch-psychometrischer Betrachtung bestimmt werden (Robitzsch, 2015). Bei der Operationalisierung der österreichischen Bildungsstandards wird argumentiert, dass es durchaus Bestandteil des zu messenden Konstrukts ist, die Fähigkeit aufzuweisen, mehrere Fragen zu einem längeren Hörtext beantworten zu können. Dies ist auch unterrichtsvalider. Somit wäre die durch Testletfaktoren eventuell erzeugte Varianz der wahren Varianz zuzuordnen.

Neben einigen Vorteilen in der Konstruktabdeckung und Validität steigt mit mehr Testitems zu einem Stimulus jedoch auch das Risiko lokaler Abhängigkeiten (Wang & Wilson, 2005). Je mehr Items zu ein und demselben Stimulus entwickelt werden, desto schwieriger wird es, dass diese Items sich auf neue Informationen in dem Stimulus beziehen und somit neue Information über die testteilnehmende Person generieren (Wainer et al., 2007, S. 44). Lokale Abhängigkeit liegt vor, wenn Itemantworten auch dann noch zusammenhängen, wenn die Schülerfähigkeit kontrolliert wird. Praktisch bedeutet das, dass Schüler/innen Teile des Textes missverstanden haben können und dadurch gleich bei mehreren Items eine geringere Wahrscheinlichkeit haben, sie richtig zu beantworten oder dass die Fragestellung eines Items bereits Hinweise auf die Beantwortung eines anderen Items enthält. Auf Letzteres wurde im Rahmen der Aufgabenentwicklung besonderes Augenmerk gelegt (siehe Kapitel 3 – Operationalisierung). Braeken (2011) argumentiert jedoch, dass die Annahme der kompletten lokalen stochastischen Unabhängigkeit in vielen Anwendungsbereichen unwahrscheinlich ist. In der Anwendung gibt es immer Subsets von Items, die nicht nur konstruktrelevante Gemeinsamkeiten teilen – sei es, dass sie das gleiche Thema behandeln, das gleiche Stimulusmaterial nutzen oder ein ähnliches Problem schildern.

Die meisten üblichen statistischen Modelle erfordern jedoch lokale stochastische Unabhängigkeit der Itemantworten, da in diesen Modellen angenommen wird, dass die Wahrscheinlichkeit einer richtigen Antwort ausschließlich von der Fähigkeit der Person und Eigenschaften des betrachteten Items abhängt (Embretson & Reise, 2000). Bei Vorliegen lokaler Abhängigkeit zwischen zwei Items hängt die Wahrscheinlichkeit einer richtigen Antwort auch von der des anderen Items ab, weswegen beide Items weniger Information

2 Als *Testlet* wird ein Stimulus mit mehreren dazugehörigen Testitems bezeichnet.

über die Schülerfähigkeit liefern können, als potenziell möglich wäre. In so einem Fall können verzerrte Modellschätzer oder eine überschätzte Testinformation resultieren (Ferrara et al., 1999; Wainer & Thissen, 1996; Yen, 1993). Daher ist eine Prüfung der lokalen stochastischen Unabhängigkeit ein wichtiger Schritt in der psychometrischen Evaluation eines Tests. Häufig wird argumentiert, dass die Reliabilität eines Tests überschätzt wird, wenn die Item-Zusammenhänge innerhalb von Testlets nicht berücksichtigt werden (Zenisky, Hambleton & Sireci, 2001). Deshalb haben sich verschiedene Ansätze zum Umgang mit solchen Testletfaktoren entwickelt, um diese lokalen Abhängigkeiten zu modellieren (z. B. Wainer et al., 2007; Yen, 1993). Es soll an dieser Stelle aber deutlich gemacht werden, dass die Frage, ob die Reliabilität bei Vorliegen solcher Testletfaktoren unter- oder überschätzt wird, mit der Annahme verbunden ist, ob diese Testletfaktoren konstruktinhärent oder konstruktirrelevant sind (Robitzsch & Lüdtke, 2015).

In Fragestellung 3 sollen den vorangegangenen Überlegungen entsprechend lokale Abhängigkeiten im Vergleich der beiden Konstruktoperationalisierungen aufgegriffen und geprüft werden.

3. Führt die größere Menge an Items je Stimulus der Konstruktoperationalisierung von 2018 zu stärkeren lokalen Abhängigkeiten und zu einer höheren Testletvarianz im Vergleich zur Konstruktoperationalisierung von 2015?

Varianz auf Klassenebene (Unterrichtsnähe)

Für jeden Schulleistungstest stellt sich die Frage nach der curricularen Validität (Robitzsch, 2015). Grundsätzlich basieren beide Konstrukte auf den Bildungsstandards und erfassen somit konkrete aus dem Curriculum abgeleitete Lernergebnisse. Im weiterentwickelten Konstrukt wurde jedoch ein gesteigerter Wert daraufgelegt, dass die Hörtexte auch tatsächlich Gegenstand von Unterricht sein könnten (siehe Kapitel 3 – Operationalisierung). Das heißt, dass die Konstruktoperationalisierung von 2018 Inhalte prüft und Kompetenzen erfasst, die sich im Allgemeinen näher am Curriculum orientieren – also unterrichtsnäher sind –, als dies bei der Konstruktoperationalisierung von 2015 der Fall war.

Je unterrichtsnäher die Testinhalte sind, desto instruktionssensitiver sollte der Test sein (Hartig & Frey, 2012), das heißt, desto mehr Unterschiede zwischen Klassen sind zu erwarten. Dies liegt daran, dass die Kompetenz, die für die erfolgreiche Bearbeitung der Testaufgaben notwendig ist, in größerem Maße im Unterricht vermittelt werden sollte, als dies bei unterrichtsferneren Testinhalten zu erwarten ist. In einem solchen Fall schlägt sich dies empirisch in den Intraklassen-Korrelationen nieder: Bei Tests über unterrichtsfernere Lerninhalte sollte der Anteil der Varianz in den gemessenen Kompetenzen, der auf Klassenebene aufgeklärt werden kann, im Vergleich zur Varianzaufklärung auf Individualebene geringer ausfallen, vergleicht man dieses Verhältnis mit den Varianzanteilen basierend auf Tests mit unterrichtsnäheren Lerninhalten.

Die größere Unterrichtsnähe der Konstruktoperationalisierung von 2018 war ein Kernziel der Weiterentwicklung. Es stellt sich die Frage, ob dieses Ziel erreicht wurde. Dies wäre ein Hinweis auf curriculare Validität und damit eine Evidenz für das Gelingen der

Überarbeitung. Daneben wäre die größere Unterrichtsnähe auch günstig für den Einsatz des Konstrukts in der zukünftigen iKM^{PLUS}. Diese wird für Lehrpersonen eine größere Bedeutung haben als die BIST-Ü, da sie sich intensiver mit den Ergebnissen ihrer Klasse auseinandersetzen sollen, um diese mit den Schülerinnen und Schülern zu besprechen. Somit ergibt sich Fragestellung 4:

4. Findet sich die angezielte größere Unterrichtsnähe der Testinhalte der Konstruktoperationalisierung von 2018 in höheren Varianzanteilen auf der Klassenebene wieder?

8.4 Methode

8.4.1 Stichprobe

Zur Beantwortung der Forschungsfragen untersuchen wir Daten der Pilotierungsstudie der österreichischen Bildungsstandardüberprüfungen der 4. Jahrgangsstufe im Fach Deutsch aus dem Jahr 2018. Dabei wurde die Stichprobe auf diejenigen Klassen eingeschränkt, die an der sogenannten Validierungsstudie teilnahmen (siehe auch den nächsten Abschnitt zum Design dieser Studie). Diese Teilstichprobe umfasst n = 950 Kinder und wurde separat als repräsentative Stichprobe für Österreich gezogen (siehe Kapitel 5 – Studiendesign).

Die teilnehmenden Kinder waren im Mittel 10,38 Jahre alt bei einer Standardabweichung von 0,49 Jahren und einem Minimum sowie Maximum von 9,50 bzw. 12,58 Jahren. Der Anteil von Mädchen in der Stichprobe beläuft sich auf 51,37 %. Die Angaben zu Alter und Geschlecht wurden systematisch erfasst und weisen keine fehlenden Werte auf. Hinsichtlich des sozioökonomischen Hintergrundes wurde der elterliche berufliche Status in Form des *International Socio-Economic Index of Occupational Status* (ISEI; Ganzeboom, De Graaf & Treiman, 1992) der Eltern erfasst. Betrachtet man dabei den höchsten elterlichen Wert (HISEI), ergab sich basierend auf den gültigen Angaben ein Mittelwert von 50,28 bei einer Standardabweichung von 20,73 und 7,79 % fehlenden Werten. Der Migrationshintergrund der Kinder wurde operationalisiert über das Geburtsland der Eltern: Sind beide Eltern des Kindes im Ausland (ausgenommen Deutschland) geboren, gilt es als Kind mit Migrationshintergrund. Ist mindestens ein Elternteil in Österreich (bzw. Deutschland) geboren, gilt es als Kind ohne Migrationshintergrund. Basierend auf den gültigen Werten wiesen in der Stichprobe nach dieser Definition 19,93 % der Kinder einen Migrationshintergrund auf, bei 4,95 % fehlenden Werten. Betrachtet man die Angaben zur Erstsprache, gaben 72,34 % der Kinder an, Deutsch als einzige Erstsprache, 7,17 % Deutsch und eine weitere Sprache und 20,49 % eine andere Sprache als Deutsch als Muttersprache zu sprechen (bei 6,00 % fehlenden Werten).

8.4.2 Messinstrumente

Kompetenztest Zuhören

Wie in Kapitel 1 – Überblick und Kapitel 5 – Studiendesign beschrieben, bestand die Pilotierungsstudie 2018 aus vier verschiedenen Substudien, denen unterschiedliche Test-

designs zugrunde lagen. Für die Validierungsstudie wurde ein *Single-Group-* bzw. *Within-Subject*-Design gewählt, d. h., ein und dieselbe Person bearbeitet Items beider Konstruktoperationalisierungen. Dorans et al. (2010) beschreiben das Single-Group-Design als das einfachste Datenerhebungsdesign, um ein Testinstrument auf der Metrik eines anderen zu verankern *(Equating),* da alle Teilnehmenden beide Tests absolvieren. Auf diese Weise können die latenten Fähigkeiten der Personen als konstant angenommen werden und Unterschiede in der Testleistung sind weitgehend auf die unterschiedlichen Operationalisierungen zurückzuführen. In dieser Studie wurde das Single-Group-Design derart umgesetzt, dass in einem Testheft Items beider Konstruktoperationalisierungen enthalten waren. Daher war die Testzeit für jede Konstruktoperationalisierung kürzer als üblich. In jedem Fall ist bei diesem Design auf eine Positionsbalancierung zu achten, d. h., dass die Hälfte der Schüler/innen zuerst Items der Konstruktoperationalisierung von 2015 und dann Items der Konstruktoperationalisierung von 2018 bearbeitet und die andere Hälfte der Schüler/innen umgekehrt. So kann im Mittel davon ausgegangen werden, dass sich Reihenfolgeeffekte wie zum Beispiel Ermüdungseffekte in der Gesamtstichprobe ausmitteln.

Die Aufgaben wurden in einem balancierten unvollständigen Blockdesign vorgegeben (Frey, Hartig & Rupp, 2009). Die acht Testhefte der Validierungsstudie enthielten jeweils acht Stimuli, von denen jeweils sechs dem Testkonstrukt von 2015 und zwei dem Testkonstrukt von 2018 entstammten. Aufgrund der größeren Itemanzahl pro Stimulus im Testkonstrukt von 2018 waren die Items beider Konstrukte annähernd zu gleichen Teilen im Testheft vertreten. Jede Schülerin/jeder Schüler bearbeitete 22–24 Items, die sich aus 11–14 Items aus dem Testkonstrukt von 2015 und 9–11 Items aus dem Testkonstrukt von 2018 zusammensetzten.

In vier der acht Testhefte wurden zuerst die Aufgaben aus der Bildungsstandardüberprüfung 2015 dargeboten, in den restlichen vier Testheften mussten zunächst die neu entwickelten Aufgaben bearbeitet werden. Insgesamt wurden aus dem Konstrukt von 2015 16 unterschiedliche Stimuli mit 34 Items, aus dem Konstrukt von 2018 acht Stimuli mit 40 Items eingesetzt.

Weitere Konstrukte

Für das nomologische Netzwerk wurden die Konstrukte Zugehörigkeitsgefühl in der Schule, Lernfreude und Selbstkonzept jeweils im Fach Deutsch verwendet. Zusätzlich wurden der Leseverstehen-Test aus der D4-Pilotierung 2018 und die im gleichen Jahr stattfindende Bildungsstandardüberprüfung in Mathematik als Hinweise für die divergente Validität herangezogen.

Das Konstrukt Zugehörigkeitsgefühl wurde mit fünf jeweils vierstufigen (0 = „stimme überhaupt nicht zu“ bis 3 = „stimme völlig zu“) Items gemessen, wie zum Beispiel „Ich habe das Gefühl, in der Schule dazuzugehören“, wobei eine interne Konsistenz von α = 0,79 erzielt wurde. Die Konstrukte Lernfreude und Selbstkonzept im Fach Deutsch wurden mit je vier Items (Beispielitem Lernfreude „Ich lerne gern Deutsch“; Beispielitem Selbstkonzept „Normalerweise bin ich gut in Deutsch“) mit jeweils einer vierstufigen Antwort-

skala (0 = „stimme überhaupt nicht zu“ bis 3 = „stimme völlig zu“) erfasst bei erreichten internen Konsistenzen von α = 0,85 für Lernfreude bzw. α = 0,78 für Selbstkonzept.

Die Kompetenz Lesen wurde mit insgesamt 294 Items erhoben, wobei ein Kind zwischen 16 und 20 Items bearbeitete. In einem Partial-Credit-Modell (PCM; Masters, 1982) wurde eine EAP-Reliabilität von 0,78 erreicht. Die Kompetenz in Mathematik wurde mit 243 Items überprüft und ein Kind bearbeitete in der Regel 68 Items bei einer internen Konsistenz von α = 0,92 (EAP).

8.4.3 Analytisches Vorgehen

Zur Beantwortung der Forschungsfrage 1 wurde ein einparametrisches logistisches Response-Modell in zwei Versionen angepasst: zum einen spezifiziert als eindimensionales Modell und zum anderen als zweidimensionales Modell, wobei die neu entwickelten Items auf der ersten Dimension und die Items aus der Erhebung 2015 auf der zweiten Dimension laden (siehe auch Abbildung 1). Um mögliche Unterschiede in den Traitvarianzen zu berücksichtigen, wurde im einparametrischen Fall für die beiden unterschiedlichen Itemgruppen je ein allgemeiner itemgruppenspezifischer Trennschärfeparameter eingeführt.

Zur Untersuchung des nomologischen Netzwerks für Forschungsfrage 2 wurde ein siebendimensionales einparametrisches Item-Response-Modell angepasst mit den Dimensionen (1) Zuhörkompetenz-2015, (2) Zuhörkompetenz-2018, (3) Lesekompetenz, (4) Mathematikkompetenz, (5) Zugehörigkeitsgefühl sowie (6) Lernfreude und (7) Selbstkonzept im Fach Deutsch.

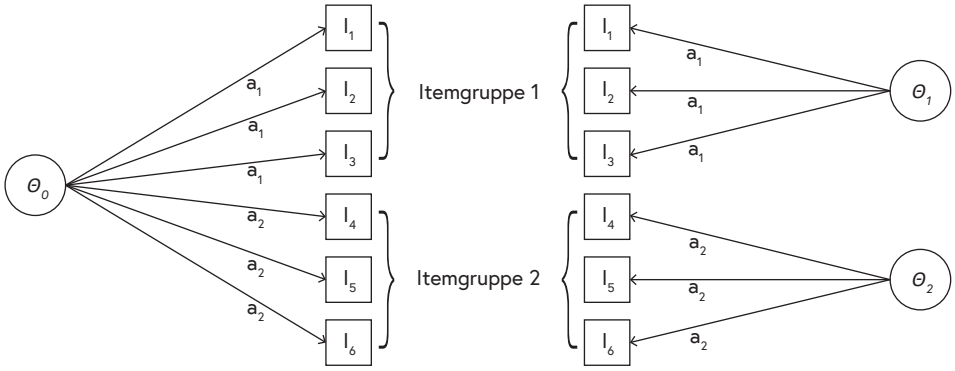


Abbildung 1: Ein- und zweidimensionales Response-Modell mit itemgruppenspezifischen Trennschärfeparametern.

Der Frage nach lokalen Abhängigkeiten innerhalb der beiden Konstruktoperationalisierungen wurde nachgegangen, indem für beide Operationalisierungen je ein eindimensionales einparametrisches Response-Modell angepasst wurde und danach die lokalen Abhängigkeiten itempaarweise mithilfe der Q3-Statistik (Yen, 1984) analysiert wurden. Für den Vergleich wurden relative Anteile von Itempaaren berechnet, die eine absolute Residual-Korrelation von größer als 0,2 aufwiesen.

Um die Varianzanteile auf Klassen- bzw. Testletebene zu bestimmen und zu vergleichen, wurden Mehrebenen-Response-Modelle angepasst. Hierfür wurden die Antwortdaten dichotomisiert (vollständig richtig = 1; sonst 0).

Die geclusterte, hierarchische Stichprobenstruktur (siehe Kapitel 5 – Studiendesign in diesem Band) wurde berücksichtigt, indem Stichprobengewichte berücksichtigt wurden und für die Ableitung unverzerrter Standardfehler entsprechende Resampling-Methoden (z. B. Kolenikov, 2010) Anwendung fanden, wie sie in Large-Scale Assessments üblich sind. Für die Berechnungen wurde die Statistik-Softwareumgebung R (R Core Team, 2020) genutzt, zusammen mit den Paketen TAM (Robitzsch, Kiefer & Wu, 2021) und lme4 (Bates, Mächler, Bolker & Walker, 2015).

8.5 Ergebnisse

Im ersten Schritt wurden für die Leistungsdaten der Validierungsstudie ein ein- sowie ein zweidimensionales Item-Response-Modell angepasst. Im Ergebnis passen beide Modelle gut zu den Daten: In beiden Spezifikationen liegen die INFIT-Werte (Wright & Panchapakesan, 1969) der Items in einem Bereich zwischen 0,85 und 1,25 und damit weit entfernt von den kritischen Schwellen von 0,5 und 1,5 (Ames & Penfield, 2015). Lediglich die Standardized-Root-Mean-Squared-Residual-Statistik (SRMSR) liegt mit jeweils 0,084 über dem kritischen Wert von 0,05 (Maydeu-Olivares, 2013). Die Reliabilitäten (siehe Tabelle 2) lagen dagegen in beiden Fällen über 0,7 und deuten auf eine gute interne Konsistenz hin (Nunally & Bernstein, 1994).

Bei einem Modellvergleich (siehe Tabelle 2) sieht man, dass das zweidimensionale Modell dem eindimensionalen Modell sowohl nach dem Akaike-Informationskriterium (AIC) als auch beim Likelihood-Ratio-Test mit einem p-Wert unter 0,01 überlegen ist. Die aus den beiden Konstruktoperationalisierungen gewonnenen Daten lassen sich entsprechend mit einem zweidimensionalen Modell besser modellieren als mit einem eindimensionalen Modell. Der Zusammenhang zwischen den beiden Dimensionen (siehe Tabelle 3) ist mit einem Wert über 0,95 (SE = 0,01) als hoch einzustufen (Nachtigall & Wirtz, 2004).

Tabelle 2: Ergebnisse der Anpassungen eines ein- und zweidimensionalen Modells an die Leistungsdaten der Validierungsstudie 2018

Modell	*Rel(EAP)*	*SRMSR*	*LL*	*N*(par)	AIC	X^2	*df*	*p*
Eindimensional	0,726	0,084	–15.288,49	111	30.799	10,6	1	< 0,01
Zweidimensional	0,715; 0,709	0,084	–15.283,19	112	30.790			

Anmerkungen: *Rel(EAP)* = Reliabilität basierend auf den expected a posteriori (EAP) als Schätzer der Personenfähigkeiten; *SRMSR* = Standardized Root Mean Square Residual; *LL* = Loglikelihood; *N*(par) = Anzahl zu schätzender Parameter im Modell; AIC = Akaike-Informationskriterium; X^2 = Chi-square-Statistik des Likelihood-Ratio-Tests; *df* = Anzahl Freiheitsgrade des Likelihood-Ratio-Tests; *p* = p-Wert des Likelihood-Ratio-Tests.

Betrachtet man die latenten Korrelationen des siebendimensionalen Modells, lassen sich keine systematischen Unterschiede in den Zusammenhängen erkennen. Beispielsweise fällt der Zusammenhang mit Mathematik bei der Konstruktoperationalisierung von 2018 höher aus als bei jener von 2015. Beim Zusammenhang mit dem Konstrukt Zugehörigkeitsgefühl verhält es sich umgekehrt. In beiden Fällen wie auch in allen anderen lassen sich diese Unterschiede jedoch nicht zufallskritisch absichern ($p > 0{,}05$). Auffallend ist, dass die Lernfreude im Fach Deutsch mit den Leistungsmaßen (bis auf Lesen) tendenziell negativ zusammenhängt. Bis auf Mathematik ($p = 0{,}02$) sind diese Zusammenhänge aber als statistisch nicht signifikant einzustufen.

Tabelle 3: Latente Korrelationen und deren Standardfehler der einzelnen Dimensionen in einem siebendimensionalen Response-Modell

Dimension	**Latente Korrelationen**					
	(1)	(2)	(3)	(4)	(5)	(6)
(1) Zuhören – 2015						
(2) Zuhören – 2018	0,954 (0,011)					
(3) Lesen	0,792 (0,076)	0,807 (0,086)				
(4) Mathematik	0,713 (0,137)	0,746 (0,078)	0,730 (0,056)			
(5) Zugehörigkeitsgefühl	0,227 (0,132)	0,209 (0,079)	0,214 (0,076)	0,101 (0,070)		
(6) Lernfreude – Deutsch	–0,028 (0,249)	–0,071 (0,074)	0,008 (0,077)	–0,160 (0,068)	0,106 (0,061)	
(7) Selbstkonzept – Deutsch	0,374 (0,045)	0,354 (0,055)	0,412 (0,084)	0,263 (0,051)	0,317 (0,050)	0,537 (0,042)

Die Ergebnisse der Analyse lokaler Abhängigkeiten mithilfe der Q3-Statistiken zeigen, dass bei der Konstruktoperationalisierung von 2015 ca. 6,2 % der Itempaare einen absoluten Wert über 0,2 aufwiesen. Bei der Operationalisierung von 2018 belief sich dieser Wert auf 14,4 %. Es liegen somit bei der Operationalisierung von 2018 mehr lokale Abhängigkeiten zwischen Items vor, als dies bei der Operationalisierung von 2015 der Fall ist.

Tabelle 4: Ergebnisse eines Mehrebenen-Response-Modells, wobei die Zufallseffekte auf Schüler-, Item- und Testletebene spezifiziert sind

	VAR (SE)	
Zufallseffekte	**Konstrukt 2015**	**Konstrukt 2018**
Schüler	1,000 (--)	1,000 (--)
Item	1,653 (0,183)	0,941 (0,137)
Testlet	0,000 (0,001)	0,017 (0,060)

Zerlegt man die Gesamtvarianz im Antwortverhalten in die Anteile, die auf Schüler-, Item- und Testletebene erzeugt werden (siehe Tabelle 4), sieht man, dass Items der Konstruktoperationalisierung von 2015 im Verhältnis zur Varianz auf Schülerebene mehr Varianz erzeugen, als dies bei der Operationalisierung von 2018 geschieht. Die Varianzen, die auf Testletebene aufgeklärt werden, sind aber in beiden Fällen nicht signifikant von null verschieden und unterscheiden sich auch untereinander nicht signifikant.

Tabelle 5: Ergebnisse eines Mehrebenen-Response-Modells, wobei die Zufallseffekte auf Schüler-, Klassen- und Itemebene spezifiziert sind

	VAR (SE)	
Zufallseffekte	**Konstrukt 2015**	**Konstrukt 2018**
Schüler	0,901 (0,053)	0,851 (0,060)
Item	0,099 (0,053)	0,149 (0,060)
Testlet	1,745 (0,408)	0,979 (0,157)

Bei der Varianzzerlegung in Anteile auf Schüler-, Klassen- und Itemebene ergeben sich Intraklassenkorrelationen von 0,10 bei der Konstruktoperationalisierung von 2015 und 0,15 bei der Operationalisierung von 2018. Der Varianzanteil auf Klassenebene ist darüber hinaus bei der Konstruktoperationalisierung von 2015 nicht signifikant von null verschieden ($p = 0{,}06$), während sich der Varianzanteil auf Klassenebene bei der Operationalisierung von 2018 signifikant von null unterscheidet ($p < 0{,}01$).

8.6 Conclusio

In dieser Studie sollten Evidenzen für eine Konstruktäquivalenz zweier Konstruktoperationalisierungen von Zuhörkompetenz in Erhebungen von 2015 und 2018 gesammelt werden. Dabei wurden im Wesentlichen vier Fragestellungen untersucht: die Anpassung der Daten an ein- und mehrdimensionale Modelle sowie der latente Zusammenhang zwischen den beiden Konstrukten (Fragestellung 1), der Zusammenhang mit weiteren kompetenzerwerbsrelevanten Konstrukten (Fragestellung 2) und das Ausmaß an Testletvarianz im Vergleich (Fragestellung 3). Weiterhin wurde analysiert, ob sich die angezielte größere Unterrichtsnähe des weiterentwickelten Konstrukts von 2018 auch in größeren Varianzanteilen auf der Klassenebene wiederfindet (Fragestellung 4).

In Fragestellung 1 wurde betrachtet, ob die Antwortdaten sich besser gemeinsam in einem eindimensionalen oder getrennt in einem zweidimensionalen Modell anpassen lassen. Weiterhin wurde der latente Zusammenhang zwischen den beiden Konstrukten modelliert. Der Modellvergleich zwischen einem ein- und zweidimensionalen Modell zeigt Hinweise darauf, dass Items der beiden Konstruktoperationalisierungen (psychometrisch) nicht auf ein- und derselben Dimension von *Zuhören* laden. Das bedeutet, dass die Antwortdaten sich durch eine getrennte Modellierung besser erklären lassen, was die Konstrukt-

äquivalenz und damit die Verlinkung der beiden Konstrukte infrage stellen könnte. Der hohe Zusammenhang dieser beiden Dimensionen von über 0,95 stellt aber die praktische Relevanz dieses Befunds in Frage und liefert eher einen Hinweis darauf, dass diese beiden Operationalisierungen als äquivalent betrachtet werden können.

Dies unterstreichen auch die Ergebnisse des in Fragestellung 2 untersuchten nomologischen Netzwerks: Auch hier zeigen sich keine statistisch bedeutsamen Unterschiede in den Zusammenhängen mit anderen Konstrukten. Insbesondere vor dem Hintergrund der Ergebnisberichterstattung der österreichischen Bildungsstandardüberprüfungen stellen diese Befunde eine valide Fortschreibung des Trends der Kompetenz Zuhören in der 4. Schulstufe nicht in Frage.

Hinsichtlich der Untersuchungsergebnisse lokaler Abhängigkeiten für Fragestellung 3 zeigt sich bei der Konstruktoperationalisierung von 2018 im Vergleich zur Operationalisierung von 2015 ein höherer Anteil von Itempaaren, die solche Abhängigkeiten aufweisen. Dies ist einerseits erwartbar bei der deutlich höheren Clusterung von Items innerhalb von Testlets durch den Einsatz von längeren Hörstimuli mit entsprechend höherer Anzahl Items. Andererseits sollte man bei der weiteren Itementwicklung im Auge behalten, solche Abhängigkeiten möglichst zu minimieren. Inwieweit der hier berichtete Anteil an betroffenen Itempaaren repräsentativ für die gesamte Itembank dieser Kompetenz ist, sollte anhand der anderen Substudien der Pilotierungsstudie überprüft werden.

Das Ergebnis des Varianzkomponentenmodells hinsichtlich der Varianzanteile auf Klassenebene bestätigte die Erwartungen. Hier deutet der höhere Varianzanteil auf Klassenebene der Konstruktoperationalisierung von 2018 im Vergleich zur Operationalisierung von 2015 an, dass man dem Ziel, die Testungen noch näher an das österreichische Curriculum zu rücken, zumindest nähergekommen ist. Auch hier könnten Untersuchungen von Daten der gesamten Pilotierungsstudie, aber auch Daten älterer Erhebungswellen weitere Einblicke gewähren, wobei man bei Letzteren vor der Herausforderung steht, Effekte von Eigenschaften des Instruments von Effekten in möglichen Veränderungen des Unterrichts zu separieren.

Alles in allem konnten Evidenzen für und gegen die Annahme von Konstruktäquivalenz gefunden werden, die aber eine valide Fortschreibung und Interpretation des Trends nicht übermäßig in Gefahr bringen. Die Evidenz für eine höhere curriculare Validität bestätigt die Bestrebungen zur Weiterentwicklung des Zuhören-Konstrukts. Außerdem ist die damit zu vermutende größere Unterrichtsnähe eine Entwicklung, die gut zur größeren Einbindung von Lehrpersonen in die zukünftigen Erhebungen passt. Dies stimmt positiv für den zukünftigen Einsatz des Konstrukts im Rahmen der iKM^{PLUS}.

Literatur

Ames, A. J. & Penfield, R. D. (2015). An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models. *Educational Measurement: Issues and Practice, 34* (3), 39–48. doi:10.1111/emip.12067

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67* (1), 6–18. doi:10.18637/jss.v067.i01

Bennett, R. E. (1991). On the Meaning of Constructed Response. *ETS Research Report Series, 1991* (2), i–46. doi:10.1002/j.2333-8504.1991.tb01429.x

Best, R. M., Floyd, R. G. & McNamara, D. S. (2008). Differential Competencies Contributing to Children's Comprehension of Narrative and Expository Texts. *Reading Psychology, 29* (2), 137–164. doi:10.1080/02702710801963951

Böhme, K., Richter, D., Weirich, S., Haag, N., Wendt, H., Bos, W. et al. (2014). Messen wir dasselbe? Zur Vergleichbarkeit des IQB-Ländervergleichs 2011 mit den internationalen Studien IGLU und TIMSS 2011. *Unterrichtswissenschaft, 42* (4), 342–365.

Braeken, J. (2011). A Boundary Mixture Approach to Violations of Conditional Independence. *Psychometrika, 76* (1), 57–76. doi:10.1007/s11336-010-9190-4

Breit, S., Bruneforth, M. & Schreiner, C. (2016). *Standardüberprüfung 2015, Deutsch/Lesen/Schreiben, 4. Schulstufe – Bundesergebnisbericht.* Salzburg: BIFIE. Verfügbar unter: https://www.iqs.gv.at/_Resources/Persistent/0d879bba1b6fb81469e09bc99aab7e4916f88bef/BiSt_UE_D4_2015_Bundesergebnisbericht.pdf

Brennan, R. L. (2001). Some Problems, Pitfalls, and Paradoxes in Educational Measurement. *Educational Measurement: Issues and Practice, 20* (4), 6–18. doi:10.1111/j.1745-3992.2001.tb00071.x

Bruneforth, M. & Höller, I. (2019). Die Kompetenzen der österreichischen Schüler/innen im Zeitvergleich. In B. Suchan, I. Höller & C. Wallner-Paschon (Hrsg.), *PISA 2018: Grundkompetenzen am Ende der Pflichtschulzeit im internationalen Vergleich* (S. 81–93). Graz: Leykam.

Dorans, N. J., Moses, T. P. & Eignor, D. R. (2010). Principles and Practices of Test Score Equating. *ETS Research Report Series, 2010* (2), i–41. doi:10.1002/j.2333-8504.2010.tb02236.x

Ehmke, T., van den Ham, A.-K., Sälzer, C., Heine, J. & Prenzel, M. (2020). Measuring Mathematics Competence in International and National Large Scale Assessments: Linking PISA and the National Educational Panel Study in Germany. *Studies in Educational Evaluation, 65* (100847), 1–10. doi:10.1016/j.stueduc.2020.100847

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: L. Erlbaum Associates.

Ferrara, S., Huynh, H. & Michaels, H. (1999). Contextual Explanations of Local Dependence in Item Clusters in a Large Scale Hands-On Science Performance Assessment. *Journal of Educational Measurement, 36* (2), 119–140. doi:10.1111/j.1745-3984.1999.tb00550.x

Frey, A., Hartig, J. & Rupp, A. A. (2009). An NCME Instructional Module on Booklet Designs in Large-Scale Assessments of Student Achievement: Theory and Practice. *Educational Measurement: Issues and Practice, 28* (3), 39–53. doi:10.1111/j.1745-3992.2009.00154.x

Ganzeboom, H. B. G., De Graaf, P. M. & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research, 21* (1), 1–56. doi:10.1016/0049-089x(92)90017-b

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and Validating Test Items.* New York, NY: Routledge. doi:10.4324/9780203850381

Hartig, J. & Frey, A. (2012). Validität des Tests zur Überprüfung des Erreichens der Bildungsstandards in Mathematik. *Diagnostica, 58* (1), 3–14. doi:10.1026/0012-1924/a000064

Henning, G. (1990). A Study of the Effects of Variation of Short-Term Memory Load, Reading Response Length and Processing Hierarchy on TOEFL Listening Comprehension Item Performance. *ETS Research Report Series, 1990* (2), i–62. doi:10.1002/j.2333-8504.1990.tb01354.x

Holland, P. W., Dorans, N. J. & Petersen, N. S. (2007). Equating Test Scores. In C. R. Rao & S. Sinharay (Hrsg.), *Handbook of Statistics: Psychometrics* (Bd. 26, S. 169–203). Amsterdam: Elsevier. doi:10.1016/S0169-7161(06)26006-1

In'nami, Y. & Koizumi, R. (2009). A Meta-Analysis of Test Format Effects on Reading and Listening Test Performance: Focus on Multiple-Choice and Open-Ended Formats. *Language Testing, 26* (2), 219–244. doi:10.1177/0265532208101006

Jude, N., Hartig, J., Schipolowski, S., Böhme, K. & Stanat, P. (2013). Definition und Messung von Lesekompetenz – PISA und die Bildungsstandards. In N. Jude & E. Klieme (Hrsg.), *PISA 2009 – Impulse für die Schul- und Unterrichtsforschung* (Bd. 59, S. 220–228). Weinheim: Beltz.

Koch, P. & Österreicher, W. (1985). Sprache der Nähe – Sprache der Distanz. *Romanistisches Jahrbuch, 36* (85), 15–43.

Kolenikov, S. (2010). Resampling Variance Estimation for Complex Survey Data. *The Stata Journal, 10* (2), 165–199. doi:10.1177/1536867X1001000201

Lohbeck, A., Hagenauer, G. & Moschner, B. (2016). Zum Zusammenspiel zwischen schulischem Selbstkonzept, Lernfreude, Konzentration und Schulleistungen im Grundschulalter. *Zeitschrift für Bildungsforschung, 6* (1), 53–69. doi:10.1007/s35834-016-0147-2

Martinez, M. E. (1999). Cognition and the Question of Test Item Format. *Educational Psychologist, 34* (4), 207–218. doi:10.1207/s15326985ep3404_2

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika, 47* (2), 149–174. doi:10.1007/BF02296272

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement: Interdisciplinary Research and Perspectives, 11* (3), 71–101. doi:10.1080/15366367.2013.831680

Nachtigall, C. & Wirtz, M. A. (2004). *Wahrscheinlichkeitsrechnung und Inferenzstatistik.* Weinheim: Juventa.

Nunally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory.* New York, NY: McGraw-Hill.

Pekrun, R. (2011). Emotion, Motivation, Selbstregulation: Gemeinsame Prinzipien und offene Fragen. In T. Götz (Hrsg.), *Emotion, Motivation und selbstreguliertes Lernen* (S. 185–205). Paderborn: Schöningh.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing.* Wien: R Foundation for Statistical Computing. Verfügbar unter: https://www.R-project.org/

Rauch, D. P. & Hartig, J. (2010). Multiple-Choice Versus Open-Ended Response Formats of Reading Test Items: A Two-Dimensional IRT Analysis. *Psychological Test and Assessment Modeling, 52* (4), 354–379.

Robitzsch, A. (2015). *Essays zu methodischen Herausforderungen im Large-Scale Assessment.* Dissertation, Humboldt-Universität, Berlin.

Robitzsch, A., Kiefer, T. & Wu, M. (2021). *TAM: Test Analysis Modules. R package version 3.6–45.*

Robitzsch, A. & Lüdtke, O. (2015). Kommentar zum Beitrag „Lokale Abhängigkeit von Items im TestDaF-Leseverstehen" von Thomas Eckes. *Diagnostica, 61* (2), 107–109. doi:10.1026/0012-1924/a000133

Rodriguez, M. C. (2002). Choosing an Item Format. In G. Tindal & T. M. Haladyna (Hrsg.), *Large-scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation* (S. 213–232). Mahwah, NJ: Lawrence Erlbaum Associates.

Rodriguez, M. C. (2003). Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement, 40* (2), 163–184. doi:10.1111/j.1745-3984.2003.tb01102.x

Schreiner, C., Gniewosz, B., Wiesner, C., Steiger, A., Kulmhofer-Bommer, A. & Egger, M. (2019). Einstellung der Schüler/innen zum Fach und zum Lernen: Freude am Fach, fachbezogenes Selbstkonzept und ihr Zusammenhang mit den fachlichen Leistungen. In A. C. George, C. Schreiner, C. Wiesner, M. Pointinger & K. Pacher (Hrsg.), *5 Jahre flächendeckende Bildungsstandardüberprüfungen in Österreich: Vertiefende Analysen zum Zyklus 2012 bis 2016* (S. 139–160). Münster: Waxmann.

Schult, J. & Sparfeldt, J. R. (2016). Reliability and Validity of PIRLS and TIMSS. *European Journal of Psychological Assessment, 34* (4), 258–269. doi:10.1027/1015-5759/a000338

Solheim, O. J. (2011). The Impact of Reading Self-Efficacy and Task Value on Reading Comprehension Scores in Different Item Formats. *Reading Psychology, 32* (1), 1–27. doi:10.1080/02702710903256601

Stedman, L. C. (2008). *The NAEP Long-Term Trend Assessment: A Review of Its Transformation, Use, and Findings.* Teaching, Learning and Educational Leadership Faculty Scholarship. Verfügbar unter: https://www.nagb.gov/content/dam/nagb/en/documents/who-we-are/20-anniversary/stedman-long-term-formatted.pdf

Traub, R. E. (1993). On the Equivalence of the Traits Assessed by Multiple-Choice and Constructed-Response Tests. In R. E. Bennett & W. C. Ward (Hrsg.), *Construction Versus Choice in Cognitive Measurement – Issues in Constructed Response, Performance Testing, and Portfolio Assessment* (S. 29–44). Hillsdale, NJ: Lawrence Erlbaum Associates.

Trendtel, M., Pham, G. & Yanagida, T. (2016). Skalierung und Linking. In S. Breit & C. Schreiner (Hrsg.), *Large-Scale Assessments mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfungen* (S. 185–224). Wien: Facultas.

van den Ham, A.-K., Ehmke, T., Nissen, A. & Roppelt, A. (2017). Assessments verbinden, Interpretationen erweitern? Lassen sich die mathematischen Kompetenzskalen im Nationalen Bildungspanel und im IQB-Ländervergleich 2012 verbinden? *Zeitschrift für Erziehungswissenschaft, 20* (1), 89–111. doi:10.1007/s11618-016-0686-2

van den Ham, A.-K., Nissen, A., Ehmke, T., Sälzer, C. & Roppelt, A. (2014). Mathematische Kompetenz in PISA, IQB-Ländervergleich und NEPS. Drei Studien, gleiches Konstrukt. *Unterrichtswissenschaft, 42* (4), 321–341.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet Response Theory and its Applications.* Cambridge: Cambridge University Press.

Wainer, H. & Thissen, D. (1996). How Is Reliability Related to the Quality of Test Scores? What Is the Effect of Local Dependence on Reliability? *Educational Measurement: Issues and Practice, 15* (1), 22–29. doi:10.1111/j.1745-3992.1996.tb00803.x

Wang, W.-C. & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement, 29* (2), 126–149. doi:10.1177/0146621604271053

Weirich, S., Bachinger, A., Trendtel, M. & Krelle, M. (2019). Listening Comprehension Tests in Germany and Austria – Research Report and Critical Review. *L1 Educational Studies in Language and Literature, 19,* 1–26. doi:10.17239/l1esll-2019.19.03.05

Wright, B. & Panchapakesan, N. (1969). A Procedure for Sample-Free Item Analysis. *Educational and Psychological Measurement, 29* (1), 23–48. doi:10.1177/001316446902900102

Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8* (2), 125–145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling Performance Assessments: Strategies for Managing Local Item Dependence. *Journal of Educational Measurement, 30* (3), 187–213. doi:10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A. L., Hambleton, R. K. & Sireci, S. G. (2001). *Effects of Local Item Dependence on the Validity of IRT Item, Test, and Ability Statistics.* Association of American Medical Colleges (AAMC).

Autorinnen und Autoren

Mag. Antonia Bachinger ist am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS) im Referat Fachdidaktik – Deutsch und an der Universität Wien tätig. Ihre Forschungsschwerpunkte sind: Diagnose von Lesekompetenz und Zuhörkompetenz, Leseunterricht und Förderung von rezeptiven Kompetenzen.

Maria Engelbert-Kocher, M.A. ist Fachkoordinatorin in den Bereichen Primarstufe Deutsch und Primarstufe Mathematik am Institut zur Qualitätsentwicklung im Bildungswesen an der Humboldt-Universität zu Berlin.

Dr. Roman Freunberger ist Leiter des Referats Psychometrie und Methoden am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS). Seine Forschungsschwerpunkte sind: Kompetenzdiagnostik, Standard Setting, Beurteilerübereinstimmung, faire Vergleiche in der Rückmeldung, nationale Systemberichterstattung, Kleinschulen in Österreich.

Dr. Stefan Gugerell, BA ist Senior Researcher am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS).

Prof. Dr. Michael Krelle ist Professor für Fachdidaktik Deutsch an der Technischen Universität Chemnitz. Seine Forschungsschwerpunkte sind: Sprachstandsdiagnose, Leistungsbeurteilung, Kompetenzentwicklung und Literalität im Fach Deutsch, Qualitäten von Deutschunterricht.

Mag. Dr. Michael Ober, BSc ist am IQS, Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen, im Referat Psychometrie und Methoden tätig.

Dipl.-Päd. Mag. Gloria Schneeweiß, BA MA, Diplomstudium für das Lehramt an Volksschulen an der Pädagogischen Akademie Salzburg, Diplomstudium Anglistik und Amerikanistik, Bachelorstudium Pädagogik sowie Masterstudium Erziehungswissenschaften an der Paris-Lodron-Universität Salzburg. Wissenschaftliche Mitarbeiterin am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS).

Dr. phil. Dipl.-Math. oec. Matthias Trendtel ist wissenschaftlicher Mitarbeiter am Institut für Schulentwicklungsforschung an der Technischen Universität Dortmund. Seine Forschungsschwerpunkte sind: Item-Response-Modellierung, Bayesianische Schätzverfahren, Diagnostik und Methodik in national und international vergleichenden Schulleistungsstudien.

Gabriele von Eichhorn, MSc BSc BA arbeitet als Psychometrikerin am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS). Ihre Arbeitsschwerpunkte sind: Testkonzeption und -entwicklung im Large-Scale Assessment, Erfassung von Lernfortschritt, kriteriale Rückmeldung und Stichprobenziehung.

Manuel Wick, uPM BA war bis 2020 am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS) in den Referaten Datenmanagement und Logistik sowie Forschungs- und Projektservices tätig.

Benedikt O. J. Winter, MSc arbeitet als Psychometriker am Institut des Bundes für Qualitätssicherung im österreichischen Schulwesen (IQS). Seine Arbeitsschwerpunkte sind: Testdesign im Large-Scale Assessment, Skalierung und Verlinkung von Leistungsdaten via Item-Response Theory, Messgenauigkeit und Fairness (DIF) von Tests, Stichprobenziehung sowie die Mitwirkung bei Item-Entwicklung und -Review.

Für Fragen zum Berichtsband *Zuhörkompetenzen messen* steht Ihnen Mag. Antonia Bachinger unter folgender E-Mail-Adresse zur Verfügung: antonia.bachinger@iqs.gv.at