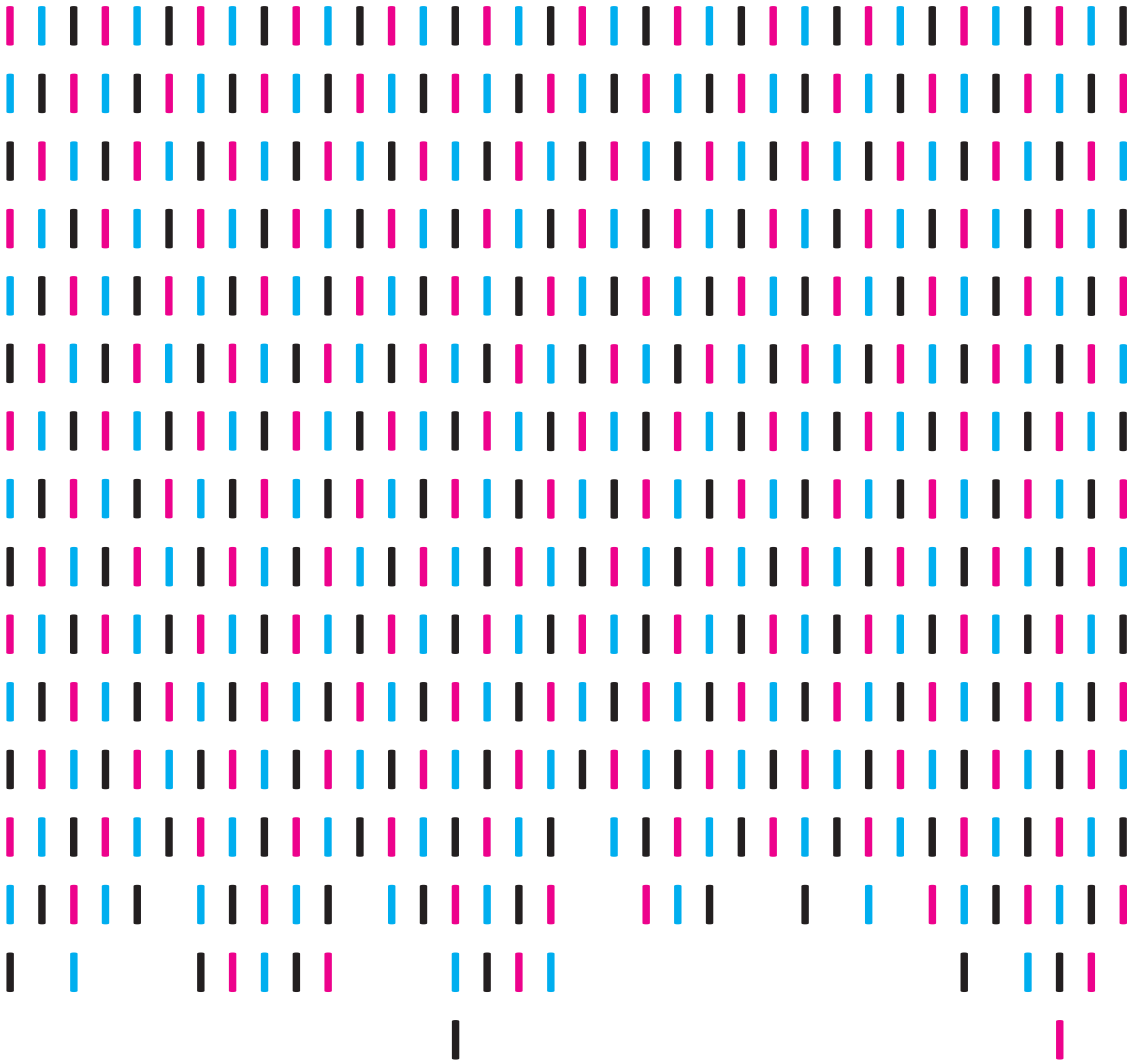

LIFBI *WORKING PAPERS*

Michael Obry, Anike Schild, Gisela Will
und Florian Kopp

DIE MESSUNG DES REZEPTIVEN WORTSCHATZES IN DER FLÜCHTLINGSSTUDIE REGES (WELLE 1)

NEPS Working Paper No. 98
Bamberg, Juli 2021

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

Die Messung des rezeptiven Wortschatzes in der Flüchtlingsstudie ReGES (Welle 1)

Michael Obry, Anike Schild, Gisela Will und Florian Kopp
Leibniz-Institut für Bildungsverläufe

E-Mail-Adresse des Erstautoren:

michael.obry@lifbi.de

Bibliographische Angaben:

Obry, M., Schild, A., Will, G. & Kopp, F. (2021). *Die Messung des rezeptiven Wortschatzes in der Flüchtlingsstudie ReGES (Welle 1)* (LIfBi Working Paper No. 98). Leibniz-Institut für Bildungsverläufe. https://doi.org/10.5157/LIfBi:WP98:1.0

Acknowledgement:

Das diesem Working Paper zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen FLUCHT03 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autorinnen und den Autoren.

Wir orientieren uns an bereits veröffentlichten NEPS Survey Papers zur Skalierung der Kompetenztests des Leibniz-Instituts für Bildungsverläufe. Abschnitte zu Analysen, die in gleicher Weise durchgeführt wurden, sind teilweise von diesen übernommen.

Die Messung des rezeptiven Wortschatzes in der Flüchtlingsstudie ReGES (Welle 1)

Zusammenfassung

Der Spracherwerb ist eine notwendige Voraussetzung für die Integration von Migrantinnen und Migranten in das Aufnahmeland. Daher wurden in der Studie Refugees in the German Educational System (ReGES) deutsche Sprachkompetenzen gemessen, da diese als Indikator für eine gelungene kulturelle Integration angesehen werden können (vgl. Esser, 2006). Zur Erfassung des rezeptiven Wortschatzes wurde eine Adaptation der deutschen Version des Peabody Picture Vocabulary Tests Version 4 (PPVT-4; Lenhard, Lenhard, Segerer & Suggatte, 2015) eingesetzt. Dieses Manuskript beschreibt den Testablauf und die erhobenen Daten sowie die Evaluierung der Qualität der Testergebnisse. Im Vordergrund steht hierbei eine Stichprobe geflüchteter Kinder und Jugendlicher, die zwischen 2014 und 2017 nach Deutschland kamen. Die Prüfung des Messmodells erfolgte mit ein- und zwei-parametrischen logistischen Testmodellen anhand von Item-Fit-Statistiken und der Untersuchung von Differential Item Functioning, Rasch-Homogenität sowie Eindimensionalität. Insgesamt war die Reliabilität des Tests sehr gut und die Anpassung des Modells für die meisten Items gut bis akzeptabel. Die Varianz des Personenparameters wies auf eine gute Differenzierung zwischen den Kindern und zwischen den Jugendlichen hin. Zudem konnte die Testfairness für alle betrachteten Untergruppen bestätigt werden. Im Großen und Ganzen hat der Test zur Erfassung des rezeptiven deutschen Wortschatzes zufriedenstellende psychometrische Eigenschaften, was die Schätzung reliabler Kompetenzwerte für die Stichprobe erlaubt.

Schlagworte

PPVT-4, Messung, Skalierung, rezeptiver Wortschatz, Sprachtest, Deutschkompetenz, Flüchtlinge

Abstract

Language acquisition is a necessary prerequisite for the integration of migrants into the host country. Therefore, German language competencies were measured in the study Refugees in the German Educational System (ReGES), as they can be considered an indicator of successful cultural integration (cf. Esser, 2006). To measure receptive vocabulary, an adaptation of the German version of the Peabody Picture Vocabulary Test Version 4 (PPVT-4; Lenhard, Lenhard, Segerer & Suggatte, 2015) was used. This manuscript describes the test procedure and the data collected as well as the evaluation of the quality of the test results. The focus here is on a sample of refugee children and adolescents who arrived in Germany between 2014 and 2017. The measurement model was tested with one- and two-parametric logistic test models using item fit statistics and examining differential item functioning, Rasch homogeneity, and unidimensionality. Overall, the reliability of the test was very good, and the fit of the model was good to acceptable for most items. The variance of the person parameter indicated good differentiation between children and between adolescents. In addition, test fairness was confirmed for all subgroups considered. Overall, the test of receptive German vocabulary has

satisfactory psychometric properties which allows the estimation of reliable proficiency scores for the sample.

Keywords

PPVT-4, Measurement, Scaling, receptive vocabulary, language test, German language competence, refugees

Inhalt

1. Einleitung

Ziel der Studie *Refugees in the German Educational System* (ReGES) ist einerseits die Beschreibung der Situation von jungen Geflüchteten im Hinblick auf verschiedene Dimensionen der Integration, insbesondere der Teilhabe am Bildungssystem, andererseits die Analyse von Bedingungsfaktoren für die Bildungsverläufe von geflüchteten Kindern und Jugendlichen (vgl. Will, Balaban, Dröscher, Homuth & Welker, 2018; Will, Gentile, Heinritz & von Maurice, 2018). Zur Erreichung dieser Ziele ist die Erfassung von Sprachkompetenzen in der Sprache des Aufnahmelandes zentral. Sprachkompetenzen in der Sprache des Aufnahmelandes können selbst als Indikator für gelungene kulturelle Integration angesehen werden. Darüber hinaus ist der Spracherwerb eine notwendige Voraussetzung für die Integration ins Aufnahmeland und sprachliche Fähigkeiten beeinflussen den Bildungserfolg maßgeblich (vgl. Esser, 2006).

In ReGES wurde unter anderem der rezeptive Wortschatz erfasst, da dieser einer der besten Indikatoren für Sprachkompetenzen, insbesondere für die Vorhersage von Hörverständnis und Textverständnis, ist (vgl. Berendes, Weinert, Zimmermann & Artelt, 2013). Die Erweiterung des Wortschatzes hängt von der Quantität und Qualität der Sprachrezeption ab (vgl. Huttenlocher, 1998) und wird daher nicht nur durch den Sprachgebrauch in der Familie (vgl. Leseman, Scheele, Mayo, & Messer, 2007) sondern auch durch die Qualität der Bildungsinstitutionen (vgl. Roßbach & Weinert, 2008) wesentlich beeinflusst (vgl. Berendes et al., 2013).

Erhoben wurde der rezeptive Wortschatz innerhalb der ReGES-Studie mit einer adaptierten, deutschen Variante der vierten Ausgabe des Peabody Picture Vocabulary Tests (PPVT-4; Lenhard, Lenhard, Segerer & Suggatte, 2015). Dieser wurde in Welle 1 zwischen Ende 2017 und Mitte 2018 sowohl in einer Kohorte im Kindergartenalter (RC1) als auch in einer Kohorte im Jugendalter (RC2) eingesetzt. Eine Wiederholungsmessung fand für einen Teil der Stichprobe in Welle 7 im Jahr 2020 statt. Aufgrund der Einschränkungen durch das Virus SARS-CoV-2 wurde die Feldphase jedoch unterbrochen und alle weiteren Interviews wurden telefonisch geführt. Die computergestützten Testungen konnten in diesen Fällen nicht durchgeführt werden (vgl. Will, Becker & Weigand, 2020).

Im vorliegenden Working Paper wird zunächst der Peabody Picture Vocabulary Test (PPVT-4) dargestellt und die Anpassungen, die für die ReGES-Studie vorgenommen wurden, beschrieben (Kapitel 2). Anschließend wird das Vorgehen zur Messung der Kompetenzen in der ersten Welle in ReGES sowie die Einbettung der Testdurchführung in die Gesamterhebung dargestellt (Kapitel 3). In diesem Kontext beschreiben wir auch die Stichprobenzusammensetzung, insbesondere im Hinblick auf den durchgeführten Test. Anschließend werden deskriptive Analysen der Testergebnisse (Kapitel 4) dargestellt. Während in Kapitel 5 die verschiedenen Analysestrategien beschrieben werden, werden in Kapitel 6 die Ergebnisse zu fehlenden Werten, Itemparametern, Test Targeting und der Qualität des Tests dargestellt. In Kapitel 7 werden die Ergebnisse und das Potenzial der Daten diskutiert. Zuletzt werden in Kapitel 8 die Variablen, die im Scientific Use File veröffentlicht werden, beschrieben.

2. Instrument und Anpassung an die ReGES-Studie

2.1 PPVT-4

Der rezeptive Wortschatz wurde mithilfe des PPVT-4 (Lenhard et al., 2015) gemessen. Im Folgenden gehen wir auf die wichtigsten Eigenschaften des Instrumentes und die Anpassungen für die Erhebung ein.

Der PPVT-4 besteht aus insgesamt 228 Items und untergliedert sich in 19 Itemsets mit jeweils 12 Items, wobei die Sets in aufsteigender Schwierigkeit angeordnet sind (vgl. Lenhard et al., 2015). Jedes Item besteht aus vier farbigen Zeichnungen und einem Teststimulus in Form eines gesprochenen Wortes. Durch Zeigen auf eine der Zeichnungen ordnet die Testperson das Wort zu. Nur eine der Zeichnungen entspricht dem Stimulus, bei den restlichen drei Bildern handelt es sich um Distraktoren. Um Effekte, die auf selektives Vorwissen zurückgehen zu minimieren, wurden die Testitems thematisch aus verschiedenen Inhaltskategorien entwickelt (vgl. Lenhard et al., 2015).

Der PPVT-4 beginnt mit einer Übungsphase. Übungsitems sollen zum einen die Testpersonen in den Test einführen und zum anderen anzeigen, welche Testpersonen die Voraussetzungen erfüllen, um mit dem PPVT-4 getestet zu werden. Die anschließende Testphase ist so aufgebaut, dass den Probandinnen und Probanden nur Itemsets angezeigt werden, die eine angemessene Schwierigkeit für sie aufweisen. Um jeder Probandin und jedem Probanden die Itemsets für ihren oder seinen Leistungsbereich vorzugeben, wird mit einem für die Altersgruppe geeigneten Set begonnen und es werden Boden- und Deckenset ermittelt. Das Bodenset ist das niedrigste vorgegebene Itemset, in dem die Person keinen oder höchstens einen Fehler macht. Das Deckenset ist das Set, bei dem die Probandin oder der Proband erstmalig mehr als sieben Fehler macht. Sind Boden- und Deckenset ermittelt, wird der Test beendet (vgl. Lenhard et al., 2015).

2.2 Computerbasierte Umsetzung und Anpassungen des Instrumentes

Um Interviewereffekte zu minimieren, wurde ein technologiebasiertes Testverfahren (TBT) eingesetzt. Diese digitale Version wurde entsprechend der Vorgaben im Testmanual programmiert, ausgenommen die unten beschriebenen Abweichungen.

Die Testungen wurden auf Tablet-PCs durchgeführt, welche sowohl wie ein Laptop, als auch als Tablet genutzt werden können, da der Bildschirm abgenommen werden kann. Die Antworten erfolgten durch Auswahl der vorgegebenen Zeichnungen über Touchscreen. Instruktionen und Stimuli wurden standardisiert als aufgezeichnete Sprachtexte präsentiert. Um mögliche Ablenkungen zu minimieren, wurden dazu Kopfhörer verwendet.

Die im Testmanual festgelegten altersgemäßen Startsets wurden anhand einer aus vorwiegend Muttersprachlerinnen und Muttersprachlern bestehenden Stichprobe normiert (vgl. Lenhard et al., 2015). Da die geflüchteten Kinder und Jugendlichen noch nicht lange Deutsch lernten, war davon auszugehen, dass die altersgemäßen Startsets für diese Stichprobe zu schwierig sein könnten. Nach Rücksprache mit Prof. Dr. Wolfgang Lenhard, einem der Autoren

der deutschen Fassung des PPVT-4, entschieden wir uns daher, die Kohorte im Kindergartenalter (RC1) von Set 1 und die Kohorte im Jugendalter (RC2) von Set 5 starten zu lassen[1]. Laut Testmanual starten mit Set 1 in der Regel Kinder, die jünger als vier Jahre sind, mit Set 5 starten in der Regel 6-jährige Kinder (vgl. Lenhard et al., 2015). Ansonsten erfolgte der Testablauf wie im Testmanual beschrieben.

Um Retraumatisierungen vorzubeugen, wurden Items aus dem PPVT-4 überprüft und gegebenenfalls durch andere ersetzt. Das Zielwort „Detonation" (Item 196) wurde durch das Zielwort „Eruption" ersetzt[2]. Zudem wurden bei drei Items aus demselben Grund jeweils einer der Distraktoren durch ein anderes Bild ersetzt[3].

Die deutschen Instruktionen des Testprogrammes wurden in drei weitere Sprachen übersetzt (Arabisch, Englisch und Kurmandschi) und ebenfalls als Audioinstruktionen angeboten.

3. Durchführung und Stichprobe (Welle 1)

Im Folgenden beschreiben wir zunächst das Vorgehen in der ersten Erhebungswelle. Dafür gehen wir auf den Ablauf, die Befragungsinhalte und die Erhebungs-Modi ein. Es folgt die Prozessbeschreibung der Testung für die Kinder und Jugendlichen. Der Ablauf der digitalen Testung entspricht zum großen Teil den Vorgaben im Testmanual. Wir erklären anschließend, wie die Auswahl an Fällen für die folgenden Analysen zustande kommt und beschreiben diese Stichprobe.

3.1 Ablauf der ersten Erhebungswelle

Im ersten Schritt der Erhebung wurden die Familien persönlich kontaktiert sowie die jeweilige Kontaktperson und Befragungssprache ermittelt. Im zweiten Schritt wurde ermittelt, welche Kinder und Jugendlichen einer Familie zur Zielgruppe gehören und die Einverständnisse zur Teilnahme und gegebenenfalls Panelbereitschaft eingeholt. Daraufhin wurden Eltern sowohl der Kohorte im Kindergartenalter als auch der im Jugendalter sowie die Jugendlichen selbst gebeten, Fragebogen am Tablet zu beantworten. In der Befragung wurden verschiedene Themenbereiche abgedeckt, wie zum Beispiel die Fluchtbiografie der Familie, die Sprache der Eltern und Kinder, die Betreuungs- und Schulgeschichte sowie Persönlichkeit, Bleibewunsch und Zufriedenheit der Eltern und Kinder. Anschließend wurden, wenn entsprechende Einverständnisse vorlagen und Kenntnisse einer der Instruktionssprachen vorhanden waren, die kognitiven Grundfähigkeiten und Sprachkenntnisse (inkl. PPVT-4) der Kinder und Jugendlichen an einem Tablet-PC getestet. Der PPVT-4 wurde als mittlerer von drei Tests durchgeführt. Die Kinder und Jugendlichen wurden jeweils einzeln im Rahmen des ReGES-Gesamtinstrumentes getestet. Zuletzt führten die Interviewerinnen und Interviewer die Befragungspersonen, wenn gewünscht, in eine Panel-App ein (vgl. Will, Gentile, et al., 2018).

[1] Wir danken Prof. Dr. Wolfgang Lenhard für die Beratung und Unterstützung.

[2] Das richtige Bild ist Bild 4 von Item 196. Die Distraktoren sind Item 201 Bild 2 und 4 sowie Item 29 Bild 4.

[3] Bei Item 111 wurde Bild 3 durch Bild 3 des Items 101 ersetzt. Bei Item 120 wurde Bild 3 durch Bild 3 des Items 59 ersetzt. Bei Item 201 wurde Bild 3 durch Bild 4 des Items 101 ersetzt.

3.2 Ablauf der Kompetenztestung

3.2.1 Ablauf in der RC1

Zunächst wurden alle Personen, außer der Zielperson und einer erziehungsberechtigten Person, gebeten, den Raum falls möglich zu verlassen. Vor der Testung wurde die erziehungsberechtigte Person nochmals über den Ablauf der Testung informiert und es bestand die Möglichkeit, die Testung an dieser Stelle zu verweigern, falls die Eltern oder das Kind nicht wollten, dass die Testung gestartet wird. Zudem konnte die Instruktionssprache geändert werden, welche auf Grundlage von Angaben zu Beginn der Erhebung voreingestellt war. Anschließend konnte diese während der gesamten Testung nicht mehr umgestellt werden. Falls sich herausstellte, dass das Kind doch keine der Instruktionssprachen ausreichend beherrschte, wurde die Testung nicht durchgeführt. Wenn die Sprache korrekt eingestellt und das Kind bereit war, wurde das Testmodul aufgerufen. Zunächst wurde dem Kind das abgelöste Tablet des Tablet-PCs gezeigt und die Lautstärke der Kopfhörer wurde eingestellt. Zuerst wurde ein Test zur Messung kognitiver Grundfähigkeiten (DGCF; Lang, Kamin, Rohr, Stünkel & Willinger, 2014) bearbeitet, welcher den Kindern gegenüber als „Zeichenrätsel“ bezeichnet wurde.

Anschließend sollte die Interviewerin oder der Interviewer das „Bilderrätsel“ (PPVT-4) starten, indem sie oder er auf den Button „Spiel starten“ klickte. Nach einer kurzen Überleitung durch die Interviewerin oder den Interviewer wurde dem Kind eine Audioaufnahme vorgespielt, in der erklärt wurde, wie das Spiel funktioniert und wie die Aufgaben gelöst werden sollen. Die Instruktion und Programmierung richtet sich nach den Vorgaben im Testmanual des PPVT-4.

Die Kinder wurden darauf hingewiesen, dass sie raten dürfen. Es folgte ein Übungsteil, in dem die Kinder Übungsaufgaben lösen sollten und das Programm Feedback gab, ob die jeweilige Aufgabe richtig gelöst wurde. Die Kinder mussten mindestens zwei der Übungsitems erfolgreich beantworten, ansonsten galten sie als nicht testbar und die Testung wurde nicht durchgeführt. In der Testung selbst erfolgte keine Rückmeldung über die Korrektheit des Items durch das Programm. Die Items wurden nach einer festgelegten Testlogik abgefragt, bis das Abbruchkriterium oder das Testende erreicht wurde (vgl. Lenhard et al., 2015).

Nach Abschluss des PPVT-4 folgte das „Suchspiel“ zur Messung der rezeptiven Grammatikkompetenzen (TROG; Fox-Boyer, 2016).

3.2.2 Ablauf in der RC2

Der Ablauf bei den Jugendlichen war dem bei den Kindern sehr ähnlich. Anders als bei den Kindern bedienten die Jugendlichen das Programm jedoch selbständig. Ihnen wurde der Tablet-PC zu Beginn der Aufgaben übergeben und die Interviewerin oder der Interviewer hielt sich im Hintergrund und war für Fragen ansprechbar. Sie verwendeten den Tablet-PC vollständig, das Tablet wurde nicht vom Rest des PCs gelöst. Die Testung wurde als „Aufgabe“ bezeichnet und die Zielpersonen wurden in den Instruktionen gesiezt. Jugendliche konnten nur von Set 5 starten, wenn sie im für sie vorgesehenen Übungsteil keine Fehler machten, ansonsten mussten sie mit denselben Übungsitems wie die Kinder unter gleichen Bedingungen fortfahren (vgl. Lenhard et al., 2015). Die Anwesenheit einer erziehungsberechtigten Person war nicht notwendig. Diese wurden, wenn möglich, wie alle anderen gegebenenfalls anwesenden Personen ebenfalls freundlich gebeten, den Raum zu verlassen.

3.3 Stichprobe

Insgesamt konnten in der ersten Welle 2.405 Interviews mit Eltern der geflüchteten Kinder und 2.415 Interviews mit geflüchteten Jugendlichen realisiert werden. Interviews wurden in acht Sprachen durchgeführt: Arabisch, Deutsch, Englisch, Französisch, Kurmandschi, Paschtu, Persisch und Tigrinya (vgl. Gentile, Heinritz & Will, 2019). Getestet wurden jedoch nur Kinder und Jugendliche, die eine der vier für die Panelbefragung vorgesehenen Sprachen nach Angaben der Eltern bzw. nach eigenen Angaben der Jugendlichen ausreichend beherrschten, um die Instruktionen zu verstehen. Die Panelsprachen waren Arabisch, Deutsch, Englisch und Kurmandschi (vgl. Will, Gentile, et al., 2018). Insgesamt wurden für 2.008 Kinder und 2.016 Jugendliche die Tests administriert. Für diese Fälle sind Kompetenzdaten im Scientific-Use-File (SUF) enthalten, unvollständige Kompetenztestungen, z.B. aufgrund von Abbrüchen, sind an entsprechender Stelle mit Missingwerten befüllt. Die für den PPVT-4 im SUF enthaltenen Variablen sind im Anhang aufgelistet. Für die Skalierung des PPVT-4 wurden jedoch nur Fälle berücksichtigt, die den PPVT-4 vollständig, d.h. bis zum Erreichen des Abbruchkriteriums, bearbeitet haben und mindestens ein Itemset bestanden haben, da das Ergebnis sonst nur auf Zufallsniveau liegt (vgl. Lenhard et al., 2015). Eine Ausnahme stellen 39 Kinder und 58 Jugendliche dar, die die Testung zwar abgebrochen, aber mindestens fünf Sets vollständig bearbeitet haben. Diese wurden ebenfalls in den Analysen berücksichtigt, da davon ausgegangen wird, dass die vorhandenen Daten ausreichen, um auf Grundlage eines Item Response Theorie-Modells einen validen Personenwert zu schätzen. Ein Summenscore wird für diese Fälle nicht angegeben, da dieser verzerrt wäre, weil alle nicht bearbeiteten Items als falsch gewertet würden. Entsprechend wurden 1.380 Testergebnisse der RC1-Kinder und 1.451 Testergebnisse der RC2-Jugendlichen in den Analysen verwendet. Für eine detaillierte Beschreibung der Ausschlussgründe und der Kodierung im Scientific Use File (SUF) sei auf das folgende Unterkapitel 3.3.1 verwiesen.

Das Alter der 1.380 Kinder lag zum Zeitpunkt der Testung im Durchschnitt bei 5,73 Jahren (*SD* = 0,94); 48,41 % der Kinder waren weiblich. Das Alter der 1.451 Jugendlichen lag zum Testzeitpunkt im Durchschnitt bei 15,94 Jahren (*SD* = 0,86); 43,97 % der Jugendlichen waren weiblich.

Bei den Kindern wurden von 1.380 Interviews 142 in Gemeinschaftsunterkünften und 1.238 in dezentralen Unterkünften durchgeführt. Von den 1.451 Interviews mit Jugendlichen wurden 122 in Gemeinschaftsunterkünften und 1.329 in dezentralen Unterkünften realisiert.

3.3.1 Ausschluss von Fällen

Im Folgenden wird detailliert beschrieben, wie viele Fälle aus welchem Grund ausgeschlossen wurden (s. auch Tabelle 1). Im Scientific Use File sind die Fälle mit einem entsprechenden Missingwert u.a. in den Variablen für den Weighted likelihood estimate WLE (t0501010_g1) und den Summenscore (t0501010_g3) kodiert[4].

Aufgrund von Abbrüchen der Kompetenztestung aus individuellen Gründen kam es insgesamt bei 286 Kindern und 569 Jugendlichen zu fehlenden Werten im PPVT-4 (Missingwert: -91). Die Abbrüche erfolgten an unterschiedlichen Stellen während der Testung: Bei 202 Kindern sowie 469 Jugendlichen wurde die Testung bereits während der Bearbeitung des Tests zur Messung

[4] Die Missingkodierung orientiert sich an der Missingkodierung in Datensätzen des Nationalen Bildungspanels (NEPS; Blossfeld, Roßbach & von Maurice, 2011) insbesondere der Startkohorte Neugeborene, doi:10.5157/NEPS:SC1:8.0.0.

der kognitiven Grundfähigkeiten oder in der Übungsphase des PPVT-4 abgebrochen. Zudem haben 45 Kinder und 42 Jugendliche den PPVT-4 während der Testphase abgebrochen und weniger als fünf Sets vollständig bearbeitet. Weitere 39 Kinder und 58 Jugendliche haben den PPVT-4 während der Testphase abgebrochen, nachdem sie mindestens fünf Sets vollständig bearbeitet haben. Letztere erhalten den Missingwert -91 in der Summenscorevariablen (t0501010_g3), nicht jedoch in der Variablen für den WLE (t0501010_g1; s. Kapitel 3.3) und sind z.B. über diese beiden Variablen im Datensatz identifizierbar.

Entsprechend dem Testmanual (Lenhard et al., 2015) wurde die Testphase des PPVT-4 nur durchgeführt, wenn die Übungsaufgaben erfolgreich bearbeitet wurden. In der RC1 bestanden 120 Fälle nicht die Übungsphase (Missingwert: -24) und 1 Fall gab gar keine Antwort bei den Übungsaufgaben (Missingwert: -25). In der Kohorte der Jugendlichen bestanden hingegen 20 Fälle nicht die Übungsphase (Missingwert: -24).

Von den Analysen ausgeschlossen wurden 140 Kinder und drei Jugendliche (Missingwert: -27), die bereits in Set 1 acht oder mehr Fehler gemacht haben, da das Ergebnis hier nur auf Zufallsniveau liegt (vgl. Lenhard et al., 2015).

Technische Systemabstürze der Testanwendung (Missingwert: -21) kamen bei 96 Kindern sowie 12 Jugendlichen vor.

Tabelle 1: Ausschlüsse von Fällen und Missingkodierung im SUF

Missingwert **deutsches Label** **(englisches Label)**	**Erläuterung**	**Anzahl Fälle Kinder** **(Anteil an administrierten Tests in %)**	**Anzahl Fälle Jugendliche** **(Anteil an administrierten Tests in %)**
-91 **Befragung abgebrochen** **(Survey aborted)**	kein Wert vorliegend, da die Testanwendung über das Testleitermenü abgebrochen wurde oder Testteile über das Testleitermenü übersprungen wurden.	weniger als 5 Sets bearbeitet: 247 (12,30) mindestens 5 Sets bearbeitet: 39 (1,94)	weniger als 5 Sets bearbeitet: 511 (25,35) mindestens 5 Sets bearbeitet: 58 (2,88)
-90 **nicht spezifizierbar fehlend** **(Unspecific missing)**	kein Wert vorliegend aus unbekanntem Grund	24 (1,20)	19 (0,94)
-27 **keine valide Aussage möglich: Abbruchkriterium in Set 1 erreicht** **(No valid statement possible: reached termination criterion in set 1)**	das Abbruchkriterium (>= 8 Fehler) wurde bereits in Set 1 erreicht. Entsprechend der Vorgabe im Testmanual wird die Person als nicht testbar mit dem PPVT betrachtet und es wird kein Summenscore/WLE angegeben	140 (6,97)	3 (0,15)
-25 **keine valide Aussage möglich: keine Antwort in allen Übungsaufgaben** **(No valid statement possible: no**	keine Testdurchführung, da in der Übungsphase keine Reaktion gezeigt wurde	1 (0,05)	0 (0,00)

response in all training tasks)			
-24 Übungsphase nicht bestanden (Test phase not passed)	keine Testdurchführung, da die Übungsphase nicht bestanden wurde	120 (5,98)	20 (0,99)
-21 technische Probleme/Absturz (Technical problems/system crash)	kein Wert vorliegend, da die Testanwendung abgestürzt ist	96 (4,78)	12 (0,60)
Gesamte Ausschlüsse ohne Fälle mit mindestens 5 bearbeiteten Sets (WLE)		628 (31,27)	565 (28,03)
Gesamte Ausschlüsse inkl. Fälle mit mindestens 5 bearbeiteten Sets (Summenwert)		667 (33,22)	623 (31,03)

Außerdem konnte bei 24 Fällen der RC1 und 19 Fällen der RC2 nicht eindeutig rekonstruiert werden, ob es sich um Testabbrüche durch die Testleitenden oder um Systemabstürze handelte, weshalb kein Summenscore oder WLE berechnet werden konnte und die Kategorie „nicht spezifizierbar fehlend“ (Missingwert: -90) vergeben wurde.

Der Anteil der Abbrüche erscheint zunächst relativ hoch. Dies kann unterschiedliche Gründe haben. Zum einen war die zeitliche Belastung durch die Befragung in den Familien teilweise sehr hoch, zum anderen ist im Fall der Jugendlichen anzumerken, dass der zuvor durchgeführte Test zu den kognitiven Grundfähigkeiten relativ lange Wartezeiten enthielt. Größtenteils wurde die Kompetenztestung bereits vor Beginn der Testphase des PPVT-4 abgebrochen (Missingwert -91). Zudem könnte eine Rolle gespielt haben, dass die Testsituation für einen Großteil der Geflüchteten ungewohnt war.

Für die Fälle, die die Übungsphase nicht bestanden oder das Abbruchkriterium bereits in Set 1 erreicht haben, ist anzunehmen, dass fehlende Deutschkompetenzen eine entscheidende Rolle gespielt haben, oder dass die Aufgabe nicht richtig verstanden wurde. Für die RC1 ist ein Vergleich mit den Daten der Startkohorte 1 des NEPS (Blossfeld et al., 2011) möglich, dort

wurde der PPVT-4 u.a. in den Wellen 4 und 6 eingesetzt[5]. Zu diesen Zeitpunkten waren die Kinder drei bzw. fünf Jahre alt. Von 2.324 Kindern, die in Welle 4 am PPVT-4 (Lenhard et al., 2015) teilgenommen haben, haben 178 Kinder (7,66%) die Übungsphase nicht bestanden und 316 Kinder (13,60%) das Abbruchkriterium bereits im ersten Set erreicht. In Welle 6 haben 2.080 Kinder an der Testung mit dem PPVT-4 teilgenommen. Davon hat nur ein Kind die Übungsphase nicht bestanden (0,05%) und 8 Kinder (0,38%) haben das Abbruchkriterium in Set 1 erreicht. Die Anteile sind nicht direkt mit den Anteilen in Tabelle 1 vergleichbar, da in der RC1 der Anteil an Fällen, die vor Beginn des PPVT-4 die Kompetenztestung abgebrochen haben, relativ groß ist. Trotzdem zeigt sich, dass die RC1-Kinder in ReGES verglichen mit den jüngeren dreijährigen Kindern im NEPS, die den PPVT-4, ebenfalls zum ersten Mal bearbeitet haben, nicht auffallend häufig die Übungsphase oder das erste Set nicht bestanden haben. Verglichen mit Welle 6, als die Kinder der Startkohorte 1 des NEPS ebenfalls fünf Jahre alt waren, sind die Missingwerte -24 und -27 der ReGES-Kinder jedoch deutlich erhöht. Hier ist allerdings zu beachten, dass es sich für die NEPS-Kinder bereits um eine Testwiederholung handelte, diese also sowohl mit der Testsituation als auch mit dem Test in der Regel bereits vertraut waren. Berücksichtigt man weiterhin, dass die Kinder der RC1 erst seit Kurzem in Deutschland sind und teilweise erst wenig Kontakt zur deutschen Sprache hatten, sind die Anteile der Missingwerte -24 und -27 als unauffällig einzustufen.

4. Deskriptive Statistiken: Summenwert

Im Folgenden gehen wir auf die Summenwertbildung sowie auf die deskriptive Analyse des Summenwertes ein.

Der Summenwert wird in Abhängigkeit der erreichten Sets gebildet. Dabei wird das Deckenset mit 12 multipliziert, da jedes Set 12 Items enthält. Davon wird die Anzahl der falschen Antworten (d.h. Fehler) subtrahiert (vgl. Lenhard et al., 2015). In Ausnahmefällen kann das Deckenset vom höchsten bearbeiteten Set abweichen (s. Kapitel 5.2; vgl. Lenhard et al., 2015).

Tabelle 2 enthält deskriptive Statistiken des Summenwerts der RC1 und der RC2. Die Summenwerte der Kinder liegen zwischen 6 und 203 (M = 56,50; SD = 29,59). Die Verteilung ist rechtsschief und bimodal, mit Modalwerten bei 30 und 58. In Abbildung 1 ist die Verteilung des Summenwerts der Kinder graphisch dargestellt.

Die Summenwerte der Jugendlichen liegen zwischen 11 und 206 (M = 89,72; SD = 32,33). Auch hier ist die Verteilung rechtsschief und bimodal. Die Modalwerte liegen bei 82 und 86. Im Mittel haben die Jugendlichen somit höhere Summenwerte erzielt als die Kinder. Die Verteilung des Summenwerts der Jugendlichen ist in Abbildung 2 graphisch dargestellt.

[5] *Diese Arbeit nutzt Daten des Nationalen Bildungspanels (NEPS): Startkohorte Neugeborene, doi:10.5157/NEPS:SC1:8.0.0. Die Daten des NEPS wurden von 2008 bis 2013 als Teil des Rahmenprogramms zur Förderung der empirischen Bildungsforschung erhoben, welches vom Bundesministerium für Bildung und Forschung (BMBF) finanziert wurde. Seit 2014 wird NEPS vom Leibniz-Institut für Bildungsverläufe e.V. (LIfBi, Bamberg) in Kooperation mit einem deutschlandweiten Netzwerk weitergeführt.*

Tabelle 2: Deskriptive Statistiken zu den Summenwerten der RC1 und RC2

	RC1	**RC2**
Anzahl an gültigen Fällen	1.341	1.393
Mittelwert	56,50	89,72
Median	54	86
Modus	{30; 58}	{82; 86}
Standardabweichung	29,59	32,33
Schiefe	0,79	0,47
Kurtosis	3,87	3,16
Minimum	6	11
Maximum	203	206

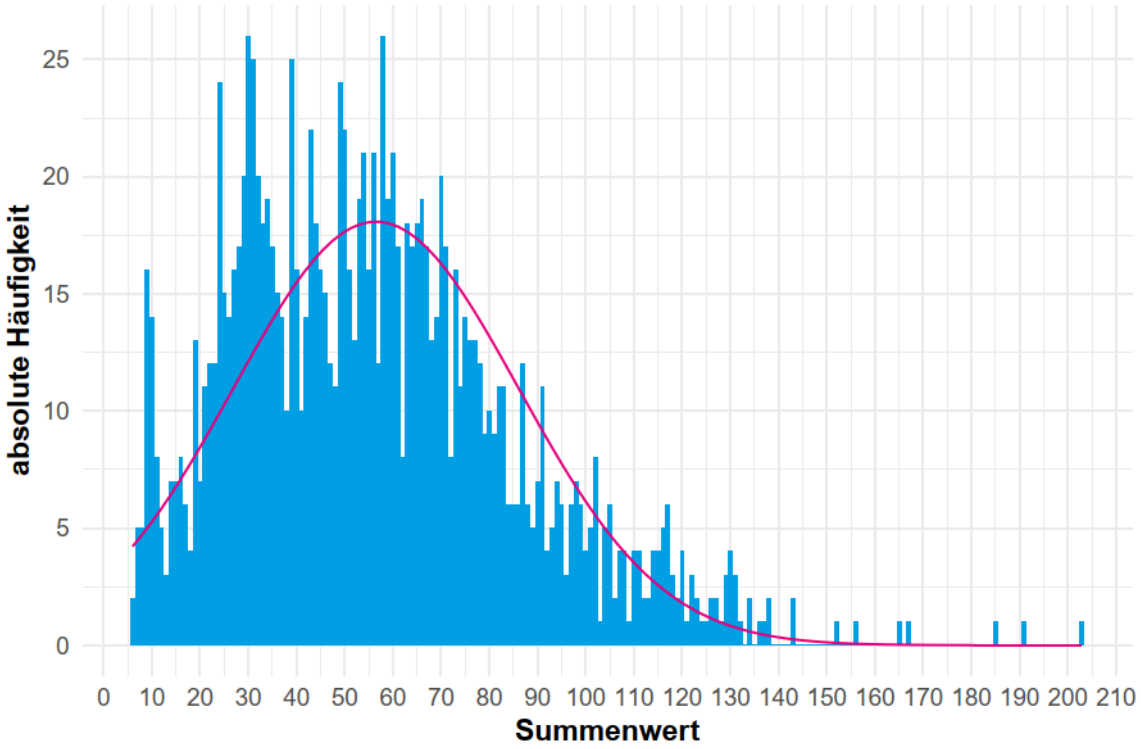

Abbildung 1. Häufigkeitsverteilung der Summenwerte bei Kindern (RC1) mit Normalverteilungskurve (in rot)

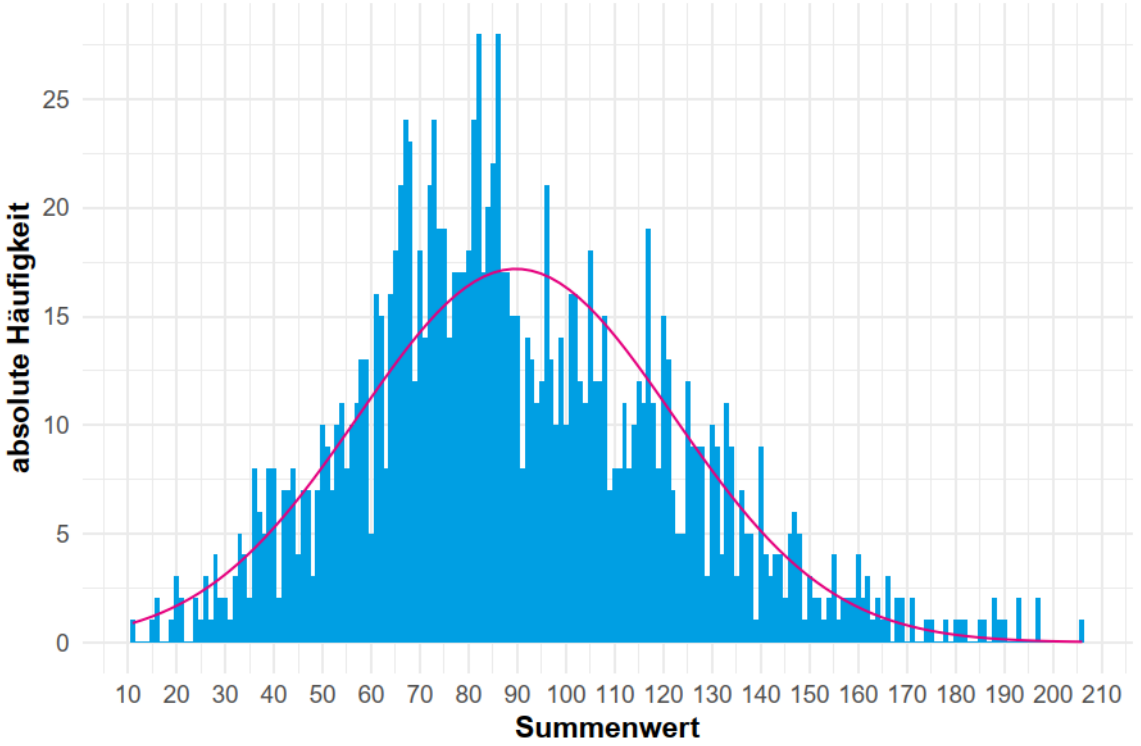

Abbildung 2. Häufigkeitsverteilung der Summenwerte bei Jugendlichen (RC2) mit Normalverteilungskurve (in rot)

5. Analyse

In diesem Kapitel beschreiben wir die verschiedenen Methoden, die wir für die Skalierung und Evaluation der Kompetenzdaten angewandt haben. Wir orientieren uns an dem von Pohl und Carstensen (2012) beschriebenen Vorgehen zur Skalierung und Überprüfung der Kompetenzdaten im Nationalen Bildungspanel (NEPS) sowie an Fischer und Durda (2020), die die Skalierung einer anderen PPVT-Version im NEPS beschreiben und ebenfalls die Skalenqualität überprüfen. Wir gehen auf die Untersuchung fehlender Werte, die Schätzung der Modelle, die Evaluation der Skalenqualität sowie die verwendete Software ein.

5.1 Fehlende Werte

Im Folgenden gehen wir zunächst auf die Missingkategorien auf Itemebene von Fällen ein, die in die Analyse einflossen. Die fehlenden Werte werden anschließend analysiert.

Auf Itemebene wurden a) „Angabe verweigert“, b) „nicht erreicht“, c) „Befragung abgebrochen“ und d) „filterbedingt fehlend“ als Missingkategorien unterschieden. Für eine Übersicht über die Missingkategorien und die dazugehörigen Missingwerte im SUF siehe Tabelle 3.

„Angabe verweigert“ wurde dann vergeben, wenn die Frage vorgegeben wurde, aber der oder die Teilnehmende nicht auf die Frage antwortete. Dies kann während der Testdurchführung verschiedene Gründe gehabt haben und ist nicht unbedingt auf eine aktive Verweigerung der Antwort zurückzuführen.

Tabelle 3: Missingkodierung auf Itemebene bei in die Analysen einbezogenen Fällen im SUF

Missingwert	englisches Label	deutsches Label	Erläuterung
-99	Filtered	filterbedingt fehlend	Missingwert für die Items, die nicht angespielt werden, da das Bodenset in höherem Set erreicht wurde, außerdem für nicht vorgegebene Übungsitems
-97	Refused	Angabe verweigert	keine Reaktion bei dem Item
-94	Not reached	nicht erreicht	Testitem wurde nicht erreicht wegen leistungsbedingtem Abbruchkriterium
-91	Survey aborted	Befragung abgebrochen	kein Wert vorliegend, da die Testanwendung über das Testleitermenü abgebrochen wurde oder Testteile über das Testleitermenü übersprungen wurden

In den meisten Fällen wurden nicht alle Itemsets durchgeführt, da das Abbruchkriterium des Tests (s. Kapitel 2) vorher erfüllt war. In diesem Fall wurde die Kategorie „nicht erreicht“ für alle darauffolgenden, nicht vorgegebenen Items vergeben.

Zudem gab es einige Fälle, die aufgrund eines frühzeitigen Testabbruches weder das Testende noch das Abbruchkriterium erreichten. Sofern aber mindestens fünf Sets vollständig beantwortet wurden, wurden diese Fälle in die Analysen miteinbezogen (s. Kapitel 3.3). Alle fehlenden Werte, die auf diesen Grund zurückgehen, bekamen das Label „Befragung abgebrochen“.

Alle Items unterhalb des Bodensets bekamen das Label „filterbedingt fehlend“.

Fehlende Werte können einen Eindruck darüber vermitteln, wie gut der Test funktioniert hat (vgl. Fischer & Durda, 2020). Die Häufigkeiten fehlender Werte wurden pro Person untersucht, um eine Vorstellung darüber zu gewinnen, wie gut Personen mit dem Test zurechtkamen und wie sich ihre Fähigkeiten voraussichtlich verteilen. Fehlende Werte pro Item wurden analysiert, um beurteilen zu können, wie gut Items funktionierten und wie gut die Sets für die Altersgruppe geeignet waren.

5.2 Modelle

Item- und Personenparameter wurden mittels Rasch-Modell (Rasch, 1980) über Marginal Maximum Likelihood Methoden geschätzt (vgl. Pohl & Carstensen, 2012). Als Personenparameter wurden Weighted maximum likelihood estimates (Warm, 1989) berechnet.

Im Testmanual des PPVT-4 wird die Aussage getroffen, dass „[…] alle Items unterhalb des Bodensets als richtig gelöst und alle Items oberhalb des Deckensets als nicht gelöst [gelten]“ (Lenhard et al., 2015, S. 46). Zudem werden auch verweigerte Items als Fehler gewertet (vgl. Lenhard et al., 2015). Um die IRT-Modelle zu schätzen, vergeben wir pro Fall daher für alle Items oberhalb des Deckensets[6] und verweigerten Items den Wert 0 und für alle filterbedingt fehlenden Items den Wert 1. Fälle, die die Befragung abgebrochen haben, nachdem sie mindestens fünf Sets vollständig bearbeitet haben, stellen hier jedoch eine Ausnahme dar. Diese wurden zusätzlich in die Analysen aufgenommen, um die entsprechenden Fälle nicht vollständig ausschließen zu müssen (s. Kapitel 3.3), was laut Testmanual jedoch so nicht vorgesehen ist. Diese Fälle haben das Abbruchkriterium nicht erreicht und hätten weitere Items bearbeitet, wenn die Testung nicht aus verschiedenen möglichen Gründen durch den Interviewer oder die Interviewerin abgebrochen worden wäre. Deshalb kann in diesen Fällen nicht davon ausgegangen werden, dass schwierigere, nicht erreichte Itemsets falsch beantwortet worden wären. Bei der Bildung von IRT-Modellen besteht im Gegensatz zur Berechnung des Summenscores die Möglichkeit, diese fehlenden Werte zu ignorieren, statt sie als falsch oder richtig zu kodieren. Fehlende Werte, die auf -91 „Befragung abgebrochen“ zurückgehen, ignorieren wir entsprechend bei der Bildung der IRT-Modelle (vgl. Pohl, Gräfe, & Rose, 2014).

Einige der höheren Itemsets wurden aufgrund des Abbruchkriteriums von sehr wenigen Teilnehmenden erreicht (s. Kapitel 6.1.1.2 für die RC1 bzw. Kapitel 6.2.1.2 für die RC2). Für die entsprechenden Items liegen demzufolge zu wenige gültige Werte vor, um sie in die Analysen zu den psychometrischen Eigenschaften des Tests (s. Kapitel 6.1.2 bzw. 6.2.2) und der Qualität

[6] In Ausnahmefällen haben Jugendliche mit Set 5 begonnen und in diesem mehr als einen aber maximal sieben Fehler gemacht und entsprechend mit Set 4 fortgefahren. Anschließend haben sie z.B. in Set 4 acht oder mehr Fehler gemacht. In einem solchen Fall ist Set 4 das Deckenset (vgl. Lenhard et al., 2015) und auch alle Items in Set 5 werden zu 0 umkodiert, obwohl Set 5 vollständig bearbeitet wurde.

des Tests (s. Kapitel 6.1.3 bzw. 6.2.3) einzubeziehen. Deshalb wurde ein kritischer Wert von 101 gültigen Werten pro Item festgelegt, damit es in die Analysen einbezogen wurde. Bevor also fehlende Werte, die auf nicht erreichte oder verweigerte Items zurückgingen, zu 0 umcodiert wurden, wurden Items, die nicht mehr als 100 gültige Werte enthielten, aus den weiteren Analysen ausgeschlossen. Dies betrifft alle Items ab Set 13 in der RC1 und alle Items ab Set 17 in der RC2.[7]

5.3 Qualität der Skala

Der PPVT-4 ist ein etablierter und ausführlich validierter Test zur Messung des rezeptiven Wortschatzes und es wurden auch die Ergebnisse von verschiedenen Personengruppen, u.a. Personen mit Migrationshintergrund betrachtet (vgl. Dunn & Dunn, 2007; Lenhard et al., 2015). Neuzugewanderte Flüchtlinge sind jedoch eine besondere Gruppe innerhalb der Gruppe der Personen mit Migrationshintergrund. Sie erwerben die deutsche Sprache erst seit relativ kurzer Zeit. Deshalb sollen zur Sicherstellung der Qualität der Skala für diese Gruppe die psychometrischen Eigenschaften des Tests untersucht werden.

Die Qualität der Items evaluierten wir zum einen mithilfe von Weighted Mean Square Statistiken (WMNSQ). Dabei handelt es sich um eine gewichtete Anpassungsgüte, die auf den standardisierten Residuen basiert und angibt, wie gut das geschätzte Modell die Beobachtungen erklärt. Die WMNSQ-Werte reichen von 0 bis unendlich und haben einen Erwartungswert von 1 (vgl. de Ayala, 2009). Werte nahe 1 bedeuten eine gute Anpassung. Werte unter 1 deuten auf eine Überanpassung hin, d.h. wenn Items stärker zwischen den Personen differenzieren, als im Modell angenommen wird. Dahingegen liegt der Wert über 1, wenn das Item schlecht zwischen Personen differenziert. Insbesondere Werte über 1 werden als eine Verletzung der Modellanpassung gewertet (vgl. Pohl & Carstensen, 2012). Items mit einem WMNSQ > 1,15 (t-Werte > |6|) gelten als Items mit einem bemerkenswerten Item-Misfit und Items mit einem WMNSQ > 1,2 (t-Werte > |8|) als Items mit einem starken Item-Misfit (vgl. Pohl & Carstensen, 2012).

Zudem analysierten wir die korrigierte Trennschärfe der Items nach der klassischen Testtheorie. Hierbei handelt es sich um die punktbiseriale Korrelation zwischen den korrekten Antworten und dem Summenwert (minus des betreffenden Items). Eine Korrelation > 0,3 wird als gut und eine Korrelation > 0,2 als akzeptabel gewertet (vgl. Pohl & Carstensen, 2012).

Ein weiterer Ansatz, um die Qualität der Skala zu untersuchen, ist die Analyse von Differential Item Functioning (DIF). Hierbei testet man die Messinvarianz zwischen verschiedenen Gruppen. Liegt Differential Item Functioning vor, unterscheiden sich bei gleicher Fähigkeit die Gruppen in der Lösungswahrscheinlichkeit bestimmter Items; d.h. in diesen Fällen funktionieren Items unterschiedlich zwischen verschiedenen Gruppen und es werden Gruppen bevorteilt (vgl. Pohl & Carstensen, 2012). Wir betrachteten Differential Item Functioning in Bezug auf Geschlecht, Herkunftsland und Bildung. Bei letzterer Variable erstellten wir eine dichotome Variable aus der Klassifikation von ISCED 97, wobei alle Werte mit den Ausprägungen „Not completed primary education“ und „Primary education“ den Wert 1 bekamen und die restlichen Ausprägungen den Wert 0. Die Merkmale der Variable ISCED 97 bezogen sich hierbei

[7] Bei den im SUF enthaltenen WLEs (t0501010_g1, Standardfehler der WLEs: t0501010_g2) handelt es sich um die in diesem Working Paper betrachteten WLEs. D.h. in deren Schätzung wurden nur die ersten zwölf Sets für die RC1 und die ersten 16 Sets für die RC2 einbezogen. Angegeben werden die WLEs im SUF auch für die Fälle, die die Befragung abgebrochen aber mindestens fünf Sets vollständig bearbeitet haben.

nicht ausschließlich auf die erziehungsberechtigte Person, sondern auf die Person mit dem höchsten Bildungsabschluss im Haushalt. Um Differential Item Functioning zu untersuchen, verwendeten wir Multigruppen-Item-Response-Theorie-Modelle (vgl. Pohl & Carstensen, 2012). Wenn der Betrag der Differenz zwischen den geschätzten Schwierigkeiten der Gruppen größer als 1 Logit ist, gilt das als sehr starker DIF-Wert. Werte zwischen 0,6 und 1 Logit gelten als stark, Werte zwischen 0,4 und 0,6 als schwach und Werte unter 0,4 als vernachlässigbar. Zusätzlich wurde die Testfairness untersucht, indem das Modell, das Differential Item Functioning zulässt, mit einem Modell verglichen wurde, das nur Haupteffekte (also Mittelwertsunterschiede zwischen den Gruppen) berücksichtigte (vgl. Pohl & Carstensen, 2012).

Um die Annahme des Rasch-Modells von gleichen Diskriminationsparametern zu untersuchen, wurden die Daten auch anhand eines zwei-parametrischen Modells (Birnbaum, 1968) skaliert und mit dem Rasch-Modell verglichen. Dazu wurden Akaike's Informationskriterium (AIC) und das Bayesianische Informationskriterium (BIC) herangezogen (vgl. Fischer & Durda, 2020; Pohl & Carstensen, 2012).

Um die Annahme der Eindimensionalität zu testen, orientierten wir uns an der Q_3-Statistik von Yen (vgl. Yen, 1984). Q_3 ist die Korrelation zwischen den Residuen des gefitteten Rasch-Modells und wird für jedes Itempaar einzeln berechnet. Da im Falle von bedingter stochastischer Unabhängigkeit die Q_3-Statistik dazu tendiert leicht negativ zu sein, stellen wir die korrigierten Q_3-Werte dar, die einen Erwartungswert von Null haben (vgl. Fischer & Durda, 2020). Wir berechneten den Mittelwert der absoluten Q_3-Werte für jedes Item. Werte unter 0,2 sind ein Hinweis auf die notwendige Eindimensionalität (vgl. de Ayala, 2009).

5.4 Software

Die Item Response Modelle wurden mithilfe des TAM-Package Version 3.5-19 (vgl. Robitzsch, Kiefer & Wu, 2020) in R Version 4.0.3 (vgl. R Core Team, 2020) geschätzt.

6. Ergebnisse

Im Folgenden gehen wir auf die Ergebnisse unserer Analysen ein. Zunächst betrachten wir die Kohorte im Kindergartenalter und im Anschluss die Kohorte im Jugendalter. Wir gehen jeweils auf die Analyse von fehlenden Werten, die Schätzung der Parameter und die Evaluation der Skalenqualität ein.

6.1 RC1

6.1.1 Fehlende Werte

An dieser Stelle wollen wir die fehlenden Werte auf Itemebene behandeln und nur die oben beschriebenen *N* = 1.380 gültigen Fälle betrachten.

6.1.1.1 *Fehlende Werte pro Person*

Fehlende Angaben können zum Beispiel entstehen, wenn Befragte Items nicht beantworten bzw. überspringen (Missingwert -97, s. Kapitel 5.1). Wie in Abbildung 3 zu sehen, beantworteten circa 73% der Befragten alle vorgegebenen Items. Circa 4% der Befragten gaben auf fünf oder mehr administrierte Items keine Antwort.

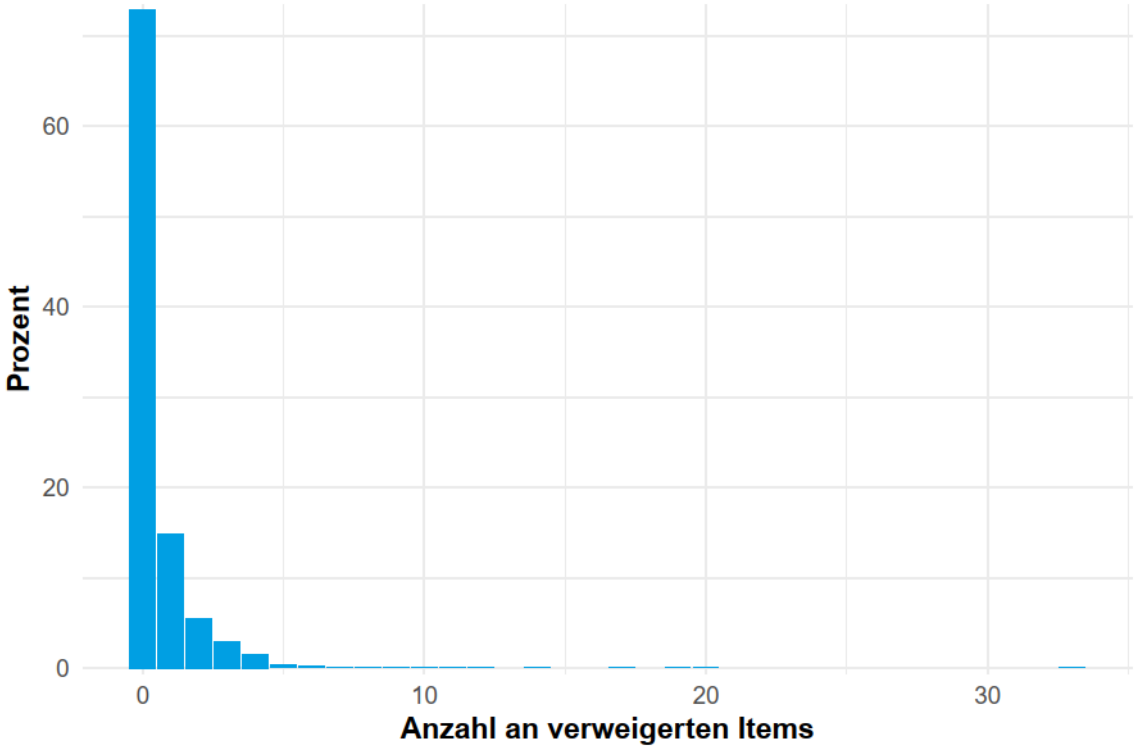

Abbildung 3. Anzahl an verweigerten Items (in Prozent) (N = 1.380)

Außerdem sind Häufigkeiten über nicht erreichte Items bzw. Sets von Interesse (Missingwert -94, s. Kapitel 5.1). In Abbildung 4 ist zu erkennen, dass 0,29% der Fälle alle Items/Sets erreicht haben. Am häufigsten wurden elf Sets nicht erreicht, wobei von den *N* = 1.380 einbezogenen Fällen 15,00% aufgrund des Abbruchkriteriums und 0,51% aufgrund des Abbruches durch die Interviewerin oder den Interviewer die letzten elf Sets nicht erreichten. 4,20% der Fälle haben das dritte Set nicht erreicht bzw. 17 Sets nicht erreicht. Im Durchschnitt haben die Kinder 11,58 (*SD* = 3,03) Sets nicht bearbeitet. Wie in Kapitel 3.3 erläutert, wurden Fälle, die im ersten Set bereits das Abbruchkriterium erreicht und somit 18 Sets nicht erreicht haben, von den Analysen ausgeschlossen.

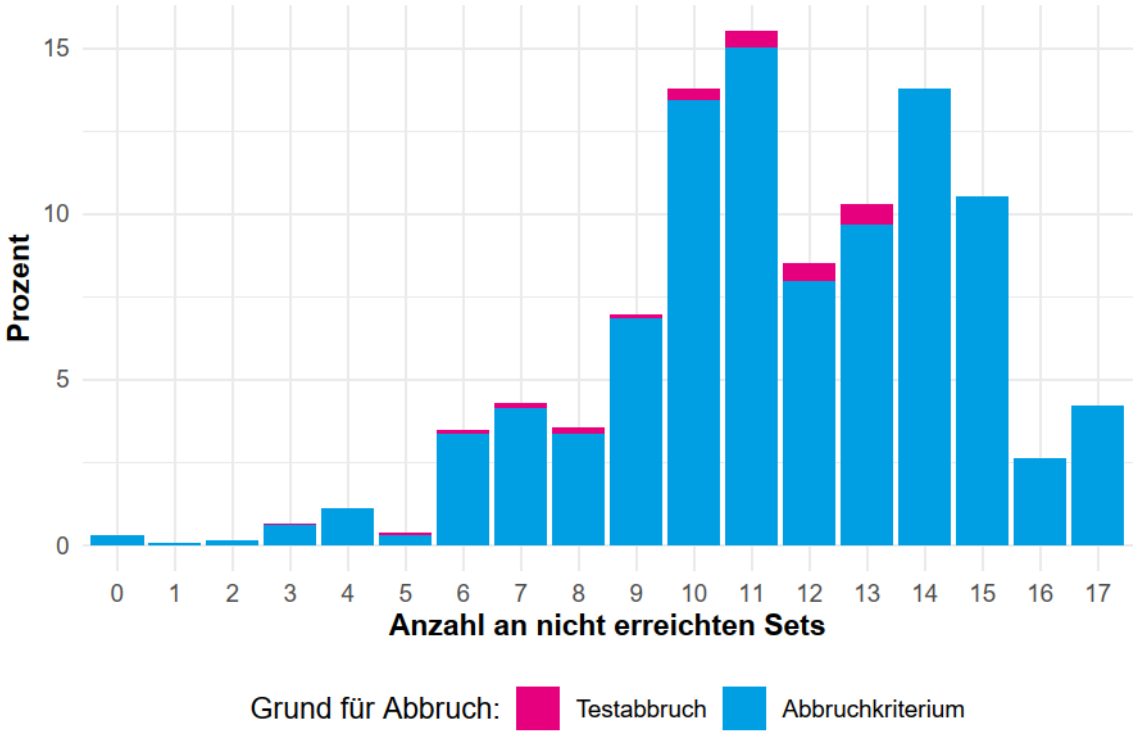

Abbildung 4. Anzahl an nicht erreichten Sets (in Prozent) (N = 1.380)

6.1.1.2 *Fehlende Werte pro Item*

In diesem Unterkapitel gehen wir auf die fehlenden Werte pro Item ein. Tabelle 4 enthält die Anteile der Missingkategorie -97 (Angabe verweigert) je Item in Prozent, wobei nur solche Items abgebildet wurden, bei denen der Anteil mindestens 1,5% beträgt. Auffällig ist, dass das Item auf Position 4 sehr häufig nicht beantwortet wurde. Der höchste Wert findet sich jedoch in den letzten Sets, wobei zu beachten ist, dass bei einer geringen Anzahl an Fällen einzelne fehlende Werte stark ins Gewicht fallen. Die Anteile verweigerter Items liegen zwischen 0,00% und 6,67% (*Mdn* = 0,31%). Wie in Kapitel 5.1 erwähnt, werden nicht beantwortete Items für die folgenden Analysen als Fehler umkodiert.

Tabelle 4: Fehlende Werte (Angabe verweigert) pro Item

Pos.	Item	Anzahl gültiger Fälle	Angabe verweigert (in %)
2	t0501021	1.359	1,52
4	t0501041	1.323	4,13
8	t0501081	1.349	2,25
10	t0501101	1.352	2,03
13	t0502011	1.349	2,25
19	t0502071	1.358	1,59
78	t0507061	792	1,61
81	t0507091	787	2,24
96	t0508121	673	1,75
120	t0510121	278	2,80
123	t0511031	189	1,56
161	t0514051	34	2,86
163	t0514071	33	5,71
164	t0514081	34	2,86
172	t0515041	30	3,23
180	t0515121	30	3,23
189	t0516091	14	6,67

Anmerkungen. N = 1.380; Pos.: Item Position im Testverlauf; nur Werte größer als 1,5 werden gelistet.

Abbildung 5 gibt an, wie viel Prozent der Kinder eine bestimmte Setposition nicht erreicht haben. Circa 86% der Teilnehmerinnen und Teilnehmer haben das 11te Set oder höher nicht erreicht.

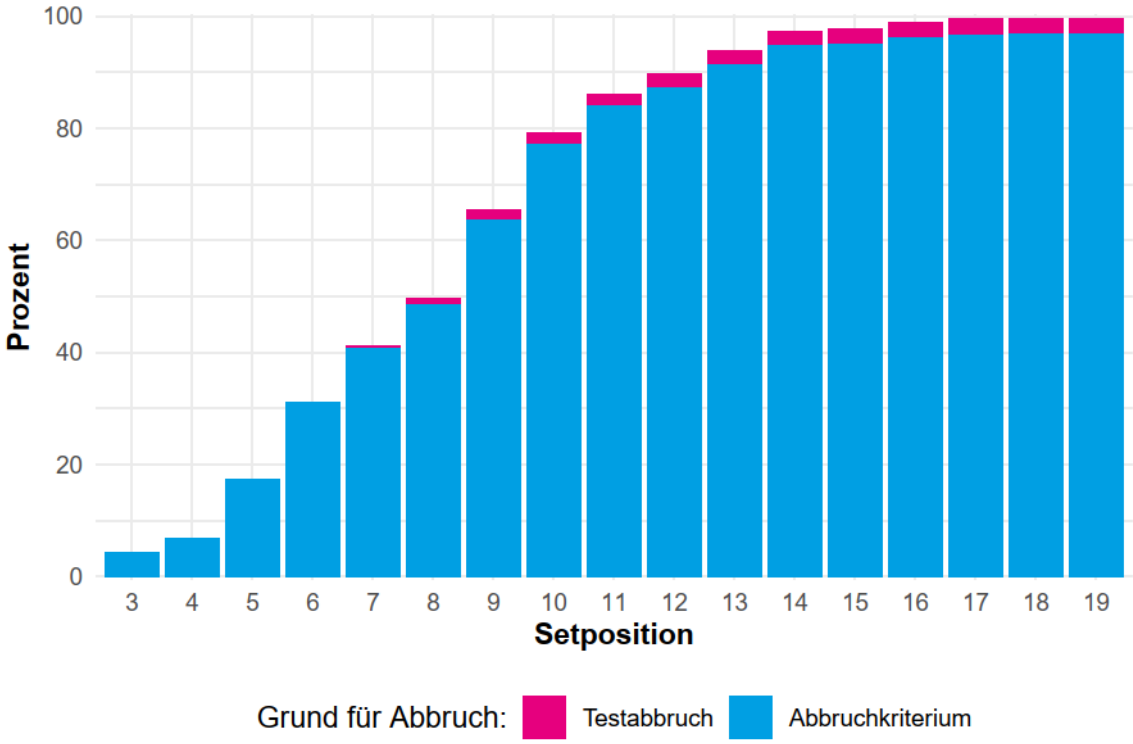

Abbildung 5. Setposition nicht erreicht (in Prozent) (N = 1.380)

6.1.2 Parameterschätzung

Im Folgenden gehen wir auf die Ergebnisse der geschätzten Itemparameter sowie auf Test Targeting und Reliabilität der Personenparameter ein. 84 Items, die nicht mehr als 100 gültige Antworten enthalten, wurden aus der Analyse ausgeschlossen (s. Kapitel 5.2).

6.1.2.1 *Itemparameter*

Die Itemparameter schätzten wir anhand des TAM-Package in R (s. Kapitel 5.4). Die fehlenden Werte, die auf verweigerte oder nicht erreichte Items zurückgehen, wurden als Fehler gewertet. Fehlende Werte, die auf abgebrochene Testungen (bei mindestens fünf vollständig bearbeiteten Itemsets) zurückgehen, wurden hingegen ignoriert (s. Kapitel 5.2).

Wir werteten unter anderem den Anteil an korrekten Antworten aus. Dieser variierte zwischen 0,82% und 98,62%. Das arithmetische Mittel lag bei 39,19% (*SD* = 29,21%). Dies legt nahe, dass die Items einen breiten Fähigkeitsbereich abdeckten.

Die Itemschwierigkeiten des Rasch-Modells wurden geschätzt, indem der Mittelwert der Personenfähigkeitsverteilung auf null fixiert wurde. Die Itemschwierigkeiten lagen zwischen -5,72 (t0502091) und 6,24 (t0512031) mit einem Mittelwert von 0,86 (*SD* = 2,48) und umfassten sowohl einfache wie auch schwierige Items. Die Standardfehler der Itemschwierigkeiten lagen zwischen 0,07 und 0,31 (*M* = 0,09; *SD* = 0,04).

6.1.2.2 *Test Targeting und Reliabilität*

Um die Angemessenheit des Tests für die Zielpopulation zu untersuchen, wurden die Itemschwierigkeiten und die Personenfähigkeiten (WLEs) gegenübergestellt (vgl. Fischer & Durda, 2020). In Abbildung 6 sind die Itemschwierigkeiten und die Personenfähigkeiten auf einer Skala abgebildet. Auf der linken Seite ist die Verteilung der Personenfähigkeiten und auf

der rechten Seite die Verteilung der Itemschwierigkeiten zu erkennen, wobei die Items nach Itemposition – mit der untersten Position auf der linken Seite beginnend – sortiert sind. Zum besseren Überblick wurden auf der X-Achse die Sets angegeben, aus welchen die Items stammen. Der Mittelwert der Fähigkeitsverteilung wurde auf null fixiert, während die Items eine durchschnittliche Schwierigkeit von *M* = 0,86 (*SD* = 2,48) aufwiesen. Das bedeutet, dass die Items im Durchschnitt etwas zu schwierig für die Stichprobe waren. Aber die Spannweite der Itemschwierigkeiten war mit 11,96 sehr groß und deckte sowohl niedrige wie auch hohe Fähigkeitsbereiche ab. Es ist dabei zu beachten, dass nicht alle Items allen Teilnehmerinnen und Teilnehmern vorgegeben wurden und dass nicht bearbeitete Items unterhalb des Bodensets als richtig, nicht bearbeitete Items oberhalb des Deckensets als falsch kodiert wurden. Über die Sets hinweg schien die Itemschwierigkeit anzusteigen, innerhalb der Sets variierte die Itemschwierigkeit jedoch teilweise stark. Alles in allem scheint die Testschwierigkeit für die untersuchte Stichprobe angemessen zu sein. Die Varianz der Personenfähigkeiten war mit 3,25 groß und erlaubt eine gute Differenzierung zwischen den Kindern. Die Reliabilität des Tests war sehr gut (EAP Reliabilität = 0,98; WLE Reliabilität = 0,98).

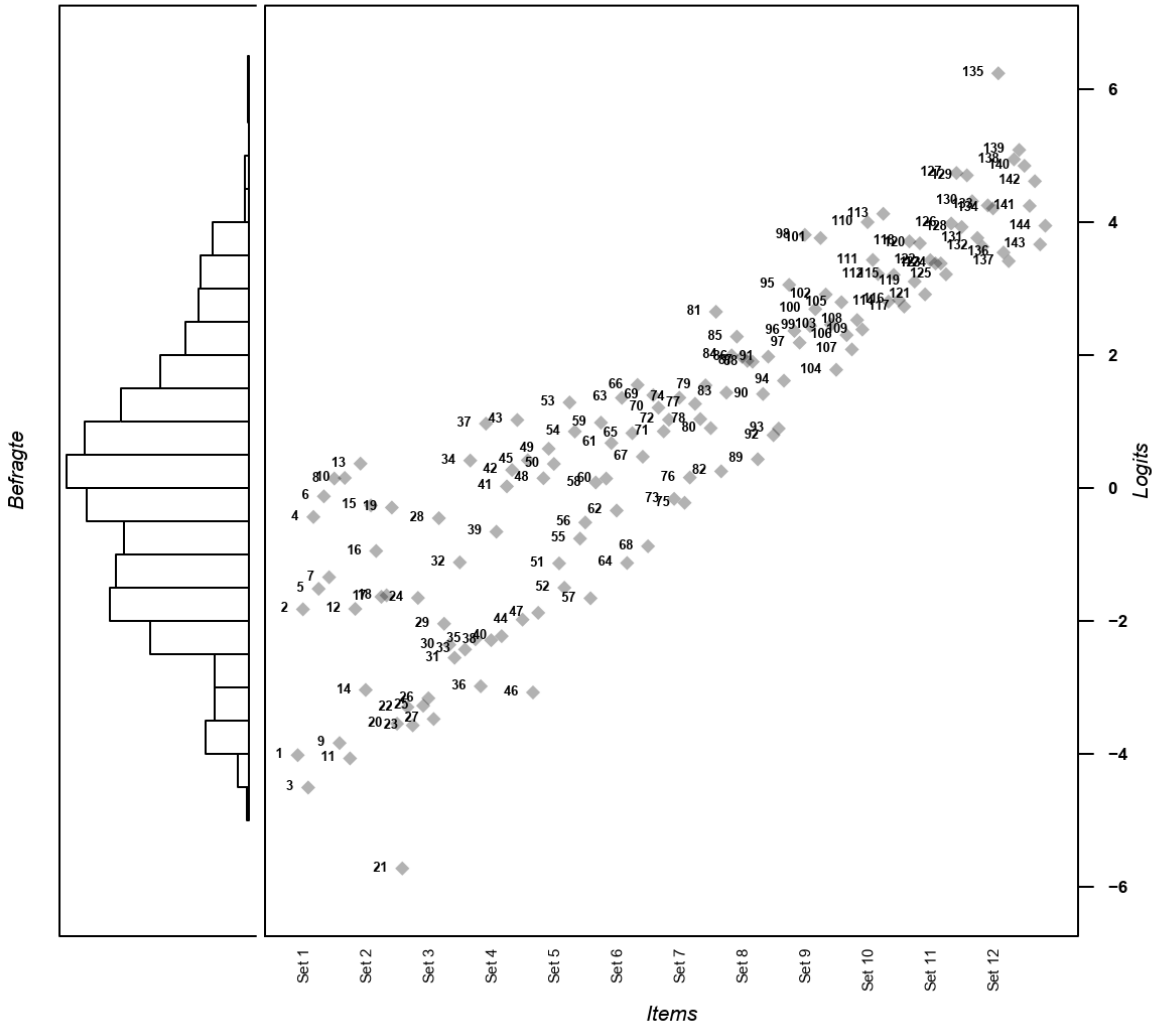

Abbildung 6. Test Targeting, RC1 (N = 1.380)

6.1.3 Qualität der Skala

In diesem Kapitel untersuchen wir die Anpassungsgüte des Rasch-Modells, gehen auf die Ergebnisse der Differential Item Functioning-Analyse ein und evaluieren Rasch-Homogenität und Eindimensionalität.

6.1.3.1 *Item-Fit*

Die Evaluation der Item-Fits wurde auf Basis des Rasch-Modells vorgenommen. Die WMNSQ Werte lagen zwischen 0,56 und 1,61. 32 Items lagen über einem WMNSQ Wert von 1,15, wiesen also mindestens einen bemerkenswerten Item-Misfit auf. 25 Items lagen über einem WMNSQ Wert von 1,2, was wir als starken Item-Misfit werteten (s. Kapitel 5.3, vgl. Pohl & Carstensen, 2012). Eine visuelle Inspektion der Itemcharakteristikkurven der auffälligen Items wies jedoch auf eine ausreichende Passung zum Rasch-Modell hin. In Tabelle 5 sind alle Items aufgelistet, die einen WMNSQ Wert größer als 1,15 hatten.

Tabelle 5: Itemparameter mit Item-Misfit, RC1

Item	kA	Schwierigkeit	SE	WMNSQ	t	r_{it}	Disk.	Q3
t0501011	94,13	-4,02	0,13	1,25	2,63	0,17	0,52	0,03
t0501021	75,94	-1,82	0,08	1,22	5,02	0,40	0,77	0,04
t0501041	55,94	-0,43	0,07	1,52	14,06	0,33	0,45	0,06
t0501051	71,96	-1,51	0,07	1,41	9,64	0,32	0,54	0,04
t0501061	51,01	-0,12	0,07	1,46	12,70	0,35	0,46	0,06
t0501071	69,57	-1,34	0,07	1,23	5,85	0,44	0,78	0,04
t0501081	46,74	0,14	0,07	1,61	16,22	0,28	0,35	0,06
t0501101	46,59	0,15	0,07	1,56	15,25	0,30	0,35	0,06
t0501111	94,35	-4,07	0,13	1,16	1,71	0,18	0,66	0,02
t0501121	75,87	-1,82	0,08	1,35	7,92	0,35	0,60	0,03
t0502011	43,19	0,37	0,07	1,34	9,58	0,41	0,62	0,05
t0502031	53,19	-0,26	0,07	1,35	9,83	0,42	0,64	0,05
t0502041	63,84	-0,95	0,07	1,28	7,55	0,42	0,74	0,04
t0502051	73,55	-1,64	0,07	1,20	5,00	0,41	0,82	0,03
t0502071	53,70	-0,29	0,07	1,15	4,63	0,50	0,94	0,04
t0502121	73,77	-1,65	0,07	1,17	4,16	0,42	0,84	0,03
t0503041	56,23	-0,45	0,07	1,23	6,67	0,47	0,83	0,05
t0503101	42,46	0,41	0,07	1,51	13,63	0,33	0,54	0,05
t0504011	34,06	0,97	0,07	1,32	8,33	0,42	0,78	0,04
t0504051	48,62	0,02	0,07	1,29	8,48	0,45	0,80	0,04
t0504061	44,71	0,27	0,07	1,23	6,72	0,48	0,88	0,04
t0504071	33,26	1,02	0,07	1,21	5,53	0,47	0,92	0,04
t0504091	42,39	0,42	0,07	1,51	13,79	0,33	0,58	0,04
t0504121	46,67	0,15	0,07	1,18	5,40	0,50	0,95	0,04

Item	kA	Schwierigkeit	SE	WMNSQ	t	r_{it}	Disk.	Q3
t0505011	39,71	0,59	0,07	1,21	6,07	0,47	1,05	0,06
t0505051	29,57	1,29	0,07	1,30	7,46	0,42	0,95	0,06
t0506031	28,62	1,36	0,07	1,17	4,20	0,49	1,24	0,04
t0506051	36,09	0,83	0,07	1,19	5,32	0,49	1,21	0,05
t0506091	28,01	1,40	0,07	1,35	8,18	0,41	1,05	0,06
t0507091	14,19	2,65	0,09	1,35	5,72	0,32	1,02	0,06
t0508011	17,67	2,28	0,09	1,29	5,32	0,41	1,43	0,05
t0508021	20,91	1,97	0,08	1,18	3,77	0,48	1,62	0,06

Anmerkungen. Item: Itembezeichnung, kA: Anteil korrekter Anworten in Prozent, Schwierigkeit: Itemschwierigkeit, SE: Standardfehler der Itemschwierigkeit, WMNSQ: Weighted Mean Square, t: *t*-Wert für Weighted Mean Square, r_{it}: korrigierte Trennschärfe, Disk.: Diskriminationsparameter des zwei-parametrischen Modells, Q_3: gemittelte absolute Residuenkorrelationen. Die Schätzung basiert auf N = 1.380 Fällen. Die Tabelle enthält nur Items mit WMNSQ > 1,15.

Die korrigierten Trennschärfen der Items lagen zwischen 0,14 und 0,74 (*M* = 0,48, *SD* = 0,13). 4 Items hatten einen Wert unter 0,20, d.h. die Korrelation zwischen dem entsprechenden Item und dem um dieses Item korrigierten Testscore war kleiner als 0,20, was als problematisch zu erachten ist (s. Kapitel 5.3, vgl. Pohl & Carstensen, 2012): t0501011, t0501031, t0501111, t0502091.

6.1.3.2 *Differential Item Functioning*

Differential Item Functioning wurde für die Variablen Geschlecht, Herkunft und Bildung der Person mit dem höchsten Bildungsabschluss im Haushalt (ISCED) betrachtet. Die DIF-Werte für die verschiedenen Gruppen haben wir in Tabelle 6 abgebildet, wobei nur Zeilen berücksichtigt wurden, die für mindestens eine Gruppe einen DIF-Wert größer als 0,4 aufwiesen. War die absolute Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen kleiner als 0,4, betrachteten wir dies als vernachlässigbar (s. Kapitel 5.3, vgl. Pohl & Carstensen, 2012). Die Spalten beinhalten die Differenzen der Schwierigkeitsparameter zwischen den Subgruppen. Vergleicht man z.B. Jungen und Mädchen, bedeutet ein Wert größer als null, dass den Mädchen das Item leichter fiel als den Jungen (vgl. Fischer & Durda, 2020). Zudem verglichen wir die Modellanpassung des DIF-Modells mit der des Haupteffekt-Modells (s. Tabelle 7).

Geschlecht

In die Analyse wurden Daten von 712 Jungen (51,59%) und von 668 Mädchen (48,41%) einbezogen. Im Durchschnitt waren die Fähigkeiten der Jungen mit denen der Mädchen vergleichbar (Haupteffekt = 0,18 Logits). Jedoch gab es von 144 Items 26 Items mit einem DIF-Wert größer als 0,4, also einer beachtlichen Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen. Nur zwei Items hatten sehr starken DIF über 1, der jedoch einmal Mädchen und einmal Jungen bevorzugte. Die 26 Items mit DIF-Wert größer als 0,4 hatten jedoch nur eine geringe Auswirkung auf den Haupteffekt. Beim Modellvergleich zwischen DIF-Modell

und Haupteffekt-Modell bevorzugte das AIC das DIF-Modell, wohingegen das BIC das Haupteffekt-Modell bevorzugte. Da das BIC auch die Anzahl zu schätzender Parameter berücksichtigt, entschieden wir uns für das sparsamere Modell und gehen davon aus, dass die DIF-Effekte insgesamt zu vernachlässigen waren.

Herkunftsland

1.037 Kinder kamen aus Syrien (75,31%) und 340 Kinder kamen nicht aus Syrien (24,69%). Zwischen den Kindern, die aus Syrien kamen und denen, die aus einem anderen Land kamen, gab es einen vernachlässigbaren Unterschied in den Fähigkeiten (Haupteffekt = -0,11 Logits). Es gab jedoch 27 Items mit DIF-Werten größer als 0,4, also einer beachtlichen Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen. Diese Items hatten jedoch nur eine geringe Auswirkung auf den Haupteffekt. Im Modellvergleich wurde sowohl durch AIC als auch BIC das Haupteffektmodell bevorzugt. Insgesamt können damit die DIF-Werte für Herkunftseffekte vernachlässigt werden.

Bildung

Bei 759 Kindern haben Erwachsene des Haushaltes mindestens einen Sekundarabschluss (57,81%). Bei 554 Kindern haben Erwachsene des Haushaltes höchstens einen Primärabschluss (42,19%). Kinder, die in Haushalten mit gebildeteren Personen lebten, wiesen kaum Unterschiede in den Fähigkeiten im Vergleich zu Kindern auf, die in Haushalten mit geringer gebildeten Personen lebten (Haupteffekt = -0,16 Logits). Es gab jedoch 14 Items, die DIF-Werte größer als 0,4 aufwiesen, d.h. für diese Items ist die Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen beachtlich. Diese Items hatten jedoch nur eine geringe Auswirkung auf den Haupteffekt. Im Modellvergleich wurde sowohl durch AIC als auch BIC das Haupteffektmodell bevorzugt. Insgesamt können damit die DIF-Werte für Bildungseffekte vernachlässigt werden.

Tabelle 6: Differential Item Functioning, RC1

Item	**Geschlecht**	**Herkunftsland**	**Bildung**
	Jungen vs. Mädchen	Syrien vs. Sonstige	höhere vs. niedrigere Bildung
t0501031	0,34	-0,59	-0,27
t0501051	0,01	0,05	-0,43
t0501081	-0,55	-0,18	-0,19
t0501091	1,67	-0,16	0,09
t0502011	-0,43	0,07	-0,19
t0502021	-0,08	0,45	0,10
t0502081	-0,40	0,61	-0,17
t0502091	0,90	0,57	0,08
t0502111	0,09	-0,65	0,00
t0502121	0,15	-0,25	-0,43
t0503031	0,87	-0,22	-0,07
t0503061	0,43	-0,12	0,02
t0503091	0,30	-0,59	-0,47
t0503101	0,57	-0,29	-0,36
t0504011	-0,47	0,02	0,21
t0504041	0,21	-0,55	-0,56
t0504051	-0,62	0,56	0,16
t0504071	-0,58	0,08	-0,16
t0504111	-0,07	-0,43	-0,04
t0504121	-0,34	0,45	-0,06
t0505011	-0,21	0,46	-0,04
t0505051	0,21	0,49	0,00
t0505111	-0,48	-0,11	-0,09
t0505121	0,80	0,19	-0,34
t0506041	0,41	-0,04	-0,15
t0507051	-0,91	0,24	0,03
t0507061	0,52	-0,23	0,02
t0507071	0,54	-0,47	0,20

Item	Geschlecht	Herkunftsland	Bildung
	Jungen vs. Mädchen	Syrien vs. Sonstige	höhere vs. niedrigere Bildung
t0508061	-0,26	0,41	0,35
t0508081	0,48	-0,03	-0,21
t0508121	0,54	-0,43	-0,14
t0509031	0,27	0,40	-0,19
t0509051	0,26	0,59	-0,59
t0509121	-0,08	-0,50	-0,08
t0510021	-0,09	0,71	0,31
t0510051	0,16	0,39	0,50
t0510091	-0,17	0,05	0,46
t0510101	-0,61	0,31	0,01
t0510121	-0,30	-0,56	0,23
t0511021	-0,23	0,07	0,48
t0511041	-0,42	-0,13	0,54
t0511061	-0,48	0,20	0,32
t0511071	-0,93	-0,52	0,37
t0511081	-0,22	-0,58	-0,28
t0511091	-0,48	-0,18	0,11
t0511121	-0,32	0,13	0,49
t0512021	-0,05	-0,76	0,05
t0512031	-1,56	-0,84	0,55
t0512061	-0,89	-0,48	0,50
t0512071	-0,20	0,35	0,67
t0512091	-0,38	-0,56	-0,07
t0512101	0,24	0,47	-0,44
t0512111	-0,42	-0,19	0,35
Haupteffekt (DIF)	0,14	-0,14	-0,11
Haupteffekt (Haupteffektmodell)	0,18	-0,11	-0,16

Anmerkungen. Differenzen zwischen Itemschwierigkeiten. Es werden nur Items berücksichtigt, die einen DIF-Wert größer als 0,4 haben.

Tabelle 7: Modellvergleich zwischen Modellen mit und ohne DIF, RC1

DIF-Variable	Modell	N	Devianz	Anzahl an Parametern	AIC	BIC
Geschlecht	DIF	1.380	134969,6	288	135547,6	137059,0
	Haupteffekt	1.380	135456,5	145	135748,5	136512,1
Herkunftsland	DIF	1.377	134982,3	288	135560,3	137071,1
	Haupteffekt	1.377	135221,2	145	135513,2	136276,5
Bildung	DIF	1.313	128777,2	288	129355,2	130852,3
	Haupteffekt	1.313	128980,9	145	129272,9	130029,2

6.1.3.3 *Rasch-Homogenität*

Um die Annahme zu testen, dass alle Diskriminationsparameter den gleichen Wert haben, wurden die Diskriminationsparameter auf Itemebene anhand eines zwei-parametrischen Testmodells geschätzt und dieses wurde mit dem Rasch-Modell verglichen (vgl. Fischer & Durda, 2020; Pohl & Carstensen, 2012). Die mit dem zwei-parametrischen Modell geschätzten Diskriminationsparameter bewegten sich zwischen 0,35 und 12,38 (*M* = 2,68; *SD* = 2,32). Die Informationskriterien implizierten eine bessere Modellanpassung für das zwei-parametrische Modell (AIC: 127.815, BIC: 129.321) im Gegensatz zum Rasch-Modell (AIC: 135.794, BIC: 136.507). Somit scheint die Annahme gleicher Diskriminationsparameter aus empirischer Sicht fraglich. Da aus theoretischer Sicht jedoch das Rasch-Modell Basis der Testkonstruktion war und auch die Verrechnung der Itemantworten als Summenscores (vgl. Lenhard et al., 2015) die Gültigkeit des Rasch-Modells impliziert, wurde für die aktuelle Stichprobe das Rasch-Modell beibehalten.

6.1.3.4 *Eindimensionalität*

Die Dimensionalität des Tests wurde mithilfe der Korrelationen zwischen den Residuen des Rasch-Modells evaluiert (vgl. Gnambs, 2017). Die angepasste Q_3-Statistik bewegte sich zwischen 0,02 und 0,11 (*M* = 0,05, *SD* = 0,02) und war damit relativ niedrig. Damit ist die notwendige Annahme der Eindimensionalität gegeben.

6.2 RC2

6.2.1 Fehlende Werte

An dieser Stelle wollen wir die fehlenden Werte auf Itemebene behandeln und nur die oben beschriebenen *N* = 1.451 gültigen Fälle betrachten.

6.2.1.1 *Fehlende Werte pro Person*

Fehlende Angaben können zum Beispiel entstehen, wenn die Jugendlichen Items nicht beantworteten bzw. übersprangen (Missingwert -97, s. Kapitel 5.1). Wie in Abbildung 7 zu sehen,

beantworteten circa 77% der Jugendlichen alle vorgegebenen Items. 2,21% der Jugendlichen gaben keine Antwort auf fünf oder mehr Items.

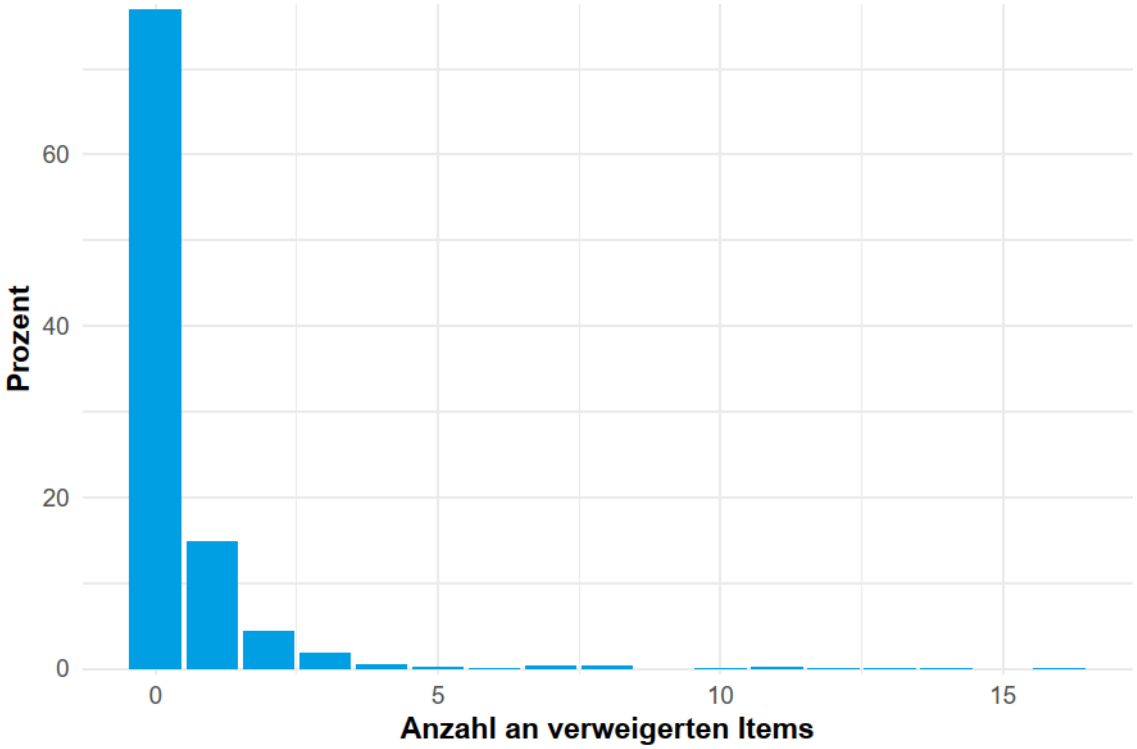

Abbildung 7. Anzahl verweigerter Items (in Prozent) (N = 1.451)

Außerdem sind Häufigkeiten über nicht erreichte Items bzw. Sets von Interesse (Missingwert -94, s. Kapitel 5.1). In Abbildung 8 ist zu erkennen, dass 1,24% der Fälle alle Items/Sets erreicht haben. Am häufigsten wurden 9 Sets nicht erreicht, wobei von den *N* = 1.451 einbezogenen Fällen 16,75% aufgrund des Abbruchkriteriums und 0,21% aufgrund des Abbruches durch die Interviewerin oder den Interviewer die letzten neun Sets nicht erreicht haben. 4,62% der Fälle haben 14 Sets nicht erreicht, haben also im für die RC2 gewählten Startset (Set 5) bereits das Abbruchkriterium erreicht. Falls im Übungsteil mindestens ein Fehler gemacht wurde, haben auch die Jugendlichen die Bearbeitung der Testphase bei Set 1 begonnen (s. Kapitel 2.2). Nur so konnten weitere 0,83% der einbezogenen Fälle 15 und weitere 0,28% der einbezogenen Fälle 16 Sets nicht erreichen. Im Durchschnitt haben die Jugendlichen 8,45 (SD = 3,28) Sets nicht bearbeitet. Wie oben erläutert, wurden Fälle, die im ersten Set bereits das Abbruchkriterium erreicht und somit 18 Sets nicht erreicht haben, von den Analysen ausgeschlossen.

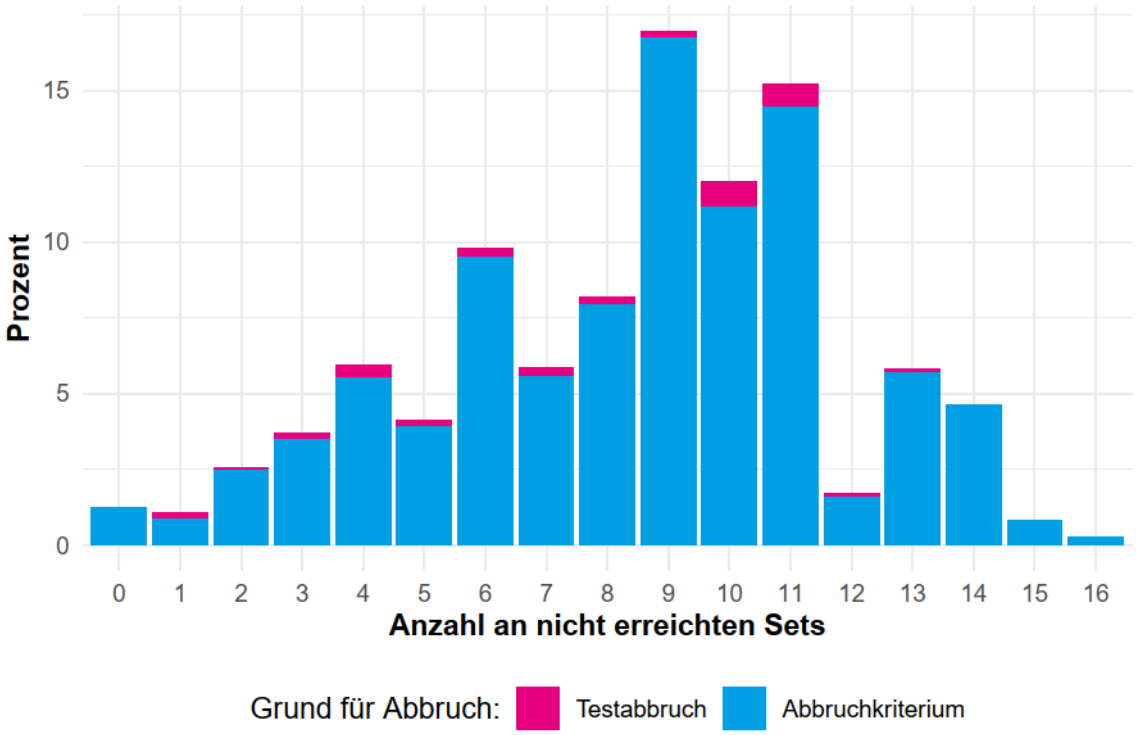

Abbildung 8. Anzahl an nicht erreichten Sets (in Prozent) (N = 1.451)

6.2.1.2 *Fehlende Werte pro Item*

In diesem Unterkapitel gehen wir auf die fehlenden Werte pro Item ein. Tabelle 8 enthält die Anteile der Missingkategorie -97 (Angabe verweigert) je Item in Prozent, wobei nur solche Items abgebildet wurden, bei denen der Anteil bei mindestens 1,5% lag. In Set 5 gibt es drei Items, auf die dieses Kriterium zutrifft (die meisten Jugendlichen haben mit diesem Set begonnen). Werte von mindestens 1,5% tauchen dann erst wieder in Set 15, Set 16 und Set 18 auf. Der höchste Anteil findet sich jedoch in Set 5 mit 3,69%. Der niedrigste Prozentwert liegt bei 0,00% (*Mdn* = 0,20%). Wie in Kapitel 5.1 erwähnt, werden nicht beantwortete Items für die folgenden Analysen als Fehler umkodiert.

Tabelle 8: Fehlende Werte (Angabe verweigert) pro Item

Pos.	Item	Anzahl gültiger Fälle	Angabe verweigert (in %)
49	t0505011	1.382	3,69
50	t0505021	1.396	2,72
53	t0505051	1.409	1,81
174	t0515061	202	2,42
188	t0516081	122	1,61
191	t0516111	120	1,64
192	t0516121	118	3,28
207	t0518031	32	3,03

Anmerkungen. N = 1.451; Pos.: Item Position im Testverlauf; nur Werte größer als 1,5 werden gelistet

Abbildung 9 gibt an, wie viel Prozent der Jugendlichen die Setposition nicht erreicht haben. Circa 85% der Teilnehmerinnen und Teilnehmer haben das 15te Set oder höher nicht erreicht.

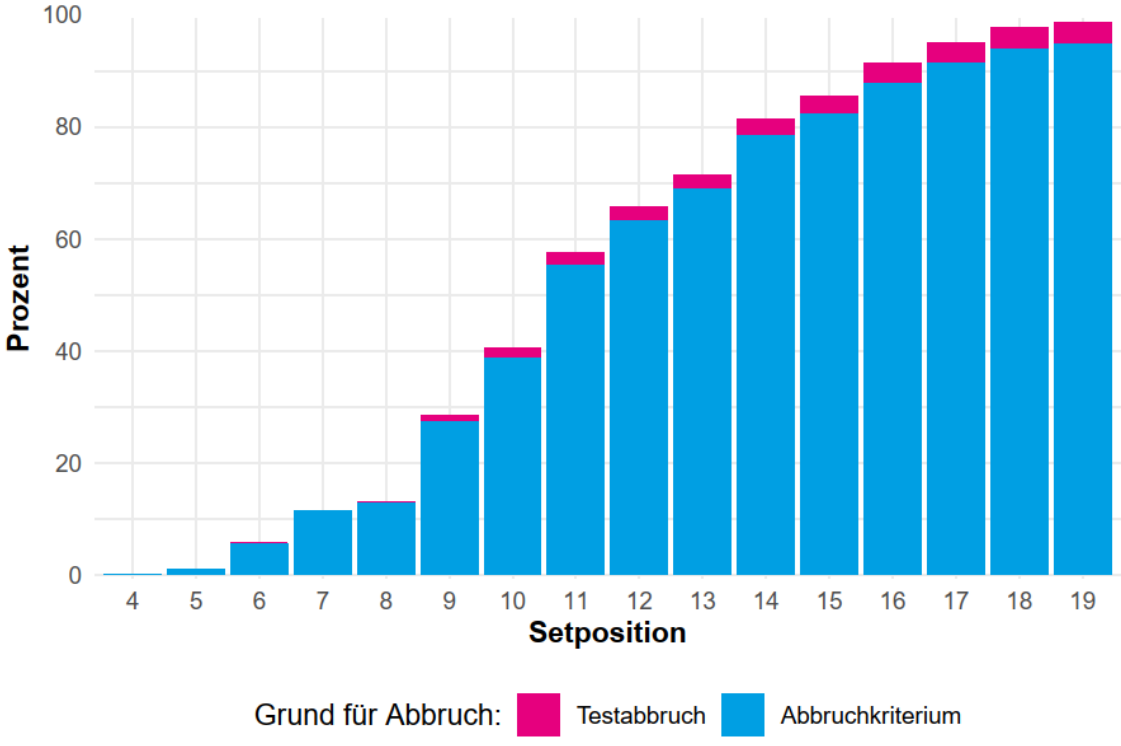

Abbildung 9. Setposition nicht erreicht (in Prozent) (N = 1.451)

6.2.2 Parameterschätzung

Im Folgenden gehen wir auf die Ergebnisse der geschätzten Itemparameter sowie auf Test Targeting und Reliabilität der Personenparameter ein. 36 Items, die nicht mehr als 100 gültige Antworten enthielten, wurden aus der Analyse ausgeschlossen (s. Kapitel 5.2).

6.2.2.1 *Itemparameter*

Die Itemparameter schätzten wir anhand des TAM-Package in R (s. Kapitel 5). Die fehlenden Werte, die auf verweigerte oder nicht erreichte Items zurückgehen, wurden als Fehler gewertet. Fehlende Werte, die auf abgebrochene Testungen (bei mindestens fünf vollständig bearbeiteten Itemsets) zurückgehen, wurden hingegen ignoriert (s. Kapitel 5.2).

Wir werteten unter anderem den Anteil an korrekten Antworten aus. Dieser variierte zwischen 1,93% und 99,93%. Das arithmetische Mittel lag bei 46,92% (*SD* = 34,15%). Dies legt nahe, dass die Items einen breiten Fähigkeitsbereich abdeckten.

Die Itemschwierigkeiten des Rasch-Modells wurden geschätzt indem der Mittelwert der Personenfähigkeitsverteilung auf null fixiert wurde. Die Itemschwierigkeiten lagen zwischen -8,78 (t0502091) und 5,40 (t0516061) mit einem Mittelwert von 0,18 (*SD* = 3,24) und umfassten sowohl einfache wie auch schwierige Items. Die Standardfehler der Itemschwierigkeiten lagen zwischen 0,06 und 1,01 (*M* = 0,12, *SD* = 0,09).

6.2.2.2 *Test Targeting und Reliabilität*

Um die Angemessenheit des Tests für die Zielpopulation zu untersuchen, wurden die Itemschwierigkeiten und die Personenfähigkeiten (WLEs) gegenübergestellt (vgl. Fischer & Durda, 2020). In Abbildung 10 sind die Itemschwierigkeiten und die Personenfähigkeiten auf einer Skala abgebildet. Auf der linken Seite ist die Verteilung der Personenfähigkeiten und auf der rechten Seite die Verteilung der Itemschwierigkeit zu erkennen, wobei die Items nach Itemposition sortiert sind. Zum besseren Überblick wurden auf der X-Achse die Sets angegeben, aus welchem die Items stammen. Der Mittelwert der Fähigkeitsverteilung wurde auf null fixiert, während die Items eine durchschnittliche Schwierigkeit von M = 0,18 (*SD* = 3,24) aufwiesen. Das bedeutet, dass die Items im Durchschnitt etwas zu schwierig für die Stichprobe waren. Aber die Spannweite der Itemschwierigkeiten war mit 14,18 sehr groß und deckte sowohl niedrige wie auch hohe Fähigkeitsbereiche ab. Es ist dabei zu beachten, dass nicht alle Items allen Teilnehmerinnen und Teilnehmern vorgegeben wurden und dass nicht bearbeitete Items unterhalb des Bodensets als richtig, nicht bearbeitete Items oberhalb des Deckensets als falsch kodiert wurden. Über die Sets hinweg schien die Itemschwierigkeit anzusteigen, innerhalb der Sets variierte die Itemschwierigkeit jedoch teilweise stark. Alles in allem scheint die Testschwierigkeit für die untersuchte Stichprobe angemessen zu sein. Die Varianz der Personenfähigkeiten war mit 3,11 groß und erlaubt eine gute Differenzierung zwischen den Jugendlichen. Die Reliabilität des Tests war sehr gut (EAP Reliabilität = 0,98; WLE Reliabilität = 0,98).

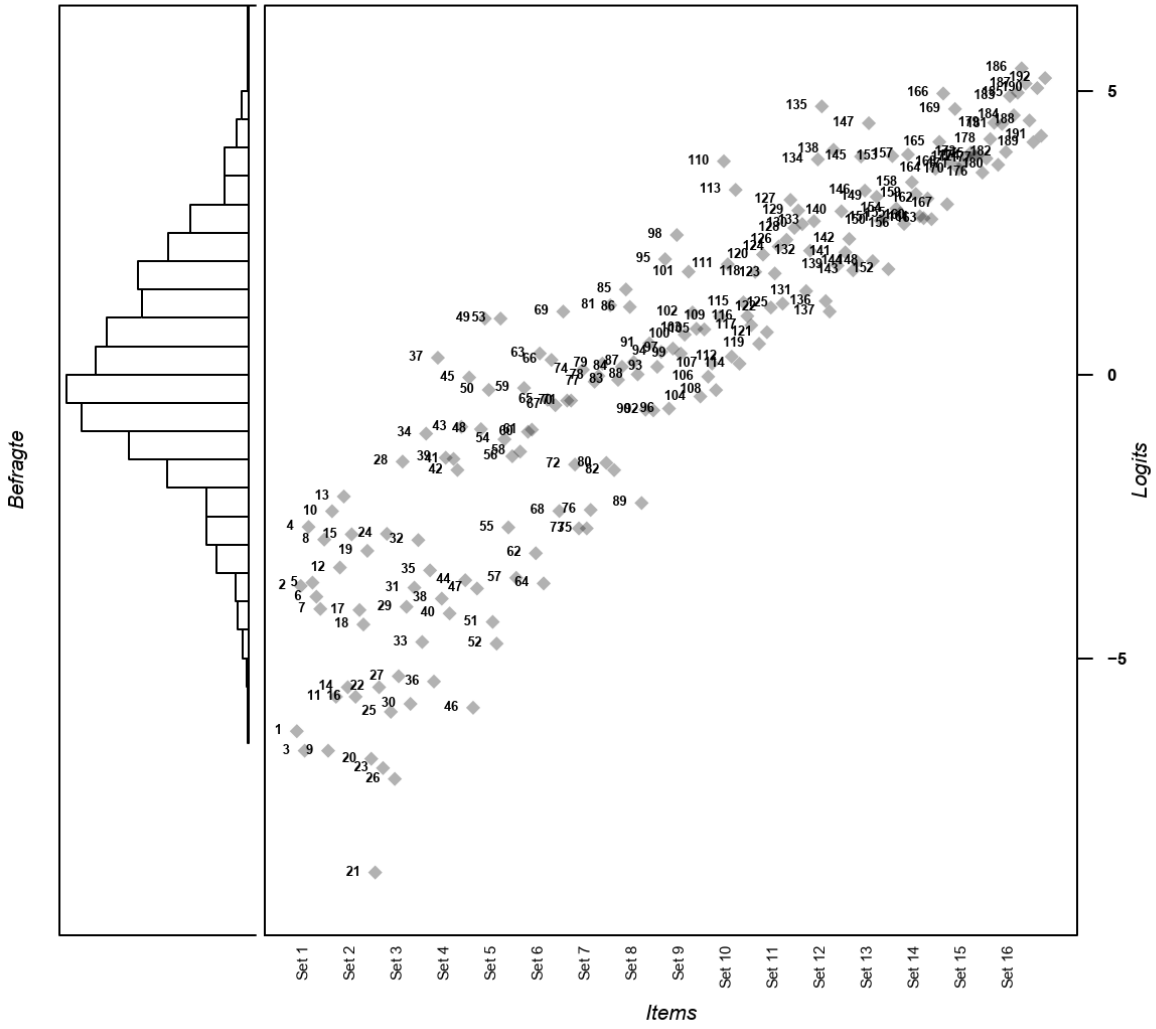

Abbildung 10. Test Targeting, RC2 (N = 1.451)

6.2.3 Qualität der Skala

In diesem Kapitel untersuchen wir die Anpassungsgüte des Rasch-Modells, gehen auf die Ergebnisse der Differential Item Functioning-Analyse ein und evaluieren Rasch-Homogenität und Eindimensionalität.

6.2.3.1 *Item-Fit*

Die Evaluation von Item-Fits wurde auf Basis des Rasch-Modells vorgenommen. Die WMNSQ Werte lagen zwischen 0,56 und 1,62. 36 Items lagen über einem WMNSQ Wert von 1,15, wiesen also mindestens einen bemerkenswerten Item-Misfit auf. 27 Items lagen über einem WMNSQ Wert von 1,2, was wir als starken Item-Misfit werteten (s. Kapitel 5.3, vgl. Pohl & Carstensen, 2012). Eine visuelle Inspektion der Itemcharakteristikkurven der auffälligen Items wies jedoch auf eine ausreichende Passung zum Rasch-Modell hin. In Tabelle 9 sind alle Items aufgelistet, die einen WMNSQ Wert größer als 1,15 haben.

Tabelle 9: Itemparameter mit Item-Misfit, RC2

Item	kA	Schwierigkeit	SE	WMNSQ	t	r_{it}	Disk.	Q3
t0503041	74,98	-1,54	0,07	1,27	6,42	0,38	0,67	0,05
t0503071	93,80	-3,76	0,12	1,18	1,98	0,25	0,78	0,03
t0503101	68,16	-1,04	0,07	1,25	6,90	0,41	0,69	0,05
t0504011	46,80	0,30	0,06	1,26	7,99	0,43	0,69	0,06
t0504021	94,62	-3,95	0,13	1,25	2,49	0,15	0,49	0,03
t0504031	74,09	-1,47	0,07	1,33	8,05	0,34	0,59	0,05
t0504051	74,36	-1,49	0,07	1,25	6,24	0,39	0,71	0,04
t0504071	66,37	-0,92	0,07	1,30	8,42	0,38	0,64	0,04
t0504091	52,52	-0,05	0,06	1,23	7,18	0,44	0,69	0,04
t0504121	66,99	-0,96	0,07	1,25	7,26	0,42	0,70	0,04
t0505011	35,84	0,99	0,07	1,62	15,82	0,24	0,37	0,06
t0505021	56,10	-0,27	0,07	1,55	16,01	0,27	0,41	0,05
t0505051	35,84	0,99	0,07	1,16	4,60	0,48	0,86	0,05
t0505071	87,18	-2,69	0,09	1,16	2,62	0,30	0,75	0,03
t0505081	73,74	-1,44	0,07	1,19	4,84	0,41	0,79	0,03
t0505101	72,64	-1,36	0,07	1,18	4,63	0,40	0,80	0,04
t0505111	55,55	-0,24	0,06	1,17	5,56	0,46	0,82	0,04
t0506031	45,55	0,37	0,07	1,27	8,21	0,43	0,75	0,05
t0506051	58,92	-0,44	0,07	1,28	8,51	0,39	0,72	0,04
t0506061	47,45	0,26	0,06	1,36	10,84	0,39	0,67	0,04
t0506091	34,02	1,11	0,07	1,33	8,92	0,39	0,71	0,04
t0506101	59,21	-0,46	0,07	1,18	5,67	0,45	0,87	0,03
t0507051	53,80	-0,13	0,06	1,27	8,60	0,41	0,74	0,06
t0507061	52,14	-0,02	0,06	1,31	9,64	0,40	0,72	0,05
t0507071	48,55	0,19	0,06	1,23	7,23	0,45	0,83	0,05
t0507111	53,28	-0,09	0,06	1,34	10,53	0,38	0,69	0,05
t0507121	49,48	0,14	0,06	1,42	12,53	0,35	0,63	0,05
t0508011	28,47	1,50	0,07	1,63	14,36	0,23	0,43	0,06
t0508021	32,85	1,19	0,07	1,42	10,70	0,35	0,66	0,05
t0508071	42,83	0,55	0,07	1,41	11,87	0,36	0,69	0,05
t0508101	45,28	0,40	0,07	1,26	7,81	0,45	0,84	0,05

Item	kA	Schwierigkeit	SE	WMNSQ	t	r_{it}	Disk.	Q3
t0509011	44,36	0,46	0,07	1,16	5,10	0,48	1,07	0,04
t0509021	17,14	2,46	0,08	1,19	3,66	0,42	1,23	0,04
t0509031	45,78	0,37	0,07	1,20	6,29	0,44	1,02	0,05
t0509051	24,51	1,81	0,07	1,24	5,56	0,43	1,00	0,04
t0510101	24,56	1,81	0,07	1,24	5,58	0,44	1,16	0,04

Anmerkungen. Item: Itembezeichnung, kA: Anteil korrekter Antworten in Prozent, Schwierigkeit: Itemschwierigkeit, SE: Standardfehler der Itemschwierigkeit, WMNSQ: Weighted Mean Square, t: *t*-Wert für Weighted Mean Square, r_{it}: korrigierte Trennschärfe, Disk.: Diskriminationsparameter des zwei-parametrischen Modells, Q_3: gemittelte absolute Residuenkorrelationen. Die Schätzung basiert auf N = 1.451 Fällen. Die Tabelle enthält nur Items mit WMNSQ > 1,15.

Die Trennschärfen der Items lagen zwischen 0,05 und 0,78 (*M* = 0,44, *SD* = 0,15). 16 Items hatten einen Wert unter 0,20, d.h. die Korrelation zwischen dem entsprechenden Item und dem um dieses Item korrigierten Testscore war kleiner als 0,20, was als problematisch zu erachten ist (s. Kapitel 5.3, vgl. Pohl & Carstensen, 2012): t0501011, t0501031, t0501091, t0501111, t0502021, t0502081, t0502091, t0502101, t0502111, t0503011, t0503021, t0503061, t0503091, t0504021, t0504101, t0505031.

6.2.3.2 *Differential Item Functioning*

Differential Item Functioning wurde für die Variablen Geschlecht, Herkunft und Bildung der Person mit dem höchsten Bildungsabschluss im Haushalt (ISCED) betrachtet. Die DIF-Werte für die verschiedenen Gruppen haben wir in Tabelle 10 abgebildet, wobei nur Zeilen berücksichtigt wurden, die mindestens einen DIF-Wert größer als 0,4 aufwiesen. Ist die absoulte Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen kleiner als 0,4, betrachten wir dies als vernachlässigbar (s. Kapitel 5.3, vgl. Pohl & Carstensen, 2012). Die Spalten beinhalten die Differenzen der Schwierigkeitsparameter zwischen den Subgruppen. Vergleicht man z.B. Jungen und Mädchen, bedeutet ein Wert größer als null, dass den Mädchen das Item leichter fiel als den Jungen (vgl. Fischer & Durda, 2020). Zudem verglichen wir die Modellanpassung des DIF-Modells mit der des Haupteffekt-Modells (s. Tabelle 11).

Geschlecht

In die Analyse wurden Daten von 813 Jungen (56,03%) und von 638 Mädchen (43,97%) einbezogen. Im Durchschnitt scheint der Test den Jungen leichter gefallen zu sein als den Mädchen (Haupteffekt = -0,33 Logit). Von 192 Items gab es 35 Items, die mindestens einen DIF-Wert von 0,4 aufwiesen, d.h. für diese Items ist die Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen beachtlich. Diese Items hatten jedoch nur eine geringe Auswirkung auf den Haupteffekt. Beim Modellvergleich zwischen DIF-Modell und Haupteffektmodell bevorzugt das AIC das DIF-Modell, wohingegen das BIC das sparsamere Haupteffektmodell bevorzugte. Damit lag wenig Evidenz für substantielles DIF für das Geschlecht vor.

Herkunftsland

964 Jugendliche kamen aus Syrien (74,90%) und 323 Jugendliche kamen nicht aus Syrien (25,10%). Zwischen den Jugendlichen, die aus Syrien kamen und denen, die aus einem ande-

ren Land kamen, gab es einen vernachlässigbaren Unterschied in den Fähigkeiten (Haupteffekt: 0,05 Logit). Es gab 36 Items mit DIF-Werten größer als 0,4, also einer beachtlichen Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen. Diese Items hatten jedoch nur eine geringe Auswirkung auf den Haupteffekt. Im Modellvergleich wurde sowohl durch AIC als auch BIC das Haupteffektmodell bevorzugt. Insgesamt können damit die DIF-Werte für Herkunft vernachlässigt werden.

Bildung

In 571 Fällen haben die Eltern mindestens einen Sekundarabschluss (39,35%). In 402 Fällen haben Erwachsene des Haushaltes höchstens einen Primärabschluss (27,71%). In 478 Fällen haben Erwachsene des Haushaltes keine Angaben zum Bildungsabschluss gemacht (32,94%). Jugendliche, die in Haushalten mit gebildeteren Personen lebten, fiel der Test leichter im Vergleich zu Jugendlichen, die in Haushalten mit geringer gebildeten Personen lebten (Haupteffekt: -0,81 Logit). Auch gegenüber Jugendlichen aus Haushalten, deren Erwachsene keine Angaben zu Bildungsabschlüssen gemacht haben, fiel Jugendlichen aus gebildeteren Haushalten der Test leichter (Haupteffekt: -0,46 Logit). Vergleicht man hingegen Jugendliche aus weniger gebildeten Haushalten mit Jugendlichen aus Haushalten ohne Bildungsangaben, zeigt sich, dass letztere Vorteile gegenüber ersteren hatten (Haupteffekt: 0,34). Im ersten Gruppenvergleich gab es 67 Items, im zweiten Gruppenvergleich 31 Items und im letzten Gruppenvergleich 57 Items, die DIF-Werte größer als 0,4 aufwiesen, d.h. für diese Items ist die Differenz der geschätzten Itemschwierigkeiten zwischen den Gruppen beachtlich. Im Modellvergleich wurde jedoch sowohl durch AIC als auch BIC das Haupteffekt-Modell bevorzugt. Insgesamt können damit die DIF-Werte für Bildung vernachlässigt werden.

Tabelle 10: Differential Item Functioning, RC2

Item	**Geschlecht**	**Herkunfts-land**	**Bildung**		
	Jungen vs. Mädchen	Syrien vs. Sonstige	höher vs. niedriger	höher vs. Missing	niedriger vs. Missing
t0501011	-0,28	-1,15	0,59	1,10	0,52
t0501021	-0,37	0,46	-0,17	-0,13	0,05
t0501031	-0,08	-0,58	2,06	1,11	-0,96
t0501041	-0,22	0,55	0,06	-0,46	-0,52
t0501061	0,73	-0,27	-0,26	-0,09	0,17
t0501091	1,10	-0,91	-0,64	0,85	1,49
t0501121	0,19	0,64	0,04	0,33	0,29
t0502011	-0,22	0,44	0,13	-0,19	-0,31
t0502041	0,26	-0,76	-0,33	0,10	0,44
t0502061	-0,47	0,18	0,62	0,70	0,09
t0502081	1,82	-1,63	-0,39	0,85	1,24
t0502091	-3,36	-4,74	12,85	3,81	-9,04
t0502111	-0,07	-1,63	-0,08	0,85	0,93
t0502121	-0,13	0,41	0,44	-0,36	-0,80
t0503011	0,68	0,51	0,13	0,35	0,22
t0503021	0,36	5,12	-0,09	4,76	4,85
t0503031	0,44	0,35	-0,82	-1,05	-0,24
t0503061	1,22	-0,59	-0,26	2,03	2,29
t0503091	0,76	-0,39	-0,27	0,91	1,17
t0503121	0,48	-0,38	0,16	-0,61	-0,77
t0504021	-0,45	0,14	0,24	-0,33	-0,56
t0504041	0,42	-0,34	-0,44	-0,66	-0,22
t0504051	-0,77	0,62	0,07	-0,12	-0,19
t0504081	-0,91	0,36	-0,28	-0,17	0,11
t0504091	-0,40	-0,11	0,16	-0,06	-0,21
t0504111	-0,21	-0,65	-0,45	-0,14	0,31
t0504121	-0,51	0,90	0,40	-0,04	-0,44
t0505011	-0,49	0,23	0,81	-0,21	-1,02

Item	Geschlecht	Herkunfts-land	Bildung		
	Jungen vs. Mädchen	Syrien vs. Sonstige	höher vs. niedriger	höher vs. Missing	niedriger vs. Missing
t0505031	0,35	-0,56	0,04	0,15	0,11
t0505041	-0,37	-0,68	-0,19	0,09	0,27
t0505071	-0,57	-0,44	0,34	0,03	-0,30
t0506101	-0,44	0,19	0,15	-0,15	-0,29
t0507051	-0,40	-0,10	-0,03	-0,13	-0,10
t0507061	1,51	-0,46	-0,04	-0,17	-0,14
t0507101	-1,34	0,44	0,14	0,04	-0,10
t0508101	0,41	0,08	0,41	-0,16	-0,58
t0509021	0,38	-0,83	0,15	-0,06	-0,21
t0509061	-0,36	0,48	-0,24	-0,29	-0,05
t0509071	-0,09	0,50	0,07	0,00	-0,07
t0509111	0,68	-0,14	0,08	0,06	-0,01
t0509121	0,42	-0,15	-0,32	-0,13	0,19
t0510011	0,74	0,34	0,15	0,01	-0,15
t0510061	0,54	0,10	-0,31	-0,05	0,26
t0510081	-0,23	0,43	-0,03	0,04	0,06
t0510101	0,09	-0,61	-0,22	-0,37	-0,15
t0511121	-0,11	0,48	0,14	-0,03	-0,17
t0512021	0,20	0,72	0,44	0,29	-0,15
t0512031	0,46	0,55	0,53	0,27	-0,26
t0512091	-0,25	-0,49	-0,21	-0,20	0,01
t0512101	0,49	-0,01	-0,49	-0,11	0,39
t0513041	-0,41	-0,28	-0,78	-0,24	0,53
t0513061	0,51	-0,10	-0,44	-0,13	0,31
t0514011	-0,40	-0,15	0,08	0,16	0,08
t0514041	-0,25	-0,49	-0,74	-0,41	0,34
t0515011	0,01	0,84	-0,16	-0,20	-0,04
t0515111	-0,19	0,40	-0,12	0,19	0,31
t0516041	-0,36	-0,46	-0,25	0,28	0,53

Item	Geschlecht	Herkunfts-land	Bildung		
	Jungen vs. Mädchen	Syrien vs. Sonstige	höher vs. niedriger	höher vs. Missing	niedriger vs. Missing
t0516051	-0,78	-0,33	-0,47	0,19	0,67
t0516081	-0,41	0,21	-0,49	-0,02	0,46
t0516101	0,59	0,35	-1,08	0,08	1,17
Haupteffekt (DIF)	-0,35	0,00	-0,75	-0,39	0,36
Haupteffekt (Haupteffektmodell)	-0,33	0,05	-0,81	-0,46	0,34

Anmerkungen. Differenzen zwischen Itemschwierigkeiten. Es werden nur Items angezeigt, bei denen Geschlecht oder Herkunft mindestens einen DIF-Wert > 0.4 haben.

Tabelle 11: Modellvergleich zwischen Modellen mit und ohne DIF, RC2

DIF-Variable	Modell	N	Devianz	Anzahl an Parametern	AIC	BIC
Geschlecht	DIF	1.451	167.817,1	384	168.587,1	170.619,9
	Haupteffekt	1.451	168.492,2	193	168.880,2	169.904,5
Nation	DIF	1.287	148.769,3	384	149.539,3	151.525,9
	Haupteffekt	1.287	149.086,8	193	149.474,8	150.475,8
Bildung	DIF	1.451	167.711,5	576	168.865,5	171.912,1
	Haupteffekt	1.451	168.441,9	194	168.831,9	169.861,5

6.2.3.3 *Rasch-Homogenität*

Um die Annahme zu testen, dass alle Diskriminationsparameter den gleichen Wert haben, wurden die Diskriminationsparameter auf Itemebene anhand eines zwei-parametrischen Testmodells geschätzt und dieses wurde mit dem Rasch-Modell verglichen (vgl. Fischer & Durda, 2020). Die mit dem zwei-parametrischen Modell geschätzten Diskriminationsparameter bewegten sich zwischen 0,37 und 9,11 (*M* = 2,31; *SD* = 1,82). Die Informationskriterien implizieren eine bessere Modellanpassung für das zwei-parametrische Modell (AIC: 159.521, BIC: 161.548) im Gegensatz zum Rasch-Modell (AIC: 168.881, BIC: 169.900). Somit scheint die Annahme gleicher Diskriminationsparameter aus empirischer Sicht fraglich. Da aus theoretischer Sicht jedoch das Rasch-Modell Basis der Testkonstruktion war und auch die Verrechnung der Itemantworten als Summenscores (vgl. Lenhard et al., 2015) die Gültigkeit des Rasch-Modells impliziert, wurde für die aktuelle Stichprobe das Rasch-Modell beibehalten.

6.2.3.4 *Eindimensionalität*

Die Dimensionalität des Tests wurde mithilfe der Korrelationen zwischen den Residuen des Rasch-Modells evaluiert (vgl. Gnambs, 2017). Die angepasste Q_3-Statistik bewegte sich zwischen 0,02 und 0,10 (M = 0,04; SD = 0,01) und war damit relativ niedrig. Damit ist die notwendige Annahme der Eindimensionalität gegeben.

7. Diskussion

Abschließend sollen die Ergebnisse unserer Analysen zusammengefasst und die Qualität der Weighted Likelihood Estimations bewertet werden sowie das Potenzial der erhobenen Testdaten diskutiert werden.

Zunächst zeigte sich, dass nur wenige der administrierten Items nicht beantwortet wurden. Die meisten fehlenden Werte gingen auf nicht erreichte Items in den höheren Itemsets zurück, da vielen Kindern und Jugendlichen die letzten Sets aufgrund des Erreichens des Abbruchkriteriums nicht vorgegeben wurden. Da der Test ein breites Leistungsspektrum abbildet und auch im hohen Fähigkeitsbereich gut differenzieren soll, erscheint es jedoch günstig, wenn die letzten Sets nicht von allen Personen erreicht werden und für diese Items damit höhere Raten fehlender Werte zu beobachten sind.

Die untersuchten Itemfit-Indizes zeigten für die meisten Items eine zufriedenstellende Passung zum gewählten Testmodell. Obwohl bei ca. 17% (Kinder) bzw. 14% (Jugendliche) der Testitems zumindest ein moderater Misfit zu beobachten war, wies eine visuelle Inspektion der Itemcharakteristikkurven auf eine ausreichende Passung zum Rasch-Modell hin. Zudem unterstützten Residuenanalysen die Eindimensionalität des Tests.

Die Spannweite der Itemschwierigkeiten war groß und deckte den Fähigkeitsbereich ab. Der PPVT-4 ist so konstruiert, dass die Itemschwierigkeiten über die Itemsets hinweg ansteigen. Das Abbruchkriterium, das eintritt, wenn das Boden- und Deckenset gefunden sind, stellt sicher, dass den Personen die Items vorgegeben werden, die weder zu leicht, noch zu schwierig sind (vgl. Lenhard et al., 2015). Jedoch zeigte sich, dass die Itemschwierigkeiten zwar tendenziell über die Sets hinweg anstiegen, sie jedoch eine große Varianz innerhalb der Sets aufwiesen. Dies könnte eine Besonderheit der untersuchten Stichprobe sein und daran liegen, dass beim Zweitspracherwerb, der in diesen Fällen vermutlich überwiegend im schulischen Kontext und weniger im familiären Kontext stattfindet, Vokabeln in einer anderen Reihenfolge erlernt werden als beim Erstspracherwerb (vgl. Appel, 1996). Wenn die Itemschwierigkeiten jedoch nicht über die Sets hinweg ansteigen, gefährdet das die Testlogik und die Annahme, dass Items oberhalb des Deckensets nicht gewusst werden. Möglicherweise hätten die Teilnehmerinnen und Teilnehmer also auch eine relevante Anzahl nicht erreichter Items, die nach der Testlogik des PPVT als falsch kodiert wurden, richtig beantwortet. Dies würde die Schätzung der Personenfähigkeiten verzerren und könnte zu dem moderaten Misfit mancher Items beigetragen haben. Diese Vermutungen müssten jedoch erst weiter untersucht werden.

Problematisch erscheint, dass die Annahme des Rasch-Modells gleicher Diskriminationsparameter empirisch eine nur unzureichende Bestätigung fand. Jedoch kommt es relativ häufig vor, dass ein zweiparametrisches Modell gegenüber dem Rasch-Modell empirisch bevorzugt wird (vgl. Fischer & Durda, 2020; Gnambs, 2017). Einige Items wiesen jedoch deutlich größere oder geringere Diskriminationsparameter auf als das Rasch-Modell impliziert. Dies liegt unter

anderem daran, dass wir uns entschieden hatten, auch Items mit geringer Anzahl an gültigen Fällen (mindestens 101) in die Bewertung einzubeziehen. Außerdem haben wir nicht erreichte Items nicht wie in anderen Kompetenztestungen im NEPS üblich ignoriert (vgl. Pohl et al., 2014), sondern wie vom Testmanual vorgesehen als falsch gewertet. Dies führte dazu, dass die Diskriminationsparameter in niedrigeren Sets bereits größere Werte annahmen. Allerdings entspricht das Rasch-Modell dem konzeptionellen Konstruktionsrational des Tests und den im Manual des PPVT-4 formulierten Verrechnungsvorschriften für den Summenscore (vgl. Lenhard et al., 2015). Deshalb haben wir uns dazu entschieden, das Rasch-Modell weiter anzuwenden. Hier orientieren wir uns auch am Vorgehen wie es im NEPS bereits gehandhabt wurde (vgl. Fischer & Durda, 2020). Dieser Aspekt sollte jedoch bei der Verwendung der Daten bedacht werden.

Die Verteilung der Personenfähigkeiten wies eine große Varianz auf. Damit erlaubt der Test die Untersuchung von interindividuellen Unterschieden im rezeptiven Wortschatz der untersuchten Stichproben.

Die Testreliabilität wurde möglicherweise etwas überschätzt, da die Items oberhalb des Deckensets alle als falsch beantwortet gewertet wurden und nicht berücksichtigt wurde, dass diese keine Antworten enthielten. Trotzdem weist die hohe Testreliabilität in beiden Kohorten auf eine gute Präzision der erhobenen Messwerte hin.

Schließlich deutet die Evaluierung der Testfairness auf nur geringes DIF hinsichtlich des Geschlechts, Herkunftslands und sozioökonomischen Hintergrunds hin. Damit sind faire Gruppenvergleiche für diese Kriterien möglich. Es muss jedoch betont werden, dass für Gruppenvergleiche in Bezug auf andere Kriterien ReGES-Datennutzerinnen und -Datennutzer vergleichbare Analysen selbst durchführen sollten, um möglichen DIF zu beurteilen.

Der PPVT-4 erfasst den rezeptiven Wortschatz und damit nur einen Teilbereich der Sprachkompetenzen. Der PPVT-4 korreliert jedoch mit Maßen für andere Bereiche der Sprachkompetenz (vgl. Lenhard et al., 2015), weshalb wir ihn als Indikator für die Deutschkompetenzen der Flüchtlinge verwenden. Im Kontext der ReGES-Studie, die die Situation der teilnehmenden Flüchtlinge im Kindes- und Jugendalter sehr umfassend darstellt, bieten die Deutschkompetenzdaten, die zusätzlich zum PPVT-4 auch mit dem TROG-D erfasst wurden, ein großes Potenzial für Analysen im Hinblick auf verschiedenste Fragestellungen. Es können beispielsweise Einflüsse verschiedenster Variablen auf die Sprachkompetenzen sowie Auswirkungen der Sprachkompetenzen auf Outcomes wie Bildungserfolg untersucht werden. Für einen Teil der Stichprobe wurden die Kompetenztests in der letzten Erhebungswelle der ReGES-Studie, die ca. zwei Jahre nach der ersten Erhebungswelle durchgeführt wurde, wiederholt, was zusätzliches Analysepotenzial birgt.

8. Daten im Scientific Use File

Die Kompetenzdaten werden wie auch die Befragungsdaten der ReGES-Studie als Scientific Use File veröffentlicht. Für die RC1 und RC2 gibt es getrennte Kompetenzdatensätze. Neben den Variablen zum PPVT-4 enthalten diese jeweils auch die Variablen zu den beiden weiteren Kompetenztests. Eine Übersicht über die Variablen im SUF zum PPVT-4 enthält Tabelle 12 und sie werden in den folgenden Abschnitten kurz beschrieben, wobei zuerst auf die Variablen auf Itemebene und anschließend auf die generierten Variablen auf Testebene eingegangen wird.

Informationen zur Missingkodierung finden sich größtenteils in den vorigen Kapiteln, deshalb ist der letzte Abschnitt kurz gehalten und verweist auf die entsprechenden Stellen.

8.1 Variablen auf Itemebene

Für jedes der 228 Items des PPVT-4 ist eine Variable enthalten, die anzeigt, ob die gegebene Antwort richtig oder falsch ist. Bei der RC1 gab es weiterhin 4 Übungsitems und bei der RC2 bis zu 6 Übungsitems, mit jeweils bis zu zwei Versuchen. Für die Übungsitems gibt es pro möglichen Versuch eine Variable, die anzeigt, ob richtig oder falsch geantwortet wurde.

8.2 Generierte Variablen auf Testebene

Neben den Variablen zu den Antworten auf einzelne Test- und Übungsitems werden im SUF außerdem verschiedene generierte Variablen zur Verfügung gestellt. Dazu gehören der Summenwert, das Boden- und das Deckenset. Die Berechnung des Summenwerts ist in Kapitel 4 beschrieben. Die Definitionen von Boden- und Deckenset finden sich in Kapitel 2.1. Alle drei Variablen wurden entsprechend den Vorgaben im Testmanual bestimmt (vgl. Lenhard et al., 2015). Für die RC1 gibt es die Variable für das Bodenset nicht, weil alle Teilnehmerinnen und Teilnehmer mit Set 1 starteten (s. Kapitel 2.2) und sie somit immer den Wert 1 annehmen würde.

Darüber hinaus wird die Anzahl der administrierten Übungsitems sowie der Übungsscore zur Verfügung gestellt.

Die für die oben beschriebenen Analysen geschätzten WLEs und dazugehörigen Standardfehler sind ebenfalls im SUF enthalten. Zu diesen sollen folgende Aspekte nochmals angemerkt sein: Zur Schätzung der WLEs wurden nur die ersten zwölf Itemsets in der RC1 und die ersten 16 Itemsets in der RC2 herangezogen (s. Kapitel 5.2). WLEs wurden für alle Kinder und Jugendliche geschätzt, die den Test vollständig bearbeitet haben und mindestens zwei Itemsets erreicht haben sowie für weitere Kinder und Jugendliche, die die Testung abgebrochen haben, nachdem sie mindestens fünf Itemsets vollständig bearbeitet haben (s. Kapitel 3.3).

Tabelle 12: Im Scientific-Use-File enthaltene Variablen

Variablenname	Variablenlabel
ID_t	
t0500<i><v>1	Wortschatz (PPVT): Übungsitem <i>, Versuch <v>
t05<bb><ii>1	Wortschatz (PPVT): Set <bb>, Item <ii>
t0501010_g1	Wortschatz (PPVT): Weighted Likelihood Estimate (WLE)
t0501010_g2	Wortschatz (PPVT): Standardfehler des Weighted Likelihood Estimate (WLE)
t0501010_g3	Wortschatz (PPVT): Summenwert
t0501010_g5	Wortschatz (PPVT): Bodenset
t0501010_g6	Wortschatz (PPVT): Deckenset
t0500110_g7	Wortschatz (PPVT): Anzahl administrierter Übungsitems
t0500110_g8	Wortschatz (PPVT): Score Übung

8.3 Missingkodierung

Die Missingkodierung im SUF ist in den Tabellen 1 und 3 der Kapitel 3.3.1 und 5.1 übersichtlich dargestellt, wobei Tabelle 1 die Missingwerte der gesamten Stichprobe enthält, die den Ausschluss von Fällen kategorisieren und Tabelle 3 die Missingwerte auf Itemebene enthält, die bei den gültigen Fällen in der Testphase vorkommen. Der Missingwert -26 „Item übersprungen durch Interviewer“ ist weder in Tabelle 1 noch in Tabelle 3 enthalten, da er nur in den Übungsitems vorkommt. Er gibt an, wenn ein Item der Übungsphase durch den Interviewer übersprungen wurde.

Literatur

Appel, R. (1996). The lexicon in second language acquisition. In P. Jordans & J. Lalleman (Hrsg.), Investigating Second Language Acquisition (S. 381–404). De Gruyter Mouton.

Berendes, K., Weinert, S., Zimmermann, S., & Artelt, C. (2013). Assessing language indicators across the life span within the German National Educational Panel Study. Journal of Educational Research Online, 5(2), 15–49.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Hrsg.), Statistical theories of mental test scores (S. 395–479). Addison-Wesley.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Hrsg.) (2011). Education as a Lifelong Process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft: Sonderheft 14. https://link.springer.com/journal/11618/14/2/suppl/page/1

de Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. The Guilford Press.

Dunn, L., & Dunn, D. (2007). Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4). Pearson.

Esser, H. (2006). Sprache und Integration: Die sozialen Bedingungen und Folgen des Spracherwerbs von Migranten. Campus.

Fischer, L., & Durda, T. (2020). NEPS Technical Report for Receptive Vocabulary: Scaling Results of Starting Cohort 2 for Kindergarten (Wave 1), Grade 1 (Wave 3) and Grade 3 (Wave 5) (No. 65; NEPS Survey Paper). https://doi.org/10.5157/NEPS:SP65:1.0

Fox-Boyer, A. V. (2016). TROG-D. Test zur Überprüfung des Grammatikverständnisses (7. Auflage). Schulz-Kirchner Verlag.

Gentile, R., Heinritz, F., & Will, G. (2019). Übersetzung von Instrumenten für die Befragung von Neuzugewanderten und Implementation einer audiobasierten Interviewdurchführung (No. 86; LIfBi Working Paper).

Gnambs, T. (2017). NEPS Technical Report for English Reading Competence : Scaling Results of Starting Cohort 4 for Grade 10 (NEPS Survey Paper No. 26). https://doi.org/10.5157/NEPS:SP26:1.0

Huttenlocher, J. (1998). Language input and language growth. Preventive Medicine, 27(2), 195–199.

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C., & Willinger, B. (2014). Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Abschlussbericht zu einer NEPS-Ergänzungsstudie (NEPS Working Paper No. 43).

Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). Peabody Picture Vocabulary Test (4. Ausgabe). Pearson.

Leseman, P. P. M., Scheele, A. F., Mayo, A. Y., & Messer, M. H. (2007). Home literacy as a special language environment to prepare children for school. Zeitschrift für Erziehungswissenschaft, 10(3), 334–355.

Pohl, S., & Carstensen, C. H. (2012). NEPS Technical Report – Scaling the Data of the Competence Tests. In NEPS Technical Report (NEPS Working Paper No. 14).

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing With Omitted and Not-Reached Items in Competence Tests: Evaluating Approaches Accounting for Missing Responses in Item Response Theory Models. Educational and Psychological Measurement, 74(3), 423–452. https://doi.org/10.1177/0013164413504926

R Core Team. (2020). R: A Language and Environment for Statistical Computing (4.0.3). R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Mesa Press.

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

Roßbach, H.-G., & Weinert, S. (Hrsg.). (2008). Kindliche Kompetenzen im Elementarbereich: Förderbarkeit Bedeutung und Messung. BMBF.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrica, 54, 427–450.

Will, G., Balaban, E., Dröscher, A., Homuth, C., & Welker, J. (2018). Integration von Flüchtlingen in Deutschland: Erste Ergebnisse aus der ReGES-Studie (LIfBi Working Paper No. 76).

Will, G., Becker, R., & Weigand, D. (2020). COVID-19 lockdown during field work. Challenges and strategies in continuing the survey. Survey Research Methods, 14(2), 247–252.

Will, G., Gentile, R., Heinritz, F., & von Maurice, J. (2018). ReGES-Refugees in the German Educational System: Überblick über Forschungsdesign, Stichprobenziehung und Ausschöpfung der ersten Welle (LIfBi Working Paper No. 75).

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