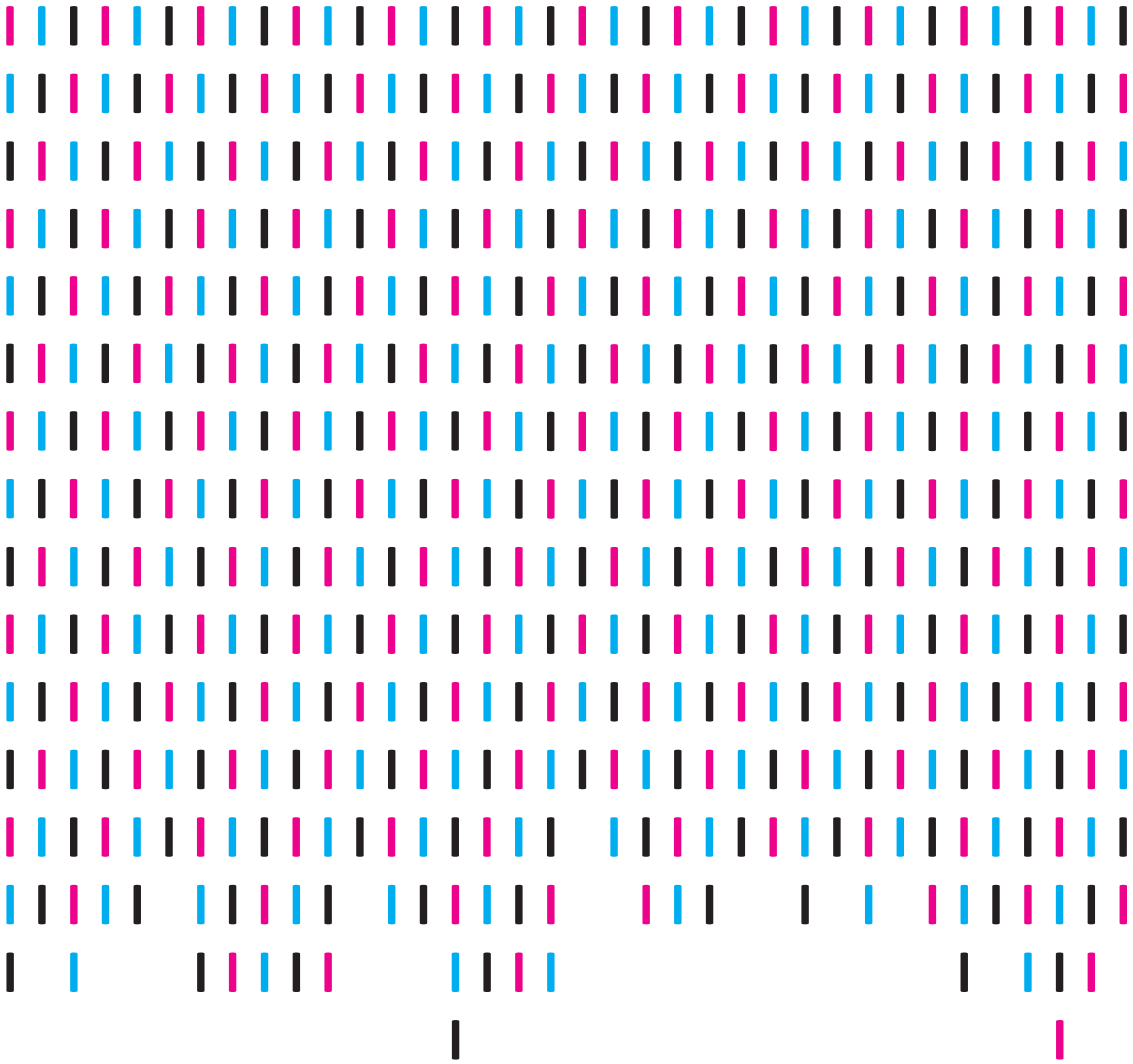

LIFBI *WORKING PAPERS*

Maja Stegenwallner-Schütz, Michael Obry, Elena Wittmann, Karin Gehrer, Lena Nusser & Katrin Böhme

INSIDE-STUDIE. DOKUMENTATION DER SKALIERUNG DER KOMPETENZ-MESSUNGEN IN DEN BEREICHEN LESEN UND MATHEMATIK DES ERSTEN MESSZEITPUNKTS IN DER JAHRGANGSSTUFE 6 (KOHORTE 1)

LIfBi *Working Paper* No. 108
Bamberg, Dezember 2022

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE

Working Papers of the Leibniz Institute for Educational Trajectories (LIfBi)
at the University of Bamberg

The LIfBi *Working Papers* series publishes articles, expert reports, and findings relating to studies and data collected by the Leibniz Institute for Educational Trajectories (LIfBi). They mainly consist of descriptions, analyses, and reports summarizing results from LIfBi projects, including the NEPS, as well as documentation of data sets other than NEPS, which are provided by the Research Data Center LIfBi.

LIfBi *Working Papers* are edited by LIfBi. The series started in 2011 under the name "NEPS *Working Papers*" and was renamed in 2017 to broaden the range of studies which may be published here.

Papers appear in this series as work in progress and may also appear elsewhere. They often present preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the LIfBi management or the NEPS Consortium.

The LIfBi *Working Papers* are available at www.lifbi.de/publications as well as at www.neps-data.de (see section "Publications").

Published by LIfBi

Contact:
Leibniz Institute for Educational Trajectories
Wilhelmsplatz 3
96047 Bamberg
Germany
contact@lifbi.de

INSIDE-Studie

Dokumentation der Skalierung der Kompetenzmessungen in den Bereichen Lesen und Mathematik des ersten Messzeitpunkts in der Jahrgangsstufe 6 (Kohorte 1)

Maja Stegenwallner-Schütz[1], Michael Obry[2], Elena Wittmann[2], Karin Gehrer[2], Lena Nusser[2] & Katrin Böhme[1]

[1]Universität Potsdam
[2]Leibniz-Institut für Bildungsverläufe

E-Mail-Adresse der Erstautorin:

stegenwa@uni-potsdam.de

Bibliographische Angabe:

Stegenwallner-Schütz, M., Obry, M., Wittmann, E., Gehrer, K., Nusser, L. & Böhme, K. (2022). *INSIDE-Studie. Dokumentation der Skalierung der Kompetenzmessungen in den Bereichen Lesen und Mathematik des ersten Messzeitpunkts in der Jahrgangsstufe 6 (Kohorte 1)* (LIfBi Working Paper No. 108). Leibniz-Institut für Bildungsverläufe. https://doi.org/10.5157/LIfBi:WP108:1.0

Acknowledgement:

Die beschriebene Studie wurde aus Mitteln des Bundesministeriums für Bildung und Forschung gefördert (Förderkennzeichen: IN1503D).

INSIDE-Studie. Dokumentation der Skalierung der Kompetenzmessungen in den Bereichen Lesen und Mathematik des ersten Messzeitpunkts in der Jahrgangsstufe 6 (Kohorte 1)

Zusammenfassung

Die Studie Inklusion in der Sekundarstufe in Deutschland (INSIDE) untersucht die Umsetzung und die Gelingensbedingungen von Inklusion an allgemeinen Schulen in Deutschland, an denen Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf gemeinsam mit Schülerinnen und Schülern ohne sonderpädagogische Förderbedarfe unterrichtet werden. Es wurden Kompetenztests für die Bereiche Leseverständnis und mathematische Kompetenzen eingesetzt, um den individuellen schulischen Kompetenzerwerb beschreiben zu können. Dazu wurden Testaufgaben der *National Educational Panel Study* (NEPS) verwendet, die teilweise adaptiert wurden und eine schwierigkeitsgestufte Zuweisung der Testversionen entwickelt. Das vorliegende Paper beschreibt das Studiendesign, die Testdurchführung sowie die Datenaufbereitung und –auswertung für die probabilistische Kompetenzmessung. An dem Lese- und dem Mathematiktest haben Schülerinnen und Schüler der Jahrgangsstufe 6 in der ersten Kohorte teilgenommen. Die Kompetenzdaten wurden mittels Partial-Credit-Modellen skaliert. Fehlende Werte, Itemfitstatistiken sowie Reliabilitätsmaße dienen der Überprüfung der Qualität der Testergebnisse. In den schwereren Testheftversionen konnten oftmals nicht alle Leseaufgaben bearbeitet werden, in leichteren Lesetestversionen und im Mathematiktest sehr wohl. Die Itemfitstatistiken und die Reliabilitätsmaße fallen insgesamt zufriedenstellend aus. Die Analysen zeigen, dass die Testpassung besser für Schülerinnen und Schüler in unteren und mittleren Leistungsbereichen als für die leistungsstärksten Schülerinnen und Schüler ausfällt. Das eingesetzte Testdesign eignet sich insgesamt zur Erfassung eines sehr heterogenen Fähigkeitsspektrums mit den erwünschten psychometrischen Eigenschaften.

Schlagworte

Kompetenzmessung, Lesekompetenz, mathematische Kompetenz, schulische Inklusion, Item-Response-Theorie

INSIDE study. Documentation of the Scaling of the Competence Tests in the Domains of Reading and Mathematics of the First Wave in Grade 6 (First Cohort)

Abstract

The INSIDE study (acronym for the German translation of inclusive education in secondary schools in Germany) investigates the implementation and critical success factors of inclusive education at regular schools in Germany, that are being attended by both, students with and without special educational needs. Reading comprehension and mathematical competence were assessed through competence tests, in order to investigate students' development in

these domains. Tests contained slightly adapted stimuli from the *National Educational Panel Study* (NEPS). Crucially, we developed a test design that takes into account students' different ability levels on a wide competence spectrum. This paper describes the study design, testing procedures, as well as data processing, and data analysis based on item response theory. Students of grade 6 participated in the reading test and the mathematical test in cohort 1. We used partial credit models for scaling the competence data. In order to evaluate the quality of the competence data, we analyzed missing values, item fits and test reliability. There was an increased number of items that were not reached toward the end of the hardest reading test versions due to time restrictions, unlike the easy and moderate reading test versions, or the mathematical test. Item fits and test reliability were satisfactory. In sum, test quality was better for students with low to moderate levels of competence in both the reading as well as the mathematics domains relative to students with the highest levels of competence. The tests yield the desired psychometric properties and therefore, prove suitable for the assessment of a heterogeneous ability spectrum.

Keywords

competence tests, reading competence, mathematical competence, inclusive education, Item Response Theory

Inhalt

1. Einleitung

Zielsetzung der INSIDE-Studie ist eine umfassende Darstellung der Umsetzung und der Gelingensbedingungen von Inklusion an allgemeinen Schulen in Deutschland, an denen Schülerinnen und Schüler mit sonderpädagogischem Förderbedarf gemeinsam mit Schülerinnen und Schülern ohne sonderpädagogische Förderbedarfe unterrichtet werden (Schmitt et al., 2020). Um die individuelle Entwicklung des schulischen Kompetenzerwerbs der Schülerinnen und Schüler der Sekundarstufe I nachvollziehen zu können, wurden sowohl die Lesekompetenz als auch die mathematische Kompetenz der Schülerinnen und Schüler erfasst. Im Folgenden beschreiben wir die Durchführung der Kompetenztestungen, sowie die Aufbereitung und Auswertung der Daten zur Lese- und mathematischen Kompetenz aus der Erhebung des ersten Messzeitpunkts in Jahrgangsstufe 6 (Kohorte 1). Sowohl für die Messung der Lesekompetenz als auch der mathematischen Kompetenz wurden Testaufgaben der *National Educational Panel Study* (NEPS) eingesetzt (vgl. Gehrer et al., 2013, Neumann et al., 2013), die teilweise für die Kompetenzmessungen von Schülerinnen und Schülern mit sonderpädagogischem Förderbedarf adaptiert wurden.

2. Methodisches Vorgehen zur Kompetenzmessung

2.1 Studiendesign

2.1.1 Gruppenzuweisung der Teilnehmenden

In der INSIDE-Stichprobe, also in Lerngruppen, in denen Schülerinnen und Schüler mit und ohne sonderpädagogische Förderbedarfe inklusiv gemeinsam lernen, ist eine hohe Leistungsheterogenität zu erwarten. Daher war für die Erfassung der Kompetenzen ein Testdesign erforderlich, das eine genaue Messung in einem sehr breiten Fähigkeitsspektrum ermöglicht. Innerhalb der beschränkten Testzeit, in diesem Fall von 28 Minuten, wie auch im NEPS üblich, können die teilnehmenden Schülerinnen und Schüler jedoch nur eine begrenzte Menge an Aufgaben bearbeiten. Um aus einer begrenzten Anzahl an Aufgaben die Kompetenzen der Teilnehmenden möglichst genau schätzen zu können, sollte das Schwierigkeitsniveau der Aufgaben möglichst gut dem Niveau der jeweiligen individuellen Kompetenzen entsprechen. Aus diesem Grund wurden eher leichte und eher schwierige Testheftversionen erstellt und die teilnehmenden Schülerinnen und Schüler Gruppen zugeordnet, deren erwartete Fähigkeitsbereiche den Schwierigkeiten der verschiedenen Testheftversionen entsprechen. Die Gruppenzugehörigkeit bildete somit die Basis für die Testheftzuweisung. Der Algorithmus, nach dem die Gruppenzuweisung erfolgte, ist im Folgenden dargestellt (siehe Tabelle 1). Die Kriterien wurden exhaustiv und seriell angewendet und wurden der Schülerdemografieliste[1] entnommen. Dies bedeutet, dass eine Schülerin oder ein Schüler, die oder der in eine Gruppe mit niedriger Ordnungszahl (z.B.

[1] Die Schülerdemografieliste ist eine Tabelle, in der Angaben zu Hintergrundvariablen für alle teilnehmenden Schülerinnen und Schüler, wie beispielsweise sonderpädagogische Förderbedarfe, Diagnosen von Teilleistungsstörungen oder Noten von Lehrkräften, gemacht werden können. Die Angaben werden von der studienkoordinierenden Lehrkraft verantwortet, die dafür schulinterne Informationen und Auskünfte weiterer Lehrkräfte heranzieht.

Gruppe 1) eingeordnet wurde, nicht zusätzlich in eine weitere Gruppe mit höherer Ordnungszahl (z.B. Gruppe 4) eingruppiert werden konnte.

Tabelle 1: Kriterien der Gruppenzuweisung

Gruppe	Zuweisungskriterien

Gruppe 1	• Festgestellter sonderpädagogischer Förderbedarf in den Bereichen Lernen, Geistige Entwicklung oder in der Sammelkategorie Lernen-Sprache-Sozial-emotionale Entwicklung (LSE)
	• Wenn Informationen zu o.g. Kriterium nicht vorhanden:
	  • Erhält sonderpädagogische Förderung
	  • Erhält zieldifferenten Unterricht
	• Diagnose einer Teilleistungsschwäche im schriftsprachlichen Bereich (isolierte oder kombinierte Lese- bzw. Rechtschreibschwäche) und im Rechnen (Rechenschwäche)

Gruppe 2	• Erhält Sprach- bzw. Leseförderung • Diagnose einer Teilleistungsschwäche im schriftsprachlichen Bereich (isolierte oder kombinierte Lese- bzw. Rechtschreibschwäche) • Festgestellter sonderpädagogischer Förderbedarf im Bereich Sprache • Notenkombination mit Deutschnote 4, 5 oder 6 und Mathematiknote 1, 2 oder 3
Gruppe 3	• Diagnose einer Teilleistungsschwäche im Bereich Rechnen • Notenkombination: Mathematiknote 4, 5 oder 6 und Deutsch 1, 2, 3
Gruppe 4	• Letzte Halbjahresnote 4, 5 oder 6 im Fach Deutsch und letzte Halbjahresnote 4, 5 oder 6 im Fach Mathematik • Kompetenzeinschätzung im unteren Drittel in den Fächern Deutsch und Mathematik bei zielgleichem Unterricht
Gruppe 5	• letzte Halbjahresnote 1-2 im Fach Deutsch und letzte Halbjahresnote 1-2 im Fach Mathematik • Kompetenzbereich im oberen Drittel in den Fächern Deutsch und Mathematik bei zielgleichem Unterricht • sofern an Schule vorhanden: Förderung für Leistungsstarke
Gruppe 6:	• Letzte Halbjahresnote 3 im Fach Deutsch und letzte Halbjahresnote 3 im Fach Mathematik • sämtliche anderen Notenkombinationen, die vorher nicht aufgeführt wurden • Kompetenzbereich im mittleren Drittel in den Fächern Deutsch und Mathematik

2.1.2 Testung der Lesekompetenz

Für die Ermittlung der Lesekompetenz der Schülerinnen und Schüler sollten diese relevante Informationen in Texten identifizieren und extrahieren, textbezogene Schlussfolgerungen herleiten sowie relevante Informationen bewerten und reflektieren. Bei der Zusammenstellung des Lesetests wurde auf eine ausgewogene Mischung verschiedener Textsorten geachtet, die Relevanz für das alltägliche Leben besitzen (siehe Literacy-Konzept des NEPS, Gehrer et al., 2013). Die eingesetzten Texte entstammten den folgenden fünf Textsorten: 1) literarische Texte, 2) Sachtexte (Texte mit Informationen zu Sachthemen), 3)

Werbe- und Anzeigentexte, 4) Anleitungstexte sowie 5) Texte mit kommentierender Funktion (vgl. Gehrer & Artelt, 2013).

Jeweils ein Lesetext und die dazugehörigen Aufgaben bilden gemeinsam eine Aufgabeneinheit, eine sogenannte *Unit.* Die Aufgaben zu den Texten können als Verständnisfragen charakterisiert werden, die das Verständnis des jeweiligen Lesetexts voraussetzen und nicht auf Vorwissen abzielen. Innerhalb einer Aufgabeneinheit waren stets mehrere Aufgaben zu einem Text zu beantworten. Einzelne Aufgaben zu Texten werden nachfolgend als *Items* bezeichnet. Die Antwortreaktionen der teilnehmenden Schülerinnen und Schüler auf diese Items wurden später kodiert und bildeten einzelne Variablen in dem resultierenden Datensatz. Für die Lesekompetenzmessung wurden insgesamt vier verschiedene Testhefte mit gestufter Schwierigkeit erstellt (vgl. Testhefte 1A-D in Tabelle 2).

Tabelle 2: Anzahl der Texte nach Textsorten im Lesetest in den Testheften für Jahrgangsstufe 6

Textsorten	**Testheft 1A**	**Testheft 1B**	**Testheft 1C**	**Testheft 1D**	**administrierte Texte**
Literarischer Text	1	1	1	1	3
Sachtext	1	1	1	1	2
Werbe- und Anzeigentext	1	1	1	1	3
Anleitungstext	1	1	1	1	2
Text mit kommentierender Funktion	-	-	1	1	1
Gesamter Lesetest	4	4	5	5	11

Die Testheftversionen enthielten jeweils nur eine Auswahl von Units und zugehörigen Items. Um dennoch eine psychometrische Verbindung zwischen allen Testheften und damit auch den Kompetenzen aller Gruppen von Schülerinnen und Schülern herstellen zu können, wurde eine Schnittmenge an identischen Units und Items in jeweils mindestens zwei Testheften eingesetzt. Hierbei handelt es sich um sogenannte Ankeritems. Sie verknüpfen die Testheftversionen miteinander, so dass die Schwierigkeiten aller Items auf einer gemeinsamen Skala geschätzt werden können. Dadurch entsteht eine querschnittliche Verlinkung (siehe Tabelle 3).

Tabelle 3: Anzahl der Items in den verschiedenen Aufgabenformaten im Lesetest für Jahrgangsstufe 6

Aufgabenformate	Testheft 1A	Testheft 1B	Testheft 1C	Testheft 1D	Anker-items	unique Items
Einfache Multiple-Choice-Aufgabe	17	19	20	24	30	50
Komplexe Multiple-Choice-Aufgabe	3	1	8	7	8	11
Zuordnungs-aufgabe	-	2	3	1	1	5
Gesamt	20	22	31	32	39	66

Anmerkung. Psychometrische Informationen zu den Ankeritems können der Tabelle 12 entnommen werden.

Die Testhefte wurden anhand vorliegender Informationen zur mittleren Schwierigkeit und zur Schwierigkeitsverteilung der Items einer Unit und den Messeigenschaften der Items aus früheren Studien des NEPS an Regel- und Förderschulen in den Klassenstufen 5, 7 und 9 zusammengestellt. Im Lesetest enthielten Testheft 1A und 1B Aufgaben, deren Schwierigkeit sich in den NEPS-Entwicklungsstudien oder NEPS-Machbarkeitsstudien an Förderschulen als eher leicht für die intendierte Jahrgangsstufe gezeigt hatten. Testheft 1A und 1B wurden mittels zehn Items aus zwei Units aus dem leichten Aufgabenspektrum verlinkt. Testheft 1C enthielt sowohl leichtere als auch schwierigere Aufgaben. Dieses Testheft 1C wurde über elf eher leichte Items aus zwei Units mit dem Testheft 1A verlinkt, um sowohl Schülerinnen und Schülern aus dem erwarteten, höheren Fähigkeitsspektrum als auch dem geringeren Fähigkeitsspektrum die Bearbeitung derselben Aufgaben zu ermöglichen. Über weitere 18 eher schwierige Items aus drei Units wurden Testheft 1C und Testheft 1D verlinkt. Letzteres enthielt insgesamt eher schwierige Aufgaben (siehe Abbildung 1).

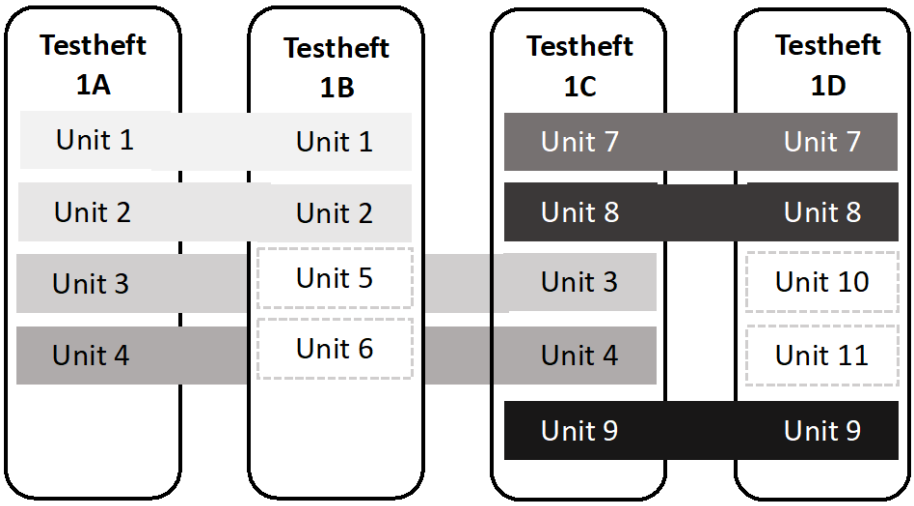

Abbildung 1. Schematische Darstellung des querschnittlichen Linkingdesigns.

In den schwierigkeitsgestuften Testheften wurden von den Schülerinnen und Schülern unterschiedlich viele Units bearbeitet. Für die Schülerinnen und Schüler mit sonderpädagogischen Förderbedarfen, z.B. in dem Bereich *Lernen,* wurden im leichten Testheft 1A, wie in den Machbarkeitsstudien des NEPS an Förderschulen, lediglich vier Units eingesetzt (Nusser et al., 2020). Um die Anforderungen der eher leichten Testhefte möglichst zu parallelisieren, war die Unitanzahl in den Testheften 1A und 1B gleich. Die eher schwierigen Testheften 1C und 1D enthielten fünf Units, wie sonst im NEPS üblich (Gehrer et al., 2013).

Items wurden in unterschiedlichen Aufgabenformaten vorgelegt (siehe Tabelle 3): einfache Multiple-Choice-Aufgaben (Engl. *Multiple Choice*, MC), komplexe Multiple-Choice-Aufgaben (Engl. *Complex Multiple Choice*, CMC) und Zuordnungsaufgaben (Engl. *Matching*, MA). Multiple-Choice-Aufgaben bestanden aus einer Frage zu einem Text, zu der die teilnehmenden Schülerinnen und Schüler aus einer Auswahlmenge von vier Möglichkeiten die korrekte Zielantwort auswählen sollten. Die verbleibenden Antwortmöglichkeiten stellten sogenannte Distraktoren, also inhaltlich falsche oder weniger gut passende Antwortoptionen dar. Komplexe-Multiple-Choice-Aufgaben enthielten mehrere Entscheidungsaufgaben, bei denen zwei Antwortoptionen (Zustimmung vs. Ablehnung) angeboten wurden, von denen eine korrekt war. Bei Aufgaben dieses Typs sollten die Schülerinnen und Schüler Aussagen zum Text dahingehend bewerten, ob diese in Bezug auf den Text zutreffend oder unzutreffend waren. In einer Zuordnungsaufgabe sollten Aussagen Textabschnitten zugeordnet und in eine Reihenfolge gebracht werden, die dem Text entsprach. In den eingesetzten Zuordnungsaufgaben gab es stets eine bis zwei überzählige Aussagen. Die Aussagen der Zuordnungsaufgaben bestanden in der Regel aus Teilüberschriften. Beispiele für die verschiedenen Aufgabenformate finden sich in Gehrer et al. (2012).

2.1.3 Testung der mathematischen Kompetenz

Entsprechend der Operationalisierung mathematischer Kompetenzen im NEPS sollten im Mathematiktest realitätsnahe Problemstellungen erkannt und bearbeitet werden. Die einbezogenen kognitiven Komponenten umfassten den Einsatz technischer Fertigkeiten, das Modellieren, Argumentieren, Kommunizieren, Repräsentieren und Problemlösen. Jede Mathematikaufgabe enthielt eine Situationsbeschreibung, auf die sich ein oder maximal zwei Items bezogen. Die im Test enthaltenen Items deckten jeweils einen der folgenden mathematischen Aufgabenbereiche ab (vgl. Tabelle 4): Quantität (Engl. *Quantity*), Raum und Form (Engl. *Space and Shape*), Veränderung und Beziehungen (Engl. *Change and Relationships*) sowie Daten und Zufall (Engl. *Data and Chance*). Die Rahmenkonzeption der Mathematiktestitems inklusive der einbezogenen kognitiven Kompetenzen wird in Neumann et al. (2013) ausführlich beschrieben. Die Testhefte wurden wie bei der Lesekompetenzmessung anhand vorliegender Informationen zur Schwierigkeit und Messeigenschaften der Items mit Aufgaben aus früheren Studien des NEPS an Regel- und Förderschulen in den Jahrgangsstufen 5, 7 und 9 zusammengestellt. Für die Messung der mathematischen Kompetenz wurden zwei schwierigkeitsgestufte Testhefte erstellt (vgl. Testhefte 2A und 2B in Tabelle 4). Testheftversion 2A enthielt eher leichte Aufgaben und Testheftversion 2B enthielt eher schwere Aufgaben. Im Mathematiktest wurden sieben Ankeritems eingesetzt, um die beiden Testheftversionen miteinander zu verlinken.

Tabelle 4: Anzahl der Items in den verschiedenen Aufgabenbereichen im Mathematiktest für Jahrgangsstufe 6

Aufgabenbereiche	**Testheft 2A**	**Testheft 2B**	**Anker-items**	**unique Items**
Quantität	7	6	1	12
Raum und Form	4	8	3	9
Veränderung und Beziehung	5	5	3	7
Daten und Zufall	3	1	-	4
Gesamter Mathematiktest	19	20	7	32

Auch hier wurden wie im Lesetest einfache Multiple-Choice-Aufgaben (Engl. *Multiple Choice*, MC) und komplexe Multiple-Choice-Aufgaben (Engl. *Complex Multiple Choice*, CMC) eingesetzt (siehe Tabelle 5). Im Unterschied zum Lesetest enthielt der Mathematiktest auch kurze Freitextaufgaben (Engl. *Short Constructed Responses,* SCR). In einer Multiple-Choice-Aufgabe sollte die teilnehmende Schülerin oder der teilnehmende Schüler aus einer Auswahlmenge von vier bis fünf Antwortmöglichkeiten immer eine Zielantwort wählen. Komplexe-Multiple-Choice-Aufgaben enthielten Unteraufgaben, die aus jeweils zwei Antwortalternativen bestanden. Von den beiden Antwortalternativen war stets eine korrekt. In einer Freitextaufgabe musste die Lösung schriftlich in ein Lösungskästchen eingetragen werden.

Tabelle 5: Anzahl der Items in den verschiedenen Aufgabenformaten im Mathematiktest für Jahrgangsstufe 6

Aufgabenformate	**Testheft 2A**	**Testheft 2B**	**Anker-items**	**unique Items**
Einfache Multiple-Choice-Aufgabe	13	18	7	24
Komplexe Multiple-Choice-Aufgabe	1	-	-	1
Freitextaufgabe	5	2	-	7
Gesamt	19	20	7	32

Anmerkung. Psychometrische Informationen zu den Ankeritems können der Tabelle 13 entnommen werden.

2.2 Testheftzuweisung

Die schwierigkeitsgestuften Testheftversionen wurden den teilnehmenden Schülerinnen und Schülern in Abhängigkeit von deren Zugehörigkeit zu den vorher definierten Gruppen (siehe Abschnitt 2.1.1) zugewiesen. Tabelle 6 stellt das Rotationsdesign für die Testheftzuweisung der INSIDE-Kompetenzmessungen dar. Die eingesetzten Testheftversionen für den Lese- und Mathematiktest wurden kombiniert und entsprechend schwierigkeitsgestuft den Gruppen vorgelegt. Bei fehlenden Angaben in der Schülerdemografieliste war keine Zuordnung zu einer Gruppe möglich, sodass eine zufällige Zuweisung zu einer der Testheftkombinationen erfolgte.

Tabelle 6: Rotationsdesign der INSIDE-Kompetenzmessung in Jahrgangsstufe 6

Testhefte 1A/2A	**Testhefte 1B/2A**	**Testhefte 1B/2B**	**Testhefte 1C/2A**	**Testhefte 1C/2B**	**Testhefte 1D/2B**
Gruppe 1	Gruppe 4	Gruppe 2	Gruppe 3	Gruppe 5	Gruppe 5
Gruppe 4	Gruppe 6 (randomisiert)	Gruppe 6 (randomisiert)	Gruppe 5	Gruppe 6 (randomisiert)	Gruppe 6 (randomisiert)
Gruppe 6 (randomisiert)			Gruppe 6 (randomisiert)		

2.3 Stichprobe

An der INSIDE-Testung haben zum ersten Messzeitpunkt in Jahrgangsstufe 6 (Kohorte 1) insgesamt 3899 Schülerinnen und Schüler teilgenommen. Davon haben 3625 Schülerinnen und Schüler ein Testheft des Lesetests und 3631 Schülerinnen und Schüler ein Testheft des Mathematiktests bearbeitet. Da eine Anzahl von weniger als drei gültigen Antworten einen zu geringen Informationsgehalt für eine zuverlässige Schätzung der Lese- oder mathematischen Kompetenz beinhaltet, wurden die Daten von Schülerinnen und Schülern, die weniger als drei gültige Antworten gaben, von der Datenauswertung ausgeschlossen. Eine Übersicht über die Anzahl der ausgeschlossenen Schülerinnen und Schüler ist in Tabelle 7 dargestellt. Der Ausschluss erfolgte getrennt für den Lese- und den Mathematiktest. Das Ausschlusskriterium betraf die Lesetestdaten von 12 Schülerinnen und Schülern und die Mathematiktestdaten von 8 Schülerinnen und Schülern.

Letztlich wurden die Daten von insgesamt 3613 Schülerinnen und Schülern für den Lesetest und die Daten von insgesamt 3623 Schülerinnen und Schülern für den Mathematiktest ausgewertet.

Tabelle 7: Anzahl der Schülerinnen und Schüler, deren Daten nicht in die Datenauswertung einfließen

Anzahl	Lesetest				
	Testheft 1A	**Testheft 1B**	**Testheft 1C**	**Testheft 1D**	**Gesamt**
ausschließlich ungültige Antworten	1	3	1	0	5
weniger als drei gültige Antworten	3	5	3	1	12
Anzahl	**Mathematiktest**				
	Testheft 2A		**Testheft 2B**		**Gesamt**
ausschließlich ungültige Antworten	3		1		4
weniger als drei gültige Antworten	3		5		8

Die Testheftversionen wurden den teilnehmenden Schülerinnen und Schülern nach Möglichkeit auf Basis ihrer Gruppenzugehörigkeit zugewiesen. Eine Darstellung der Kriterien für die Zuweisung befindet sich in Unterkapitel 2.1.1.

Tabelle 8: Übersicht zur Anzahl der Schülerinnen und Schüler pro Gruppe (auf Grundlage der Gruppenzugehörigkeitskriterien) und pro tatsächlich bearbeiteter Testheftversion des Lesetests

Gruppe	Testheft 1A	Testheft 1B	Testheft 1C	Testheft 1D	Gesamtanzahl pro Gruppe
1	**321**	24	17	3	321
2	14	**528**	35	15	528
3	14	13	**349**	11	349
4	**145**	**180**	10	9	325
5	25	33	**454**	**212**	666
6	**192**	**342**	**366**	**145**	1045
Ohne Gruppe	25	56	62	13	156
Korrekt (%)	93%	94%	95%	91%	94%
Gesamt	736	1176	1293	408	3390

Anmerkungen. Es werden alle Schülerinnen und Schüler berücksichtigt, deren Daten mit ausreichend vielen gültigen Fällen in die Lesetestauswertung einfließen. Die bearbeiteten Testheftversionen der Schülerinnen und Schüler ohne Gruppenzuordnung zählen als korrekt. In Fettdruck hervorgehoben sind die korrekt zugewiesenen Testheftversionen pro Gruppe.

Die Tabellen 8 und 9 enthalten für den Lese- und den Mathematiktest eine Übersicht, wie sich die Anzahl der Schülerinnen und Schüler der verschiedenen Gruppen auf die bearbeiteten Testheftversionen verteilt. Diese Tabellen beinhalten ferner eine Darstellung der Fehlzuweisungen. In diesen Fällen haben die Schülerinnen und Schüler nicht die beabsichtigte Testheftversion mit dem für sie vorgesehenen Schwierigkeitsgrad bearbeitet. Solche Fehlzuweisungen entstanden beispielsweise bei der Testadministration. Auch gab es zum Zeitpunkt der Testheftzuweisung Schülerinnen und Schüler mit fehlenden oder nicht aufbereiteten Angaben in der Schülerdemografieliste, so dass diesen Schülerinnen und Schülern eine Testheftversion zufällig zugewiesen wurde. Diese Schülerinnen und Schüler werden in den Tabellen 8 und 9 ohne Gruppenzuweisung aufgeführt. Für die Gesamtanzahl pro Gruppe, die in Tabelle 8 für den Lesetest und in Tabelle 9 für den Mathematik dargestellt ist, wurden jedoch nur diejenigen Schülerinnen und Schüler berücksichtigt, die die jeweiligen Kriterien für die Gruppenzugehörigkeit auf Basis der vorliegenden soziodemografischen Angaben aus der Schülerdemografieliste erfüllten und auch das beabsichtigte Testheft bearbeitet haben. Diese Gruppen eignen sich beispielsweise für spätere Gruppenvergleiche.

Tabelle 9: Übersicht zur Anzahl der Schülerinnen und Schüler pro Gruppe (auf Grundlage der Gruppenzugehörigkeitskriterien) und pro tatsächlich bearbeiteter Testheftversion des Mathematiktests

Gruppe	Testheft 2A	Testheft 2B	Gesamtanzahl pro Gruppe
1	**348**	21	348
2	65	**526**	526
3	**367**	23	367
4	**323**	22	323
5	**285**	**442**	727
6	**552**	**494**	1046
Ohne Gruppe	81	74	155
Korrekt (%)	97%	96%	96%
Gesamt	2021	1602	3492

Anmerkungen. Es werden alle Schülerinnen und Schüler berücksichtigt, deren Daten mit ausreichend vielen gültigen Fällen in die Mathematiktestauswertung einfließen. Die bearbeiteten Testheftversionen der Schülerinnen und Schüler ohne Gruppenzuordnung zählen als korrekt. In Fettdruck hervorgehoben sind die korrekt zugewiesenen Testheftversionen pro Gruppe.

2.4 Durchführung

Die Kompetenzmessungen erfolgten durch geschulte Testleiterinnen und Testleiter zu einem Testtermin an den Schulen der teilnehmenden Schülerinnen und Schüler zwischen Mai und Juli 2019. Zum jeweiligen Testtermin wurden die folgenden Tests nacheinander administriert: ein Lesegeschwindigkeitstest (Zimmermann et al., 2012), ein Lesetest, ein Mathematiktest und im Anschluss ein kognitiver Fähigkeitstest (BEFKI - Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. bis 7. Jahrgangsstufe; Schroeders et al., 2020). Nach einer Pause erfolgte die Bearbeitung eines Schülerfragebogens. Die folgende Tabelle 10 stellt den genauen Testablauf einschließlich der Test- und Pausenzeiten dar.

Tabelle 10: Übersicht zum Testablauf mit Zeitangaben

Testablauf	**Zeitangaben**
Einweisung in die Bearbeitung des Lesegeschwindigkeitstests	ca. 8 min
Bearbeitung des Lesegeschwindigkeitstests	2 min
Einweisung in die Bearbeitung der Lesetests	ca. 8 min
Bearbeitung des Lesetests	28 min
Pause	10 min
Einweisung in die Bearbeitung des Mathematiktests	ca. 5 min
Bearbeitung des Mathematiktests	28 min
Einweisung in die Bearbeitung des BEFKI	ca. 5 min
Bearbeitung des BEFKI	14 min
Pause	15 min
Einweisung in die Bearbeitung des Schülerfragebogens	ca. 5 min
Bearbeitung des Schülerfragebogens	ca. 25 min
Reine Bearbeitungszeit	ca. 97 min
Gesamtdauer der Testsitzung	ca. 153 min

3. Datenaufbereitung und -auswertung

In einem ersten Datenaufbereitungsschritt wurden die Testhefte von der *International Association for the Evaluation of Educational Achievement* (IEA) eingescannt und entsprechend vorliegender Kodieranweisungen maschinell kodiert. Die Kodieranweisungen wurden als sogenannte *Codebooks* im Zuge der Testheftzusammenstellung erstellt und enthalten systematische Kodes für sämtliche Antwortreaktionen. In einem weiteren Schritt wurden die kodierten Antwortreaktionen anhand der in den Kodieranweisungen hinterlegten Zielantworten als richtig oder falsch gewertet. Da für Freitextaufgaben im Mathematiktest keine Kodieranweisungen vorhanden waren, wurden sie von den Autorinnen und Autoren nachträglich ergänzt und die maschinell gescannt vorliegenden, offenen Angaben als korrekt oder inkorrekt bewertet.

Die Beschaffenheit der Daten war dabei abhängig vom Aufgabenformat. Einfache Multiple-Choice-Aufgaben und Freitextaufgaben erzeugten eine dichotome Datenstruktur, da gültige Antworten entweder korrekt oder inkorrekt ausfallen konnten. Komplexe Multiple-Choice-Aufgaben und Zuordnungsaufgaben hingegen erzeugten eine polytome Datenstruktur, da die Anzahl der korrekt gelösten Teilantworten gewertet wurde.

Im Falle der polytomen Items musste für eine verlässliche Auswertung sichergestellt werden, dass in jeder Teilantwortkategorie eine ausreichend hohe Anzahl an Datenpunkten vorlag. Zu diesem Zweck war es teilweise notwendig Teilantwortkategorien mit einer zu geringen Anzahl an Datenpunkten mit einer benachbarten Teilantwortkategorie zusammenzulegen. Die minimale Anzahl gültiger Fälle in einer Teilantwortkategorie betrug 50 Datenpunkte (bei beispielsweise 4 Teilantworten gibt es 4+1 Teilantwortkategorien: 4 korrekte Antworten, 3 korrekte Antworten, 2 korrekte Antworten, 1 korrekte Antwort oder 0 korrekte Antworten). Im Vergleich zu den Vorgaben des NEPS (Pohl & Carstensen, 2012) von 200 Datenpunkten wurde somit ein weniger strenges Kriterium für die vorliegende Datenauswertung gewählt. Diese Anpassung wurde durch den geringeren Stichprobenumfang der INSIDE-Studie notwendig und wurde anhand der gegebenen Interpretierbarkeit (z.B. ausreichend kleiner Standardfehler der Koeffizienten) des Outputs des Messmodells überprüft.

Die Auswertung der aufbereiteten Testdaten erfolgte anschließend auf Grundlage der Item-Response-Theorie. Diese gehört zu den probabilistischen Testtheorien. Dabei wurde für jede Aufgabe eine Lösungswahrscheinlichkeit geschätzt. Dieses Vorgehen setzt sich zusammen aus einer Schätzung der relativen Schwierigkeit einer jeden Aufgabe, i.e., die Schätzung des Itemparameters, als auch einer Schätzung der Fähigkeit bzw. Kompetenzen der Schülerinnen und Schüler, i.e., die Schätzung des Personenparameters. Diese Art der Datenauswertung wird üblicherweise als *Skalierung* bezeichnet, weil sie zum Ziel hat, die Schwierigkeit einer Aufgabe und die Personenfähigkeit, die mit dem Test erfasst werden soll, auf derselben Skala abzubilden. Die Skalierung der Kompetenztests orientierte sich an der Vorgehensweise des NEPS, die in Pohl und Carstensen (2012) beschrieben ist. Für die Parameterschätzung wurde das Paket TAM (Robitzsch et al., 2020) in R (R Core Team, 2020), Version 4.0.3, verwendet. Die Parameterschätzung erfolgte in einem zweistufigen Verfahren: Zunächst wurden die Itemparameter (sogenannte *Itemkalibrierung*) und in einem weiteren Schritt die Personenparameter, die die Fähigkeiten einzelner Schülerinnen und Schüler abbilden, geschätzt. Die Schätzung der Itemparameter basierte dabei auf den Testdaten der Schülerinnen und Schüler der INSIDE-Studie und schloss somit sowohl Schülerinnen und

Schüler ohne sonderpädagogischen Förderbedarf als auch mit sonderpädagogischem Förderbedarf und unterschiedlichen Förderschwerpunkten ein. Als Messmodell diente das eindimensionale und einparametrische Partial-Credit-Modell (Masters, 1982), das sich für die dichotome wie polytome Datenstruktur eignet. Dichotome Items gingen in die Schätzung einwertig ein. Die Itemparameter für dichotome Items sind sogenannte Schwierigkeitsparameter. In Anlehnung an das Vorgehen im NEPS (vgl. Pohl & Carstensen, 2012), wurden die Teilantworten polytomer Items halbwertig gewichtet. Die Itemparameter polytomer Items wurden als Lageparameter und Schwellenparameter modelliert. Da ein polytomes Item aus mehreren Teilantwortmöglichkeiten besteht, schätzt das Modell bei diesem Vorgehen einen allgemeinen Lageparameter für das gesamte Item und Schwellenparameter in Abhängigkeit von der Anzahl der Teilantwortmöglichkeiten (bei beispielsweise vier Teilantwortmöglichkeiten wurden ebenfalls vier Schwellenparameter modelliert, d.h. ein Parameter pro Übergang, beispielsweise von null korrekten Antworten zu einer korrekten Antwort etc. bis hin zum Übergang von drei zu vier korrekten Antworten). Diese Schwellenparameter geben in Relation zum Lageparameter eines polytomen Items Aufschluss über die Schnittstellen der Lösungshäufigkeitsverteilungen der benachbarten Teilantwortkategorien. An diesen Stellen übersteigt die Lösungswahrscheinlichkeit der nächsthöheren Kategorie 50% im Vergleich zur unteren Kategorie. Alle Parameter wurden auf einer Logitskala modelliert. Für die Itemparameterschätzung wurde das marginale Maximum-Likelihood-Schätzverfahren (Engl. *Marginal Maximum Likelihood*) verwendet. Die Personenparameter wurden mittels *Weighted Maximum Likelihood Estimates (WLEs;* Warm, 1989) geschätzt. WLEs sind Punktschätzer, die die Kompetenz auf Grundlage der beobachteten Daten einer Person im Messmodell repräsentieren.

3.1 Fehlende Werte

Antwortreaktionen wurden aus den folgenden Gründen als fehlende Werte klassifiziert: (a) Das Item wurde aufgrund der begrenzten Testzeit nicht erreicht (Engl. *Missing Not Reached*), (b) das Item wurde ausgelassen und nicht beantwortet (Engl. *Omitted Response oder Missing By Intention*), (c) die Antwort war ungültig (Engl. *Invalid Response*) oder (d) das Item wurde nicht vorgelegt (Engl. *Not Administered/Missing By Design*). Reaktionen auf polytome Items wurden komplett als fehlend gewertet, wenn bei mindestens einer Teilantwort ein fehlender Wert vorlag. Gemäß dem Vorgehen im NEPS (Pohl & Carstensen, 2012) wurden sämtliche Arten fehlender Werte als fehlend und nicht als falsche Antwort gewertet.

3.2 Itemselektion für die Skalierung

Vor der endgültigen Skalierung der Kompetenztestdaten, die später berichtet wird, wurden alle administrierten Items auf ihre psychometrischen Eigenschaften überprüft. Psychometrisch nicht geeignete Items wurden im Zuge der Itemselektion ausgeschlossen. Die Überprüfung eines Items erfolgte zunächst in der Gesamtschau aller Kompetenztestitems und anschließend separat für die jeweiligen Testheftversionen. Dieses Vorgehen ermöglichte die differenzierte Überprüfung der psychometrischen Eigenschaften der Items und stellte zusätzliche diagnostische Informationen bereit, wenn einzelne Items beispielweise ausschließlich in bestimmten Subgruppen der Studie auffällig waren. Bei der Itemselektion wurde vorzugsweise auf einen ausreichenden Informationsgehalt der Items geachtet. Items wurden in Anlehnung an das Vorgehen im NEPS als problematisch eingestuft, wenn

1) Die Lösungshäufigkeit dichotomer Items >= 0,975 betrug.
2) Der WMNSQ (Engl. *Weighted Mean Square Error*), der sogenannte Infit, eines Items > 1,20 betrug.
3) Der t-Wert des WMNSQ > |8| ausfiel.
4) Die Item-Skalen-Korrelation, als Maß der Trennschärfe eines Items, < 0,2 ausfiel.

Bei als auffällig eingestuften Items erfolgte zusätzlich eine visuelle Prüfung der itemcharakteristischen Funktionen, sogenannte *Item Characteristic Curves.* Dies war eine zusätzliche Überprüfung des grafischen Itemfits, indem anhand der sogenannten *Expected Scores Curves* die itemcharakteristischen Funktionen für die beobachteten und erwarteten Werte verglichen wurden. Im Idealfall liegen beide Funktionen in der grafischen Darstellung sehr nahe beieinander. Auch wurden für die Gesamttests für Lese- und mathematische Kompetenzen und die einzelnen Testheftversionen die Verteilungen der Itemschwierigkeiten und der Personenfähigkeiten in einer grafischen Darstellung, einer sogenannten *Wright-Map*, veranschaulicht und diese Informationen bei der Abwägung zum Für und Wider eines Ausschlusses einzelner Items berücksichtigt. Tabelle 11 enthält eine detaillierte Übersicht zu den auffälligen Items und der letztlich getroffenen Selektionsentscheidung.

Der Prozess der Itemselektion einschließlich der Entscheidungsfindung zum Ausschluss erfolgte separat für den Lese- und Mathematiktest. Nach Abwägungen der genannten Kriterien wurden drei Leseitems von der Gesamtskalierung und der Verlinkung ausgeschlossen. Zwei weitere auffällige Items blieben jedoch aus inhaltlichen Gründen für die testheftweise Skalierung erhalten. Ein Mathematikitem wurde von der Gesamtskalierung und der Verlinkung des Mathematiktests ausgeschlossen.

Tabelle 11: Übersicht zu auffälligen Items

Item	Eigenschaften	Entscheidung
Lesen		
Item REG60210_I_C (aus Testheft 1A und 1C)	• auffälliger Infit in Testheft 1A und Gruppe 1 (Testheft 1A) • auffällige Trennschärfe	• Ausschluss vom Gesamttest und der Längsschnittverlinkung • Erhalt in Testheft 1C
Item REG60350_I_C (aus Testheft 1A und 1C)	• auffällige Infitmaße (WMNSQ und teilweise auch t-Wert des WMNSQ) im gesamten Test, auf Testheft- und Gruppenebene	• Ausschluss vom Gesamttest und der Längsschnittverlinkung • Erhalt in Testheft 1C
Item REG60560_I_C (aus Testheft 1B)	• auffälliger Infit im gesamten Test und auf Gruppenebene • auffälliges Kurvenbild der itemcharakteristischen Funktion	• Ausschluss vom Gesamttest
Mathematik		
MAG6D121_I_C (aus Testheft 2A)	• extrem geringer Informationsgehalt in Gruppe 5, in Testheft 2A perfekt gelöst von Gruppe 5 • Kurvenbild der itemcharakteristischen Funktion und Wright-Map verweisen auf niedrigen Informationsgehalt, weil zu leicht	• Ausschluss vom Gesamttest
MAG6Q16_I_S_C (aus Testheft 2A)	• extreme Werte für Lageparameter in Gruppe 5/Testheft 2A • unauffällige Itemfitmaße und Trennschärfe auf Gesamttestebene und auf Testheftebene	• Keine Änderung
MAG6R251_I_C (aus Testheft 2B)	• geringer Informationsgehalt und geringe Trennschärfe in Gruppe 5 • unauffällige Itemfitmaße und Trennschärfe auf Gesamttestebene und auf Testheftebene • Kurvenbild der itemcharakteristischen Funktion und Wright-Map verweisen auf niedrigen Informationsgehalt, jedoch weniger extrem als bei MAG6D121_I_C	• Keine Änderung

4. Ergebnisse

4.1 Fehlende Werte

4.1.1 Fehlende Werte pro Schülerin oder Schüler

Nicht erreichte Items

Fehlende Werte, die auf nicht erreichte Items zurückgehen, entstehen, wenn Schülerinnen und Schüler den Test aufgrund der begrenzten Testzeit nicht bis zum Ende bearbeiten können. Der Anteil an Schülerinnen und Schülern, die Items nicht erreicht haben, nimmt also mit aufsteigender Itemposition zu. Für die einzelnen Testhefte des Lesetests sind in den Abbildungen 2 bis 5 die Anteile an Schülerinnen und Schülern für die Itemanzahl, die sie nicht erreicht haben, dargestellt. In den Lesetestheften 1A und 1B haben mindestens 99% bzw. 97% der Schülerinnen und Schüler die Hälfte der Items erreicht. Mindestens 78% der Schülerinnen und Schüler erreichten alle vorgegebenen Items in den Testheften 1A und 1B (siehe Abbildungen 2 und 3), d.h. der Großteil der Schülerinnen und Schüler erreichte das Testende. Dies ist ein Hinweis darauf, dass der Test einen adäquaten Umfang aufwies.

In den Lesetestheften 1C und 1D erreichten ungefähr 93% bzw. 91% der Schülerinnen und Schüler jeweils mindestens die Hälfte der vorgegebenen Items. Jeweils nur 35% bzw. 39% der Schülerinnen und Schüler erreichten alle Items in den Testheften 1C und 1D (siehe Abbildungen 4 und 5), was darauf hindeutet, dass der Test eher zu schwer bzw. eher zu lang war. In den Testheften 1C und 1D waren – wie sonst in Lesetests im Rahmen des NEPS üblich – jeweils fünf und nicht nur vier Units enthalten, wie in den Testheften 1A und 1B. Dadurch erklärt sich der höhere Anteil an nicht erreichten Items in den beiden Testheftversionen 1C und 1D. Zum Vergleich, in der NEPS-Erhebung in der Jahrgangsstufe 5 erreichten 48% der Schülerinnen und Schüler alle Leseitems (Pohl et al., 2012). In der NEPS-Erhebung in der Jahrgangsstufe 7, in der es eine leichte und eine schwere Testheftversion mit jeweils fünf Units gab, erreichten 81% der Schülerinnen und Schüler das Ende der leichten Testheftversion und 54% der Schülerinnen und Schüler das Ende der schweren Testheftversion (Krannich et al., 2017). Im Vergleich zu den NEPS-Erhebungen fällt der Anteil an nicht erreichten Items in den schweren Testheftversionen 1C und 1D somit ähnlich hoch aus wie in der schweren Testheftversion in Jahrgangsstufe 7. Die Anzahl an nicht erreichten Items wirkt sich nicht negativ auf die Kompetenzschätzung aus, da fehlende Werte ignoriert und nicht als falsch gewertet werden.

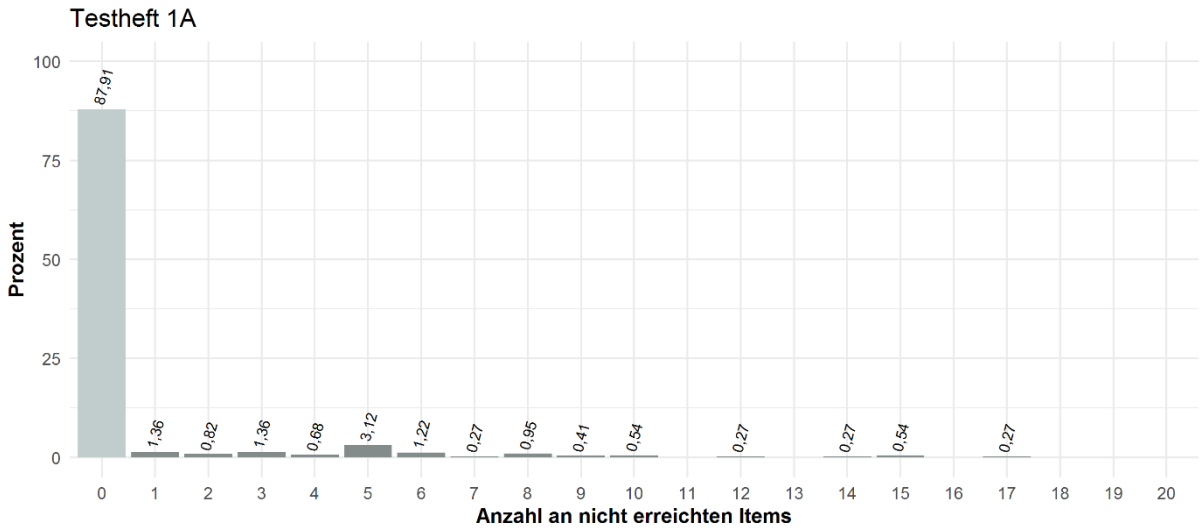

Abbildung 2. Anzahl an nicht erreichten Items für Testheft 1A (Lesekompetenz).

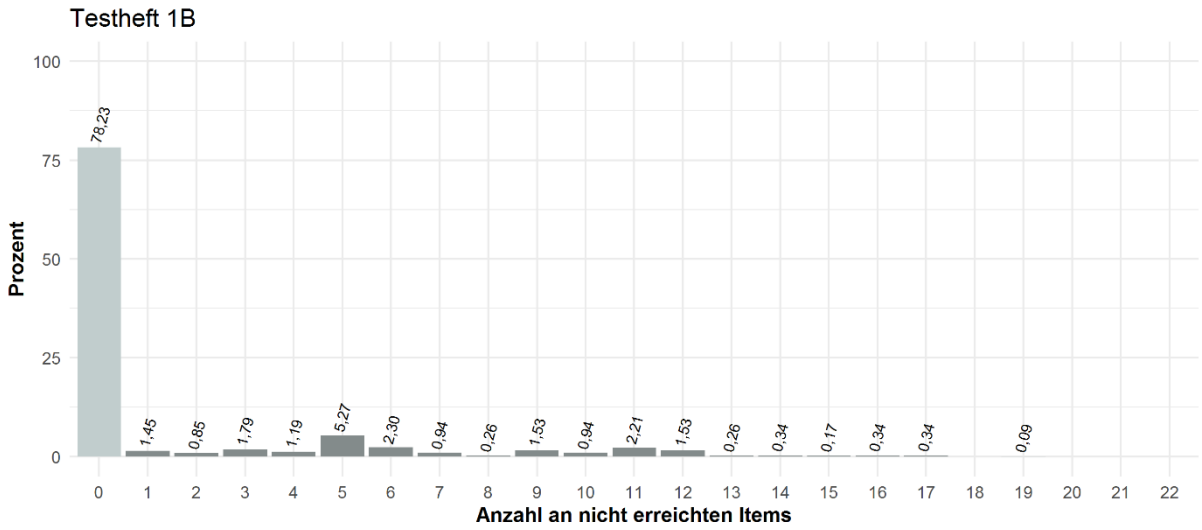

Abbildung 3. Anzahl an nicht erreichten Items für Testheft 1B (Lesekompetenz).

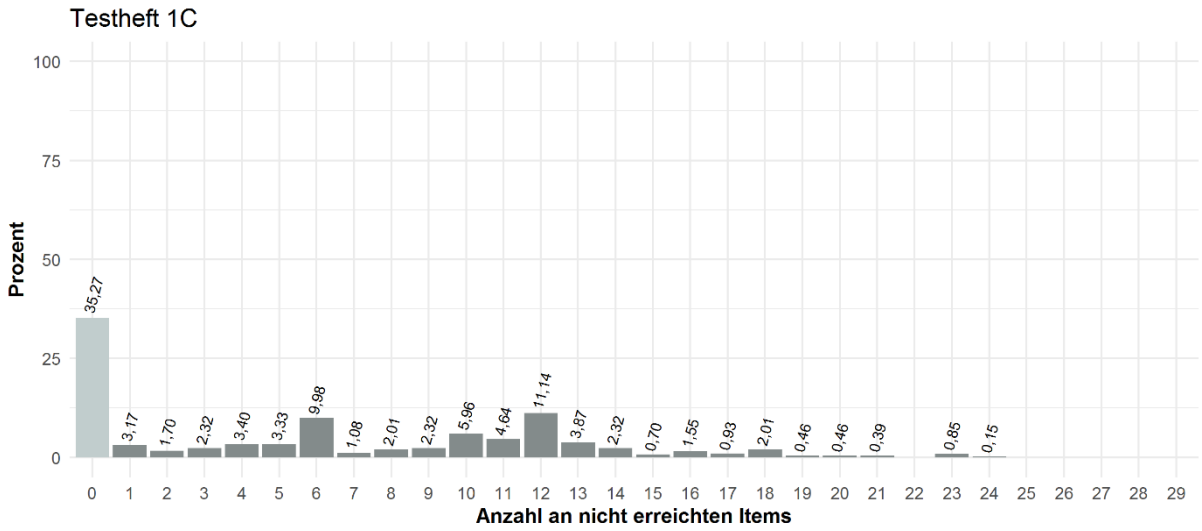

Abbildung 4. Anzahl an nicht erreichten Items für Testheft 1C (Lesekompetenz).

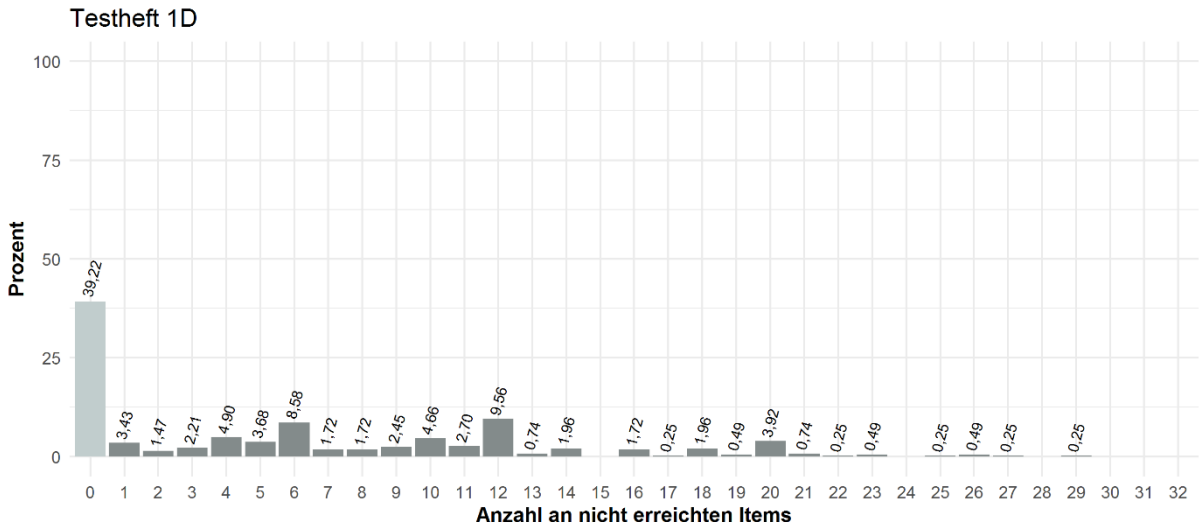

Abbildung 5. Anzahl an nicht erreichten Items für Testheft 1D (Lesekompetenz).

Für die einzelnen Testhefte des Mathematiktests sind in den Abbildungen 6 und 7 die Anteile an Schülerinnen und Schülern für die Anzahl an nicht erreichten Items dargestellt. In den Mathematiktestheften 2A und 2B erreichten fast alle Schülerinnen und Schüler mindestens zehn Items und damit mindestens die Hälfte der enthaltenen Items. In Testheft 2A (siehe Abbildung 6) erreichten circa 99% der Schülerinnen und Schüler das Testende. In Testheft 2B (siehe Abbildung 7) erreichten circa 92% der Schülerinnen und Schüler alle vorgegebenen Items. Beide Tests sind daher nicht zu umfangreich gewesen für die teilnehmenden Schülerinnen und Schüler.

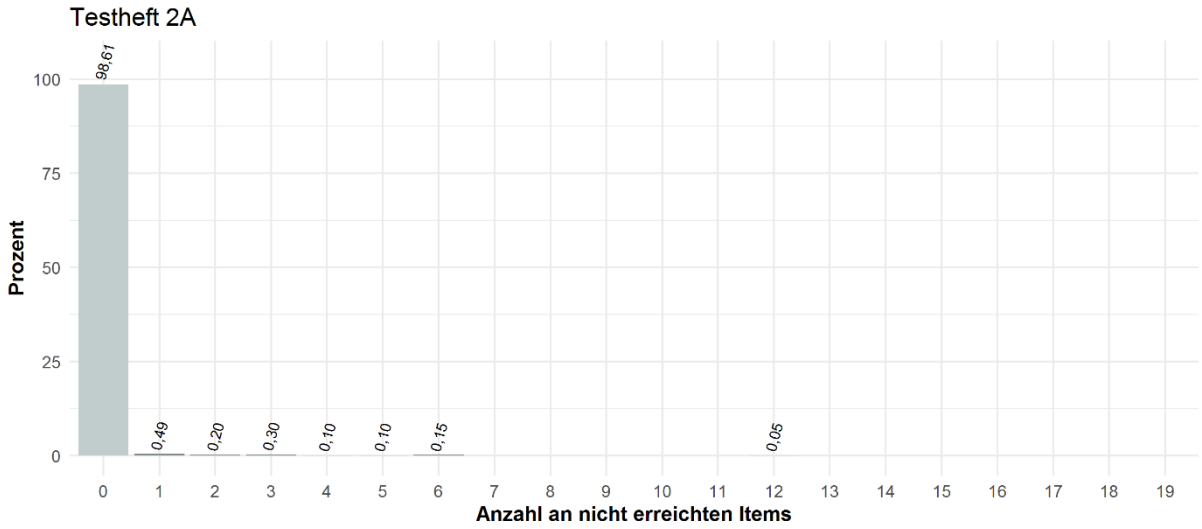

Abbildung 6. Anzahl an nicht erreichten Items für Testheft 2A (mathematische Kompetenz).

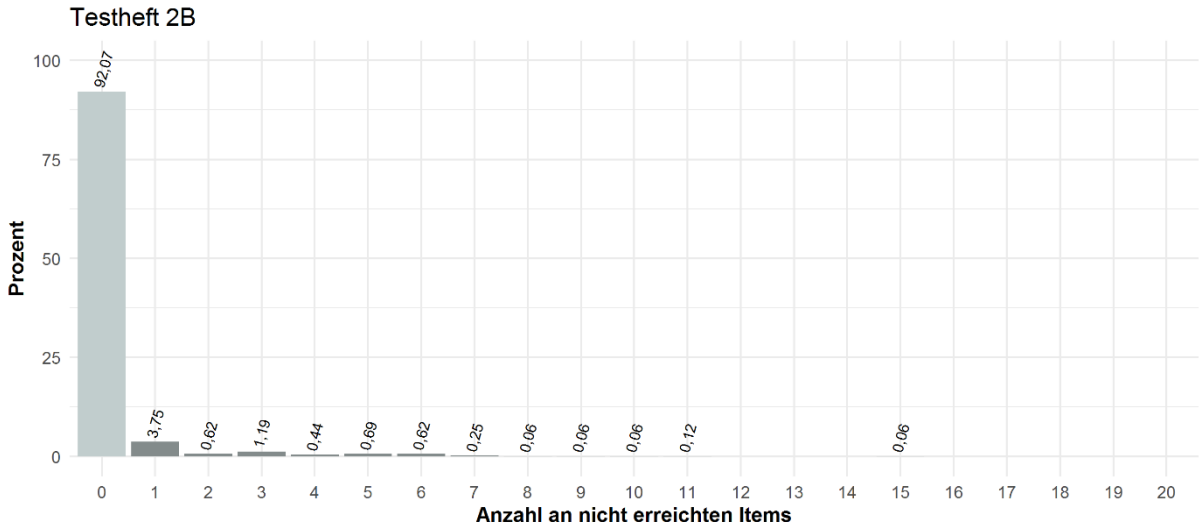

Abbildung 7. Anzahl an nicht erreichten Items für Testheft 2B (mathematische Kompetenz).

Ausgelassene Items

In Tests, die keine Antwortauswahl erzwingen, können Aufgaben oder Fragen übersprungen und dadurch ausgelassen werden. Dies kann eine Bearbeitungsstrategie darstellen, wenn eine Schülerin oder ein Schüler zuerst die als leicht empfundenen Aufgaben bzw. Items erledigen

möchte und als schwierig empfundene Aufgaben bzw. Items (zunächst) zurückstellt. Auch das Itemformat kann einen Einfluss haben, so dass beispielsweise komplex erscheinende Aufgabenformate, z.B. Zuordnungsaufgaben im Lesetest, (zunächst) ausgelassen werden. Für die einzelnen Testhefte des Lesetests ist in den Abbildungen 8 bis 11 die Anzahl an ausgelassenen Items pro Schülerin oder Schüler dargestellt. In den Testheften 1A und 1B wurden tendenziell seltener Items ausgelassen als in den Testheften 1C und 1D. In allen Lesetestheften übersprangen mindestens 62% der Schülerinnen und Schüler keine Items. Je nach Testheftversion haben zwischen circa 16% und 19% der Schülerinnen und Schüler ein Item ausgelassen, zwischen circa 4% und 10% der Schülerinnen und Schüler zwei Items und zwischen circa 5% und 10% der Schülerinnen und Schüler mehr als zwei Items. Insgesamt traten eher vereinzelt Auslassungen auf.

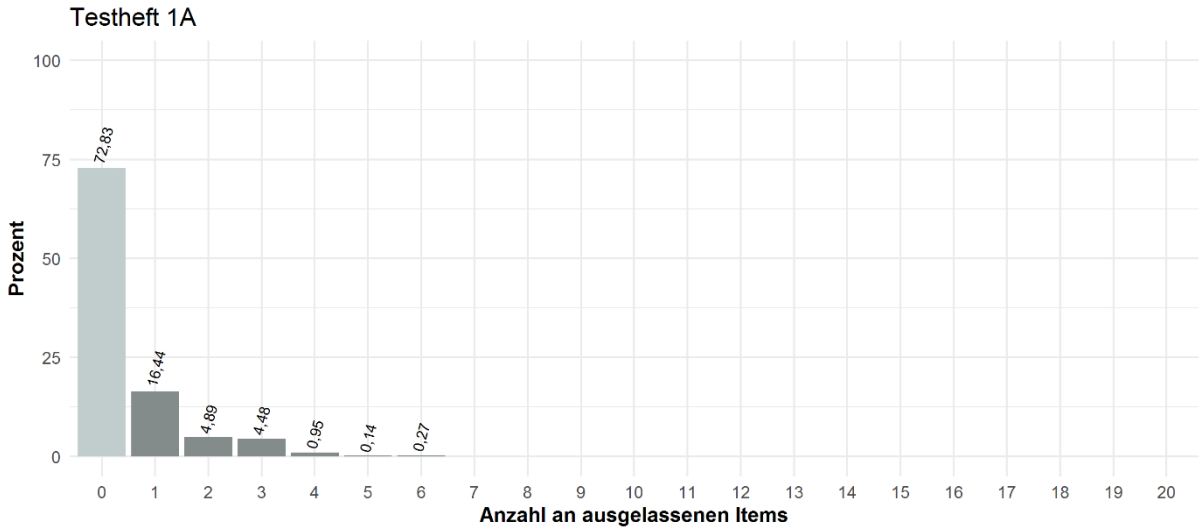

Abbildung 8. Anzahl an ausgelassenen Items für Testheft 1A (Lesekompetenz).

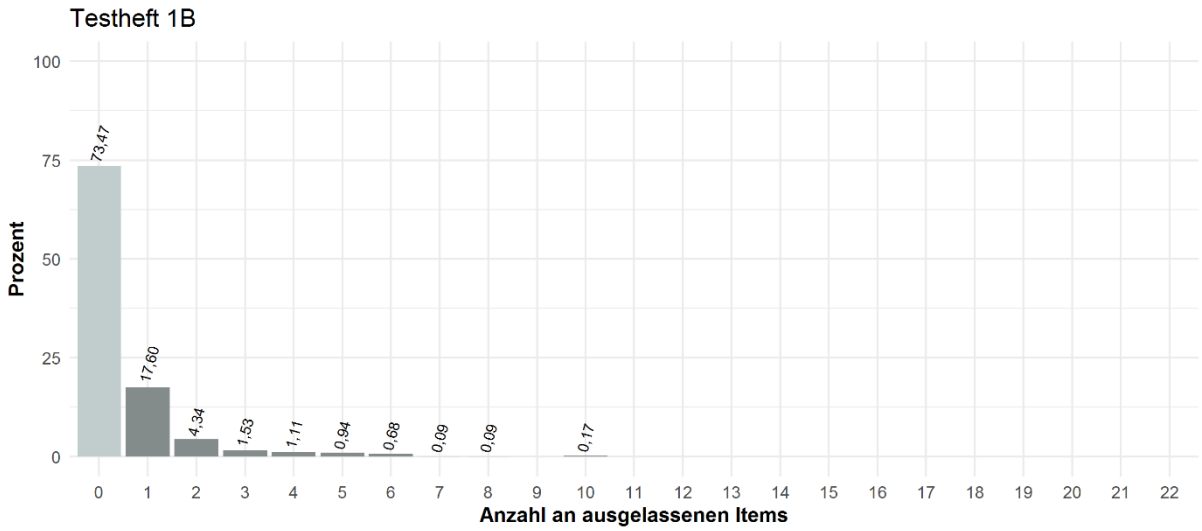

Abbildung 9. Anzahl an ausgelassenen Items für Testheft 1B (Lesekompetenz).

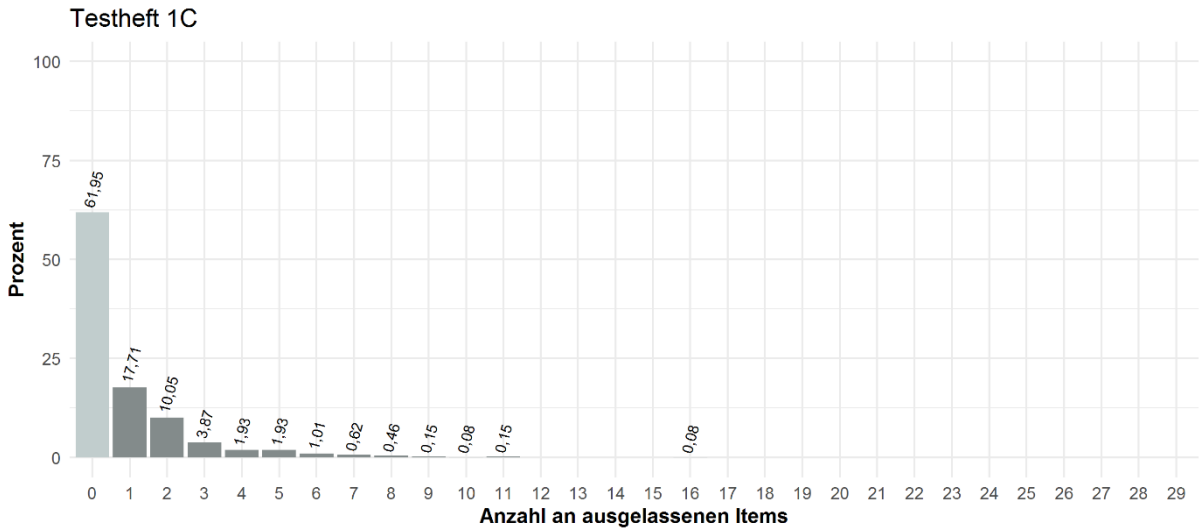

Abbildung 10. Anzahl an ausgelassenen Items für Testheft 1C (Lesekompetenz).

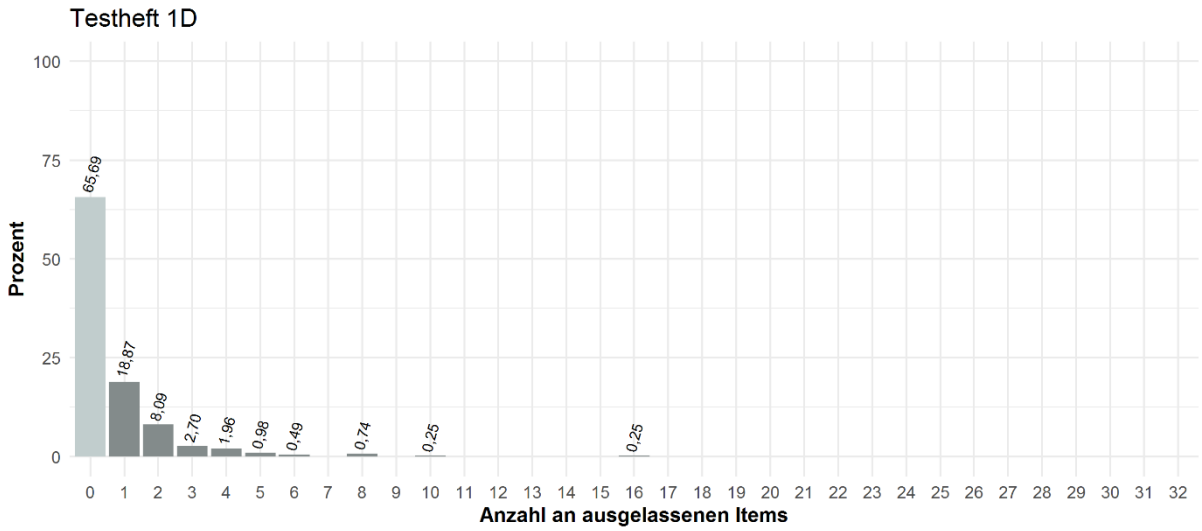

Abbildung 11. Anzahl an ausgelassenen Items für Testheft 1D (Lesekompetenz).

Für die beiden Testhefte zur Erfassung der mathematischen Kompetenz ist in den Abbildungen 12 und 13 die Anzahl an ausgelassenen Items pro Schülerin oder Schüler dargestellt. Die Verteilungen der ausgelassenen Items in den Testhefte 2A und 2B ähneln sich sehr. In beiden Mathematiktestheften haben ähnlich wie in den Lesetestheften mindestens 64% der Schülerinnen und Schüler kein Item ausgelassen. Je nach Testheftversion haben zwischen circa 15% und 16% der Schülerinnen und Schüler ein Item, zwischen circa 7% und 9% der Schülerinnen und Schüler zwei Items und zwischen circa 10% und 13% der Schülerinnen und Schüler mehr als zwei Items ausgelassen. Insgesamt traten auch in den Testheften des Mathematiktests eher vereinzelt Auslassungen auf.

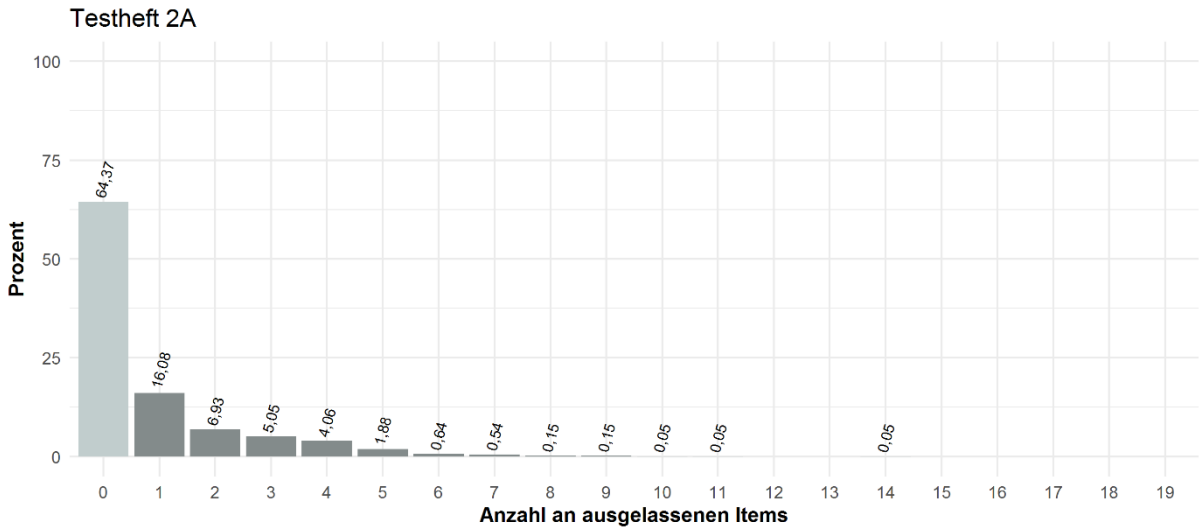

Abbildung 12. Anzahl an ausgelassenen Items für Testheft 2A (mathematische Kompetenz).

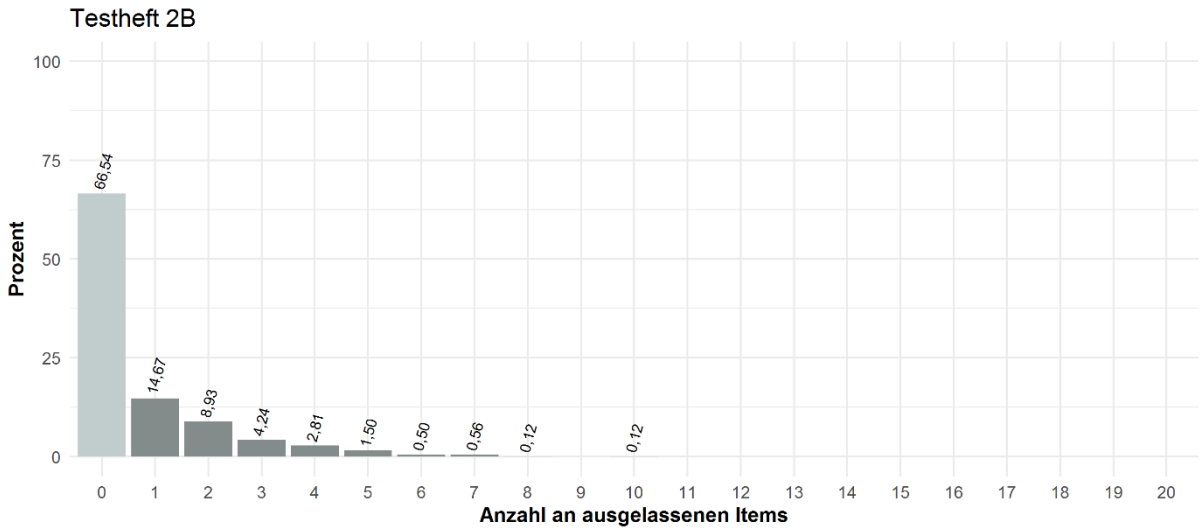

Abbildung 13. Anzahl an ausgelassenen Items für Testheft 2B (mathematische Kompetenz).

Ungültige Antworten

Neben nicht erreichten Items und Auslassungen gab es fehlende Werte, die auf ungültige Antworten zurückgingen. Eine Antwort war zum Beispiel ungültig (d.h., nicht valide), wenn mehrere Antwortoptionen angekreuzt wurden, obwohl die Auswahl nur einer Antwortoption gefordert war (z.B. bei einfachen und komplexen Multiple-Choice-Fragen). In den Abbildungen 14 bis 17 ist die Anzahl an ungültigen Antworten pro Schülerin oder Schüler für die einzelnen Testhefte des Lesetests dargestellt. In allen Lesetestheften gaben mindestens 85% der Schülerinnen und Schüler keine ungültigen Antworten. Am häufigsten kam es in Testheft 1B zu einer ungültigen Antwort und zwar bei knapp 9% der Schülerinnen und Schüler. Mehr als eine ungültige Angabe wurde am häufigsten in Testheft 1D von circa 8% der Schülerinnen und Schüler erzeugt. Insgesamt kam diese Art fehlender Werte eher selten und wenn, dann nur vereinzelt vor.

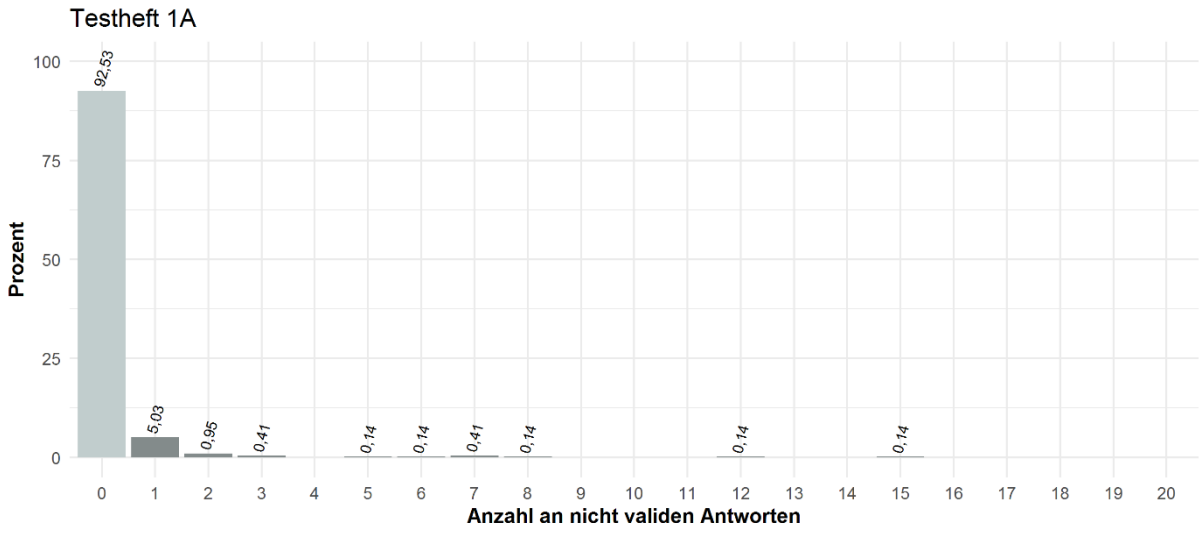

Abbildung 14. Anzahl an ungültigen Antworten für Testheft 1A (Lesekompetenz).

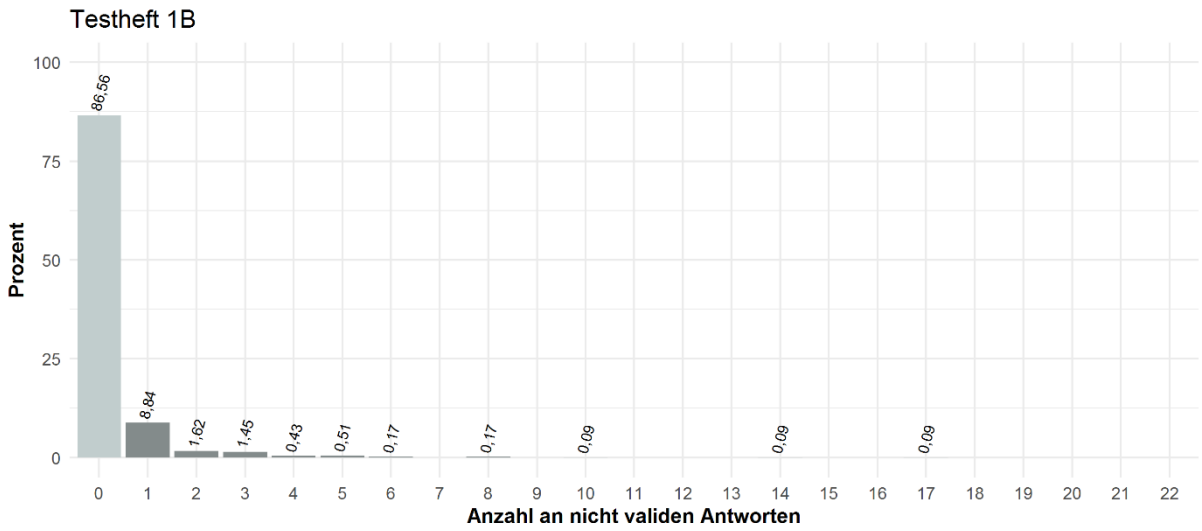

Abbildung 15. Anzahl an ungültigen Antworten für Testheft 1B (Lesekompetenz).

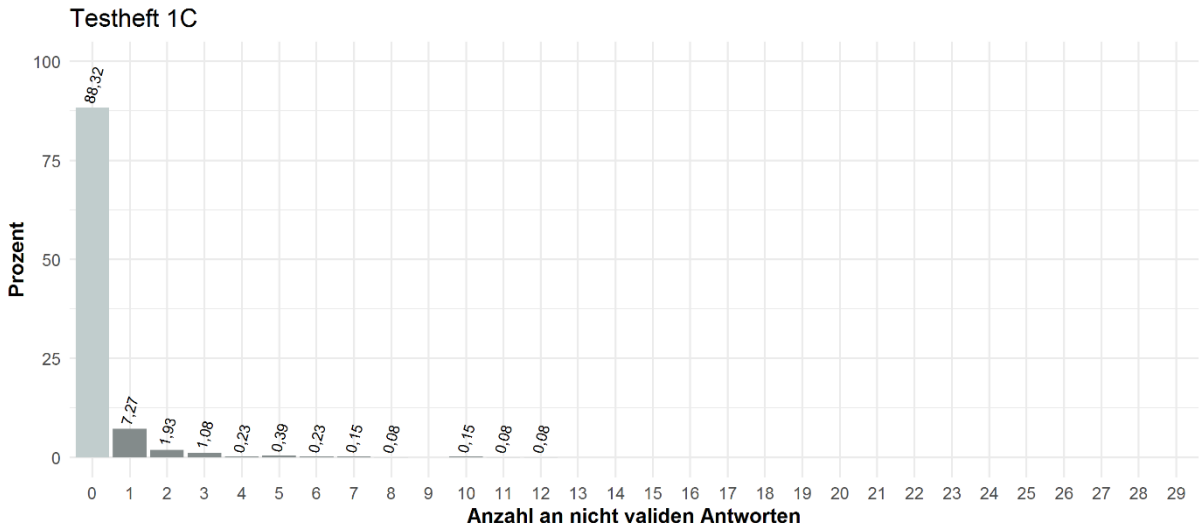

Abbildung 16. Anzahl an ungültigen Antworten für Testheft 1C (Lesekompetenz).

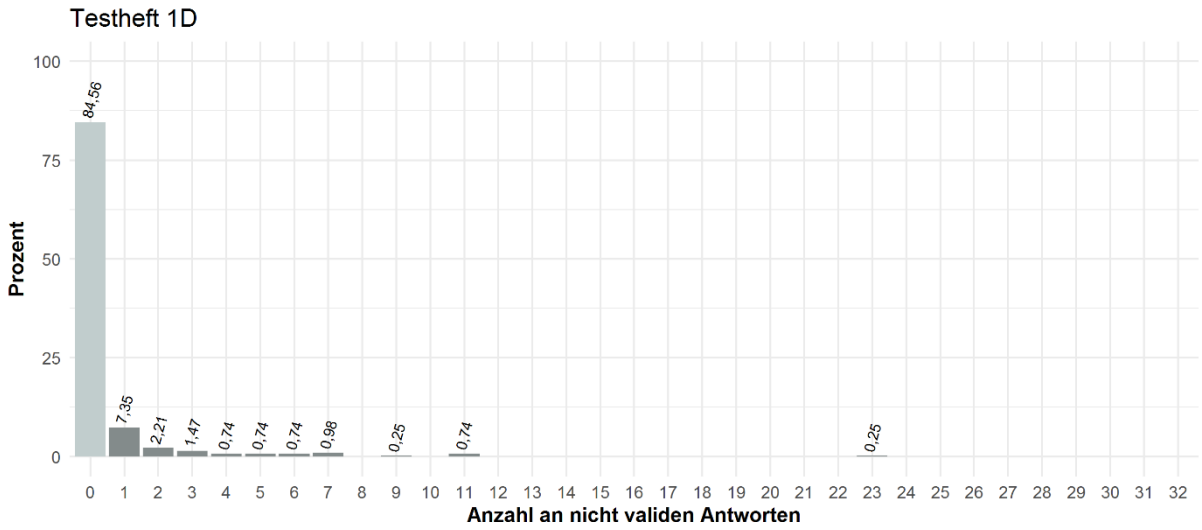

Abbildung 17. Anzahl an ungültigen Antworten für Testheft 1D (Lesekompetenz).

In den Abbildungen 18 und 19 ist die Anzahl an ungültigen Antworten pro Schülerin oder Schüler für die einzelnen Testhefte des Mathematiktests dargestellt. In allen Mathematiktestheften gaben mindestens 78% der Schülerinnen und Schüler keine ungültigen Antworten. Am häufigsten kam es in Testheft 2B und zwar bei circa 18% der Schülerinnen und Schüler zu einer ungültigen Antwort. In beiden Testheften gaben jeweils circa 4% der Schülerinnen und Schüler zwei oder mehr ungültige Antworten. Insgesamt kam auch im Mathematiktest diese Art fehlender Werte eher selten und wenn, dann nur vereinzelt vor.

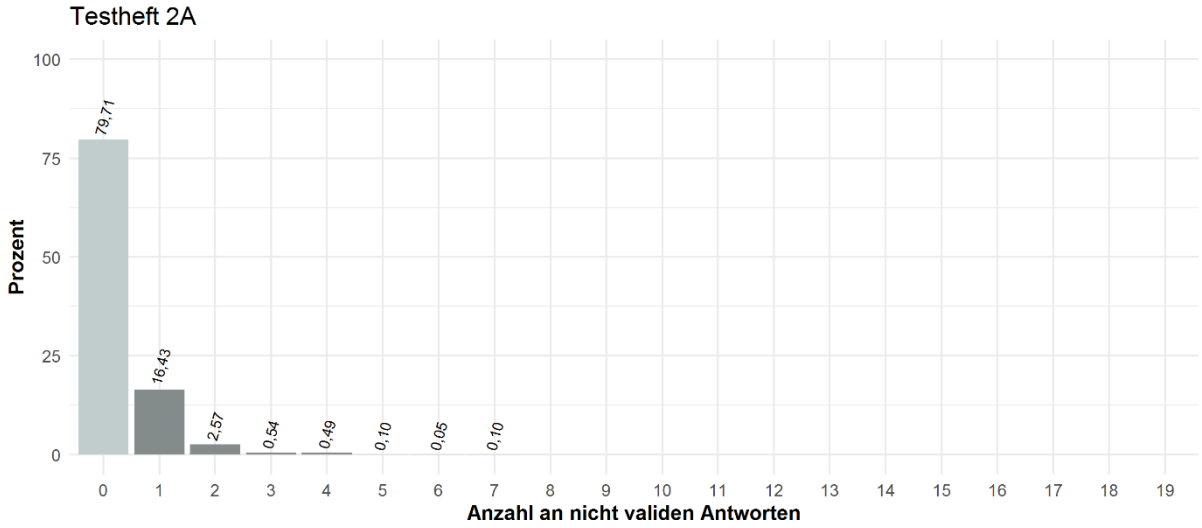

Abbildung 18. Anzahl an ungültigen Antworten für Testheft 2A (mathematische Kompetenz).

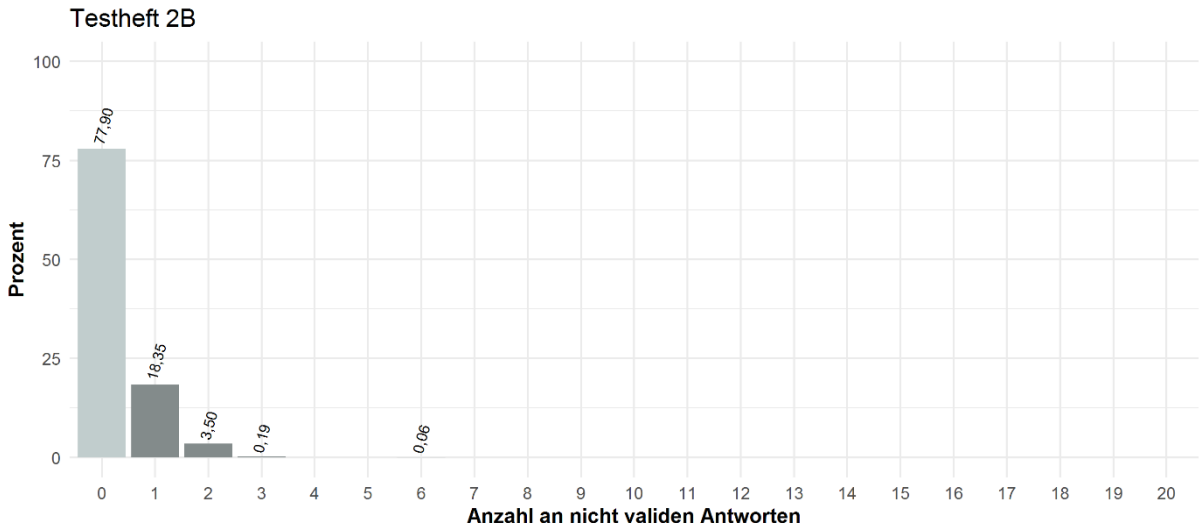

Abbildung 19. Anzahl an ungültigen Antworten für Testheft 2B (mathematische Kompetenz).

Gesamtanzahl fehlender Werte

Die Abbildungen 20 bis 23 zeigen die Gesamtanzahl an fehlenden Werten pro Schülerin oder Schüler für die einzelnen Testhefte des Lesetests. In den Abbildungen wird also nicht mehr nach der Art der fehlenden Werte unterschieden. In den Testheften 1A und 1B hatten circa 61% bzw. 51% der Schülerinnen und Schüler gar keine fehlenden Werte. Circa 10% und 20% der Schülerinnen und Schüler hatten in den Testheften 1A und 1B bei fünf oder mehr Items fehlende Werte. In den Testheften 1C und 1D hatten hingegen nur circa 19 bzw. 23% der Schülerinnen und Schüler gar keine fehlenden Werte erzeugt und circa 59% und 55% der Schülerinnen und Schüler hatten bei fünf oder mehr Items fehlende Werte. Diese hohe Anzahl ging insbesondere auf den hohen Anteil an nicht erreichten Items zurück.

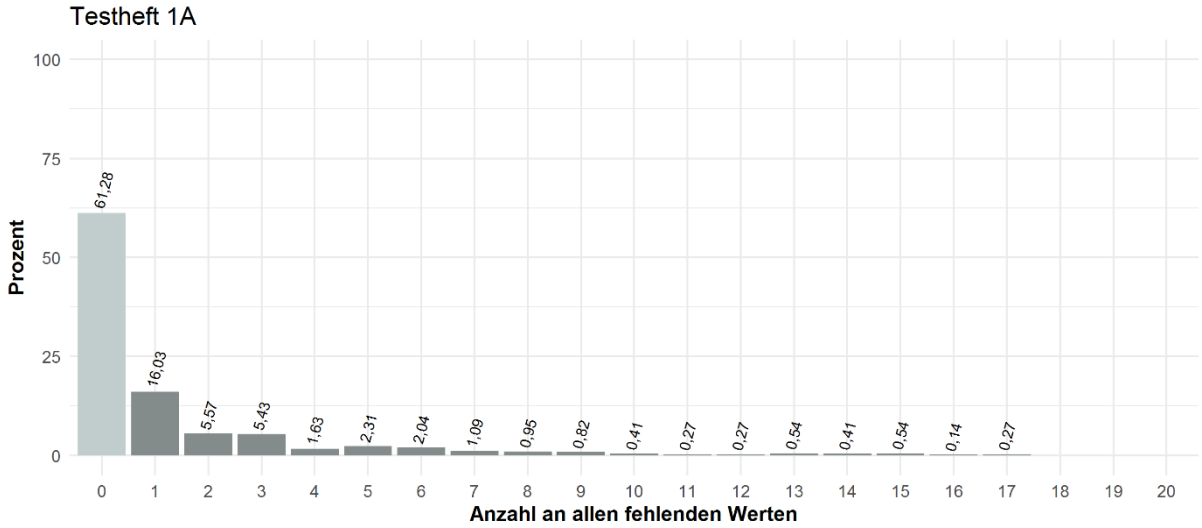

Abbildung 20. Anzahl an fehlenden Werten insgesamt für Testheft 1A (Lesekompetenz).

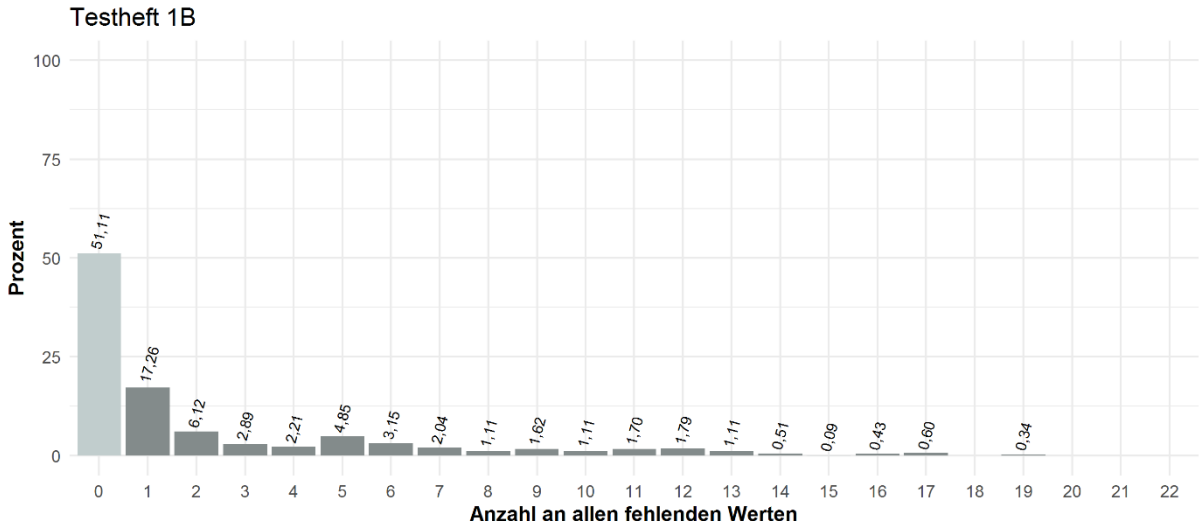

Abbildung 21. Anzahl an fehlenden Werten insgesamt für Testheft 1B (Lesekompetenz).

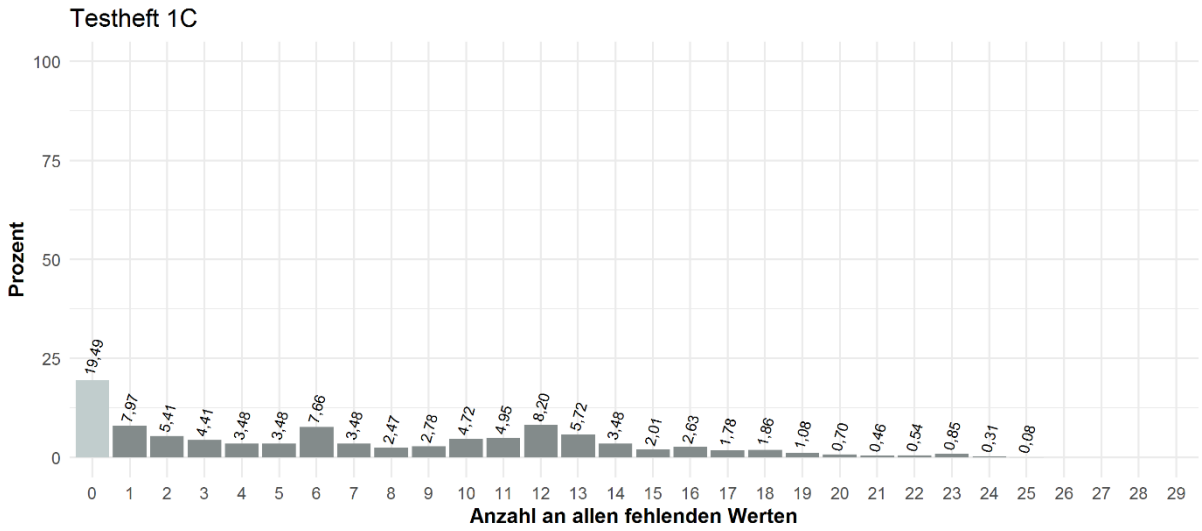

Abbildung 22. Anzahl an fehlenden Werten insgesamt für Testheft 1C (Lesekompetenz).

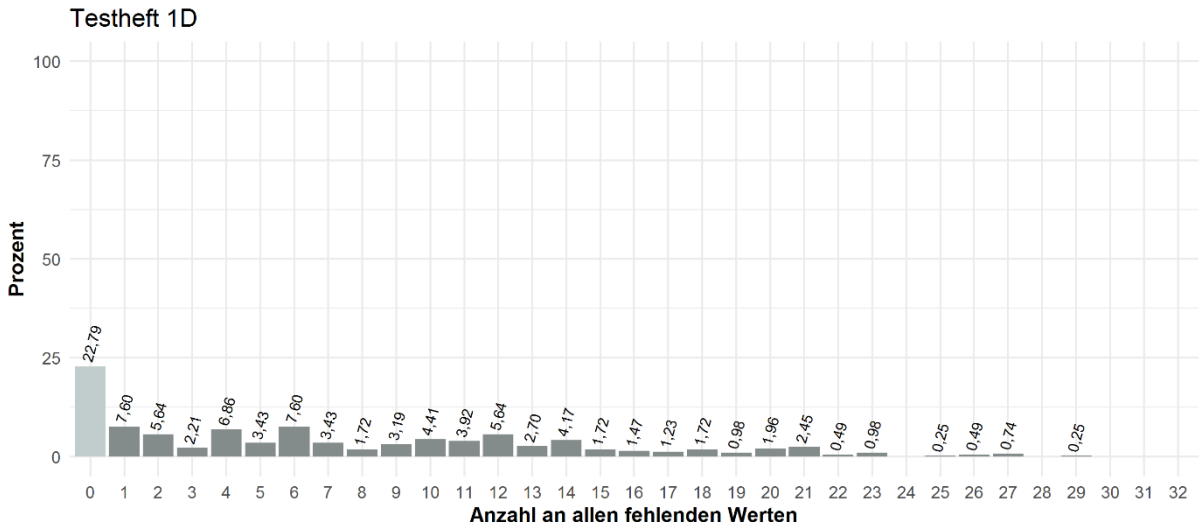

Abbildung 23. Anzahl an fehlenden Werten insgesamt für Testheft 1D (Lesekompetenz).

Die Abbildungen 24 und 25 zeigen die Gesamtanzahl an fehlenden Werten pro Schülerin oder und Schüler für die einzelnen Testhefte des Mathematiktests. In den Testheften 2A und 2B hatten circa 51 bzw. 50% der Schülerinnen und Schüler keine fehlenden Werte. Circa 5 bzw. 7% der Schülerinnen und Schüler hatten für fünf oder mehr Items fehlende Werte.

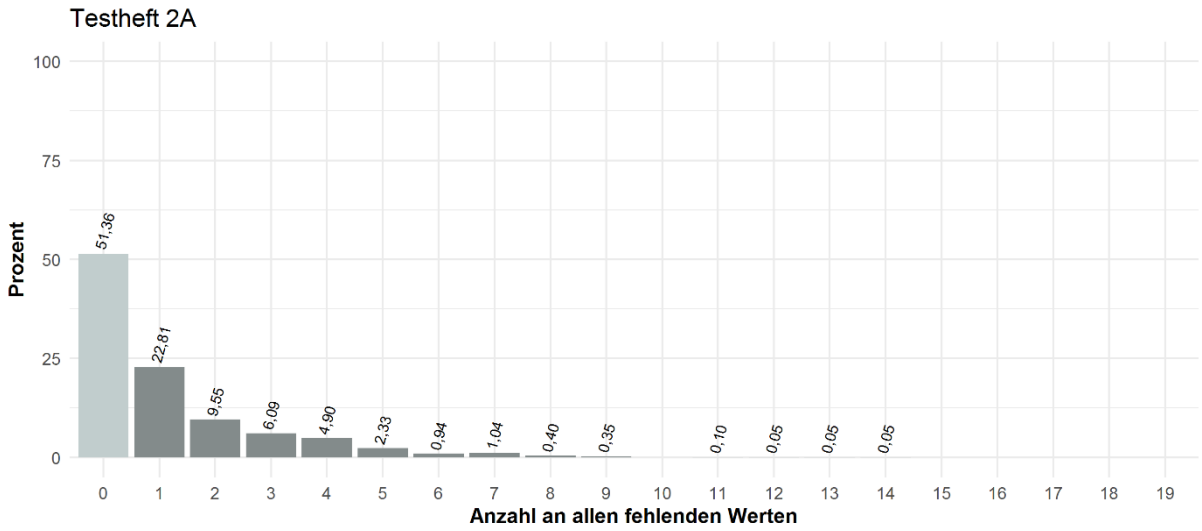

Abbildung 24. Anzahl an fehlenden Werten insgesamt für Testheft 2A (mathematische Kompetenz).

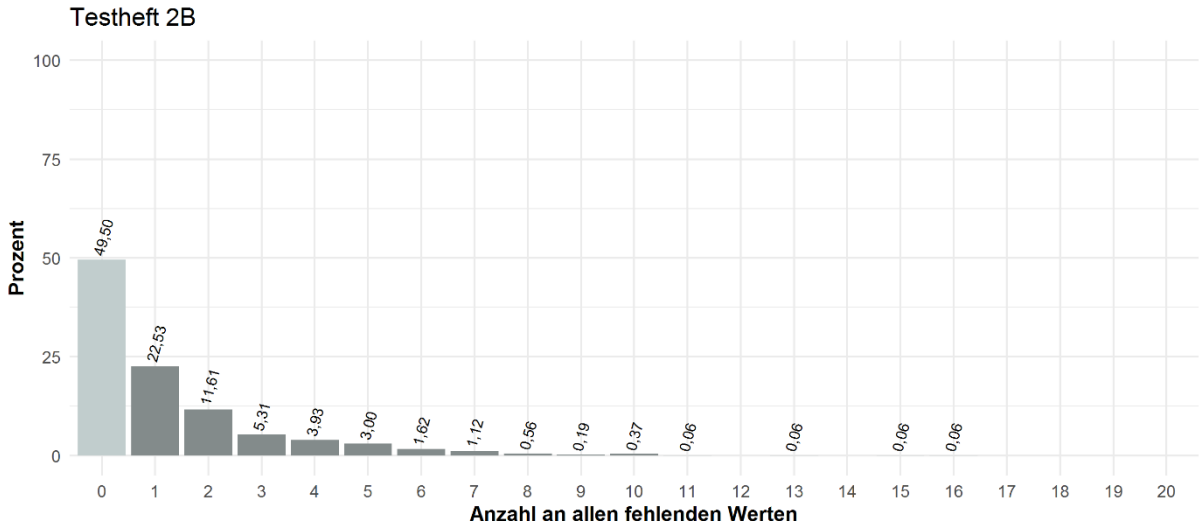

Abbildung 25. Anzahl an fehlenden Werten insgesamt für Testheft 2B (mathematische Kompetenz).

4.1.2 Fehlende Werte pro Item

Die prozentualen Anteile an fehlenden Werten pro Item werden in Tabelle 12 für den Lesetest aufgelistet. Der Anteil an ausgelassenen Angaben lag zwischen 0,00% und 13,05% (*Mdn*=1,45%). Wesentlich häufiger entstanden fehlende Werte, wenn Items nicht erreicht wurden. Einzelne Items wurden von zwischen 0,00% und 63,79% (*Mdn*=5,79%) der Schülerinnen und Schüler nicht erreicht. Ungültige Angaben hatten einen Anteil zwischen 0,00% und 6,62% (*Mdn*=0,84%).

Tabelle 12: Prozentuale Anteile für die verschiedenen Arten fehlender Werte pro Item für den Lesetest

Item	Pos. 1A	Pos. 1B	Pos. 1C	Pos. 1D	N_g	ausgelassen	nicht erreicht	ungültig
REG60110_I_C	1	1			1866	0,16	0,00	2,25
REG60120_I_C	2	2			1886	0,37	0,00	0,99
REG60130_I_C	3	3			1841	2,25	0,00	1,46
REG60140_I_C	4	4			1849	1,88	0,16	1,26
REG60150_I_C	5	5			1841	2,20	0,16	1,36
REG60220_I_C	7		8		1954	2,37	0,94	0,39
REG6023_I_S_C	8		9		1872	6,70	1,03	0,00
REG60240_I_C	9		10		1974	1,03	1,28	0,39
REG6025_I_S_C	10		11		1875	5,91	1,68	0,00
REG60260_I_C	11				717	0,82	1,36	0,41
REG6027_I_S_C			12		1190	4,49	2,32	0,93
REG60310_I_C	12		13		1919	0,94	3,45	1,03
REG6032_I_S_C	13		14		1809	6,41	4,19	0,15
REG60330_I_C	14		15		1847	2,96	5,52	0,49
REG60340_I_C	15		16		1881	0,74	6,06	0,49
REG60360_I_C	17		18		1811	1,38	7,98	1,38
REG6037_I_S_C			19		1060	3,71	13,69	0,46
REG60410_I_C	18	19			1636	0,58	13,18	0,68
REG60420_I_C	19	20			1608	0,89	14,17	0,84
REG60430_I_C	20	21			1584	0,10	15,79	1,26
REG60440_I_C	21	22			1562	0,58	16,63	1,10
REG60450_I_C	22	23			1542	0,00	18,04	1,31
REG60510_I_C		6			1144	1,45	0,43	0,85
REG60520_I_C		7			1066	4,08	0,77	4,51

Item	Pos. 1A	Pos. 1B	Pos. 1C	Pos. 1D	N_g	ausgelassen	nicht erreicht	ungültig
REG6053_I_S_C		8			1070	7,99	0,94	0,09
REG60540_I_C		9			1098	3,83	1,28	1,53
REG60550_I_C		10			1085	4,85	1,53	1,36
REG6057_I_S_C		12			1070	5,19	3,06	0,43
REG60610_I_C		13			1104	0,43	5,27	0,43
REG60620_I_C		14			1087	0,77	6,21	0,60
REG60630_I_C		15			1058	1,62	7,74	0,68
REG60640_I_C		16			1068	0,94	7,99	0,26
REG60650_I_C		17			1017	1,45	8,93	3,15
REG6066_I_S_C		18			991	4,25	11,22	0,26
REG60710_I_C			1	1	1680	0,47	0,00	0,76
REG60720_I_C			2	2	1633	2,82	0,00	1,18
REG6073_I_S_C			3	3	1540	9,47	0,00	0,00
REG6074_I_S_C			4	4	1474	13,05	0,06	0,18
REG60750_I_C			5	5	1603	2,59	0,06	3,12
REG6076_I_S_C			6	6	1522	9,47	0,24	0,53
REG60810_I_C			20	21	1271	0,41	24,46	0,41
REG60820_I_C			21	22	1159	2,23	28,63	1,00
REG6083_I_S_C			22	23	1027	4,35	34,27	0,65
REG60840_I_C			23	24	1013	1,82	36,63	2,00
REG60850_I_C			24	25	989	1,23	38,57	2,06
REG60860_I_C			25	26	956	1,18	39,80	2,82
REG60910_I_C			26	27	839	0,59	49,44	0,65
REG6092_I_S_C			27	28	769	1,82	52,85	0,06
REG60930_I_C			28	29	725	0,24	56,61	0,53

Item	Pos. 1A	Pos. 1B	Pos. 1C	Pos. 1D	N_g	ausgelassen	nicht erreicht	ungültig
REG60940_I_C			29	30	672	0,59	58,91	1,00
REG6095_I_S_C			30	31	655	0,94	60,55	0,00
REG60960_I_C			31	32	604	0,00	63,79	0,71
REG61010_I_C				7	399	0,49	0,98	0,74
REG61020_I_C				8	380	0,49	1,23	5,15
REG61030_I_C				9	372	0,98	1,23	6,62
REG61040_I_C				10	372	1,72	1,72	5,39
REG6105_I_S_C				11	376	5,88	1,96	0,00
REG6106_I_S_C				12	385	2,94	2,70	0,00
REG61110_I_C				13	372	0,74	6,62	1,47
REG61120_I_C				14	361	0,98	7,11	3,43
REG61130_I_C				15	354	1,47	9,07	2,70
REG61140_I_C				16	363	0,98	9,31	0,74
REG61150_I_C				17	354	1,23	11,03	0,98
REG61160_I_C				18	347	1,47	11,03	2,45
REG61170_I_C				19	346	1,23	12,99	0,98
REG61180_I_C				20	334	1,96	13,73	2,45

Anmerkungen. Gesamtanzahl = 3613 Schülerinnen und Schüler; Pos. = Itemposition im Testheftverlauf, wird mehr als eine Position angegeben, handelt es sich um ein Ankeritem; N_g = Anzahl an gültigen Fällen; ausgelassen = Prozent an Schülerinnen und Schülern, die das Item ausgelassen haben; nicht erreicht = Prozent an Schülerinnen und Schülern, die das Item nicht erreicht haben; ungültig = Prozent an Schülerinnen und Schülern, die ungültige Angaben bei dem Item gegeben haben.

In Tabelle 13 sind die prozentualen Anteile an fehlenden Werten pro Item für den Mathematiktest aufgelistet. Der Anteil an ausgelassenen Angaben lag zwischen 0,00 und 14,65% (*Mdn*=1,77%). Vergleichsweise wenige fehlende Werte entstanden, indem Items nicht erreicht wurden. Der Anteil lag hier zwischen 0,00 und 7,93% (*Mdn*=0,09%). Ungültige Angaben hatten einen Anteil zwischen 0,00 und 14,13% (*Mdn*=0,34%).

Tabelle 13: Prozentuale Anteile für die verschiedenen Arten fehlender Werte pro Item für den Mathematiktest

Item	Pos. 2A	Pos. 2B	N_g	ausgelassen	nicht erreicht	ungültig
MAG6D011_I_C	1		1987	1,14	0,00	0,54
MAG6D131_I_S_C	15		1851	6,93	0,20	1,24
MAG6D151_I_C	17		1988	0,89	0,40	0,35
MAG6D282_I_C		19	1497	1,94	4,18	0,44
MAG6Q021_I_C	2		2014	0,20	0,00	0,15
MAG6Q031_I_C	3	3	3564	1,52	0,00	0,11
MAG6Q071_I_C	7		1775	11,13	0,00	1,04
MAG6Q081_I_C	8		1753	11,43	0,05	1,78
MAG6Q082_I_C	9		1745	11,53	0,05	2,08
MAG6Q101_I_C	12		1682	14,65	0,05	2,08
MAG6Q16_I_S_C	18		1961	1,83	0,69	0,35
MAG6Q191_I_C		1	1591	0,62	0,00	0,06
MAG6Q231_I_C		8	1581	1,12	0,06	0,12
MAG6Q261_I_C		14	1569	1,37	0,62	0,06
MAG6Q281_I_C		18	1499	2,75	3,56	0,12
MAG6Q291_I_C		20	1471	0,00	7,93	0,25
MAG6R041_I_C	4		2009	0,30	0,00	0,30
MAG6R051_I_C	5	5	3585	0,47	0,00	0,58
MAG6R061_I_C	6	6	3048	1,71	0,03	14,13
MAG6R111_I_C	13	12	3439	1,96	0,17	2,95
MAG6R221_I_C		7	1465	8,43	0,06	0,06
MAG6R241_I_C		11	1588	0,56	0,25	0,06
MAG6R251_I_C		13	1591	0,31	0,37	0,00
MAG6R271_I_S_C		16	1417	8,05	1,94	1,19

Item	Pos. 2A	Pos. 2B	N_g	ausgelassen	nicht erreicht	ungültig
MAG6R273_I_S_C		17	1350	12,36	2,37	0,69
MAG6V091_I_C	10	9	3431	4,97	0,06	0,28
MAG6V092_I_C	11	10	3283	8,94	0,11	0,33
MAG6V141_I_C	16	15	3434	4,25	0,72	0,25
MAG6V171_I_C	19		1977	0,99	0,89	0,30
MAG6V181_I_C	20		1979	0,00	1,39	0,69
MAG6V201_I_C		2	1454	8,49	0,00	0,75
MAG6V211_I_C		4	1589	0,56	0,00	0,25

Anmerkungen. Gesamtanzahl = 3623 Schülerinnen und Schüler; Pos. = Itemposition im Testheftverlauf, wird mehr als eine Position angegeben, handelt es sich um ein Ankeritem; N_g = Anzahl an gültigen Fällen; ausgelassen = Prozent an Schülerinnen und Schülern, die das Item ausgelassen haben; nicht erreicht = Prozent an Schülerinnen und Schülern, die das Item nicht erreicht haben; ungültig = Prozent an Schülerinnen und Schülern, die ungültige Angaben bei dem Item gegeben haben.

4.2 Parameterschätzung

4.2.1 Itemparameter

Die geschätzten Itemschwierigkeitsparameter für dichotome Items und Lageparameter (Engl. *Location Parameter*) für polytome Items sowie Itemfitstatistiken werden in Tabelle 14 für eine Skalierung des gesamten Lesetests dargestellt. Für die Schätzung der Itemschwierigkeit wurde die Personenfähigkeitsverteilung auf den Mittelwert (*M*) = 0 Logit gesetzt. Geschätzte Schwellenparameter für Teilantwortkategorien polytomer Items aus der Skalierung des gesamten Lesetests sind in Tabelle 15 enthalten. Es folgt eine Darstellung der geschätzten Itemparameter und Itemstatistiken auf Testheftebene für Testheft 1A in den Tabellen 16 und 17, für Testheft 1B in den Tabelle 18 und 19, für Testheft 1C in den Tabellen 20 und 21 sowie für Testheft 1D in den Tabelle 22 und 23.

Lösungshäufigkeiten für dichotome Leseitems werden als prozentualer Anteil korrekter Antwortreaktionen angegeben. Diese Angaben sollten aufgrund des hohen Anteils fehlender Werte und der hohen Ratewahrscheinlichkeit (bei zwei Antwortalternativen) nicht als Lösungswahrscheinlichkeiten interpretiert werden. Im gesamten Lesetest haben die Schülerinnen und Schüler Multiple-Choice-Aufgaben zwischen zu 14,55% und 93,39% korrekt gelöst (*M* = 73,07%, *SD* = 18,09%).

Die geschätzten Itemschwierigkeiten (für dichtome Items) und Lageparameter (für polytome Items) liegen im gesamten Lesetest zwischen 2,22 und -3,05 Logit (*M* = -1,20 Logit). Die Größe des Standardfehlers der Itemschwierigkeiten oder Lageparameter variiert und nimmt tendenziell bei Items mit Itemschwierigkeiten an den Extrema der Logitskala und bei Items mit geringer Anzahl an gültigen Fällen zu.

In Tabelle 24 werden die geschätzten Itemschwierigkeitsparameter für dichotome Items und Lageparameter für polytome Items sowie Itemstatistiken für eine Skalierung des gesamten Mathematiktests dargestellt. Geschätzte Schwellenparameter für Teilantwortkategorien polytomer Items aus der Skalierung des gesamten Mathematiktests sind in Tabelle 25 enthalten. Es folgt eine Darstellung der geschätzten Itemparameter und Itemstatistiken auf Testheftebene für Testheft 2A in den Tabellen 26 und 27 und für Testheft 2B in Tabelle 28. (Da es in diesem Testheft keine polytomen Items gab, fehlt hier das inhaltliche Äquivalent zu Tabelle 27).

Lösungshäufigkeiten für dichotome Mathematikitems werden ebenfalls als prozentualer Anteil korrekter Antwortreaktionen angegeben. Analog zum Lesetest, sollten diese Angaben aufgrund des hohen Anteils fehlender Werte nicht als Lösungswahrscheinlichkeiten interpretiert werden. Im gesamten Mathematiktest haben die Schülerinnen und Schüler Multiple-Choice-Aufgaben und Freitextaufgaben zwischen zu 22,00% und 96,35% korrekt gelöst (*M* = 65,57%, *SD* = 20,49%).

Die geschätzten Itemschwierigkeiten (für dichtome Items) und Lageparameter (für polytome Items) liegen im gesamten Mathematiktest zwischen 1,64 und -3,67 Logit (*M* = -0,98 Logit). Die Größe des Standardfehlers der Itemschwierigkeiten oder Lageparameter nimmt tendenziell bei Items mit Itemschwierigkeiten an den Extrema der Logitskala zu.

Tabelle 14: Itemparameter für den gesamten Lesetest

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60110_I_C	1866	90,25	-2,92	0,09	0,99	-0,11	1,22	0,23	1,21
REG60120_I_C	1886	91,25	-3,04	0,09	0,93	-1,09	0,68	0,31	1,89
REG60130_I_C	1841	80,99	-1,97	0,07	1,02	0,46	1,06	0,32	1,21
REG60140_I_C	1849	71,50	-1,32	0,06	0,98	-0,53	0,96	0,40	1,34
REG60150_I_C	1841	78,65	-1,81	0,07	1,11	3,06	1,29	0,26	0,84
REG60220_I_C	1954	86,80	-2,40	0,07	1,01	0,15	1,12	0,24	1,20
REG6023_I_S_C	1872	k.A.	-1,38	0,03	0,96	-1,02	0,92	0,33	0,77
REG60240_I_C	1974	90,37	-2,81	0,08	0,92	-1,42	0,79	0,29	1,80
REG6025_I_S_C	1875	k.A.	-0,95	0,03	1,07	1,86	1,13	0,23	0,47
REG60260_I_C	717	88,98	-2,97	0,13	1,12	1,34	1,34	0,18	0,79
REG6027_I_S_C	1190	k.A.	-0,12	0,03	1,12	3,15	1,13	0,39	0,45
REG60310_I_C	1919	85,20	-2,24	0,07	1,03	0,61	1,14	0,22	1,08
REG6032_I_S_C	1809	k.A.	-0,98	0,03	0,90	-2,79	0,84	0,40	0,98
REG60330_I_C	1847	33,24	0,89	0,06	1,06	2,38	1,19	0,31	0,99

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG60340_I_C	1881	75,01	-1,44	0,06	0,92	-2,81	0,82	0,40	1,79
REG60360_I_C	1811	63,67	-0,75	0,06	1,10	4,11	1,12	0,27	0,92
REG6037_I_S_C	1060	k.A.	-0,41	0,03	0,95	-1,31	0,95	0,47	0,73
REG60410_I_C	1636	66,75	-1,02	0,06	0,87	-5,06	0,82	0,52	2,14
REG60420_I_C	1608	66,17	-0,99	0,06	0,94	-2,06	0,92	0,46	1,53
REG60430_I_C	1584	74,49	-1,50	0,07	0,99	-0,31	0,96	0,41	1,33
REG60440_I_C	1562	77,59	-1,72	0,07	0,88	-3,54	0,74	0,50	2,17
REG60450_I_C	1542	75,03	-1,53	0,07	0,92	-2,41	0,82	0,47	1,80
REG60510_I_C	1144	72,90	-1,26	0,08	1,01	0,18	0,97	0,34	1,25
REG60520_I_C	1066	40,81	0,49	0,07	1,10	3,29	1,23	0,28	0,80
REG6053_I_S_C	1070	k.A.	-0,87	0,04	1,02	0,37	0,98	0,34	0,58
REG60540_I_C	1098	30,87	1,04	0,08	1,14	4,12	1,34	0,21	0,61
REG60550_I_C	1085	26,64	1,30	0,08	1,11	2,89	1,34	0,23	0,69
REG6057_I_S_C	1070	k.A.	0,15	0,03	1,08	1,98	1,12	0,35	0,43
REG60610_I_C	1104	83,70	-2,06	0,09	0,97	-0,48	0,98	0,35	1,37

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG60620_I_C	1087	80,77	-1,82	0,09	0,89	-2,35	0,75	0,45	1,98
REG60630_I_C	1058	84,40	-2,11	0,09	0,91	-1,64	0,76	0,43	1,89
REG60640_I_C	1068	80,90	-1,82	0,09	0,93	-1,37	0,99	0,38	1,59
REG60650_I_C	1017	14,55	2,22	0,10	1,07	1,09	1,87	0,18	0,68
REG6066_I_S_C	991	k.A.	0,10	0,04	1,07	1,91	1,10	0,35	0,45
REG60710_I_C	1680	92,08	-2,83	0,10	0,97	-0,35	0,94	0,17	1,48
REG60720_I_C	1633	80,04	-1,61	0,07	1,09	2,36	1,19	0,19	0,86
REG6073_I_S_C	1540	k.A.	-0,89	0,04	1,03	0,95	1,04	0,24	0,53
REG6074_I_S_C	1474	k.A.	-0,87	0,03	1,00	-0,07	0,98	0,30	0,67
REG60750_I_C	1603	66,81	-0,75	0,06	1,02	0,74	1,00	0,32	1,20
REG6076_I_S_C	1522	k.A.	-0,39	0,02	0,94	-1,87	0,92	0,46	0,73
REG60810_I_C	1271	88,36	-2,33	0,10	1,04	0,72	1,30	0,20	0,98
REG60820_I_C	1159	72,04	-1,05	0,08	1,11	2,95	1,23	0,29	0,86
REG6083_I_S_C	1027	k.A.	-0,64	0,04	0,90	-2,18	0,88	0,46	0,90
REG60840_I_C	1013	67,82	-0,79	0,08	0,88	-3,41	0,79	0,49	1,85

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG60850_I_C	989	77,55	-1,40	0,09	0,89	-2,56	0,74	0,46	1,93
REG60860_I_C	956	63,60	-0,52	0,08	1,06	1,76	1,08	0,39	1,10
REG60910_I_C	839	71,63	-1,06	0,09	1,04	0,99	1,03	0,39	1,16
REG6092_I_S_C	769	k.A.	-0,56	0,05	1,02	0,40	0,99	0,37	0,62
REG60930_I_C	725	47,45	0,30	0,09	0,82	-4,90	0,75	0,55	2,04
REG60940_I_C	672	57,14	-0,26	0,09	1,00	0,12	1,02	0,46	1,27
REG6095_I_S_C	655	k.A.	-0,61	0,05	1,03	0,72	0,95	0,37	0,61
REG60960_I_C	604	60,10	-0,50	0,10	1,05	1,18	1,05	0,43	1,15
REG61010_I_C	399	84,21	-1,92	0,15	0,96	-0,38	0,89	0,27	1,50
REG61020_I_C	380	87,89	-2,29	0,17	1,05	0,45	1,20	0,18	1,02
REG61030_I_C	372	62,10	-0,40	0,13	1,07	1,31	1,12	0,35	1,04
REG61040_I_C	372	83,06	-1,81	0,16	1,01	0,19	0,98	0,27	1,26
REG6105_I_S_C	376	k.A.	-0,26	0,08	0,97	-0,58	0,96	0,32	0,72
REG6106_I_S_C[a]	385	88,57	-2,36	0,18	1,13	1,10	1,57	0,14	0,65
REG61110_I_C	372	90,05	-2,55	0,19	0,87	-1,07	0,55	0,31	2,58

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG61120_I_C	361	84,76	-1,97	0,16	1,03	0,38	1,27	0,22	1,08
REG61130_I_C	354	82,20	-1,72	0,16	0,91	-1,06	0,87	0,36	1,81
REG61140_I_C	363	93,39	-3,05	0,23	0,88	-0,72	0,51	0,27	2,72
REG61150_I_C	354	87,57	-2,23	0,18	0,91	-0,82	0,62	0,32	2,13
REG61160_I_C	347	72,62	-1,03	0,14	1,02	0,32	1,08	0,37	1,20
REG61170_I_C	346	89,88	-2,51	0,19	0,92	-0,63	0,75	0,29	1,86
REG61180_I_C	334	64,07	-0,49	0,13	1,07	1,26	1,11	0,32	1,05

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step“. Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

[a]Item REG6106_I_S_C ist ein ursprünglich polytomes Item, das nach der Zusammenlegung von Teilkategorien nur noch zwei Ausprägungen beibehielt.

Tabelle 15: Schwellenparameter für die polytomen Leseitems im gesamten Lesetest

Item	1. Schwelle	(SE)	2. Schwelle	(SE)	3. Schwelle	(SE)	4. Schwelle	(SE)	5. Schwelle
REG6023_I_S_C	-0,54	(0,05)	0,56	(0,06)	-0,03				
REG6025_I_S_C	1,32	(0,09)	-1,32						
REG6027_I_S_C	-0,73	(0,06)	-0,15	(0,06)	1,80	(0,12)	-0,92		
REG6032_I_S_C	0,23	(0,06)	0,13	(0,06)	-0,36				
REG6037_I_S_C	-0,54	(0,07)	-0,02	(0,06)	0,51	(0,08)	0,05		
REG6053_I_S_C	-0,05	(0,06)	-0,29	(0,07)	0,34				
REG6057_I_S_C	-1,01	(0,07)	0,14	(0,07)	1,04	(0,11)	-0,17		
REG6066_I_S_C	-0,87	(0,07)	0,69	(0,08)	0,18				
REG6073_I_S_C	-0,24	(0,06)	0,24						
REG6074_I_S_C	-0,95	(0,05)	0,09	(0,06)	0,86				
REG6076_I_S_C	-0,84	(0,06)	-0,03	(0,05)	0,31	(0,06)	1,50	(0,10)	-0,93
REG6083_I_S_C	-0,11	(0,06)	-0,84	(0,06)	0,95				
REG6092_I_S_C	-0,27	(0,08)	0,27						
REG6095_I_S_C	0,77	(0,11)	-0,77						
REG6105_I_S_C	-0,63	(0,11)	0,63						

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step“ geschätzt.

Tabelle 16: Itemparameter für Testheft 1A (Lesekompetenz)

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60110_I_C	716	87,71	-2,53	0,13	0,95	-0,62	1,09	0,33	1,52
REG60120_I_C	728	89,42	-2,72	0,13	0,95	-0,54	0,76	0,35	1,70
REG60130_I_C	713	77,56	-1,62	0,10	1,02	0,44	1,09	0,37	1,23
REG60140_I_C	717	68,34	-1,02	0,09	0,98	-0,45	0,93	0,47	1,41
REG60150_I_C	711	77,64	-1,64	0,10	1,12	2,13	1,34	0,30	0,87
REG60220_I_C	709	80,68	-1,86	0,11	1,01	0,25	0,99	0,35	1,33
REG6023_I_S_C	648	k.A.	-0,60	0,06	0,98	-0,41	0,97	0,38	0,71
REG60240_I_C	715	83,92	-2,13	0,11	0,92	-1,29	0,80	0,43	1,81
REG6025_I_S_C	669	k.A.	-0,82	0,05	1,06	1,11	1,07	0,35	0,56
REG60260_I_C	717	88,98	-2,66	0,13	1,13	1,45	1,38	0,18	0,77
REG60310_I_C	710	79,72	-1,78	0,11	1,03	0,50	1,21	0,31	1,10
REG6032_I_S_C	644	k.A.	-0,76	0,05	0,91	-1,69	0,86	0,54	0,91
REG60330_I_C	690	26,09	1,41	0,10	1,12	2,40	1,49	0,23	0,74
REG60340_I_C	698	65,76	-0,84	0,09	0,97	-0,72	0,94	0,49	1,43

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60360_I_C	682	58,94	-0,44	0,09	1,13	3,30	1,19	0,34	0,86
REG60410_I_C	668	62,72	-0,65	0,09	0,88	-2,95	0,86	0,54	2,06
REG60420_I_C	658	62,16	-0,62	0,10	0,96	-1,06	0,93	0,48	1,53
REG60430_I_C	655	70,53	-1,12	0,10	1,04	0,85	1,02	0,41	1,17
REG60440_I_C	646	76,47	-1,52	0,11	0,89	-2,04	0,77	0,49	2,07
REG60450_I_C	637	73,16	-1,28	0,10	0,90	-1,99	0,79	0,51	1,93

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step". Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

Tabelle 17: Schwellenparameter für polytome Items in Testheft 1A (Lesekompetenz)

Item	1. Schwelle	(SE)	2. Schwelle	(SE)	3. Schwelle
REG6023_I_S_C	0,34	(0,10)	-0,34		
REG6025_I_S_C	1,43	(0,15)	-1,43		
REG6032_I_S_C	0,01	(0,09)	0,20	(0,10)	-0,20

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step" geschätzt.

Tabelle 18: Itemparameter für Testheft 1B (Lesekompetenz)

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60110_I_C	1150	91,83	-2,94	0,12	1,03	0,35	1,30	0,18	0,98
REG60120_I_C	1158	92,40	-3,02	0,12	0,92	-0,92	0,62	0,29	2,03
REG60130_I_C	1128	83,16	-1,97	0,09	1,01	0,20	1,03	0,30	1,21
REG60140_I_C	1132	73,50	-1,28	0,08	0,98	-0,52	0,97	0,36	1,32
REG60150_I_C	1130	79,29	-1,68	0,08	1,09	1,87	1,22	0,23	0,85
REG60410_I_C	968	69,52	-1,03	0,08	0,86	-4,08	0,78	0,52	2,28
REG60420_I_C	950	68,95	-0,99	0,08	0,93	-1,80	0,92	0,45	1,54
REG60430_I_C	929	77,29	-1,52	0,09	0,95	-1,05	0,91	0,41	1,48
REG60440_I_C	916	78,38	-1,60	0,09	0,85	-3,16	0,71	0,51	2,37
REG60450_I_C	905	76,35	-1,45	0,09	0,92	-1,67	0,83	0,45	1,79
REG60510_I_C	1144	72,90	-1,23	0,08	1,00	0,00	0,96	0,34	1,26
REG60520_I_C	1066	40,81	0,50	0,07	1,09	3,02	1,20	0,28	0,81
REG6053_I_S_C	1070	k.A.	-0,85	0,04	1,01	0,21	0,98	0,34	0,59
REG60540_I_C	1098	30,87	1,04	0,08	1,13	3,85	1,31	0,21	0,62

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60550_I_C	1085	26,64	1,30	0,08	1,10	2,65	1,31	0,23	0,70
REG6057_I_S_C	1070	k.A.	0,15	0,03	1,07	1,74	1,11	0,35	0,44
REG60610_I_C	1104	83,70	-2,03	0,09	0,97	-0,58	0,97	0,35	1,39
REG60620_I_C	1087	80,77	-1,79	0,09	0,89	-2,43	0,75	0,45	2,01
REG60630_I_C	1058	84,40	-2,08	0,09	0,91	-1,71	0,75	0,43	1,92
REG60640_I_C	1068	80,90	-1,79	0,09	0,93	-1,47	0,97	0,38	1,61
REG60650_I_C	1017	14,55	2,22	0,10	1,06	0,97	1,80	0,18	0,68
REG6066_I_S_C	991	k.A.	0,11	0,04	1,06	1,68	1,08	0,35	0,45

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step". Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

Tabelle 19: Schwellenparameter für polytome Items in Testheft 1B (Lesekompetenz)

Item	1. Schwelle	(SE)	2. Schwelle	(SE)	3. Schwelle	(SE)	4. Schwelle
REG6053_I_S_C	-0,04	(0,06)	-0,29	(0,07)	0,33		
REG6057_I_S_C	-1,00	(0,07)	0,14	(0,07)	1,04	(0,11)	-0,18
REG6066_I_S_C	-0,86	(0,07)	0,69	(0,08)	0,17		

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step" geschätzt.

Tabelle 20: Itemparameter für Testheft 1C (Lesekompetenz)

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60210_I_C[a]	1234	50,24	-0,01	0,07	1,17	6,51	1,23	0,21	0,52
REG60220_I_C	1245	90,28	-2,66	0,10	1,01	0,21	1,24	0,16	0,91
REG6023_I_S_C	1224	k.A.	-1,21	0,05	0,92	-1,70	0,83	0,29	0,91
REG60240_I_C	1259	94,04	-3,23	0,12	0,96	-0,38	0,82	0,20	1,46
REG6025_I_S_C	1206	k.A.	-0,97	0,04	1,05	1,06	1,09	0,22	0,40
REG6027_I_S_C	1190	k.A.	-0,18	0,03	1,09	2,38	1,10	0,39	0,43
REG60310_I_C	1209	88,42	-2,42	0,10	1,02	0,32	1,03	0,17	0,94
REG6032_I_S_C	1165	k.A.	-1,05	0,04	0,90	-1,89	0,83	0,37	0,93
REG60330_I_C	1157	37,51	0,65	0,07	1,00	0,12	1,02	0,34	1,07
REG60340_I_C	1183	80,47	-1,73	0,08	0,91	-2,17	0,76	0,38	1,91
REG60350_I_C[a]	1137	52,59	-0,12	0,07	1,15	5,63	1,20	0,22	0,59
REG60360_I_C	1129	66,52	-0,84	0,07	1,05	1,60	1,03	0,30	0,94
REG6037_I_S_C	1060	k.A.	-0,46	0,03	0,93	-1,92	0,92	0,47	0,71
REG60710_I_C	1275	91,53	-2,82	0,11	0,96	-0,52	0,90	0,19	1,47

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG60720_I_C	1236	79,77	-1,69	0,08	1,07	1,69	1,13	0,19	0,79
REG6073_I_S_C	1166	k.A.	-0,93	0,05	1,03	0,80	1,04	0,22	0,46
REG6074_I_S_C	1119	k.A.	-0,34	0,04	0,97	-0,94	0,96	0,29	0,66
REG60750_I_C	1213	66,45	-0,85	0,07	1,01	0,47	1,00	0,30	1,07
REG6076_I_S_C	1159	k.A.	-0,42	0,02	0,91	-2,49	0,88	0,48	0,73
REG60810_I_C	963	87,85	-2,36	0,11	1,03	0,43	1,20	0,21	0,93
REG60820_I_C	872	72,13	-1,17	0,09	1,14	3,37	1,33	0,23	0,63
REG6083_I_S_C	766	k.A.	-0,66	0,05	0,89	-2,23	0,86	0,47	0,89
REG60840_I_C	754	66,58	-0,84	0,09	0,89	-2,93	0,81	0,49	1,66
REG60850_I_C	732	76,37	-1,42	0,10	0,85	-3,10	0,68	0,50	2,05
REG60860_I_C	712	64,61	-0,70	0,09	1,04	1,04	1,03	0,37	1,02
REG60910_I_C	617	72,29	-1,20	0,10	1,03	0,56	0,97	0,38	1,09
REG6092_I_S_C	569	k.A.	-0,62	0,06	1,01	0,18	0,99	0,35	0,57
REG60930_I_C	539	45,64	0,26	0,10	0,80	-5,03	0,73	0,57	2,01
REG60940_I_C	500	56,00	-0,32	0,11	0,97	-0,56	0,96	0,47	1,24

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG6095_I_S_C	488	k.A.	-0,63	0,06	1,07	1,26	0,99	0,33	0,50
REG60960_I_C	447	57,27	-0,46	0,11	1,01	0,23	0,99	0,47	1,15

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step“. Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

[a] Die Skalierung für Testheft 1C enthält zwei Items (REG60210_I_C und REG60350_I_C), die von der Gesamtskalierung ausgeschlossen wurden, für die Skalierung in Testheft 1C jedoch nicht als problematisch auffielen.

Tabelle 21: Schwellenparameter für polytome Items in Testheft 1C (Lesekompetenz)

Item	1. Schwelle	(SE)	2. Schwelle	(SE)	3. Schwelle	(SE)	4. Schwelle	(SE)	5. Schwelle
REG6023_I_S_C	0,43	(0,08)	-0,43						
REG6025_I_S_C	1,27	(0,10)	-1,27						
REG6027_I_S_C	-0,70	(0,06)	-0,14	(0,06)	1,79	(0,12)	-0,95		
REG6032_I_S_C	0,43	(0,07)	0,04	(0,08)	-0,47				
REG6037_I_S_C	-0,50	(0,06)	0,00	(0,06)	0,50	(0,08)	0,01		
REG6073_I_S_C	-0,24	(0,07)	0,24						
REG6074_I_S_C	-0,31	(0,06)	0,31						
REG6076_I_S_C	-0,74	(0,07)	-0,01	(0,06)	0,21	(0,07)	1,46	(0,12)	-0,93
REG6083_I_S_C	-0,07	(0,07)	-0,80	(0,07)	0,87				
REG6092_I_S_C	-0,28	(0,09)	0,28						
REG6095_I_S_C	0,74	(0,13)	-0,74						

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step“ geschätzt.

Tabelle 22: Itemparameter für Testheft 1D (Lesekompetenz)

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60710_I_C	405	93,83	-3,35	0,22	0,99	-0,03	1,00	0,16	1,35
REG60720_I_C	397	80,86	-1,87	0,14	1,08	1,02	1,14	0,24	1,01
REG6073_I_S_C[a]	374	60,96	-0,60	0,13	1,12	2,27	1,19	0,32	0,88
REG6074_I_S_C	355	k.A.	-0,43	0,08	0,94	-1,08	0,91	0,37	0,85
REG60750_I_C	390	67,95	-0,98	0,13	0,97	-0,46	0,93	0,43	1,51
REG6076_I_S_C	363	k.A	-0,35	0,06	0,91	-1,53	0,88	0,47	0,91
REG60810_I_C	308	89,94	-2,74	0,21	1,03	0,27	1,57	0,22	0,98
REG60820_I_C	287	71,78	-1,21	0,15	0,94	-0,88	0,87	0,50	1,59
REG6083_I_S_C[a]	261	39,85	0,66	0,15	1,15	2,15	1,19	0,38	0,87
REG60840_I_C	259	71,43	-1,18	0,16	0,82	-2,45	0,70	0,58	2,27
REG60850_I_C	257	80,93	-1,85	0,18	0,96	-0,35	0,90	0,42	1,48
REG60860_I_C	244	60,66	-0,52	0,16	1,04	0,62	1,10	0,46	1,21
REG60910_I_C	222	69,82	-1,13	0,17	1,01	0,20	0,99	0,45	1,27
REG6092_I_S_C[a]	200	52,00	-0,03	0,17	1,12	1,46	1,33	0,40	0,96

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
REG60930_I_C	186	52,69	-0,14	0,18	0,84	-2,04	0,78	0,61	1,89
REG60940_I_C	172	60,47	-0,60	0,19	1,03	0,41	1,10	0,48	1,24
REG6095_I_S_C[a]	167	68,26	-1,05	0,20	0,98	-0,19	0,94	0,50	1,36
REG60960_I_C	157	68,15	-1,14	0,21	1,13	1,35	1,27	0,39	0,99
REG61010_I_C	399	84,21	-2,14	0,15	0,95	-0,53	0,87	0,31	1,63
REG61020_I_C	380	87,89	-2,51	0,17	1,05	0,49	1,24	0,20	1,05
REG61030_I_C	372	62,10	-0,61	0,13	1,09	1,58	1,15	0,35	1,04
REG61040_I_C	372	83,06	-2,03	0,16	1,01	0,16	0,99	0,29	1,29
REG6105_I_S_C	376	k.A.	-0,37	0,08	0,97	-0,53	0,96	0,34	0,72
REG6106_I_S_C[a]	385	88,57	-2,58	0,18	1,13	1,12	1,60	0,16	0,69
REG61110_I_C	372	90,05	-2,76	0,19	0,86	-1,08	0,55	0,35	2,78
REG61120_I_C	361	84,76	-2,19	0,16	1,03	0,34	1,25	0,25	1,11
REG61130_I_C	354	82,20	-1,94	0,16	0,91	-0,98	0,90	0,39	1,82
REG61140_I_C	363	93,39	-3,27	0,23	0,88	-0,72	0,50	0,31	2,82
REG61150_I_C	354	87,57	-2,45	0,18	0,90	-0,86	0,63	0,36	2,24

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
REG61160_I_C	347	72,62	-1,24	0,14	1,02	0,33	1,09	0,39	1,20
REG61170_I_C	346	89,88	-2,72	0,19	0,92	-0,61	0,76	0,33	1,83
REG61180_I_C	334	64,07	-0,70	0,13	1,08	1,31	1,11	0,33	1,03

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step". Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

[a] REG6073_I_S_C, REG6083_I_S_C, REG6092_I_S_C, REG6095_I_S_C, REG6106_I_S_C sind ursprünglich polytome Items, die nach der Zusammenlegung von Teilkategorien nur noch zwei Ausprägungen beibehielten.

Tabelle 23: Schwellenparameter für polytome Items in Testheft 1D (Lesekompetenz)

Item	1. Schwelle	(SE)	2. Schwelle	(SE)	3. Schwelle
REG6074_I_S_C	-0,14	(0,12)	0,14		
REG6076_I_S_C	0,03	(0,11)	0,14	(0,13)	-0,18
REG6105_I_S_C	-0,63	(0,11)	0,63		

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step" geschätzt.

Tabelle 24: Itemparameter für den gesamten Mathematiktest

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
MAG6D011_I_C	1987	47,91	0,00	0,05	1,03	1,45	1,04	0,36	0,98
MAG6D131_I_S_C[a]	1851	72,88	-1,26	0,06	0,96	-1,46	0,90	0,40	1,45
MAG6D151_I_C	1988	90,14	-2,77	0,08	0,91	-1,57	0,71	0,39	1,95
MAG6D282_I_C	1497	74,68	-1,19	0,07	1,05	1,55	1,05	0,30	0,89
MAG6Q021_I_C	2014	94,59	-3,50	0,10	0,96	-0,46	1,02	0,25	1,35
MAG6Q031_I_C	3564	72,11	-1,18	0,04	1,01	0,63	1,02	0,33	1,06
MAG6Q071_I_C	1775	69,35	-1,02	0,06	0,96	-1,54	0,96	0,42	1,43
MAG6Q081_I_C	1753	63,15	-0,68	0,06	0,98	-0,86	0,97	0,40	1,30
MAG6Q082_I_C	1745	66,19	-0,84	0,06	0,98	-0,83	0,98	0,39	1,30
MAG6Q101_I_C	1682	62,78	-0,64	0,06	1,00	0,13	1,01	0,36	1,14
MAG6Q16_I_S_C	1961	k.A.	-1,42	0,04	0,85	-3,37	0,69	0,51	1,68
MAG6Q191_I_C	1591	90,82	-2,60	0,09	1,02	0,26	1,15	0,22	0,93
MAG6Q231_I_C	1581	86,15	-2,08	0,08	0,92	-1,54	0,77	0,38	1,72
MAG6Q261_I_C	1569	67,56	-0,77	0,06	0,98	-0,83	0,95	0,41	1,25

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
MAG6Q281_I_C	1499	45,70	0,37	0,06	0,98	-0,74	1,00	0,42	1,14
MAG6Q291_I_C	1471	22,98	1,64	0,07	1,05	1,34	1,28	0,28	0,80
MAG6R041_I_C	2009	93,38	-3,27	0,10	0,98	-0,24	1,00	0,25	1,25
MAG6R051_I_C	3585	81,79	-1,84	0,05	0,98	-0,63	0,96	0,33	1,23
MAG6R061_I_C	3048	52,20	-0,10	0,04	0,95	-2,97	0,94	0,41	1,33
MAG6R111_I_C	3439	47,46	0,12	0,04	1,09	5,58	1,14	0,26	0,77
MAG6R221_I_C	1465	47,71	0,29	0,06	1,00	0,07	1,02	0,39	1,06
MAG6R241_I_C	1588	50,19	0,13	0,06	1,04	2,02	1,06	0,34	0,93
MAG6R251_I_C	1591	96,35	-3,67	0,14	1,00	0,02	0,93	0,17	1,14
MAG6R271_I_S_C[a]	1417	62,03	-0,43	0,06	0,87	-5,19	0,82	0,52	2,29
MAG6R273_I_S_C[a]	1350	57,48	-0,17	0,06	0,89	-4,57	0,86	0,50	2,07
MAG6V091_I_C	3431	41,68	0,42	0,04	0,99	-0,95	1,05	0,36	1,14
MAG6V092_I_C	3283	31,71	0,97	0,04	1,05	2,52	1,17	0,28	0,84
MAG6V141_I_C	3434	41,58	0,44	0,04	1,11	6,94	1,20	0,23	0,67
MAG6V171_I_C	1977	84,27	-2,16	0,07	0,99	-0,18	1,00	0,34	1,21

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
MAG6V181_I_C	1979	87,92	-2,51	0,08	1,03	0,64	1,04	0,26	1,03
MAG6V201_I_C	1454	41,61	0,62	0,06	1,05	2,03	1,07	0,32	0,89
MAG6V211_I_C	1589	88,29	-2,30	0,08	0,93	-1,33	0,75	0,36	1,80

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step". Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

[a] MAG6D131_I_S_C, MAG6D271_I_S_C, MAG6D273_I_S_C sind dichotome Items, die das Antwortformat der Freitextaufgaben aufweisen.

Tabelle 25: Schwellenparameter für das polytome Item im gesamten Mathematiktest

Item	1. Schwelle	(SE)	2. Schwelle
MAG6Q16_I_S_C	0,59	(0,07)	-0,59

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step" geschätzt.

Tabelle 26: Itemparameter für Testheft 2A (mathematische Kompetenz)

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
MAG6D011_I_C	1987	47,91	0,11	0,05	1,03	1,46	1,04	0,36	0,97
MAG6D131_I_S_C[a]	1851	72,88	-1,15	0,06	0,96	-1,43	0,90	0,40	1,42
MAG6D151_I_C	1988	90,14	-2,66	0,08	0,91	-1,57	0,72	0,39	1,91
MAG6Q021_I_C	2014	94,59	-3,39	0,10	0,96	-0,45	1,03	0,25	1,33
MAG6Q031_I_C	1984	70,46	-1,07	0,06	1,00	-0,02	1,00	0,36	1,11
MAG6Q071_I_C	1775	69,35	-0,90	0,06	0,96	-1,52	0,96	0,42	1,41
MAG6Q081_I_C	1753	63,15	-0,56	0,06	0,98	-0,84	0,97	0,40	1,28
MAG6Q082_I_C	1745	66,19	-0,73	0,06	0,98	-0,80	0,98	0,39	1,28
MAG6Q101_I_C	1682	62,78	-0,52	0,06	1,00	0,16	1,01	0,36	1,13
MAG6Q16_I_S_C	1961	k.A.	-1,36	0,04	0,85	-3,38	0,69	0,51	1,65
MAG6R041_I_C	2009	93,38	-3,15	0,10	0,98	-0,23	1,00	0,25	1,23

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
MAG6R051_I_C	2002	80,47	-1,73	0,06	0,98	-0,52	0,97	0,36	1,21
MAG6R061_I_C	1733	45,82	0,21	0,06	0,97	-1,55	0,99	0,42	1,24
MAG6R111_I_C	1966	43,90	0,31	0,05	1,10	4,93	1,17	0,25	0,69
MAG6V091_I_C	1927	37,73	0,63	0,05	1,01	0,27	1,12	0,34	1,01
MAG6V092_I_C	1846	27,79	1,19	0,06	1,03	1,13	1,19	0,29	0,86
MAG6V141_I_C	1924	38,20	0,62	0,05	1,11	5,01	1,22	0,23	0,62
MAG6V171_I_C	1977	84,27	-2,04	0,07	0,99	-0,15	1,01	0,34	1,18
MAG6V181_I_C	1979	87,92	-2,39	0,08	1,03	0,66	1,04	0,26	1,01

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; k.A. = keine Angabe; Keine Angabe der Lösungshäufigkeit für polytome Items; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem Partial-Credit-Modell mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step“. Der Diskriminationsparameter wurden mittels eines Generalized-Partial-Credit-Modells geschätzt, in dem neben den Itemschwierigkeits- und Lageparametern auch die Trennschärfe der Items geschätzt wird.

[a] MAG6D131_I_S_C ist ein dichotomes Item, das das Antwortformat der Freitextaufgaben aufweist.

Tabelle 27: Schwellenparameter für das polytome Item in Testheft 2A (mathematische Kompetenz)

Item	1. Schwelle	(SE)	2. Schwelle
MAG6Q16_I_S_C	0,58	(0,07)	-0,58

Anmerkungen. SE = Standardfehler. Die Schwellenparameter wurden mit der sogenannten ConQuest-Parameterisierung „Item + Item:Step“ geschätzt.

Tabelle 28: Itemparameter für Testheft 2B (mathematische Kompetenz)

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskriminations-parameter 2PL
MAG6D282_I_C	1497	74,68	-1,33	0,07	1,05	1,43	1,04	0,30	0,90
MAG6Q031_I_C	1580	74,18	-1,30	0,06	1,03	0,88	1,03	0,33	0,98
MAG6Q191_I_C	1591	90,82	-2,74	0,09	1,01	0,22	1,13	0,22	0,93
MAG6Q231_I_C	1581	86,15	-2,21	0,08	0,92	-1,58	0,76	0,38	1,73
MAG6Q261_I_C	1569	67,56	-0,91	0,06	0,98	-0,94	0,95	0,41	1,26
MAG6Q281_I_C	1499	45,70	0,22	0,06	0,98	-0,84	0,99	0,42	1,15
MAG6Q291_I_C	1471	22,98	1,49	0,07	1,05	1,31	1,27	0,28	0,79
MAG6R051_I_C	1583	83,45	-1,97	0,07	0,98	-0,40	0,94	0,32	1,23
MAG6R061_I_C	1315	60,61	-0,51	0,06	0,94	-2,25	0,91	0,46	1,38
MAG6R111_I_C	1473	52,21	-0,12	0,06	1,07	2,89	1,10	0,32	0,83

Item	N_g	% korrekt	Itemschwierigkeits-/Lageparameter	SE	WMNSQ	t	Outfit	Item-Skalen-Korrelation	Diskrimi-nations-parameter 2PL
MAG6R221_I_C	1465	47,71	0,14	0,06	1,00	-0,03	1,02	0,39	1,07
MAG6R241_I_C	1588	50,19	-0,01	0,06	1,04	1,90	1,06	0,34	0,93
MAG6R251_I_C	1591	96,35	-3,80	0,14	1,00	0,00	0,92	0,17	1,15
MAG6R271_I_S_C[a]	1417	62,03	-0,57	0,06	0,87	-5,28	0,82	0,52	2,31
MAG6R273_I_S_C[a]	1350	57,48	-0,32	0,06	0,89	-4,68	0,85	0,50	2,08
MAG6V091_I_C	1504	46,74	0,16	0,06	0,96	-1,62	0,98	0,43	1,25
MAG6V092_I_C	1437	36,74	0,69	0,06	1,07	2,78	1,16	0,31	0,77
MAG6V141_I_C	1510	45,89	0,22	0,06	1,11	4,86	1,17	0,27	0,68
MAG6V201_I_C	1454	41,61	0,47	0,06	1,05	1,93	1,06	0,32	0,89
MAG6V211_I_C	1589	88,29	-2,44	0,08	0,93	-1,35	0,75	0,36	1,81

Anmerkungen. N_g = Anzahl gültiger Antworten; % korrekt = Angabe der Lösungshäufigkeit in Prozent; SE = Standardfehler; WMNSQ = Weighted Mean Square Error als Maß für den Infit; t = t-Wert des Weighted Mean Square Error; Itemparameter und Fitstatistiken basieren auf einem eindimensionalen Rasch-Modell (1PL-Modell), da das Testheft keine polytomen Items enthält. Der Diskriminationsparameter wurden mittels eines zweidimensionalen Rasch-Modells (2PL-Modell) geschätzt, in dem neben den Itemschwierigkeitsparametern auch die Trennschärfe der Items geschätzt wird.

[a] MAG6D271_I_S_C, MAG6D273_I_S_C sind dichotome Items, die das Antwortformat der Freitextaufgaben aufweisen.

4.2.2 Personenparameter

Personenfähigkeiten wurden für den Lese- und den Mathematiktest als WLEs geschätzt (vgl. Pohl & Carstensen, 2012). Die WLEs, die einmal auf Basis des gesamten Lesetests und einmal auf Basis des gesamten Mathematiktests geschätzt wurden, werden in einem Scientific Usefile bereitgestellt.

5. Testqualität

5.1 Passung und Reliabilität

Die Abbildungen 26 bis 33 sind sogenannte Wright-Maps und zeigen eine Gegenüberstellung der geschätzten Personenfähigkeitsverteilungen auf der linken Seite und der Itemschwierigkeiten auf der rechten Seite auf derselben Logitskala. Unter dem Aspekt der Passung wird nachfolgend untersucht, inwieweit diese beiden Verteilungen deckungsgleich sind. Der Mittelwert der Personenfähigkeitsverteilung wurde bei der Skalierung auf 0 gesetzt. Die Wright-Maps für die Testheftversionen bilden die Fähigkeiten aller Schülerinnen und Schüler ab, die die jeweilige Version bearbeitet haben, also auch derjenigen, denen das Testheft nicht aufgrund einer Gruppenzugehörigkeit, sondern zufällig zugewiesen wurde. Es ist zu erwarten, dass die Passung zwischen den Fähigkeiten der Schülerinnen und Schüler und den Itemschwierigkeiten in den zufällig zugewiesenen Testheftversionen geringer ist, da die Gruppenzugehörigkeit nicht als Proxy für das Fähigkeitsniveau genutzt werden konnte. Bei der Interpretation der Passung und Reliabilität werden dennoch die Daten aller Schülerinnen und Schüler berücksichtigt.

Die WLE-Varianz für den gesamten Lesetest beträgt 1,90, für Testheft 1A 2,03, für Testheft 1B 1,74, für Testheft 1C 1,59 und für Testheft 1D 1,94. Für den gesamten Mathematiktest beträgt die WLE-Varianz 1,66, für Testheft 2A 1,69 und für Testheft 2B 1,58. Anhand der Wright-Maps kann die Passung der Testschwierigkeit für die vorliegende Stichprobe grafisch eingeschätzt werden. Sie zeigen, inwiefern und in welcher Anzahl die Testaufgaben das Spektrum der Personenfähigkeiten abdecken. Grundsätzlich gilt, dass ein Test besonders gut in Fähigkeitsbereichen differenziert, in denen das Niveau der Itemschwierigkeiten ungefähr dem Niveau der Personenfähigkeiten entspricht.

Es wird deutlich, dass die Passung der Personenfähigkeiten zu den Itemschwierigkeiten sowohl für die Lesekompetenzmessung als auch für die Messung der mathematischen Kompetenz angemessen ist (siehe Abbildungen 26 und 31). Es liegt weder eines der 66 Items des Lesetests noch der 32 Items des Mathematiktests außerhalb der Personenfähigkeitsverteilung der Stichprobe. Die Verteilungen der Personenfähigkeiten im obersten Leistungsbereich übersteigen jedoch etwas die Itemschwierigkeiten der schwersten Items in beiden Gesamttests und in den jeweils schwereren Testheftversionen (Testhefte 1C und 1D des Lesetests und Testheft 2B des Mathematiktests) (siehe Abbildungen 26, 29 bis 31, und 33). Dies bedeutet, dass die Messgenauigkeit sowohl des Lese- als auch des Mathematiktests im obersten Leistungsbereich leicht eingeschränkt ist und in diesem Fähigkeitsbereich eine Differenzierung zwischen den leistungsstärksten Schülerinnen und Schüler (Schülerinnen und Schüler mit „überdurchschnittlich ausgeprägter Kompetenz“ vs. „weit überdurchschnittlich ausgeprägte Kompetenz“) nur bedingt möglich ist. Sowohl der Lesetest als auch der Mathematiktest differenzieren sehr gut im unteren und mittleren

Leistungsbereich. Die Testpassung fällt somit insgesamt besser für Schülerinnen und Schüler in unteren und mittleren Leistungsbereichen als für die leistungsstärksten Schülerinnen und Schüler aus. Dieses Ergebnis entspricht einem zentralen Ziel der INSIDE-Kompetenzmessung, nämlich die Lesekompetenz und die mathematische Kompetenz von Schülerinnen und Schülern mit sonderpädagogischen Förderbedarfen gemeinsam mit denen von Schülerinnen und Schülern ohne sonderpädagogischen Förderbedarf differenziert zu erfassen.

Die EAP (*Expected A Posteriori*)-Reliabilität und WLE-Reliabilität sind zwei Maße zur Einschätzung der Konsistenz der Personenfähigkeitsschätzung. Zur Einschätzung der Reliabilitätsmaße werden die Ergebnisse der NEPS-Lesestudien herangezogen. Für die Jahrgangsstufe 5 lag die EAP-Reliabilität bei 0,81 (bei einem Testheft) und die WLE-Reliabilität bei 0,77 (Pohl et al., 2012). Für die Jahrgangsstufe 7 der NEPS-Erhebung lag die EAP-Reliabilität bei 0,83 (bei zwei schwierigkeitsgestuften Testheften) und die WLE-Reliabilität bei 0,79 (Krannich et al., 2017). Da die verwendeten Items aus dem Itempool des NEPS stammen, wird erwartet, dass die Reliabilitätswerte ähnlich zu denen der NEPS-Lesestudien in den Jahrgangsstufen 5 und 7 ausfallen. Auch sollte das komplexe querschnittliche Linkingdesign mit Ankeritems über vier schwierigkeitsgestufte Testheftversionen die Reliablitätswerte stabilisieren.

Die EAP-Reliabilität für alle Leseitems der INSIDE-Studie beträgt in der gemeinsamen freien Skalierung 0,80 (für vier Testheftversionen), für die Leseitems in Testheft 1A 0,79, in Testheft 1B 0,79, in Testheft 1C 0,80 und in Testheft 1D 0,81. Die WLE-Reliabilität beträgt für die Personenfähigkeitsschätzung aller Schülerinnen und Schüler für den gesamten Lesetest 0,75 (für vier Testheftversionen), für die Schülerinnen und Schüler in Testheft 1A 0,71, in Testheft 1B 0,75, in Testheft 1C 0,76 und in Testheft 1D 0,74. Die Reliabilitätswerte für den gesamten Lesetest liegen somit nah am erwarteten Bereich. Für einzelne Testhefte fallen die Werte etwas niedriger aus.

Auch für die Bewertung der Reliabilität der Personenfähigkeitsschätzung auf Basis des Mathematiktests dienen NEPS-Ergebnisse als Orientierung. Für die Jahrgangsstufe 5 lag die EAP- Reliabilität bei 0,80 und die WLE-Reliabilität bei 0,78 (Duchhardt & Gerdes, 2012). Für die Jahrgangsstufe 7 lag die EAP- Reliabilität bei 0,76 und die WLE-Reliabilität bei 0,72 (Schnittjer & Gerken, 2017).

Da die verwendeten Items im Mathematiktest ebenfalls aus dem Itempool des NEPS stammen, wird erwartet, dass die Reliabilitätswerte ähnlich zu denen der NEPS-Mathematikstudien in den Jahrgangsstufen 5 und 7 ausfallen.

Die EAP-Reliabilität beträgt für alle Mathematikitems in der gemeinsamen freien Skalierung 0,77 (für zwei Testheftversionen), in Testheft 2A 0,76 und in Testheft 2B 0,78. Die WLE-Reliabilität beträgt für die Personenfähigkeitsschätzung aller Schülerinnen und Schüler für den gesamten Mathematiktest 0,74 (für zwei Testheftversionen), für die Schülerinnen und Schüler in Testheft 2A 0,72 und in Testheft 2B 0,75. Die Reliabilitätswerte für den gesamten Mathematiktest und für die beiden einzelnen Testheftversionen liegen somit innerhalb des erwarteten Bereichs.

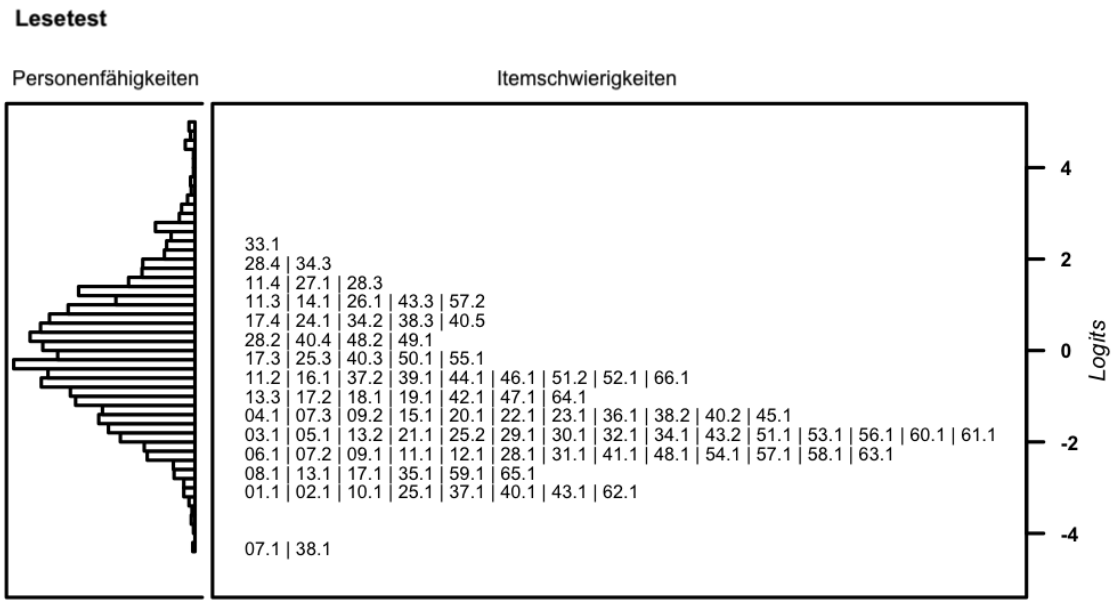

Abbildung 26. Gesamter Lesetest.

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 14. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 243 Schülerinnen und Schüler.

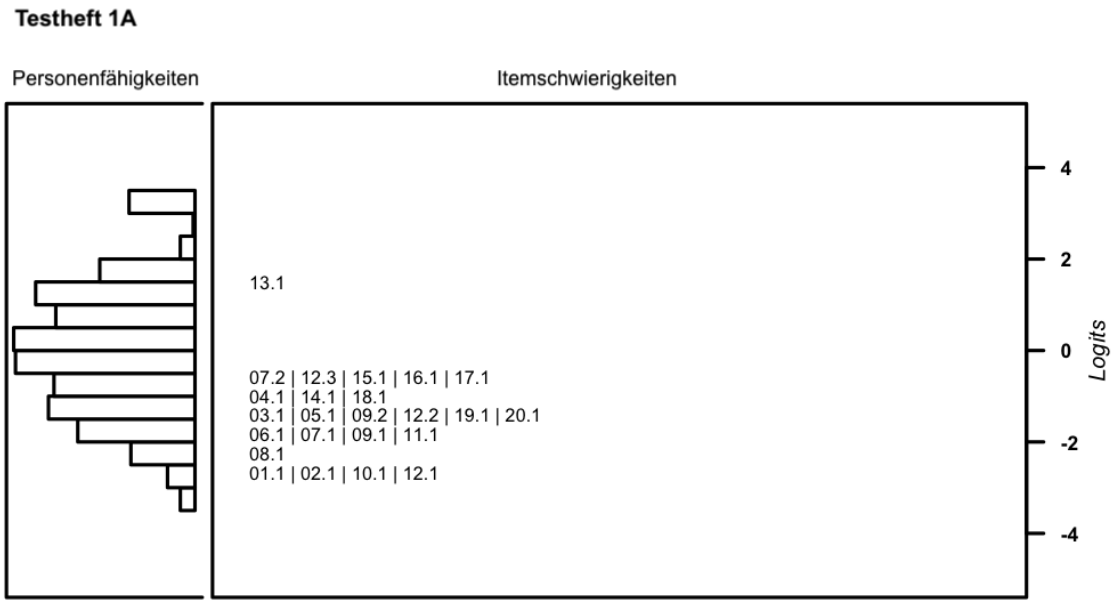

Abbildung 27. Testheft 1A (Lesekompetenz).

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 16. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 99 Schülerinnen und Schüler.

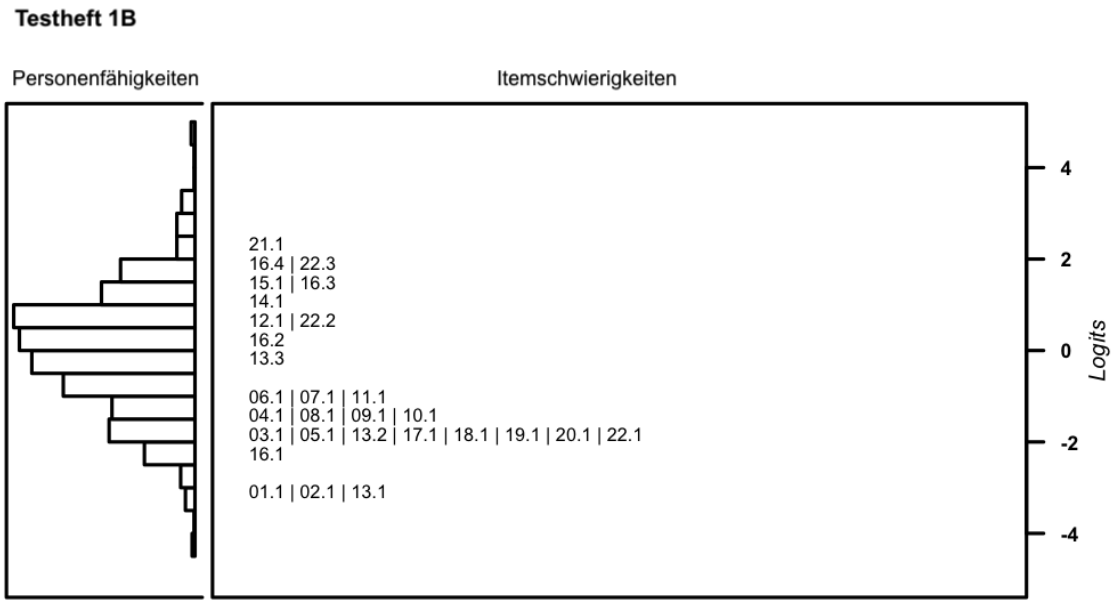

Abbildung 28. Testheft 1B (Lesekompetenz).

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 18. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 190 Schülerinnen und Schüler.

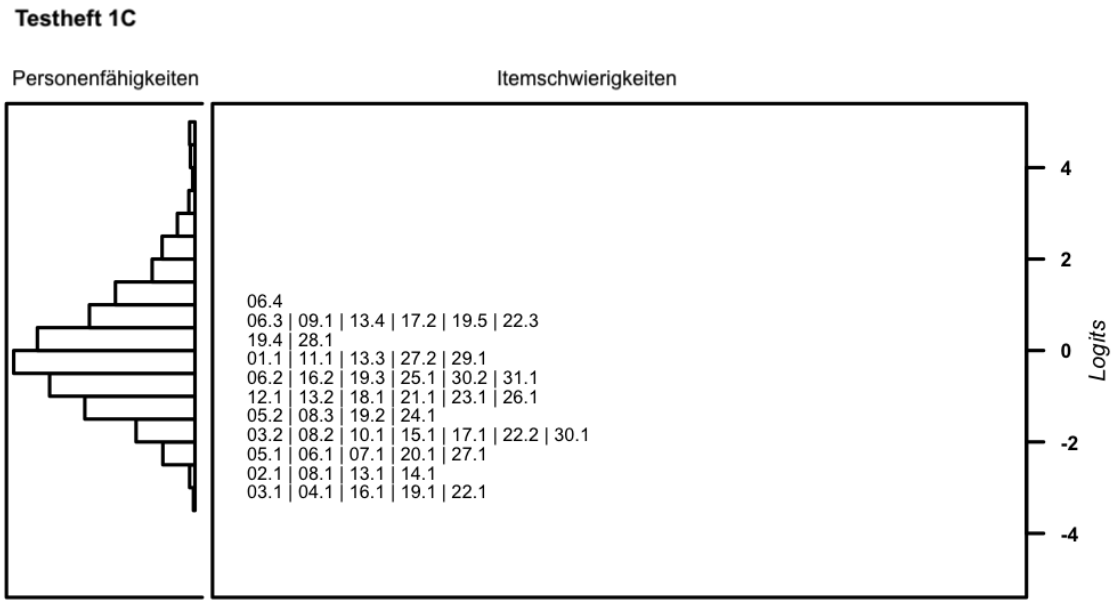

Abbildung 29. Testheft 1C (Lesekompetenz).

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 20. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 237 Schülerinnen und Schüler.

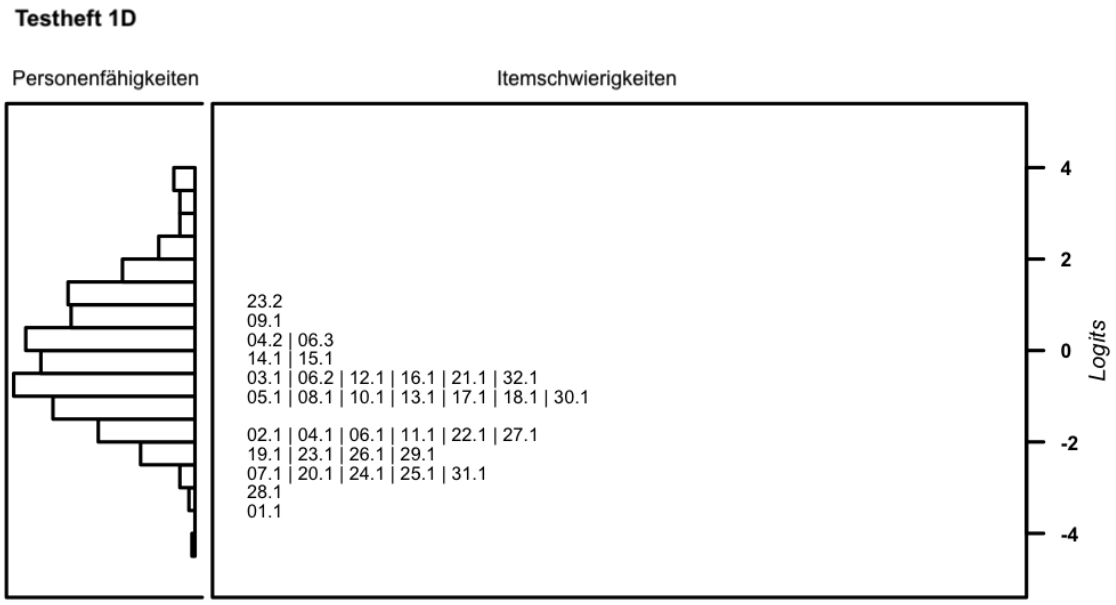

Abbildung 30. Testheft 1D (Lesekompetenz).

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 22. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 60 Schülerinnen und Schüler.

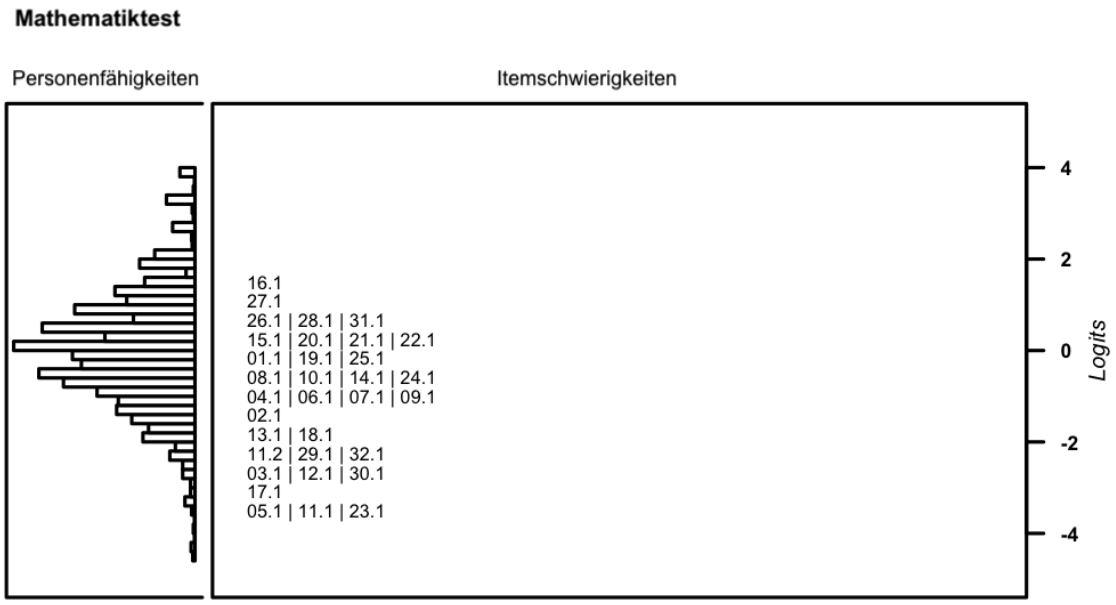

Abbildung 31. Mathematiktest.

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 24. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 324 Schülerinnen und Schüler.

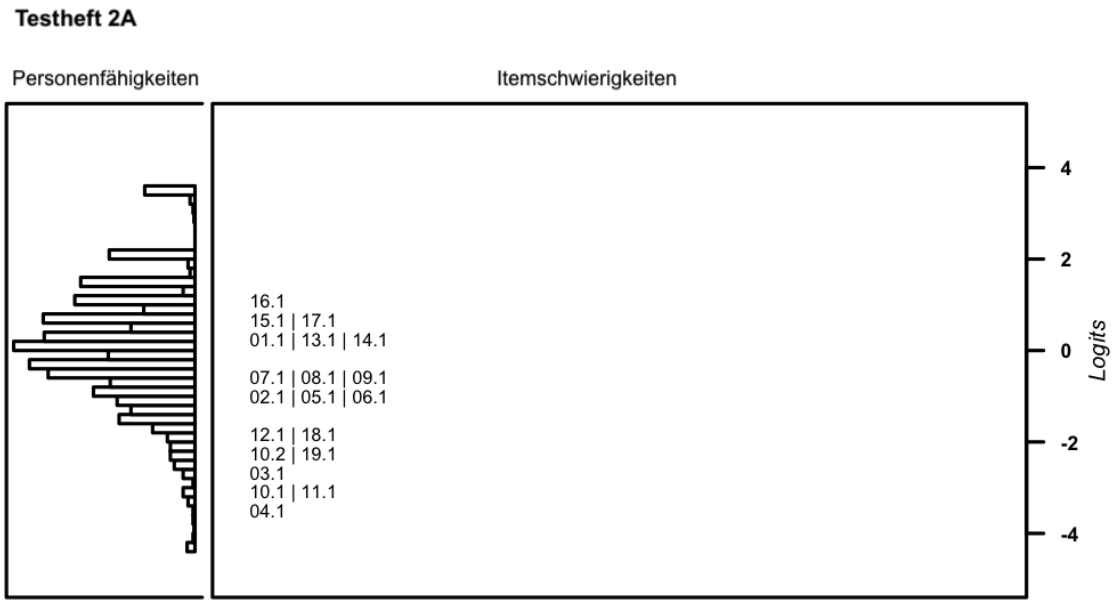

Abbildung 32. Testheft 2A (mathematische Kompetenz).

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 26. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 184 Schülerinnen und Schüler.

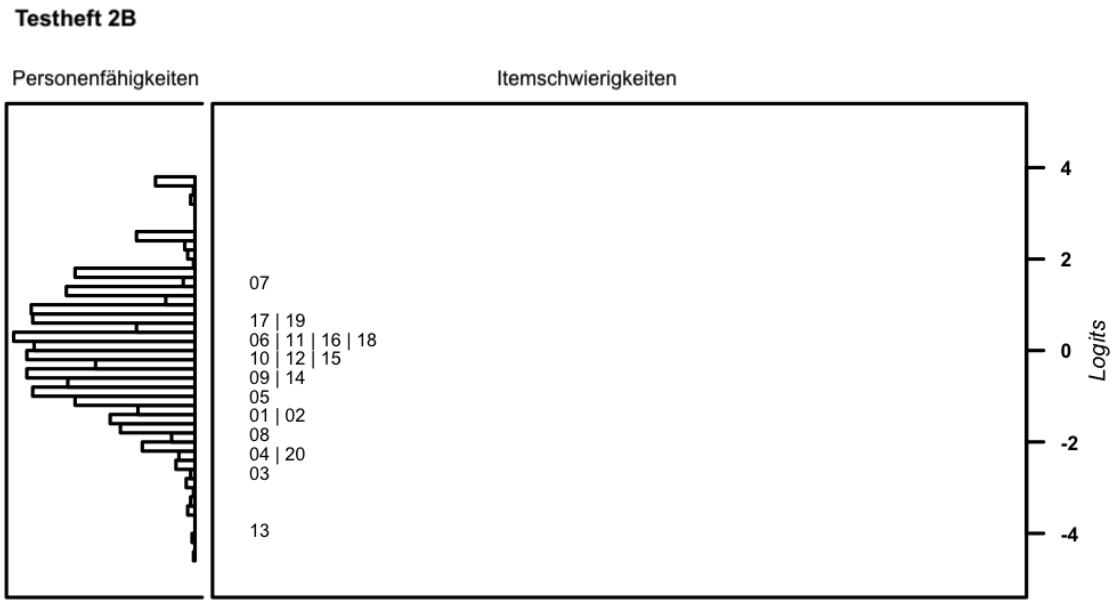

Abbildung 33. Testheft 2B (mathematische Kompetenz).

Anmerkungen. Die Fähigkeitsverteilung der Schülerinnen und Schüler ist auf der linken Seite und die ermittelten Itemschwierigkeiten sind auf der rechten Seite abgebildet. Die Itemziffern entsprechen der Reihenfolge der Items in Tabelle 28. Der längste Balken auf dem Abschnitt der Personenfähigkeiten repräsentiert 124 Schülerinnen und Schüler.

5.2 Itemfit

Der Itemfit wird auf Basis des Partial-Credit-Modells für die Skalierung des gesamten Lesetests bzw. Mathematiktests überprüft. Ein Infit (WMNSQ) eines Items nahe dem Wert 1 signalisiert einen guten Itemfit. Werte deutlich über 1 sind als problematischer einzuschätzen als Werte deutlich unter 1, weil sie einen sog. Underfit und damit mangelnde Messgüte signalisieren. Dies gilt für die Itemfits beider Tests. Der Großteil der Items sowohl des Lesetests als des Tests für die mathematische Kompetenz liegt innerhalb des erwarteten Bereichs. Der geringste WMNSQ eines Leseitems liegt bei 0,82 (Item REG60930_I_C) und der höchste WMNSQ liegt bei 1,14 (Item REG60540_I_C) (vgl. Tabelle 14). Der geringste WMNSQ eines Mathematikitems liegt bei 0,85 (Item MAG6Q16_I_S_C) und der höchste WMNSQ liegt bei 1,11 (Item MAG6V141_I_C) (vgl. Tabelle 24). Schon bei der Itemselektion wurden ausschließlich Items in die Gesamtskalierung für alle Leseitems bzw. alle Mathematikitems aufgenommen, die einen WMNSQ < 1,20 auf Test- und Testheftebene und damit einen unproblematischen Itemfit aufwiesen (vgl. Pohl & Carstensen, 2012).

Residuale Korrelationen zwischen Itempaaren ermöglichen den lokalen Itemfit einzuschätzen. Mithilfe der residualen Korrelationen können potenzielle lokale Abhängigkeiten zwischen Itempaaren aufgedeckt werden, die vermieden werden sollten. Damit wird eine Vorannahme der Skalierung mittels Modellen der Item-Response-Theorie überprüft, die besagt, dass die Wahrscheinlichkeit ein Item korrekt zu beantworten nicht von der Antwortgenauigkeit für ein

anderes Item abhängen darf. Die residualen Korrelationen zwischen Itempaaren werden mittels Yen's Q3-Statistik berichtet (Yen, 1984). Als Daumenregel gilt, dass die residualen Korrelationen zwischen Itempaaren nicht kleiner als -0,2 oder größer als 0,2 ausfallen sollten (Chen & Thissen, 1997). Der Betrag der residualen Korrelationen zwischen Itempaaren (aQ3) im gesamten Lesetest liegt im Mittel bei 0,04 (SD = 0,04) und ist somit als sehr gut zu bewerten. Die für einzelne Leseitems gemittelten residualen Korrelationen liegen im Betrag zwischen 0,02 und 0,08. Der Betrag der residualen Korrelationen zwischen Itempaaren (aQ3) im gesamten Mathematiktest liegt im Mittel bei 0,03 (SD = 0,04). Die für einzelne Mathematikitems gemittelten residualen Korrelationen liegen im Betrag zwischen 0,02 und 0,07. Insgesamt fallen die Maße des lokalen Itemfits in beiden Tests unproblematisch aus und es konnten auch keine unerwünschten, lokalen Abhängigkeiten zwischen Itempaaren festgestellt werden.

6. Zitation

Stegenwallner-Schütz, M., Obry, M., Wittmann, E., Gehrer, K., Nusser, L. & Böhme, K. (2022). INSIDE-Studie. *Dokumentation der Skalierung der Kompetenzmessungen in den Bereichen Lesen und Mathematik des ersten Messzeitpunkts in der Jahrgangsstufe 6 (Kohorte 1)* (NEPS Working Paper Nr. 108). Leibniz-Institut für Bildungsverläufe.

7. Beiträge zur Analyse der hier dokumentierten Daten

Die Entwicklung des INSIDE-Lesetests und -Mathematiktests sowie des Linkdesigns in Jahrgangsstufe 6 erfolgte (in alphabetischer Reihenfolge) durch Katrin Böhme, Karin Gehrer, Oksana Kerbs, Lena Nusser, und Elena Wittmann, mit Beratung durch Claus Carstensen.

Die Skalierung des INSIDE-Lese- und Mathmatiktests in der Jahrgangsstufe 6 erfolgte (in alphabetischer Reihenfolge) durch Michael Obry, Maja Stegenwallner-Schütz und Elena Wittmann, mit Beratung durch Timo Gnambs.

8. Referenzen

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265–289.

Duchhardt, C., & Gerdes, A. (2012). *NEPS technical report for mathematics - Scaling results of starting cohort 3 in fifth grade* (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Gehrer, K., & Artelt, C. (2013). Literalität und Bildungslaufbahn: Das Bildungspanel NEPS. [Literacy and educational career: The NEPS educational panel]. In A. Bertschi-Kaufmann & C. Rosebrock (Eds.), *Literalität erfassen: bildungspolitisch, kulturell, individuell.* (pp. 168–187). Juventa.

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2012). The assessment of reading competence (Including sample items for Grade 5 and 9) Status: 2012. *NEPS Research Data Paper*. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_re_2012_en.pdf

Gehrer, K., Zimmermann, S., Artelt, C., & Weinert, S. (2013). NEPS framework for assessing reading competence and results from an adult pilot study. *Journal for Educational Research Online*, *5*, 50–79.

Krannich, M., Jost, O., Rohm, T., Koller, I., Pohl, S., Haberkorn, K., Carstensen, C. H., Fischer, L., & Gnambs, T. (2017). *NEPS technical report for reading – Scaling results of starting cohort 3 for grade 7* (Update NEPS Survey Paper No. 14). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP14:2.0.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica*, *47*, 149–174.

Neumann, I., Duchhardt, C., Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal of Educational Research Online*, *5*(2), 80–109.

Nusser, L., Weinert, S., Artelt, C., & Carstensen, C. H. (2020). Machbarkeitsstudien an Förderschulen mit dem Schwerpunkt Lernen im Rahmen des Nationalen Bildungspanels (2010 bis 2013) – Ergebnisse und Resümee. In C. Gresch, P. Kuhl, M. Grosche, C. Sälzer, & P. Stanat (Eds.), *Schüler*innen mit sonderpädagogischem Förderbedarf in Schulleistungserhebungen - Einblicke und Entwicklungen* (pp. 147–175). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27608-9

Pohl, S., & Carstensen, C. H. (2012). *NEPS technical report - scaling the data of the competence tests* (NEPS Working Paper No. 14). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). *NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade* (NEPS Working Paper No. 15). Otto-Friedrich-Universität, Nationales Bildungspanel.

R Core Team. (2015). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.*

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules (R package version 3.5-19)*. https://cran.r-project.org/package=TAM

Schmitt, M., Roßbach, H.-G., Gresch, C., Stanat, P., Böhme, K., Grosche, M., Labsch, A., Külker, L., Michel, A., Stegenwallner-Schütz, M., & Schledjewski, J. (2020). Projekt: Inklusion in der Sekundarstufe I in Deutschland (INSIDE). *Erziehungswissenschaft*, *31*, 199–202. https://doi.org/10.3224/ezw.v31i1.30

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS technical report for mathematics: Scaling results of starting cohort 3 in grade 7* (NEPS Survey Paper No. 16). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2020). *BEFKI 5-7. Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 5. bis 7. Jahrgangsstufe*. Hogrefe.

Warm, T. A. (1989). Weighted likelihood estimation of ability in Item Response Theory. *Psychometrica*, *54*, 427–450.

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. https://doi.org/10.1177/014662168400800201

Zimmermann, S., Gehrer, K., Artelt, C., & Weinert, S. (2012). *The assessment of reading speed in grade 5 and grade 9 Status: 2012*. https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC4/1-0-0/com_rs_2012_en.pdf