SOEP The Socio-Economic Panel

899

SOEP Survey Papers

Series B - Survey Reports (Methodenberichte)

SOEP – The German Socio-Economic Panel at DIW Berlin 2020

Dokumentation der Kompetenztestung im Rahmen der IAB-BAMF-SOEP-Befragung von Geflüchteten 2017 und 2018, Stichproben M3-M5

Stefan Schipolowski und Aileen Edele

DIW SOEP

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research,DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:

Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at **http://www.diw.de/soepsurveypapers**

Please cite this paper as follows:

Stefan Schipolowski und Aileen Edele. 2020. Dokumentation der Kompetenztestung im Rahmen der IAB-BAMF-SOEP-Befragung von Geflüchteten 2017 und 2018, Stichproben M3-M5. SOEP Survey Papers 899: Series B. Berlin: DIW/SOEP.

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin, Germany

Contact: soeppapers@diw.de

Dokumentation der Kompetenztestung im Rahmen der IAB-BAMF-SOEP-Befragung von Geflüchteten 2017 und 2018, Stichproben M3-M5

Stefan Schipolowski und Aileen Edele

Dokumentation der Kompetenztestung im Rahmen der IAB-BAMF-SOEP-Befragung von Geflüchteten 2017 und 2018, Stichproben M3-M5

Erfassung schulrelevanten Vorwissens und kognitiver Grundfähigkeiten von Schülerinnen und Schülern mit Fluchtbiographie

Stefan Schipolowski[1] & Aileen Edele[2]

[1] Institut zur Qualitätsentwicklung im Bildungswesen (IQB) an der Humboldt-Universität zu Berlin; [2] Berliner Institut für empirische Integrations- und Migrationsforschung (BIM), Humboldt-Universität zu Berlin

Inhalt

Kontakt:

stefan.schipolowski@iqb.hu-berlin.de (Stefan Schipolowski)
aileen.edele@hu-berlin.de (Aileen Edele)

1 Ziel der Kompetenztestung und Auswahl zu erfassender Konstrukte

1.1 Zielstellung der Kompetenzmessung

Aufgrund der großen Zahl der in den letzten Jahren als Schutzsuchende nach Deutschland gekommenen Kinder und Jugendlichen stehen viele Schulen und Lehrkräfte vor der Herausforderung, neuzugewanderte Heranwachsende mit geringen Kenntnissen der deutschen Sprache zu integrieren. Da sich die Schulsysteme in den Herkunftsländern der Geflüchteten substanziell vom deutschen System unterscheiden und die jungen Geflüchteten häufig unterbrochene Bildungsbiografien aufweisen (Brücker et al., 2016; Brücker, Schewe, & Sirries, 2016; Schupp et al., 2017), ist schwer einschätzbar, über welche schulrelevanten Kompetenzen diese Population verfügt und welche Bedeutung diese im Sinne eines „Bildungspotenzials" für den weiteren Bildungsverlauf haben. Ziel der Kompetenzmessung in der IAB-BAMF-SOEP-Befragung von Geflüchteten war es daher, belastbare Informationen über schulrelevantes Vorwissen Heranwachsender mit Fluchtbiographie zu generieren.

1.2 Auswahl der zu erfassenden Konstrukte

Die meisten Domänen schulisch vermittelter Kompetenzen, etwa im Schulfach Deutsch oder in sozialwissenschaftlichen Fächern, sind in hohem Maß sprach- und kulturspezifisch. Daher dürften Schülerinnen und Schüler, die im Ausland beschult wurden, geringe Kompetenzen in diesen Domänen aufweisen, selbst wenn sie die Lernziele des Bildungssystems ihres Herkunftslandes vollumfänglich erreicht hätten. Für die Domänen Mathematik und Naturwissenschaften/Technik kann hingegen angenommen werden, dass entsprechende Kompetenzen bzw. deren Erwerb weniger kulturabhängig sind. Eine reliable und valide Erfassung mathematischer Kompetenzen jenseits grundlegender Rechenoperationen – etwa im Sinne der Bildungsstandards der Kultusministerkonferenz (KMK, 2004, 2005) – ist jedoch relativ zeitaufwändig (vgl. etwa Roppelt, Blum, & Pöhlmann, 2013) und wäre im Rahmen der IAB-BAMF-SOEP-Befragung Geflüchteter aufgrund der begrenzten Erhebungszeit nicht sinnvoll realisierbar gewesen. Um abzuschätzen, über welches schulrelevante Vorwissen geflüchtete Kinder und Jugendliche verfügen, die erst vergleichsweise kurz in Deutschland leben, wurde daher deklaratives Wissen in der Domäne Naturwissenschaften/Technik erfasst. Die hohe prädiktive Validität deklarativen Wissens für den Bildungs- und Berufserfolg ist empirisch gut belegt (Baumert, Lüdtke, Trautwein, & Brunner, 2009; Dye, Reck, & McDaniel, 1993; McGrew & Hessler, 1995).

Des Weiteren sollten in Anlehnung an groß angelegte Schulleistungsuntersuchungen individuelle Unterschiede in kognitiven Grundfähigkeiten erfasst werden, die bei weiterführenden Analysen häufig als Kontrollvariable genutzt werden. Aus Zeitgründen konnte nur ein einzelner Indikator zum Einsatz kommen. Die Wahl fiel dabei auf die Domäne des figuralen (nonverbalen) schlussfolgernden Denkens. Indikatoren des figuralen schlussfolgernden Denkens gelten als prototypische Indikatoren für die Fähigkeit zum schlussfolgernden Denken (*Reasoning*; vgl. etwa Carroll, 1993; Cattell, 1987) sowie als weitgehend kultur- und sprachunabhängig (Cattell, 1940; McCallum, 2003; siehe jedoch DeShon, Chan, & Weissbein, 1995). Von kognitiven Grundfähigkeiten kann weiterhin angenommen werden, dass sie im Vergleich zu schulisch vermittelten Kompetenzen weniger sensitiv für Bildungsprozesse sind (Hartig & Klieme, 2006).

2 Itembeschreibung, Itemauswahl und Beschreibung der finalen Erhebungsinstrumente

Zur Erfassung von Wissen im Bereich Naturwissenschaften/Technik sowie der Fähigkeit zum figuralen schlussfolgernden Denken wurde auf den umfangreichen Itempool des BEFKI-Projekts (*Berliner Test zur Erfassung fluider und kristalliner Intelligenz*; Schroeders, Schipolowski & Wilhelm, 2015, 2020; Wilhelm, Schroeders, & Schipolowski, 2014) zurückgegriffen, der mehrere hundert empirisch erprobte

Items zu 16 Wissensbereichen sowie zum schlussfolgenden Denken für verschiedene Altersgruppen (ab ca. 8 Jahre) umfasst. Die Testinstrumente der BEFKI-Testreihe wurden für Schülerinnen und Schüler der 3. bis 12. Jahrgangsstufe an allgemeinen Schulen in Deutschland normiert.

2.1 Wissen im Bereich Naturwissenschaften/Technik

Die Domäne Naturwissenschaften/Technik umfasst Fragen zu den Wissensbereichen Physik, Chemie, Biologie, Medizin, Geographie und Technologie. Die Items wurden auf Basis von deutschsprachigen Lehr- und Nachschlagewerken konstruiert und haben ein Multiple-Choice-Format mit vier Antwortoptionen (vgl. Abbildung 1), wobei immer genau eine Antwort zutreffend ist. Um die erwartete große Varianz im Vorwissen der befragten Population einschließlich zu vermutender geringer Kompetenzstände aufgrund unterbrochener Bildungsbiographien abzudecken, wurde eine Vorauswahl von Items getroffen, die von geringer Schwierigkeit (Entwicklung für die Grundschule/Klassenstufen 3 und 4) bis hin zu mittlerer Schwierigkeit (Entwicklung für die Sekundarstufe I bis einschließlich Klasse 10) reichte. Diese Vorauswahl von Items sowie die Instruktionen wurden zunächst ins Arabische übersetzt und in eine Darstellungsform übertragen, bei der die deutsche und die arabische Version nebeneinandergestellt wurden, so dass die Befragten auswählen konnten, in welcher Sprache sie die Aufgaben bearbeiteten.

In welchem Organ erfolgt die Anreicherung des Blutes mit Sauerstoff?

A. Leber
B. Lunge
C. Gehirn
D. Niere

Abbildung 1: Beispielitem Deklaratives Wissen

Anschließend wurden die Aufgaben einer Gruppe von Expertinnen und Experten vorgelegt, die die Angemessenheit der Aufgaben für die Zielpopulation beurteilen sollten. Dabei handelte es sich um 9 Lehrkräfte syrischer Herkunft, die zwischen 3 und 20 Jahren Berufserfahrung in ihrem Herkunftsland hatten und über gute Deutschkenntnisse verfügten (mindestens Niveaustufe B2 *Selbstständige Sprachverwendung* gemäß dem Gemeinsamen Europäischen Referenzrahmen). Sie wurden gebeten, zu beurteilen, ob die Übersetzung der Instruktion und Aufgaben ins Arabische angemessen und eindeutig war, ob die Wissensfragen insofern kulturunabhängig und in ihrer Schwierigkeit angemessen waren, als dass sie von einer exzellenten Schülerin oder einem exzellenten Schüler, der das syrische Schulsystem durchlaufen hat, beantwortet werden könnten, sowie ob die Inhalte der Wissensfragen kulturell und biographisch angemessen waren. Aufgrund der großen Zahl von Aufgaben wurde jeder Expertin bzw. jedem Experten nur ein Teil der Aufgaben vorgelegt; jede Aufgabe wurde von mindestens zwei Personen beurteilt.

Darüber hinaus wurden die Instruktionen und Aufgaben 10 Schülerinnen und Schülern syrischer bzw. irakischer Herkunft im Alter von 10 bis 16 Jahren vorgelegt, die erst seit vergleichsweise kurzer Zeit eine deutsche Schule besuchten (zwischen 4 Monaten und 2,5 Jahren). Es wurde geprüft, ob sie die Instruktionen, Fragen und Antwortmöglichkeiten verstehen konnten, in welcher Sprache sie die Aufgaben bearbeiteten, ob sie über ausreichende Lesefähigkeiten in zumindest einer der Sprachen verfügten und ob sie die Aufgaben lösen konnten. In beiden Situationen war eine Übersetzerin anwesend, um bei eventuell auftretenden Sprachproblemen zu vermitteln.

Die kognitive Befragung der Lehrkräfte ergab, dass die Aufgabenstellung als prinzipiell geeignet für den Einsatz in der Zielpopulation eingeschätzt wurde. Auch wurde die Mehrheit der vorausgewählten 117 Wissensitems als uneingeschränkt geeignet für den Einsatz in der Zielpopulation erachtet. 13 Items wurden als ungeeignet eingeschätzt, da sie als entweder zu stark kulturabhängig und somit für die

Zielpopulation unverhältnismäßig schwer eingeschätzt wurden (z.B. „In welcher Stadt steht das Kolosseum?“) oder als potenziell retraumatisierend (z.B. Thematisierung von Sprengstoffen im Bereich Chemie). Bei einigen weiteren Items wurde die Übersetzung moniert; diese wurden im Anschluss wenn möglich ausgebessert. Items mit nicht behebbaren Übersetzungsproblemen wurden von der Erhebung ausgeschlossen. Am Ende ergab sich ein Pool von 100 Items, die für den Einsatz in der Zielpopulation geeignet erschienen.

Zu den zentralen Ergebnissen der kognitiven Interviews mit den Heranwachsenden mit Fluchtbiographie zählte, dass sie die Aufgabenstellung gut verstanden, dass sie über sehr heterogenes Wissen im Bereich Naturwissenschaften und Technik verfügten, dass die Aufgaben teils auf Deutsch und teils auf Arabisch bearbeitet wurden und dass nicht alle Heranwachsenden über ausreichende Lesefähigkeiten in Deutsch oder Arabisch verfügten, um die Wissensfragen selbstständig bearbeiten zu können.

Aus dem vorerprobten Itempool wurden anschließend zwei Testformen gebildet: eine leichtere Testform für Testteilnehmende im Alter bis einschließlich 15 Jahre und eine schwerere Testform für Teilnehmende im Alter von mindestens 16 Jahren. Beide Testformen umfassen jeweils 36 Items, die sich ausgeglichen auf die 6 Inhaltsbereiche verteilen. 24 Items sind in beiden Testformen enthalten (sog. Ankeritems), so dass die beiden Testformen auf einer gemeinsamen Metrik dargestellt werden können (siehe Abschnitt 3.3).

Zusätzlich zur Übersetzung ins Arabische wurden die Instruktionen und Items in fünf weitere Sprachen übersetzt (Englisch, Farsi, Paschtu, Urdu, Kurmandschi). Der Test wurde den Teilnehmenden auf Deutsch sowie auf einer von ihnen gewählten weiteren Sprache vorgelegt, wobei die Items gleichzeitig in beiden Sprachen dargeboten wurden. Um auch Schülerinnen und Schülern mit geringen Lesefähigkeiten die Testteilnahme zu ermöglichen, konnten für die Instruktionen und Aufgaben einschließlich der Antwortmöglichkeiten zusätzlich in allen Herkunftssprachen Audiodateien abgerufen werden. Die Items wurden den Teilnehmenden in individuell vollständig randomisierter Reihenfolge vorgelegt, sodass bei einer Auswertung auf Gruppenebene Reihenfolgeeffekte ausgeschlossen werden können. Die Bearbeitungszeit für diesen Testteil exklusive Instruktionszeit und Beispielitem betrug 13 Minuten. Die Bemessung des Bearbeitungszeitlimits orientierte sich an den Normierungsstudien zum BEFKI.

2.2 Figurales schlussfolgerndes Denken

Bei den Items zum figuralen schlussfolgernden Denken wurde jeweils eine Reihe geometrischer Figuren vorgegeben, deren Elemente sich regelhaft verändern und schlüssig aufeinander aufbauend entwickeln. Die Aufgabe der Testteilnehmenden bestand darin, die Regelmäßigkeiten oder Entwicklungsverläufe zu erkennen und die fehlenden Glieder in der Kette zu erschließen, indem unter drei möglichen Antwortalternativen die passende Lösung ausgewählt wird. Bei allen Aufgaben mussten jeweils zwei Figuren bestimmt werden, die die mit einem Fragezeichen markierten fehlenden Glieder darstellen (vgl. Abbildung 2). Ein Item wird nur dann als gelöst bewertet, wenn beide Figuren korrekt gewählt wurden (Wilhelm et al., 2014).

Die Instruktionen zum figuralen schlussfolgernden Denken wurden ins Arabische übersetzt und die Instruktionen und Items wurden ebenfalls den 10 syrischen Lehrkräften sowie 9 Heranwachsenden mit Fluchtbiographie vorgelegt (vgl. Abschnitt 2.1). Die Lehrkräfte sollten die generelle Eignung der Aufgaben für die Zielpopulation beurteilen und die Übersetzung prüfen. Die Schülerinnen und Schüler wurden gebeten, einige der Aufgaben gemäß der Instruktion zu lösen, um sicherzustellen, dass sie die Instruktion verstehen und das Aufgabenformat bearbeiten konnten. Es ergaben sich keine Hinweise auf Schwierigkeiten für den Einsatz des Tests in der Zielpopulation.

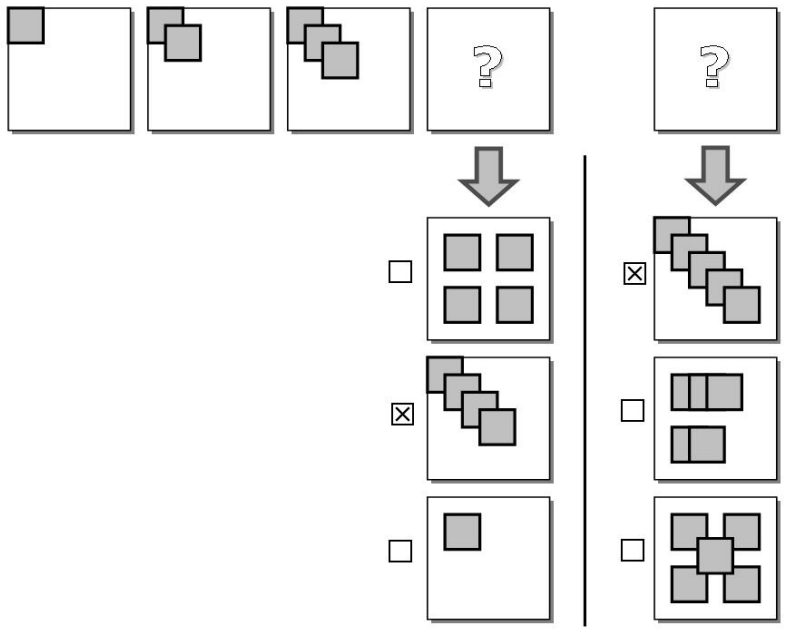

Abbildung 2: Beispielitem Figurales schlussfolgerndes Denken

Analog zum Vorgehen bei der Erfassung des Wissens im Bereich Naturwissenschaften/Technik wurden zwei Testformen eingesetzt: eine leichtere Testform für Testteilnehmende im Alter bis einschließlich 15 Jahre und eine schwerere Testform für Teilnehmende im Alter ab 16 Jahren. Die Testformen umfassten jeweils 16 Items, wovon 8 Items in beiden Testformen enthalten waren, so dass auch die Kompetenzwerte dieser Skala auf einer gemeinsamen Metrik für alle Altersgruppen dargestellt werden können.

Die schwerere Form des Tests zum schlussfolgernden Denken erwies sich in der ersten Erhebungswelle auch für die Jugendlichen ab 16 Jahren als zu schwierig und wurde daher in der zweiten Erhebungswelle nicht mehr eingesetzt. Dort kam stattdessen für alle Testteilnehmenden die leichtere Testform zum Einsatz (siehe Abschnitt 3).

Analog zum Wissenstest waren die Instruktionen dieses Testteils in den Sprachen Arabisch, Englisch, Farsi, Paschtu, Urdu und Kurmandschi verfügbar und wurden auf Deutsch sowie zusätzlich in der gewählten Sprache dargestellt. Die Bearbeitungszeit für diesen Testteil exklusive Instruktionszeit und Beispielitem betrug – wie in den Normierungserhebungen zum BEFKI – 14 Minuten.

3 Methoden

3.1 Datenerhebung

Die Daten wurden in zwei Wellen erhoben, die zeitlich um ein Jahr versetzt mit jeweils drei Geburtsjahrgängen durchgeführt wurden. Die erste Erhebungswelle fand im Zeitraum September 2017 bis März 2018 statt, die zweite im Zeitraum September 2018 bis Februar 2019. Alle Erhebungen wurden durch geschulte Interviewerinnen und Interviewer durchgeführt, die dazu die Testpersonen an ihrem Wohnort aufsuchten. Insgesamt wurden die Testverfahren 679 Kindern und Jugendlichen (48,2% weiblich) im Alter zwischen 11 und 18 Jahren (M = 14,5 Jahre) vorgelegt. Die Testpersonen verteilten sich auf die Geburtsjahrgänge 2000/2001 (n = 167), 2003/2004 (n = 228) und 2005/2006 (n = 284). Die Personen aus den genannten Doppeljahrgängen hatten zum jeweiligen Erhebungszeitpunkt der Da-

tenerhebung nahezu denselben Altersdurchschnitt. Zum Beispiel betrug das Alter des Geburtsjahrganges 2000 zum Testzeitpunkt (erste Erhebungswelle) im Mittel 17,5 Jahre und das Durchschnittsalter des Geburtsjahrganges 2001 zum Testzeitpunkt (zweite Erhebungswelle) 17,6 Jahre.

Wie oben erwähnt (vgl. Abschnitte 2.1 und 2.2) wurden sowohl zum Wissen als auch zum figuralen schlussfolgernden Denken zwei unterschiedlich schwere Testformen eingesetzt. In der ersten Erhebungswelle erhielten die Geburtsjahrgänge 2003 und 2005 jeweils die leichtere Testform, während den Jugendlichen des Geburtsjahrgangs 2000 die schwerere Version der Tests vorgelegt wurde. Eine Ausnahme bilden 35 Testpersonen des Geburtsjahrgangs 2003, die ebenfalls die schwereren Testformen erhielten. In der zweiten Erhebungswelle wurde beim Wissenstest analog verfahren, d. h. die Geburtsjahrgänge 2004 und 2006 erhielten die leichtere Testform und der Geburtsjahrgang 2001 die schwerere. Zur Erfassung des schlussfolgernden Denkens wurde in der zweiten Welle jedoch allen Geburtsjahrgängen die leichtere Testform vorgelegt.

Die Bearbeitung der Testaufgaben erfolgte technologiegestützt (*Computer Assisted Personal Interview*, CAPI). Ein Weitergehen zum nächsten Testitem war erst möglich, nachdem zum aktuellen Item eine Antwortoption ausgewählt wurde (*Forced Choice*). Innerhalb der verfügbaren Bearbeitungszeit konnten die Testpersonen jederzeit zu bereits bearbeiteten Items desselben Tests zurückgehen und ihre Antworten korrigieren. Nach Ablauf der vorgegebenen Bearbeitungszeit brach der jeweilige Test ab, sobald das nächste Testitem aufgerufen wurde. Bei Bedarf konnten die Testpersonen eine akustische Wiedergabe (Audiodateien) der Instruktion und des jeweiligen Items auslösen.

3.2 Datenaufbereitung

Zunächst wurden die Rohdaten für die verschiedenen Testformen anhand eines Lösungsschlüssels in richtige versus falsche Antworten rekodiert. Fehlende Werte finden sich in den Rohdaten aufgrund der unterschiedlich schweren Testformen sowie der erzwungenen Antwortauswahl nur in zwei Fällen:

a) Das Item wurde nicht zur Bearbeitung vorgelegt.
b) Das Item wurde nicht bearbeitet, weil die vorgesehene Bearbeitungszeit überschritten wurde.

Fehlende Werte aufgrund von a) wurden für die Datenauswertungen als fehlend (*Missing*) behandelt und gehen nicht in Item- und Teststatistiken ein. Bei den Skalierungen wurde ein modellbasierter Ansatz zum Umgang mit fehlenden Werten gewählt (siehe Abschnitt 3.3).

Fehlende Werte aufgrund von b) traten beim Test zum figuralen schlussfolgernden Denken nur sehr selten auf. Insgesamt überschritten 4,7% der Testpersonen das Zeitlimit für die Bearbeitung. Da dies vergleichbar ist mit den Ergebnissen und Bedingungen der papierbasierten Normierungsstudien (vgl. etwa Wilhelm et al., 2014), wurden diese fehlenden Werte wie in den Normierungen als Falschantworten gewertet.

Ein anderes Bild ergibt sich für fehlende Werte aufgrund von b) beim Wissenstest: Insgesamt überschritten 35,5% der Testpersonen das Zeitlimit für die Bearbeitung. Da dieser Anteil deutlich über dem Anteil in der Normierungsstudie liegt und die Wissenstestung konzeptuell nicht als zeitlimitierte Testung angelegt sein sollte (es handelt sich im Hinblick auf das zu erfassende Konstrukt um einen *Power Test* und keinen *Speed Test*), wurden auch die aufgrund der Zeitbegrenzung fehlenden Werte im Wissenstest im Folgenden als fehlend behandelt und nicht als Falschantworten gewertet. Dies gilt sowohl für die Itemstatistiken als auch für die Skalierungen, bei denen ein modellbasierter Ansatz zum Umgang mit fehlenden Werten gewählt wurde (siehe Abschnitt 3.3).

Aus allen folgenden Auswertungsschritten ausgeschlossen wurden Personen, die bei einem Test weniger als 25% der Testitems bearbeitet hatten, da in diesen Fällen nicht von einer instruktionsgemäßen

Bearbeitung ausgegangen werden kann. Der Ausschluss erfolgte hierbei nur für den jeweils betroffenen Test. Ferner wurden die Kommentare der Interviewerinnen und Interviewer gesichtet und Personen für einen oder beide Tests ausgeschlossen, wenn die Kommentierung auf eine instruktionswidrige Bearbeitung schließen ließ (z. B. Hilfestellung bei der Beantwortung der Aufgaben durch Familienangehörige oder „Durchklicken“ der Items ohne inhaltliche Bearbeitung).

Nach allen genannten Ausschlüssen gingen in die folgenden Auswertungen zum Wissenstest die Daten von insgesamt 622 Kindern und Jugendlichen ein. Für den Test zum figuralen schlussfolgenden Denken wurden Daten von insgesamt 636 Testpersonen einbezogen.

3.3 Analytisches Vorgehen

Auf der Itemebene wurden für alle eingesetzten Aufgaben Itemschwierigkeit und Itemtrennschärfe berechnet. Der Schwierigkeitsparameter ergibt sich hierbei als Anteil der Testpersonen, die das Item korrekt bearbeitet haben. Die Trennschärfe wird zum einen als biseriale Korrelation zwischen Item und Gesamttestwert berechnet, wobei berücksichtigt wird, dass die einzelnen Items als künstlich-dichotome Indikatoren einer kontinuierlichen latenten Dimension aufzufassen sind. Zum anderen wird die Trennschärfe als punktbiseriale Korrelation angegeben, die geringer ausfällt, da hierbei die künstliche Dichotomisierung nicht in Rechnung gestellt wird. Dieser Koeffizient wird dennoch berichtet, da er in der Literatur häufig zur Berechnung der Trennschärfe herangezogen wird. Auf der Testebene werden für die einzelnen Testformen Verteilungsparameter (Mittelwerte und Streuungen) der Gesamttestwerte (*Scores*) sowie Angaben zur Reliabilität berichtet.

Im nächsten Schritt wurden die Testitems unter Verwendung des 1pl-Modells skaliert (Rasch, 1960/1980). Die Skalierungen erfolgten getrennt für den Wissenstests und den Test zum figuralen schlussfolgernden Denken (d. h. eindimensional) unter Einbezug aller Items zum jeweiligen Konstrukt. Die verwendete Testform wurde über eine Dummy-Variable im Hintergrundmodell berücksichtigt, um unverzerrte Schätzungen der Itemschwierigkeiten zu erhalten. Bei der Skalierung werden auch Fälle berücksichtigt, für die ein Teil der Testdaten fehlt, wobei das Vorgehen dem modellbasierten *Full Information Maximum Likelihood* (FIML)-Ansatz entspricht (Lüdtke, Robitzsch, Trautwein & Köller, 2007).

Die Skalierungen ermöglichen es im Zusammenhang mit der Verwendung von Ankeritems, die sowohl in der leichteren als auch in der schwereren Testform enthalten sind, alle Testpersonen unabhängig von der verwendeten Testform auf einer einheitlichen Metrik abzubilden. Hierzu können die aus der Skalierung resultierenden Personenparameter (*Weighted Likelihood Estimates*; WLEs) herangezogen werden. Die WLEs wurden in T-Werte transformiert, die in der hier untersuchten Stichprobe einen Mittelwert von 50 Punkten und eine Standardabweichung von 10 Punkten aufweisen.

4 Ergebnisse

4.1 Nutzung der Audiodateien und Übersetzungen sowie Besonderheiten in der Testsituation

Wie oben beschrieben, wurden die Instruktionen und die Items des Wissenstests zweisprachig vorgegeben (Deutsch plus eine von sechs weiteren Sprachen). Zudem konnten sich die Testpersonen die Instruktion und die Items in der gewählten Fremdsprache vorlesen lassen (Abspielen einer entsprechenden Audiodatei). In den Tabellen 1 und 2 sind die Statistiken zur Nutzung der Sprachfassungen und Audiodateien dargestellt.

Tabelle 1: Wahl der angebotenen Sprachfassungen

Sprachfassung	Absolute Häufigkeit (*n*)	Anteil [%]
Deutsch / Englisch	74	10,9
Deutsch / Arabisch	492	72,6
Deutsch / Farsi	87	12,8
Deutsch / Paschtu	3	0,4
Deutsch / Urdu	3	0,4
Deutsch / Kurmandschi	19	2,8

Anmerkung: *N* = 678.

Dass die überwiegende Mehrheit der Teilnehmenden die Deutsch-Arabische Testversion gewählt hat, deutet darauf hin, dass Arabisch die in der Stichprobe verbreitetste Herkunftssprache ist. Etwas mehr als die Hälfte der Teilnehmenden nutzte laut Interviewereinschätzung zumindest teilweise die Übersetzungen. Demnach traute sich im Umkehrschluss fast die Hälfte der Teilnehmenden die Bearbeitung der Tests auf Deutsch zu, während die andere Hälfte die Bearbeitung mindestens einiger Fragen in der Herkunftssprache bzw. Englisch vorzog. Die Audiodateien wurden vergleichsweise selten genutzt, was auf eine hohe Alphabetisierungsquote der Stichprobe hindeutet.

Tabelle 2: Nutzung der Übersetzungen und der Audiodateien

Häufigkeit	Übersetzungen		Audiodateien[1]	
	Abs. H. (*n*)	Anteil [%]	Abs. H. (*n*)	Anteil [%]
Bei jeder Frage	140	20,6	9	2,9
Bei etwa zwei Dritteln der Fragen	78	11,5	4	1,3
Bei etwa der Hälfte der Fragen	65	9,6	5	1,6
Bei weniger als der Hälfte der Fragen	87	12,8	13	4,2
Gar nicht	308	45,4	278	90,0

Anmerkung: *N* = 678. Abs. H. = Absolute Häufigkeit. Es handelt sich jeweils um Einschätzungen der Interviewerinnen und Interviewer zur Testung. [1] Angaben zur Nutzung der Audiodateien liegen nur aus der ersten Erhebungswelle vor (*n* = 309).

Zudem machten die Interviewerinnen und Interviewer Angaben zu eventuellen Störungen während der Testsitzung (z. B. durch andere Personen im Raum oder durch zeitliche Unterbrechungen). Zur Frage „Gab es während des Kompetenztests Störungen, die die Konzentration des/der Befragten beeinflusst haben?“ gaben die Interviewerinnen und Interviewer an, dass dies in rund 65 Prozent der Testsitzungen „überhaupt nicht“ der Fall war; in 29 Prozent der Sitzungen traten „hin und wieder“ Störungen auf und in ca. 7 Prozent der Sitzungen „unentwegt“. Ein ähnliches Bild ergibt sich für die Frage „Gab es während des Kompetenztests zeitliche Unterbrechungen?“. Hier gaben die Interviewerinnen und Interviewer an, dass dies in rund 77 Prozent der Testsitzungen „überhaupt nicht“ auftrat. In 15 Prozent der Testungen kam es „einmal“ zu einer Unterbrechung und in knapp 9 Prozent der Sitzungen „mehrmals“. Insgesamt scheint die Testsituation somit in der überwiegenden Mehrzahl der Fälle gut oder akzeptabel gewesen zu sein.

4.2 Itemstatistiken

Im Folgenden werden die Itemschwierigkeiten und -trennschärfen für die einzelnen Testformen wiedergegeben (vgl. Tabellen 3 und 4). Bei der Interpretation der Itemschwierigkeit ist die Ratewahrscheinlichkeit für die eingesetzten Multiple-Choice-Items zu beachten, die für die Items des Wissenstests bei .25 liegt und für die Items zum figuralen schlussfolgernden Denken bei rund .11. Ferner ist zu berücksichtigen, dass für die schwerere Testform jeweils nur eine relativ geringe Fallzahl vorliegt, wodurch die Aussagekraft der entsprechenden Kennwerte für diese Testform eingeschränkt ist.

Tabelle 3: Itemstatistiken für den Wissenstest

Item	Leichtere Testform (Geburtsjahrgänge 2003 bis 2006)				Schwerere Testform (Geburtsjahrgänge 2000/2001)			
	N	*p*	$r_{it(bis)}$	$r_{it(pbis)}$	*N*	*p*	$r_{it(bis)}$	$r_{it(pbis)}$
gcA_G_bio5	382	.69	.58	.45	-	-	-	-
gcB_G_bio5	377	.55	.35	.27	146	.58	.46	.33
gcB_U_bio1	377	.57	.48	.37	-	-	-	-
gcB_U_bio4	375	.67	.55	.42	144	.72	.35	.26
gcA_UM_bio3	374	.51	.60	.47	142	.62	.77	.59
gcA_UMO_bio2	385	.49	.50	.38	141	.50	.39	.30
gcA_G_che5	372	.57	.61	.47	-	-	-	-
gcB_G_che5	386	.62	.33	.25	142	.65	.64	.49
gcA_U_che3	378	.53	.60	.47	-	-	-	-
gcB_U_che1	375	.70	.53	.40	147	.85	.47	.31
gcA_UM_che1	380	.66	.45	.34	148	.75	.41	.29
gcA_UMO_che2	383	.39	.43	.32	145	.48	.46	.35
gcA_G_geo5	388	.81	.65	.47	-	-	-	-
gcB_G_geo5	393	.73	.53	.39	147	.83	.44	.30
gcB_U_geo3	363	.87	.61	.42	141	.95	.86	.42
gcA_UM_geo2	386	.50	.40	.30	146	.58	.36	.28
gcB_UM_geo4	384	.56	.37	.29	-	-	-	-
gcA_UMO_geo1	380	.42	.32	.24	147	.54	.42	.32
gcB_G_med5	386	.79	.70	.52	-	-	-	-
gcB_G_med6	379	.83	.69	.50	142	.91	.67	.40
gcA_U_med2	379	.82	.78	.57	143	.93	.69	.41
gcB_U_med3	384	.71	.62	.48	-	-	-	-
gcA_UM_med1	388	.72	.57	.43	138	.81	.42	.31
gcA_UMO_med3	379	.45	.56	.43	140	.53	.54	.41
gcA_G_phy5	388	.82	.54	.38	144	.83	.47	.32
gcA_G_phy6	376	.59	.62	.48	-	-	-	-
gcA_U_phy2	384	.60	.56	.43	-	-	-	-
gcB_U_phy3	379	.63	.66	.51	146	.69	.57	.43
gcB_UM_phy1	382	.41	.41	.32	148	.53	.66	.50
gcB_UMO_phy4	371	.61	.59	.46	145	.74	.22	.16
gcA_G_tec5	375	.81	.65	.46	140	.87	.25	.16
gcB_G_tec6	371	.81	.59	.43	-	-	-	-
gcA_U_tec1	384	.47	.48	.37	144	.60	.35	.28
gcB_U_tec2	376	.39	.38	.28	-	-	-	-
gcA_UM_tec4	377	.79	.59	.45	146	.90	.59	.36
gcB_UMO_tec3	386	.55	.56	.43	142	.65	.39	.30
gcA_M_bio1	-	-	-	-	142	.39	.14	.09
gcB_MO_bio4	-	-	-	-	141	.56	.32	.25
gcA_M_che4	-	-	-	-	149	.62	.68	.53
gcA_MO_che3	-	-	-	-	142	.44	.32	.24
gcA_MO_geo4	-	-	-	-	145	.59	.36	.27
gcB_MO_geo2	-	-	-	-	148	.78	.26	.17
gcA_M_med4	-	-	-	-	143	.65	.44	.33
gcA_MO_med2	-	-	-	-	146	.78	.57	.40
gcA_UMO_phy4	-	-	-	-	139	.60	.48	.37
gcB_MO_phy2	-	-	-	-	142	.41	.06	.04
gcA_M_tec1	-	-	-	-	145	.53	.47	.34
gcB_MO_tec2	-	-	-	-	147	.53	.31	.24

Fortsetzung Tabelle 3: Itemstatistiken für den Wissenstest

	Leichtere Testform (Geburtsjahrgänge 2003 bis 2006)				Schwerere Testform (Geburtsjahrgänge 2000/2001)			
	N	*p*	$r_{it(bis)}$	$r_{it(pbis)}$	*N*	*p*	$r_{it(bis)}$	$r_{it(pbis)}$
M	380	.63	.54	.41	144	.66	.45	.32
Min	363	.39	.32	.24	138	.39	.06	.04
Max	393	.87	.78	.57	149	.95	.86	.59

Anmerkungen: Aufgrund fehlender Werte variiert die Fallzahl je nach Item. 35 Testpersonen des Geburtsjahrgangs 2003, die die schwerere Testform erhielten, blieben bei den Berechnungen unberücksichtigt. In die Berechnung der Itemtrenn-schärfen wurden nur Personen mit vollständigen Daten einbezogen (n = 257 für die leichtere Testform bzw. n = 116 für die schwerere Testform). p = Itemschwierigkeit; $r_{it(bis)}$ = Itemtrennschärfe als biserialer Korrelationskoeffizient; $r_{it(pbis)}$ = Item-trennschärfe als punktbiserialer Korrelationskoeffizient.

Tabelle 4: Itemstatistiken für den Test zum figuralen schlussfolgernden Denken

Item	Leichtere Testform (Geburtsjahrgänge 2001 bis 2006)				Schwerere Testform (Geburtsjahrgang 2000)			
	N	*p*	$r_{it(bis)}$	$r_{it(pbis)}$	*N*	*p*	$r_{it(bis)}$	$r_{it(pbis)}$
gffB_U1	530	.30	.43	.31	-	-	-	-
gffB_U2	530	.49	.61	.46	-	-	-	-
gffB_U3	530	.23	.23	.16	-	-	-	-
gffB_U4	530	.27	.64	.46	-	-	-	-
gffB_U5	530	.43	.59	.44	-	-	-	-
gffB_U6	530	.33	.44	.32	-	-	-	-
gffB_U7	530	.12	.40	.23	-	-	-	-
gffB_U8	530	.18	.17	.11	-	-	-	-
gffB_UM1	530	.85	.49	.29	71	.76	.74	.40
gffB_UMO3	530	.58	.42	.30	71	.56	.58	.37
gffB_UMO1	530	.62	.57	.41	71	.59	.92	.56
gffB_UMO2	530	.18	.45	.30	71	.24	.59	.41
gffB_UM2	530	.51	.31	.22	71	.52	.50	.32
gffB_UM3	530	.30	.64	.47	71	.39	.54	.36
gffB_UMO4	530	.32	.51	.37	71	.30	.53	.38
gffB_UM4	530	.33	.49	.35	71	.45	.67	.45
gffB_M1	-	-	-	-	71	.13	-.18	.00
gffB_MO3	-	-	-	-	71	.10	.38	.26
gffB_MO1	-	-	-	-	71	.23	.36	.22
gffB_M2	-	-	-	-	71	.07	-.25	.05
gffB_M4	-	-	-	-	71	.08	-.05	.11
gffB_M3	-	-	-	-	71	.11	.42	.29
gffB_MO2	-	-	-	-	71	.18	.05	.10
gffB_MO4	-	-	-	-	71	.08	-.92	-.27
M	530	.38	.46	.33	71	.30	.30	.25
Min	530	.12	.17	.11	71	.07	-.92	-.27
Max	530	.85	.64	.47	71	.76	.92	.56

Anmerkungen: 35 Testpersonen des Geburtsjahrgangs 2003, die die schwerere Testform erhielten, blieben bei den Berechnungen unberücksichtigt. In die Berechnung der Itemtrennschärfen wurden nur Personen mit vollständigen Daten einbezogen (n = 530 für die leichtere Testform bzw. n = 71 für die schwerere Testform). p = Itemschwierigkeit; $r_{it(bis)}$ = Item-trennschärfe als biserialer Korrelationskoeffizient; $r_{it(pbis)}$ = Itemtrennschärfe als punktbiserialer Korrelationskoeffizient.

Für den Wissenstest zeigt sich sowohl für die leichtere Testform als auch für die schwerere Testform eine angemessene Spannweite an Itemschwierigkeiten. Auch die Itemtrennschärfen fallen für den Wissenstest in beiden Testformen überwiegend hoch aus. Die Items erwiesen sich demnach aus psychometrischer Sicht als gut geeignet, um das Wissen der untersuchten Stichprobe im Bereich Naturwissenschaften und Technik zu erfassen.

Ein anderes Bild ergibt sich für den Test zum figuralen schlussfolgenden Denken. Zwar ist auch hier eine große Spannweite an Itemschwierigkeiten zu beobachten. Es überwiegen jedoch Items mit geringen bis sehr geringen Lösungshäufigkeiten. Dies gilt insbesondere für die schwerere Testform, die in der ersten Erhebungswelle beim Geburtsjahrgang 2000 eingesetzt wurde; hier weisen insgesamt 7 Items Lösungshäufigkeiten in Höhe der Ratewahrscheinlichkeit auf. Für diese Items zeigen sich in der untersuchten Stichprobe teilweise sehr niedrige oder negative Trennschärfen.

4.3 Teststatistiken

Die Testwerte (*Scores*) für die Testpersonen wurden zunächst als Summenscores gebildet, die der Anzahl der durch eine Testperson korrekt gelösten Items in der vorgelegten Testform entsprechen. Tabelle 5 gibt die Verteilungskennwerte der Summenscores sowie die Reliabilität bzw. interne Konsistenz der Skalen wieder. Die Reliabilität wurde als Koeffizient α (Cronbach, 1951) berechnet, da dieser Koeffizient in der Literatur trotz verschiedener Einschränkungen (vgl. etwa Sijtsma, 2009) weit verbreitet ist. Bei der Interpretation der Angaben ist zum einen zu beachten, dass bei der Bildung der Summenscores fehlende Werte de facto als Falschantworten eingehen (vgl. Abschnitt 3.2). Des Weiteren muss berücksichtigt werden, dass die Reliabilität unter anderem von der Anzahl der Testitems einer Skala abhängt, die für den Test zum figuralen schlussfolgernden Denken deutlich geringer ist als beim Wissenstest.

Tabelle 5: Kennwerte für die Verteilungen der Skalenwerte (Summenscores)

	Leichtere Testform		Schwerere Testform	
	Wissen	Schlussf. Denken	Wissen	Schlussf. Denken
Min	4	0	1	0
Max	36	16	35	11
M	19.96	6.02	22.10	4.80
SD	7.61	3.21	7.09	2.67
Cronbach α	.89	.74	.83	.67

Anmerkungen: Wissen = Wissenstest; Schlussf. Denken = Test zum figuralen schlussfolgernden Denken. Min = Minimal erreichte Anzahl richtig gelöster Items; Max = Maximal erreichte Anzahl richtig gelöster Items; *M* = Mittelwert; *SD* = Standardabweichung; Cronbach α = Koeffizient α (Cronbach, 1951). Die unterschiedlich schwierigen Testformen wurden teilweise in verschiedenen Geburtsjahrgängen eingesetzt (siehe Abschnitt 3.1).

In den Abbildungen 3 und 4 sind die Verteilungen der Summenscores für die einzelnen Testformen dargestellt. Diese geben Aufschluss über das Vorliegen von Boden- bzw. Deckeneffekten.

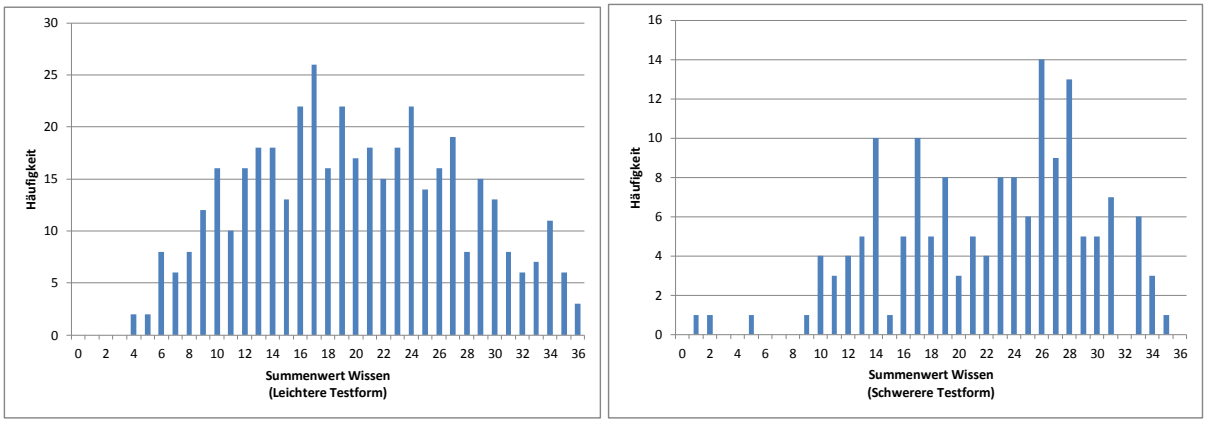

Abbildung 3: Verteilung für den Summenscore zum Wissenstest in der leichteren Testform (Geburtsjahrgänge 2003 bis 2006; links) bzw. schwereren Testform (Geburtsjahrgänge 2000/2001; rechts).

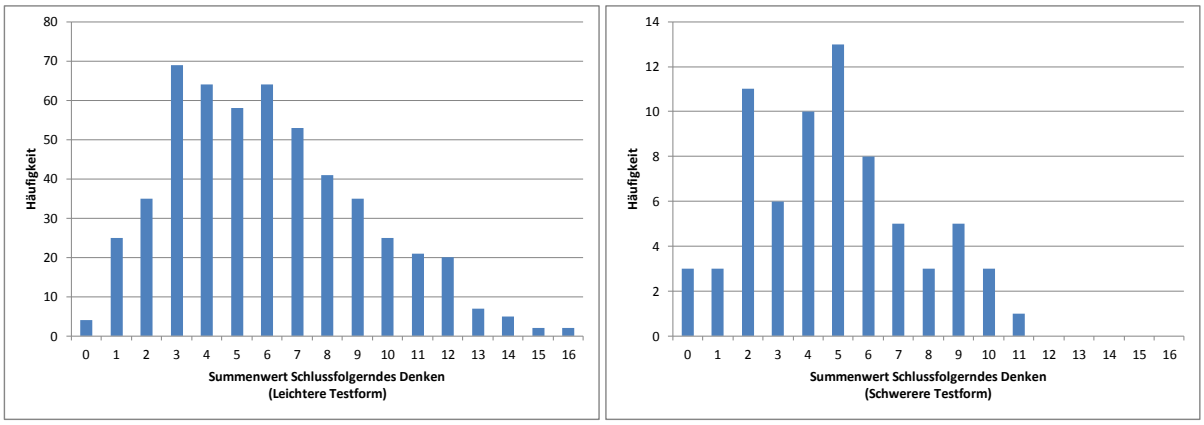

Abbildung 4: Verteilung für den Summenscore zum Test des figuralen schlussfolgernden Denkens in der leichteren Testform (links) bzw. schwereren Testform (rechts).

Die Verteilungen für die Summenscores zeigen, dass die Testformen zum Wissen eine angemessene Schwierigkeit aufweisen. Hier wurden von 36 vorgelegten Items im Mittel rund 20 Items (leichtere Testform in den Geburtsjahrgängen 2003 bis 2006) bzw. gut 22 Items (schwerere Testform in den älteren Geburtsjahrgängen 2000 und 2001) richtig gelöst und es liegen keine Boden- oder Deckeneffekte vor. Die Reliabilität der Summenscores ist als hoch einzuordnen.

Bei den Testformen zum figuralen schlussfolgernden Denken wurden von 16 vorgelegten Items im Mittel rund 6 Items (leichtere Testform) bzw. knapp 5 Items (schwerere Testform) richtig gelöst. Für beide Testformen sind dabei leichte bis moderate Bodeneffekte zu verzeichnen: So lösten etwa 12 Prozent (leichtere Testform) bzw. rund 24 Prozent (schwerere Testform) der Testpersonen lediglich maximal 2 der vorgelegten Items und erzielten damit ein Ergebnis, das mit der Annahme von Rateverhalten konform ist. Die Reliabilität der Summenscores ist dementsprechend insbesondere für die schwerere Testform im Geburtsjahrgang 2000 eingeschränkt.

4.4 Skalierungsergebnisse

Wie in Abschnitt 3.3 erläutert, wurden die Antworten der Testpersonen im Wissenstest sowie im Test zum figuralen schlussfolgenden Denken unter Verwendung des Rasch-Testmodells skaliert, um alle Personen unabhängig von der verwendeten Testform auf einer einheitlichen Metrik abzubilden und die Ergebnisse somit über Testformen und Geburtsjahrgänge hinweg vergleichen zu können.

Bei der Skalierung der insgesamt 48 Wissensitems konvergierte die Parameterschätzung nach 42 Iterationen (Konvergenzkriterien: *Parameter Change* = 0.0001, *Deviance Change* = 0.0001). Die Passung der Items zum Testmodell (*Weigthted Fit* / *Infit*) variierte zwischen 0.84 und 1.20, wobei lediglich 2 Items (gcA_M_bio1, gcB_MO_phy2) einen Infit > 1.15 aufwiesen. Somit zeigte sich für nahezu alle Items eine Passung im akzeptablen Bereich (Köhler & Hartig, 2017). Die Reliabilität der Personenparameter (WLEs) beträgt .82.

Die Skalierung der insgesamt 24 Items zum figuralen schlussfolgernden Denken führte nach 25 Iterationen zur Konvergenz, wobei die gleichen Konvergenzkriterien verwendet wurden wie bei den Wissensitems. Die Passung der Items variierte für 23 Items in einem akzeptablen Bereich zwischen 0.90 und 1.15; lediglich für 1 Item (gffB_MO4) ergab sich höherer *Infit*-Werte von 1.2, der darauf hinweist, dass dieses Item eine geringere Trennschärfe aufweist als vom Testmodell prognostiziert. Die Reliabilität der Personenparameter (WLEs) beträgt .67.

Wie in Abschnitt 3.3 erläutert, wurden die Personenparameter im letzten Schritt durch eine lineare Transformation in T-Werte umgerechnet. Dabei ergeben sich für die hier untersuchten Kinder und Jugendlichen die in Tabelle 6 aufgeführten Verteilungskennwerte.

Tabelle 6: Kennwerte für die Verteilungen der Personenparameter aus den Skalierungen (T-Werte)

Geburts-jahrgang	Alters-durch-schnitt [Jahre]	Wissenstest			Test zum figuralen schlussfolgernden Denken		
		n	*M*	*SD*	*n*	*M*	*SD*
2005/2006	12.6	252	47.38	9.72	267	49.22	9.41
2003/2004	14.6	214	51.20	10.53	213	50.63	10.82
2000/2001	17.5	156	52.60	8.68	156	50.47	9.79
alle Jg.	14.5	622	50.00	10.00	636	50.00	10.00

Anmerkungen: *M* = Mittelwert; *SD* = Standardabweichung; alle Jg. = alle Geburtsjahrgänge insgesamt.

Für den Wissenstest zeigen sich erwartungskonform höhere Testwerte mit steigendem Alter. Die mittleren Personenfähigkeitsschätzer des Tests zum figuralen schlussfolgernden Denken unterscheiden sich hingegen kaum zwischen den Altersgruppen.

5 Diskussion

Die beiden Testversionen zur Erfassung deklarativen Wissens im Bereich Naturwissenschaften/Technik erwiesen sich aus psychometrischer Sicht als gut geeignet, um das Vorwissen in der Zielpopulation zu erfassen. Insbesondere waren die Tests in ihrer Schwierigkeit angemessen und erlauben es, individuelle Unterschiede im naturwissenschaftlich-technischen Wissen differenziert abzubilden.

Die Ergebnisse weisen darauf hin, dass sich Heranwachsende mit Fluchtbiographie, die in den letzten Jahren nach Deutschland gekommen sind, in ihrem Wissen erheblich unterscheiden. So fällt die Streuung der Testwerte in der hier untersuchten Stichprobe für alle Geburtsjahrgänge substanziell aus. Dieses Ergebnis erscheint vor dem Hintergrund unterschiedlicher Bildungsbiographien und Bildungssysteme in den Herkunftsländern plausibel.

Im Hinblick auf den Test zur Erfassung des figuralen schlussfolgernden Denkens fällt auf, dass sich keine nennenswerten Altersunterschiede zwischen den untersuchten Geburtsjahrgängen zeigen. Dies widerspricht früheren Befunden (vgl. etwa Schroeders et al., 2015) und könnte ein Hinweis auf eine eingeschränkte Validität der Testergebnisse sein. Darüber hinaus lösten die Testteilnehmenden im Durchschnitt nur wenige Items dieses Tests korrekt. Hierfür sind verschiedene Erklärungsansätze denkbar:

Erstens wäre es möglich, dass die Kinder und Jugendlichen mit Fluchtbiografie die Aufgabenstellung bzw. das Itemformat – etwa aufgrund einer geringen Vertrautheit mit pädagogisch-psychologischen Leistungstests – nicht ausreichend verstanden haben. Dies ist jedoch eher unwahrscheinlich, da zum einen die kognitiven Interviews mit Expertinnen und Experten und Heranwachsenden keine Hinweise auf Probleme mit der Aufgabenstellung ergeben hatten. Zum anderen löste die überwiegende Mehrheit der Testteilnehmenden das erste vorgelegte Item, das von geringer Schwierigkeit war, korrekt, was nahelegt, dass die Instruktion verstanden wurde und korrekt angewendet werden konnte.

Plausibler scheint die Erklärung, dass ungünstige Testbedingungen (Störungen durch andere Personen im Haushalt, Nebengeräusche, Durchführung der Tests am Ende des gesamten Befragungsprogramms und z. T. zu sehr später Uhrzeit) die Testleistung negativ beeinflusst haben, wobei diese Faktoren die Leistung im schlussfolgernden Denken aufgrund der hohen Arbeitsgedächtnisbelastung stärker beeinträchtigt haben dürften als den Abruf deklarativen Wissens im Wissenstest.

Außerdem leiden Heranwachsende mit Fluchtbiographie überproportional häufig an psychischen Belastungen bis hin zu posttraumatischen Belastungsstörungen (Mannhart & Freisleder, 2017; Metzner, Reher, Kindler & Pawils, 2016; Nowotny, Mall & Langer, 2018), womit häufig Schlafstörungen und Konzentrationsprobleme einhergehen (Ruf, Schauer & Elbert, 2010). Somit könnte eine verminderte Konzentrationsfähigkeit die Performanz in Tests zur Erfassung der kognitiven Grundfähigkeiten eingeschränkt haben.

Schließlich ist es auch möglich, dass die Testteilnehmenden wenig motiviert für die Bearbeitung der Aufgaben zum figuralen schlussfolgenden Denken waren, etwa, da die Testung am Ende der Befragung erfolgte oder da sich ihnen der Sinn der Testung nicht erschloss. Außerdem erfordert die Bearbeitung dieser Aufgaben einen vergleichsweise hohen kognitiven Aufwand, was sich zusätzlich negativ auf die Testteilnahmemotivation niedergeschlagen haben könnte.

Insgesamt ist auf der Grundlage der vorliegenden Erhebungs- und Befragungsdaten kein abschließendes Urteil zur Validität der Testwerte zum figuralen schlussfolgernden Denken möglich. Hierzu sind weitere Untersuchungen erforderlich, bei denen auch andere Indikatoren mit unterschiedlichen Itemformaten in einer besser kontrollierten Testumgebung zum Einsatz kommen sollten. Die bisher vorliegenden Testwerte zum schlussfolgernden Denken sollten vor diesem Hintergrund nur mit Vorsicht inhaltlich interpretiert werden.

Für die Tests zur Erfassung des deklarativen Wissens im Bereich Naturwissenschaften/Technik ergaben sich hingegen keine Hinweise auf eine eingeschränkte Validität beim Einsatz in der hier untersuchten Population. Sie können daher als Indikatoren individueller Unterschiede in schulrelevantem Wissen von Kindern und Jugendlichen mit Fluchtbiographie herangezogen werden.

Literatur

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review, 4*(3), 165-176. doi: 10.1016/j.edurev.2009.04.002

Brücker, H., Schewe, P., & Sirries, S. (2016). *Eine vorläufige Bilanz der Fluchtmigration nach Deutschland. Aktuelle Berichte 19*. Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung (IAB).

Brücker, H., Kunert, A., Mangold, U., Kalusche, B., Siegert, M., & Schupp, J. (2016). *Geflüchtete Menschen in Deutschland - eine qualitative Befragung. SOEP-Survey Papers, 313*. Berlin: DIW.

Cattell, R. B. (1940). A culture-free intelligence test. I. *Journal of Educational Psychology, 31*(3), 161-179. doi: 10.1037/h0059043

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action*. Amsterdam: Elsevier.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. doi: 10.1007/BF02310555

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence, 21*(2), 135-155. doi: 10.1016/0160-2896(95)90023-3

Dye, D. A., Reck, M., & McDaniel, M. A. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment, 1*(3), 153-157. doi: 10.1111/j.1468-2389.1993.tb00103.x

Hartig, J., & Klieme, E. (2006). Kompetenz und Kompetenzdiagnostik. In K. Schweizer (Hrsg.), *Leistung und Leistungsdiagnostik* (S. 127-143). Berlin: Springer.

Köhler, C. & Hartig, J. (2017). Practical significance of item misfit in educational assessments. *Applied Psychological Measurement, 41,* 388-400. doi: 10.1177/0146621617692978

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

Mannhart, A., & Freisleder, F. J. (2017). Traumatisierung bei unbegleiteten minderjährigen Flüchtlingen. Behandlung in der kinder- und jugendpsychiatrischen Klinik. *Monatsschrift Kinderheilkunde, 165*(1), 38-47. doi: 10.1007/s00112-016-0199-3

McCallum, R. S. (Hrsg.). (2003). *Handbook of Nonverbal Assessment.* Boston: Springer.

McGrew, K. S., & Hessler, G. L. (1995). The relationship between the WJ-R Gf-Gc cognitive clusters and mathematics achievement across the life-span. *Journal of Psychoeducational Assessment, 13*(1), 21-38. doi: 10.1177/073428299501300102

Metzner, F., Reher, C., Kindler, H., & Pawils, S. (2016). Psychotherapeutische Versorgung von begleiteten und unbegleiteten minderjährigen Flüchtlingen und Asylbewerbern mit Traumafolgestörung in Deutschland. *Bundesgesundheitsblatt, 59*, 642-651. doi: 10.1007/s00103-016-2340-9

Nowotny, T., Mall, V., & Langer, T. (2018). Medizinische Versorgung von Kindern und Jugendlichen mit Fluchthintergrund. In B. Stier, N. Weissenrieder & K. O. Schwab (Hrsg.), *Jugendmedizin* (S. 385-398). Berlin: Springer.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press (Original work published 1960).

Roppelt, A., Blum, W., & Pöhlmann, C. (2013). Beschreibung der untersuchten mathematischen Kompetenzen. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Hrsg.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (S. 23-37). Münster: Waxmann.

Ruf, M., Schauer, M. & Elbert, T. (2010). Prävalenz von traumatischen Stresserfahrungen und seelischen Erkrankungen bei in Deutschland lebenden Kindern von Asylbewerbern. *Zeitschrift für Klinische Psychologie und Psychotherapie, 39*, 151-160. doi: 10.1026/1616-3443/a000029

Schipolowski, S., Haag, N., & Böhme, K. (2016). Anlage und Durchführung. In P. Stanat, S. Schipolowski, C. Rjosk, S. Weirich, & N. Haag (Hrsg.), *IQB-Bildungstrend 2016. Kompetenzen in den Fächern Deutsch und Mathematik am Ende der 4. Jahrgangsstufe im zweiten Ländervergleich* (S. 95-119). Münster: Waxmann.

Schroeders, U., Schipolowski, S., & Wilhelm, O. (2015). Age-related changes in the mean and covariance structure of fluid and crystallized intelligence in childhood and adolescence. *Intelligence, 48*, 15-29. doi: 10.1016/j.intell.2014.10.006

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

Schupp, J., Brücker, H., Brenzel, H., Jacobsen, J., Jaworski, J.,…, Siegert, M. (2017). Bildung, Sprache und kognitive Potentiale. In H. Brücker, N. Rother & J. Schupp (Hrsg.), *IAB-BAMF-SOEP-Befragung von Geflüchteten 2016: Studiendesign, Feldergebnisse sowie Analysen zu schulischer wie beruflicher Qualifikation, Sprachkenntnissen, sowie kognitiven Potentialen* (S. 19-80). Berlin: Deutsches Institut für Wirtschaftsforschung (DIW).

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. doi: 10.1007/s11336-008-9101-0

Wilhelm, O., Schroeders, U., & Schipolowski, S. (2014). *Berliner Test zur Erfassung fluider und kristalliner Intelligenz für die 8. bis 10. Jahrgangsstufe (BEFKI 8-10)*. Göttingen: Hogrefe.