



NEPS WORKING PAPERS

Frieder R. Lang, Stefan Kamin, Margund Rohr, Conrad  
Stünkel und Bettina Williger

ERFASSUNG DER FLUIDEN KOGNITI-  
VEN LEISTUNGSFÄHIGKEIT ÜBER DIE  
LEBENSSPANNE IM RAHMEN DER NA-  
TIONAL EDUCATIONAL PANEL STUDY:  
ABSCHLUSSBERICHT ZU EINER NEPS-  
ERGÄNZUNGSSTUDIE

NEPS Working Paper No. 43  
Bamberg, Juni 2014

**Working Papers of the German National Educational Panel Study (NEPS)**

at the Leibniz Institute for Educational Trajectories (LifBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).

The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent 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 NEPS Consortium.

The NEPS Working Papers are available at

<https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers>

**Editorial Board:**

Jutta Allmendinger, WZB Berlin

Cordula Artelt, University of Bamberg

Jürgen Baumert, MPIB Berlin

Hans-Peter Blossfeld, EUI Florence

Wilfried Bos, University of Dortmund

Edith Braun, DZHW Hannover

Claus H. Carstensen, University of Bamberg

Henriette Engelhardt-Wölfler, University of Bamberg

Frank Kalter, University of Mannheim

Corinna Kleinert, IAB Nürnberg

Eckhard Klieme, DIPF Frankfurt

Cornelia Kristen, University of Bamberg

Wolfgang Ludwig-Mayerhofer, University of Siegen

Thomas Martens, DIPF Frankfurt

Manfred Prenzel, TU Munich

Susanne Rässler, University of Bamberg

Marc Rittberger, DIPF Frankfurt

Hans-Günther Roßbach, LifBi

Hildegard Schaeper, DZHW Hannover

Thorsten Schneider, University of Leipzig

Heike Solga, WZB Berlin

Petra Stanat, IQB Berlin

Volker Stocké, University of Kassel

Olaf Struck, University of Bamberg

Ulrich Trautwein, University of Tübingen

Jutta von Maurice, LifBi

Sabine Weinert, University of Bamberg

**Contact:** German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – [contact@lifbi.de](mailto:contact@lifbi.de)

# **Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study**

Abschlussbericht zu einer NEPS-Ergänzungsstudie<sup>1</sup>

*Frieder R. Lang, Stefan Kamin, Margund Rohr, Conrad Stünkel & Bettina Williger  
Institut für Psychogerontologie, Universität Erlangen-Nürnberg*

## **Korrespondenzanschrift:**

Prof. Dr. Frieder R. Lang  
Institut für Psychogerontologie der Universität Erlangen-Nürnberg  
Kobergerstraße 62  
90408 Nürnberg  
Tel.: 0911 5302-96100  
Fax: 0911 5302-96101  
E-Mail: ipg-sekretariat@fau.de

## **Bibliographische Angaben:**

Lang, F. R., Kamin, S., Rohr, M., Stünkel, C. & Williger, B. (2014). Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen des Nationalen Bildungspanels: Abschlussbericht zu einer NEPS-Ergänzungsstudie (NEPS Working Paper No. 43). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

---

<sup>1</sup> Das vorliegende Working Paper wurde 2012 als Abschlussbericht einer Ergänzungsstudie zu NEPS zur Entwicklung und Validierung eines large-scale geeigneten Kurztests zur Erfassung allgemeiner kognitiver Grundfähigkeiten (nonverbal) verfasst.

## Inhalt

1. Zielsetzung und Rahmenbedingungen des CAL-Projekts .....	3
2. Methodisches Vorgehen .....	3
2.1 Stichprobe .....	3
2.2 Projektdurchführung .....	4
2.3 Eingesetzte Testverfahren.....	5
2.3.1 Neue Verfahren NEPS-BZT und NEPS-MAT .....	5
2.3.2 Validierungsverfahren .....	6
3. Erhebungsinstrumente .....	7
3.1 Prämissen und Zielkriterien.....	8
3.2 Bilder-Zeichen-Test des NEPS (BZT-NEPS).....	8
3.2.1 Testaufbau und Durchführung .....	8
3.2.2 Auswertung .....	9
3.3 Matrizen-Test des NEPS (NEPS-MAT).....	10
3.3.1 Testaufbau und Durchführung .....	10
3.3.2 Auswertung .....	11
4. Ergebnisse zu den Gütekriterien der ultrakurzen kognitiven Tests .....	12
4.1 Verteilung der Kennwerte .....	12
4.1.1 NEPS-BZT .....	12
4.1.2 NEPS-MAT.....	12
4.2 Reliabilität.....	13
4.2.1 Interne Konsistenz.....	13
4.2.2 Testhalbierungsreliabilität.....	14
4.2.3 Retest-Reliabilität .....	15
4.3 Konvergente und divergente Validität .....	16
4.4 Darstellung der Messäquivalenz .....	18
4.1.1 Einzelgruppenanalysen.....	18
4.3.3 Multigruppenanalysen .....	20
4.5 Durchführungs- und Auswertungsobjektivität .....	21
5. Diskussion und Fazit .....	21
6. Literatur .....	23

## 1. Zielsetzung und Rahmenbedingungen des CAL-Projekts

Ziel des Projekts war die Entwicklung von robusten ultrakurzen Tests zur Messung der fluiden kognitiven Leistung über die Lebensspanne, die sich dafür eignen sollen, im Rahmen der Panel-Erhebungen des Nationalen Bildungspanels (NEPS) je nach Bedarf in Einzel- oder Gruppensitzungen, eingesetzt zu werden. Zu diesem Zweck wurden zwei kognitive Kurztests entwickelt, die aufgrund einer über alle Altersgruppen gleichen Aufgabenstruktur eine kontinuierliche Messung bei Kindern, Jugendlichen und Erwachsenen erlauben. In ihrer Expertise empfehlen Brunner, Lang und Lüdtke (2009), dass als Grundlage für eine robuste, non-verbale Erfassung fluider Leistungspotenziale über die Lebensspanne bereits bestehende, nicht-kommerzielle Tests zur Erfassung der Wahrnehmungsgeschwindigkeit (WG, vgl. dazu den SOEP-Zeichentest; Lang, 2005; Lang, Weiss, Stocker & von Rosenblatt, 2007) und des Schlussfolgernden Denkens (SF, vgl. dazu Gerotest-Matrizen-Test: GerMaT, IPG Erlangen, [www.gerotest.de](http://www.gerotest.de)) dienen. In der zweijährigen Projektlaufzeit von Oktober 2009 bis September 2011 wurden auf Basis der Expertise von Brunner, Lang und Lüdtke (2009/2014)<sup>2</sup> zwei ultrakurze Instrumente entwickelt, die sich hinsichtlich der genannten Kriterien als valide, messinvariant und über verschiedene Erhebungsbedingungen als robust erweisen.

## 2. Methodisches Vorgehen

### 2.1 Stichprobe

An allen Erhebungen zusammen nahmen insgesamt 409 Probanden teil. Der Anteil männlicher ( $n = 216$ , 53 %) und weiblicher ( $n = 193$ , 47 %) Teilnehmer ist in der Gesamtstichprobe und in den einzelnen Teilstichproben weitestgehend ausgewogen. Das Alter der getesteten Probanden reicht von fünf bis 78 Jahren ( $M = 14.1$  Jahre,  $SD = 11.1$ ). Tabelle 1 gibt einen Überblick der Zusammensetzung der Stichproben nach Bildungsabschluss bzw. Schulform.

---

<sup>2</sup> Diese 2009 verfasste Expertise wird 2014 als NEPS Working Paper erscheinen.

## 2.2 Projektdurchführung

Im Zeitraum von Oktober 2009 bis September 2011 wurde eine Validierungsstudie mit insgesamt 409 Teilnehmer(n)/-innen aus fünf Altersgruppen (Vorschule, 2. Klasse, 5. und 9. Klasse sowie Erwachsene) durchgeführt. Die Probanden wurden über verschiedene Zugangswege rekrutiert, wobei darauf geachtet wurde, eine hinsichtlich sozio-ökonomischer Bedingungen ausgeglichene Stichprobe zu generieren.

Tabelle 1: Merkmale der Stichproben

	Gesamt	Vorschule	2. Klasse	5. Klasse	9. Klasse	Erwachsene
	<i>N</i> = 409	<i>n</i> = 88	<i>n</i> = 99	<i>n</i> = 60	<i>n</i> = 92	<i>n</i> = 70
Geschlecht (weiblich in %)	193 (47)	44 (50)	42 (42)	26 (43)	38 (41)	43 (61)
Alter (in Jahren)						
<i>M</i>	14.1	6.3	8.3	11.2	15.3	30.4
<i>SD</i>	11.1	0.5	0.4	0.7	0.9	16.7

Anm.: *M* = Mittelwert, *SD* = Standardabweichung

**Vorschule und 2. Klasse.** Das Projekt wurde an sechs Kindergärten und sechs Grundschulen unter Einbezug des schulpsychologischen Dienstes im Raum Erlangen-Nürnberg vorgestellt. Eltern und Erzieher/-innen bzw. Lehrer/-innen wurden mittels Informationsschreiben über das Projekt und dessen Ziele und Ablauf informiert. Die Testung erfolgte in Abstimmung mit den Ansprechpartnern in den Kindergärten und Schulen und mit Zustimmung der Eltern und Kinder. Die Testung fand in Sitzungen mit jeweils ein bis drei Kindern (Kindergarten) bzw. mit Kleingruppen mit zehn Kindern (Grundschule) unter Anleitung von geschulten Testleitern und Testleiterinnen in der Papier-Bleistift-Version statt. Die Erhebung nahm bei der Ersttestung 60 Minuten (Kindergarten) bzw. 90 Minuten (Grundschule) in Anspruch. Eine Wiederholungstestung mit einem reduzierten Testheft (vgl. Abschnitt 4.2.3), für die maximal 45 Minuten veranschlagt wurden, erfolgte nach durchschnittlich zehn Tagen (Range von sieben bis zwölf Tagen).

**5. und 9. Klasse.** Die Rekrutierung der Schulen erfolgte nach Genehmigung des Schulamts Mittelfranken. Lehrer/-innen, Eltern und Schüler/-innen wurden mittels Informationsschreiben über das Projekt informiert. Auch in diesen beiden Altersgruppen erfolgte die Testung in Abstimmung mit Ansprechpartnern in den Schulen und mit Zustimmung der Eltern und Schüler/-innen. Die Erhebung in den Schulen fand in Gruppentestungen mit jeweils 20 bis 40 Teilnehmer/-innen in der Papier-Bleistift-Version unter Anleitung von zwei geschulten Testleiter/-innen statt und dauerte durchschnittlich 70 Minuten.

**Erwachsene.** Die Erwachsenen wurden über Aushänge, Informationen in lokalen Medien und Informationsveranstaltungen rekrutiert. Die Testung fand in Einzel- wie auch Gruppensitzungen unter Anleitung eines/einer geschulten Testleiter/-in in der Papier-Bleistift-Version statt. Die Erhebung nahm bei der Ersttestung 60 Minuten in Anspruch. Die Wiederholungstestun-

gen mit einem reduzierten Testheft fanden jeweils nach durchschnittlich 41 Tagen (27 bis 58 Tage) statt.

Unabhängig vom Alter der Probanden war jeweils ein/e Testleiter/-in anwesend, der/die die Testverfahren instruierte und auf Einhaltung der Instruktion achtete. In den Kindergärten und an den Schulen wurden zusätzlich für jeden Testtermin ein bis zwei mit den Testverfahren vertraute Aufsichten eingesetzt. Die Probanden saßen mit ausreichend Abstand zum nächsten Probanden jeweils an einem Schreibtisch und bearbeiteten nach Maßgabe des Testleiters/der Testleiterin der Reihe nach die in einem Testheft zusammengefassten Aufgaben. Lösungen waren mit einem Stift einzutragen.

## 2.3 Eingesetzte Testverfahren

Die im Rahmen des Projekts neu entwickelten Testverfahren NEPS-BZT (Bilder-Zeichen-Test) und NEPS-MAT (Matrizen-Test) wurden in einer Validierungsstudie erprobt. Hierzu wurde neben diesen beiden Tests eine Batterie aus konvergenten und divergenten Verfahren erhoben.

### 2.3.1 Neue Verfahren NEPS-BZT und NEPS-MAT

Ausgangspunkt der Entwicklung der neuen Erhebungsinstrumente waren die Vorarbeiten des Instituts für Psychogerontologie zur Entwicklung eines Zeichentests für das Sozio-Ökonomische Panel (SOEP) auf Basis des Symbol Digit Modalities Test (SDMT; Smith, 1995) sowie die Arbeiten zur Entwicklung des Gerotest-Matrizen-Tests (GerMaT). Im SDMT gilt es, bestimmte Zahlen nach einem vorgegebenen Zuordnungsschlüssel bestimmten Symbolen zuzuordnen. Die Aufgaben der Gerotest-Matrizen erfordern, aus einer Reihe von Alternativen diejenige Lösung zu identifizieren, welche ein vorgegebenes Muster optimal bzw. passend ergänzt.

*NEPS-BZT.* Der NEPS-BZT erfasst primär die Wahrnehmungsgeschwindigkeit (speed), mit der die einem bildhaften Symbol zugeordnete Zahl bzw. ein Zeichen erkannt und niedergeschrieben wird. Erfasst wird somit auch die Geschwindigkeit der Auge-Hand-Koordination. Der zur Konstruktion des NEPS-BZT herangezogene Test ist wie der SOEP-Zeichentest (Lang et al., 2007) an den *Symbol Digit Modalities Test* (SDMT) und den Zahlen-Symbol-Tests der Wechsler-Testfamilie (ZST; englisch: Digit Symbol Test, DST) angelehnt, unterscheidet sich jedoch in der Umsetzung der konkreten Aufgabenstellung. Die im SOEP-Zeichentest noch abstrakten Symbole wurden zugunsten konkreter, bildlicher Darstellungen alltäglicher Objekte ersetzt. Damit sollen die Aufgaben insbesondere für Kinder unter zehn Jahren attraktiver werden. Für diesen Altersbereich wurden außerdem die Zahlen durch einfache Zeichen ersetzt, da Kinder Zahlen vor der Einschulung noch nicht ausreichend sicher schriftlich beherrschen. Dadurch sollte ermöglicht werden, dass der Test auch bei Kindern unter zehn Jahren einsetzbar ist. Zudem steigert der hohe Alltagsbezug die Relevanz und die Attraktivität der Tests für die Probanden.

*NEPS-MAT.* Dieser Test erfasst als traditioneller Matrizen-Test das schlussfolgernde Denken (*reasoning*). Für die Matrizen wurde ein Itempool von Aufgaben bzw. Items entwickelt, die insbesondere in Typ und Komplexität variierten. Die bereits früher entwickelten Gerotest-Matrizen (GerMaT, IPG Erlangen, [www.gerotest.de](http://www.gerotest.de)) dienten als Ausgangspunkt der Testkon-

struktion und wurden in Anlehnung und auf Grundlage einschlägiger Arbeiten zur Matrizenkonstruktion (vgl. Carpenter, Just & Shell, 1990; Jacobs & Vandeventer, 1972; Primi, 2001) sowie der RAVEN-Matrizen (*Standard Progressive Matrices SPM-C*, Raven, 1977) im Hinblick auf die grundlegenden Funktionsprinzipien und die einzusetzenden Stimulusbestandteile weiterentwickelt bzw. ergänzt. Die resultierende Hierarchie von Regeln unterschiedlichen Typs und unterschiedlicher Komplexität diente zusammen mit einem Kanon möglichst einfacher so genannter Geome zur Konstruktion eines Pools von 105 Matrizenaufgaben, die in ihrer grundlegenden Struktur äquivalent sind (z. B. bezüglich der Regeln bei der Generierung der Matrizen und Stimuli) und somit bei verschiedenen Altersgruppen eingesetzt werden können. Die so generierten Matrizen sind also in ihrem Aufbau und der ihnen zugrunde liegenden Konstruktionslogik strukturgleich. Aus dem Itempool wurden 18 Matrizen für die Validierungsstudie zum NEPS-MAT mit Vorschüler/-innen und Zweitklässler/-innen eingesetzt. Ein weiteres Matrizen-Subset aus 38 altersadäquaten Aufgaben wurde für die Validierungsstudie mit Fünft- und Neuntklässlern sowie Erwachsenen ausgewählt.

### 2.3.2 Validierungsverfahren

Nach Brunner und Kollegen (2009/2014) spiegeln die Wechsler Testverfahren einen aktuellen Gold-Standard zur Erfassung fluider kognitiver Leistungspotenziale wider. Aufgrund ihrer weiten Verbreitung und Bekanntheit wurden die hier neu entwickelten Testverfahren an ausgewählten Tests der Wechsler-Familie (z. B. HAWIVA, HAWIK, WIE) validiert.

Wie bei den Tests der Wechsler-Familie vorgesehen, wurde für die beiden jüngsten Altersgruppen als Speed-Test der Test „Symbole Kodieren“ des HAWIVA-III (Petermann & Petermann, 2008) eingesetzt, der für Probanden bis zum Alter von 7;11 Jahren adäquat ist. Ab der 5. Klasse wurde der ab 8;0 Jahren geeignete Zahlen-Symbol-Test B (ZST-B) des HAWIK-IV (Ricken, Fritz, Schuck & Preuß, 2007) eingesetzt. Da die Aufgabenstellung dieses Tests weitgehend mit der ZST-Fassung für Erwachsene (WIE - Wechsler Intelligenztest für Erwachsene; von Aster, Neubauer & Horn, 2006) übereinstimmt und aus weniger Items besteht (119 statt 131 im WIE), wurde der Test für die Validierungsstudie bevorzugt. Zudem konnte so die Vergleichbarkeit der konvergenten Validierung des NEPS-BZT zwischen den unterschiedlichen Altersgruppen hergestellt werden.

Wegen ihrer herausgehobenen Etablierung als Matrizen-tests für Reasoning wurde einem Subset der *RAVEN Standard Progressive Matrices*, klassische Ausführung (SPM-C; Raven, 1977), als konvergentem Validierungsmaß des NEPS-MAT der Vorzug gegenüber den Matrizenaufgaben der Wechsler-Tests gegeben. Die Vorschüler/-innen bearbeiteten Set A der SPM-C, während allen Probanden ab der 2. Klasse Set C der SPM-C vorgelegt wurde.

Tabelle 2: Übersicht der Validierungsmaße

Validität	Maß	Validierungsstudie 1		Validierungsstudie 2	
		Vorschule	2. Klasse	5. & 9. Klasse	Erwachsene
Konvergent ( <i>Speed</i> )	KO	✓	✓		
	ZST			✓	✓
Konvergent ( <i>Reasoning</i> )	RAVEN A	✓			
	RAVEN C		✓	✓	✓
Diskriminanz	CFT	✓	✓		
	BK	✓	✓		
	MWT			✓	✓
	LPS		✓	✓	✓
	IST-V			✓	✓
	IST-F			✓	✓

Anm.: KO = Symbole Kodieren (HAWIVA-III); ZST = Zahlen-Symbol-Test B (HAWIK-IV bzw. HAWIE); RAVEN = Standard Progressive Matrices, Set A bzw. C; BK = Bildkonzepte (HAWIK-IV); CFT = Klassifikationen (CFT-1); MWT = MWT-B; LPS = Wortflüssigkeit (L-P-S); IST-V = Merkfähigkeit Verbal (I-S-T 2000 R); IST-F = Merkfähigkeit Figural (I-S-T 2000 R)

In der Vorschule und 2. Klasse wurden als diskriminante Verfahren der Subtest 3, Form B des CFT („Klassifikationen“; Cattell, Weiß & Osterland, 1997) sowie der Subtest „Bildkonzepte“ des HAWIK-IV (Ricken et al., 2007) eingesetzt. Die „Klassifikationen“ erfassen das beziehungsstiftende Denken bei figuralem Material. Der Subtest „Bildkonzepte“ (BK) misst die Wahrnehmungsorganisation und das logische Denken. Die Vorschüler/-innen bearbeiteten eine Fassung mit den ersten zwölf Items des BK, während die Zweitklässler/-innen den kompletten Subtest aus 26 Items bearbeiteten. Ab der 2. Klasse kam zusätzlich der Subtest 6 des L-P-S (Horn, 1983) zur Erfassung von Wortflüssigkeit und verbalem Wissen zum Einsatz. Die beiden Merkfähigkeitstests des I-S-T 2000 R (Amthauer, Brocke, Liepmann & Beauducel, 2001) erfassen darüber hinaus zum einen die verbale (hier als IST-V geführt) und zum anderen die figurale Merkfähigkeit (hier als IST-F geführt). Sie wurden ab der fünften Klasse bei allen Probanden erhoben. Zusätzlich kam ab der 5. Klasse der Mehrfachwahl-Wortschatz-Intelligenztest MWT-B (Lehrl, 1977) zum Einsatz. Dieser erhebt eine Wissensstichprobe, wird aber auch als Schätzer der allgemeinen Intelligenz eingesetzt. Tabelle 2 gibt einen Überblick aller eingesetzten Erhebungsinstrumente zur Validierung der beiden NEPS-Tests.

### 3. Erhebungsinstrumente

Im Folgenden werden die neu entwickelten Erhebungsinstrumente zur Erfassung fluider kognitiver Leistungspotenziale im Einzelnen beschrieben. Dabei gehen wir zunächst auf die Ausgangsprämissen und Zielkriterien der zu entwickelnden Tests ein. Anschließend werden die beiden neu entwickelten Leistungstests hinsichtlich Aufbau, Durchführung und Auswertung beschrieben. Dabei werden drei strukturäquivalente, aber altersangepasste Testversionen differenziert: (1) Eine Version ist für Vorschüler/-innen, (2) eine für Schüler/-innen der 2. Jahrgangsstufe und (3) eine für Personen ab zehn Jahren (5. Klasse plus) vorgesehen. Da die

Tests derselben Testlogik folgen, wird einleitend jeweils das allgemeine Prinzip des Tests erläutert und dann auf Unterschiede zwischen den einzelnen Testversionen hingewiesen.

Vom ursprünglichen Konzept eines für alle Altersgruppen identischen Tests (inkl. Vorschulalter und 2. Klasse) musste abgerückt werden, nachdem sich der Einsatz der beiden ab der Klassenstufe 5 aufwärts bewährten kognitiven Kurztests in der Vorschule und Grundschule als nicht praktikabel erwies. Als Konsequenz hieraus wurden die ursprünglichen Tests vereinfacht, wobei darauf geachtet wurde, dass die vereinfachten Testversionen im Hinblick auf Aufgabenstellung und –struktur zu den ursprünglichen Versionen der Tests äquivalent sind. Die psychometrische Messäquivalenz der entwickelten Tests über alle Altersgruppen hinweg war dann auch Gegenstand empirischer Prüfungen (vgl. 4.4).

### **3.1 Prämissen und Zielkriterien**

Eine Zielsetzung bei der Entwicklung der beiden neuen Ultrakurztests war eine ökonomische Erhebung von kognitiven Leistungspotenzialen, die auch im Rahmen von Survey- und Panelstudien unter den dort üblichen Erhebungsbedingungen gut geeignet ist. Ein weiteres Zielkriterium der beiden Tests war es, eine möglichst kontinuierliche Erhebung über verschiedene Altersgruppen hinweg zu ermöglichen.

### **3.2 Bilder-Zeichen-Test des NEPS (BZT-NEPS)**

Beim Bilder-Zeichen/Zahlen-Test (BZT, englisch: *Picture-Digit-Test*, PDT) geht es darum, in einer vorgegebenen Zeit einen visuellen Reiz wahrzunehmen und in eine motorische Reaktion umzusetzen. Die zeitliche Begrenzung der Bearbeitung führt dazu, dass bei schnellerer Wahrnehmungsgeschwindigkeit eine höhere Zahl korrekter Antworten produziert wird.

#### **3.2.1 Testaufbau und Durchführung**

Auf jeder Testseite ist eine Doppelzeile aus Stimulus-Target-Verknüpfungen vorgegeben, die den Zuordnungsschlüssel darstellt: Während in der oberen Zeile die Testreize abgebildet sind, sind in der unteren Zeile diejenigen Reize zu entnehmen, die im eigentlichen Test fehlen und ergänzt werden müssen (Abbildungen 1 und 2). Unter dem Zuordnungsschlüssel folgen mehrere Testzeilen, in denen in ungeordneter Folge jeweils mehrere Stimuli (oberes Kästchen) vorgegeben werden, das jeweilige Target (unteres Kästchen) aber fehlt. Aufgabe des Probanden ist es, unter Zeitdruck für möglichst viele der Stimuli die entsprechenden Targets zu ergänzen.

Am Anfang des Tests steht eine Beispielseite zum Einüben des Testprinzips, danach folgen zwei bzw. drei Testdurchgänge. Da für jede dieser Testseiten (außer der Beispielseite) vom Testleiter/von der Testleiterin ein strenges Zeitlimit vorgegeben wird, werden diese im Folgenden als „Blöcke“ bezeichnet. Es wird jeweils klar instruiert, wann mit dem nächsten Aufgabenblock fortzufahren ist.

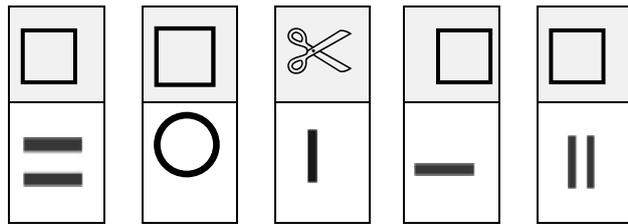


Abbildung 1: Beispielzeile des NEPS-BZT (Vorschüler/-innen und 2. Klasse)

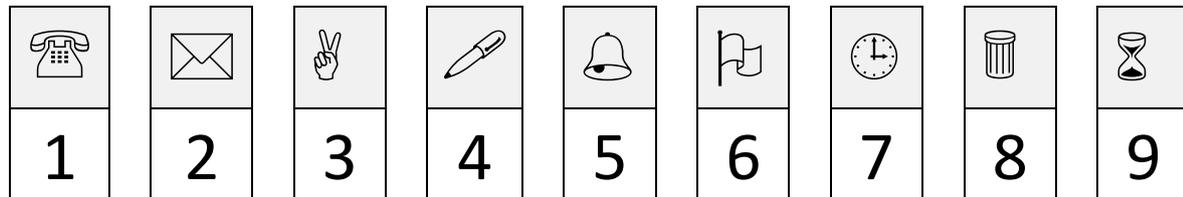


Abbildung 2: Beispielzeile des NEPS-BZT (5. Klasse bis Erwachsene)

*Vorschüler/-innen und 2. Klasse.* Für die beiden jüngsten Altersgruppen werden einfache geometrische Formen als Stimuli und einfache Zeichen als Targets gewählt (Abbildung 1). Jeder Aufgabenblock umfasst in drei Zeilen insgesamt 21 Aufgaben. Es gibt zwei Aufgabenblöcke, also insgesamt 42 Items. Für die Bearbeitung eines Aufgabenblocks haben Vorschüler/-innen 45, Grundschüler/-innen 30 Sekunden Zeit. Der Test umfasst für beide Altersgruppen fünf Seiten (inklusive zwei Zwischenblättern, um Vorausarbeiten in Gruppensituationen zu erschweren) und dauert in der Anwendung entsprechend 90 Sekunden für die Vorschüler/-innen und 60 Sekunden für die Grundschüler/-innen. Hinzu kommen ca. drei Minuten Instruktions- und Übungsdauer vor Testbeginn. Der Test ist sprachfrei und mit den Zwischenblättern so konzipiert, dass er auch in Kleingruppen zum Einsatz kommen kann.

*5. Klasse bis Erwachsene.* Probanden ab der 5. Klasse bekommen ein Stimulus-Set aus alltagsnahen, konkret-bildlichen Darstellungen vorgegeben. Targets sind die Zahlen von 1 bis 9 (Abbildung 2). Jeder Aufgabenblock umfasst in drei Zeilen insgesamt 31 Aufgaben. Es gibt drei Aufgabenblöcke mit insgesamt 93 Items. Die Bearbeitungszeit liegt bei 30 Sekunden pro Block. Der Test umfasst sechs Seiten (inkl. zwei Zwischenblättern) und dauert in der Anwendung 90 Sekunden plus ca. drei Minuten für Instruktion und Übung. Der Test ist so konzipiert, dass er in Einzel- wie auch in Gruppensitzungen eingesetzt werden kann.

### 3.2.2 Auswertung

Zur Auswertung des NEPS-BZT wird die Anzahl richtiger Zuordnungen für jeden der Testblöcke als Blockscore aufsummiert. Die Summenwerte bilden ein gemeinsames Konstrukt. Schwankungen in der Messgenauigkeit der zwei bzw. drei BZT-Subtests können über Kovarianzen der Subtests dargestellt und korrigiert werden.

*Vorschüler/-innen und 2. Klasse.* Der theoretische Wertebereich je Block liegt zwischen 0 und 21. Die zwei Summenwerte (Blockscores) können addiert und als Gesamtwert des BZT ausgegeben werden. Die theoretisch möglichen Werte liegen damit zwischen 0 und 42.

*5. Klasse bis Erwachsene.* Der theoretische Wertebereich je Block liegt zwischen 0 und 31. Die drei resultierenden Summenwerte (Blockscores) können addiert und als Gesamtwert des BZT ausgegeben werden. Die möglichen Werte liegen damit zwischen 0 und 93.

### 3.3 Matrizen-Test des NEPS (NEPS-MAT)

Der NEPS-MAT erfasst in der Tradition der RAVEN-Matrizen (Raven, 1977) schlussfolgerndes Denken (*reasoning*). Die Regeln, nach denen die Aufgaben erstellt wurden, sind für alle Altersgruppen gleich. Die Erstellung altersangepasster Testvarianten war aufgrund der unterschiedlichen Leistungsfähigkeit der Altersgruppen notwendig.

#### 3.3.1 Testaufbau und Durchführung

Jede Matrizenaufgabe besteht aus mehreren waagrecht und senkrecht angeordneten Feldern, in denen verschiedene geometrische Elemente abgebildet sind, wobei ein Feld frei bleibt. Die logischen Regeln, denen die Anordnung der geometrischen Elemente folgt, müssen erschlossen werden, um aus den angebotenen Lösungen die richtige Ergänzung für das frei gebliebene Feld auswählen zu können (Abbildung 3). Der Großteil der Matrizen besteht aus neun Matrizenfeldern, daneben gibt es einige weniger komplexe Matrizen mit vier Aufgabenfeldern. Aus dem vorliegenden Itempool von insgesamt 105 Matrizen wurde ein Test-Set von zwölf Aufgaben erstellt.

Am Anfang des Tests ist eine Seite mit zwei Beispielaufgaben enthalten, anhand derer der/ Testleiter bzw. die Testleiterin die Aufgabenstellung erklärt und einüben lässt. Da der NEPS-MAT auf die maximale Leistungsfähigkeit im Bereich des schlussfolgernden Denkens zielt und nicht auf die Schnelligkeit, mit der Lösungen produziert werden können, sind die Zeitlimits bei diesem Test möglichst großzügig bemessen. Erst nach deren jeweiligem Ablauf fordert der Testleiter bzw. die Testleiterin zur Bearbeitung des nächsten Aufgabenblocks auf. Es wird klar instruiert, wann mit der Aufgabe begonnen werden soll bzw. wann mit dem nächsten Aufgabenblock fortzufahren ist.

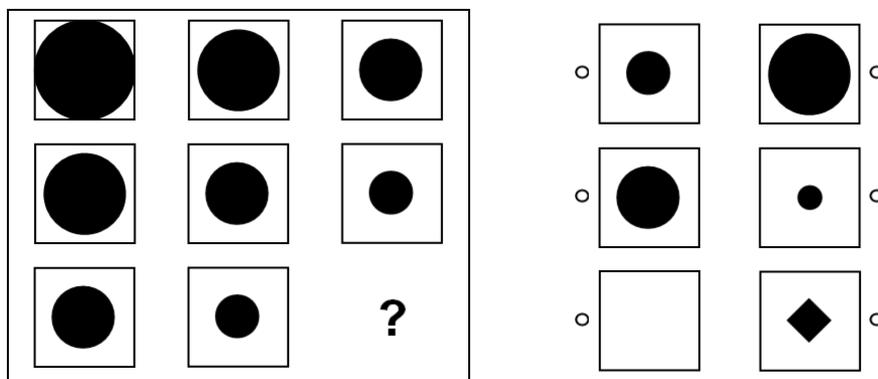


Abbildung 3: Beispielaufgabe des NEPS-MAT

*Vorschüler/-innen und 2. Klasse.* Den Vorschüler/-innen und Zweitklässler/-innen werden insgesamt zwölf Matrizen in zwei Blöcken vorgegeben. Der Test umfasst insgesamt sieben Seiten (inklusive Beispielseite und zwei Zwischenblättern). Für die Bearbeitung eines Aufgabenblocks haben die Probanden drei Minuten Zeit. Die Anwendungsdauer liegt damit bei sechs Minuten plus etwa vier Minuten Instruktion.

*5. Klasse bis Erwachsene.* Im NEPS-MAT für Erwachsene sind die Aufgaben auf drei Blöcke mit jeweils vier Items verteilt. Der Test umfasst inklusive Beispielseite und einem Zwischenblatt acht Seiten. Für die Bearbeitung jedes Aufgabenblocks haben die Testteilnehmer/-

innen drei Minuten Zeit; inklusive ca. zwei Minuten Instruktions- und Übungsdauer liegt die Anwendungsdauer damit bei insgesamt elf Minuten.

### **3.3.2 Auswertung**

Die Auswertung erfolgt in allen Testversionen durch Auszählen der richtigen Lösungen. Diese werden pro Aufgabenblock bzw. für die Gesamtskala als Summenwerte angegeben.

*Vorschüler/-innen und 2. Klasse.* In jedem der beiden Blöcke sind Testwerte von 0 bis 6 möglich. Der Gesamtscore liegt entsprechend zwischen 0 und 12.

*5. Klasse bis Erwachsene.* Bei jedem der drei Aufgabenblöcke sind 0 bis maximal 4 Punkte zu erreichen. Für den Gesamtscore lassen sich damit Werte zwischen 0 und 12 ermitteln.

## 4. Ergebnisse zu den Gütekriterien der ultrakurzen kognitiven Tests

Im folgenden Abschnitt werden die ermittelten Gütekriterien und Kennwerte für die beiden Instrumente berichtet. Zugunsten der Übersichtlichkeit werden diese, wo immer möglich, für alle Altersgruppen in einer Tabelle zusammengefasst.

### 4.1 Verteilung der Kennwerte

Im ersten Schritt werden die Verteilungen der Kennwerte (Mittelwerte, Streuungen) für den NEPS-BZT und die NEPS-MAT, im zweiten Schritt die Strukturäquivalenz der Tests über die verschiedenen Altersgruppen aufgezeigt. Abschließend werden die Instrumente hinsichtlich der Testgütekriterien evaluiert.

#### 4.1.1 NEPS-BZT

In Tabelle 3 sind die deskriptiven Verteilungen für die zwei bzw. drei Aufgabenblöcke des BZT zusammengefasst für die einzelnen Altersgruppen zu entnehmen. Erwartungsgemäß finden sich zwischen den Altersgruppen Leistungszuwächse (sowohl bei den Aufgabenblöcken als auch bzgl. der Gesamtwerte).

Tabelle 3: Deskriptive Statistiken des NEPS-BZT nach Altersgruppen und Aufgabenblöcken

	Vorschule		2. Klasse		5. Klasse		9. Klasse		Erwachsene	
	<i>n</i> = 88		<i>n</i> = 99		<i>n</i> = 60		<i>n</i> = 92		<i>n</i> = 70	
	<i>M</i>	<i>SD</i>								
NEPS-BZT-1	7.4	3.1	11.9	3.6	16.6	4.7	19.9	3.7	21.0	4.1
NEPS-BZT-2	9.3	3.5	14.0	2.4	12.9	3.0	16.4	3.1	20.2	3.7
NEPS-BZT-3	-	-	-	-	15.1	4.3	19.3	3.9	21.2	4.3
NEPS-BZT Gesamt	16.7	6.2	25.9	5.2	44.1	10.1	55.5	8.9	62.5	11.0

Anm.: M = Mittelwert, SD = Standardabweichung

Tabelle 3 zeigt die Mittelwerte für die drei (bzw. zwei bei Vorschüler/-innen und Zweitklässler/-innen) innerhalb einer Sitzung bearbeiteten Aufgabenblöcke von jeweils 30 Sekunden (bzw. 45 Sekunden für die Vorschule). Erkennbar wird, dass es in allen Altersgruppen eine weitgehende Gleichverteilung der je Aufgabenblock (= BZT-1, BZT-2, BZT-3) bearbeiteten Zahl von Bild-Zeichen-Elementen gab. Eine Ausnahme bilden die Ergebnisse für Vorschule und zweite Klasse, bei denen der zweite Aufgabenblock (45 Sekunden) leicht höhere Werte ergab. Diese Schwankungen in den Leistungen sind jedoch unerheblich, da eine akzeptable interne Konsistenz der Aufgabenblöcke besteht (vgl. Tab. 5)

#### 4.1.2 NEPS-MAT

In den Tabellen 4a und 4b sind die Verteilungen der Kennwerte für die NEPS-MAT getrennt nach Altersgruppen zusammengefasst.

*Vorschüler/-innen und 2. Klasse.* Die Schwierigkeiten der Matrizenaufgaben liegen mit  $P_{MAT-Kiga} = .45$  und  $P_{MAT-2.KI} = .60$  im erwünschten mittleren Bereich (Tabelle 4a).

5. Klasse bis Erwachsene. Die Schwierigkeiten der Matrizen liegen für die einzelnen Altersgruppen zwischen  $P_{\text{MAT-5.KI}} = .49$  und  $P_{\text{MAT-Erw}} = .76$  (Tabelle 4b).

Tabelle 4a: Deskriptive Statistiken für NEPS-MAT für Vorschüler/-innen und Zweitklässler/-innen

	Vorschule				2. Klasse			
	<i>n</i> = 88				<i>n</i> = 99			
	<i>M</i>	<i>SD</i>	<i>P</i>	<i>r<sub>it</sub></i>	<i>M</i>	<i>SD</i>	<i>P</i>	<i>r<sub>it</sub></i>
NEPS-MAT-1	2.9	1.4	.49	.91	3.6	1.4	.61	.91
NEPS-MAT-2	2.5	1.4	.42	.91	3.5	1.3	.59	.90
NEPS-MAT	5.3	2.5	.45	-	7.1	2.4	.60	-
Gesamt								

Anm.: *M* = Mittelwert, *SD* = Standardabweichung, *P* = Schwierigkeit, *rit* = Trennschärfe der Blöcke

Tabelle 4b: Deskriptive Statistiken für NEPS-MAT für Fünftklässler/-innen bis Erwachsene

	5. Klasse plus				5. Klasse				9. Klasse				Erwachsene			
	<i>N</i> = 222				<i>n</i> = 60				<i>n</i> = 92				<i>n</i> = 70			
	<i>M</i>	<i>SD</i>	<i>P</i>	<i>r<sub>it</sub></i>	<i>M</i>	<i>SD</i>	<i>P</i>	<i>r<sub>it</sub></i>	<i>M</i>	<i>SD</i>	<i>P</i>	<i>r<sub>it</sub></i>	<i>M</i>	<i>SD</i>	<i>P</i>	<i>r<sub>it</sub></i>
NEPS-MAT-1	2.6	1.1	.64	.86	1.7	1.1	.43	.78	2.6	1.0	.66	.78	3.1	0.9	.79	.78
NEPS-MAT-2	2.6	1.0	.65	.89	2.0	1.0	.51	.65	2.7	0.9	.67	.69	3.0	1.0	.74	.81
NEPS-MAT-3	2.7	1.1	.68	.91	2.2	1.2	.55	.79	2.8	1.0	.70	.79	3.0	1.0	.76	.83
Gesamtwert	7.8	2.6	.65	-	5.9	2.4	.49	-	8.1	2.2	.68	-	9.1	2.3	.76	-

Anm.: *M* = Mittelwert, *SD* = Standardabweichung, *P* = Schwierigkeit, *rit* = Trennschärfe der Blöcke

## 4.2 Reliabilität

Im Folgenden wird die Reliabilität der beiden Instrumente im Hinblick auf deren interne Konsistenz, die Testhalbierungsreliabilität sowie die Retest-Reliabilität dargestellt.

### 4.2.1 Interne Konsistenz

**NEPS-BZT.** Die innere Konsistenz für den geschwindigkeitsbezogenen Test beruht auf dem Vergleich der Anzahl richtig bearbeiteter Bild-Zeichen-Elemente in jedem der drei (oder zwei) 30 (bzw. 45) Sekunden-Intervalle (= Zeitblöcke). Für den NEPS-BZT ergaben sich in den einzelnen Altersgruppen somit interne Konsistenzen zwischen den jeweils zwei bzw. drei Zeitblöcken von jeweils 30 bzw. 45 Sekunden Bearbeitungszeit, die im Bereich von  $\alpha = .64$  bis  $.90$  (siehe Tabelle 5) liegen.

**NEPS-MAT.** Für den NEPS-MAT wurden in den einzelnen Altersgruppen interne Konsistenzen zwischen  $\alpha = .59$  und  $.90$  ermittelt (Tabelle 5). Bei heterogenen Tests wie dem NEPS-MAT, der teilweise Items stark unterschiedlicher Schwierigkeit beinhaltet, unterschätzt Cronbachs Alpha tendenziell deren Reliabilität.

Tabelle 5: Interne Konsistenzen des NEPS-MAT nach Altersgruppen

Altersgruppe	Interne Konsistenz (Cronbachs $\alpha$ )	
	NEPS-BZT	NEPS-MAT
Vorschüler/-innen	.84	.70
2. Klasse	.64	.69
5. Klasse	.75	.60
9. Klasse	.76	.71
Erwachsene	.90	.59

Anm.: Standardisierte Alpha-Werte. Alpha beruht auf der Konsistenz der jeweiligen Anzahl richtig bearbeiteter Bild-Zeichen-Elemente je 30 (bzw. 45) Sekunden-Zeitblock

#### 4.2.2 Testhalbierungsreliabilität

In den Altersgruppen der Vorschüler/-innen und der Zweitklässler/-innen können die beiden Aufgabenblöcke des NEPS-MAT wie auch des NEPS-BZT direkt als Testhälften aufgefasst und korreliert werden. Wie bei der Ermittlung der Testhalbierungsreliabilität üblich, werden die mit der Spearman-Brown-Formel korrigierten Koeffizienten berichtet. Da in Vorschule und 2. Klasse nur jeweils zwei Zeit- bzw. Aufgabenblöcke betrachtet wurden, sind die hier berichteten Koeffizienten bedeutungsgleich (aber nicht identisch) mit den in Tabelle 5 berichteten Alpha-Werten.

Analog wurde mit den Testversionen für Schüler/-innen der 5. und 9. Klasse und Erwachsene verfahren. Da hier eine Unterteilung in jeweils drei statt zwei Aufgabenblöcke vorgenommen wurde, ergeben sich getrennte Korrelationen für die drei möglichen Blockkombinationen: Block 1 mit Block 2, Block 1 mit Block 3 und Block 2 mit Block 3. Hier stellt jeder Block ein Testdrittel dar, daher wird bei der Spearman-Brown-Korrektur der Faktor  $k=3$  eingesetzt. Zusätzlich werden für die Altersgruppen von der 5. Klasse bis ins Erwachsenenalter Gesamtwerte berichtet, da hier dieselbe Testversion zum Einsatz kam. Alle genannten Ergebnisse zur Testhalbierungsreliabilität des NEPS-BZT finden sich in Tabelle 6 und die Ergebnisse für den NEPS-MAT in Tabelle 7.

Tabelle 6: Testhalbierungsreliabilitäten des NEPS-BZT nach Altersgruppen

	NEPS-BZT				
	Vorschule	2. Klasse	5. Klasse	9. Klasse	Erwachsene
Block1 – Block2	.84**	.69**	.70**	.66**	.92**
Block1 – Block3	-	-	.74**	.79**	.90**
Block2 – Block3	-	-	.89**	.81**	.88**

Anm.: \*\* p < .001; berichtet werden nach der Spearman-Brown korrigierte Pearson-Korrelationen

Tabelle 7: Testhalbierungsreliabilitäten des NEPS-MAT nach Altersgruppen

	NEPS-MAT				
	Vorschule	2. Klasse	5. Klasse	9. Klasse	Erwachsene
Block1 – Block2	.78**	.77**	.68**	.56**	.73**
Block1 – Block3	-	-	.69**	.70**	.70**
Block2 – Block3	-	-	.50**	.59**	.74**

Anm.: \*\* p < .001; berichtet werden nach der Spearman-Brown korrigierte Pearson-Korrelationen

#### 4.2.3 Retest-Reliabilität

Mit insgesamt 198 Probanden aus unterschiedlichen Altersgruppen wurden Retest-Reliabilitäten ermittelt. Dazu wurde eine verkürzte Testbatterie mit NEPS-BZT, NEPS-MAT, KO, ZST-B und RAVEN-Matrizen (wiederum SPM-C, Set A für Vorschule und SPM-C, Set C für 2. Klasse) bei einer Teilstichprobe von  $n = 94$  Zweitklässler/-innen in einem Zeitabstand von acht Tagen ( $M = 7.7$ ,  $SD = 2.3$ ) und bei einer Teilstichprobe von  $n = 58$  Vorschüler/-innen im Abstand von zehn Tagen ( $M = 10.8$ ,  $SD = 0.8$ ) erhoben. Außerdem wurde für eine Teilstichprobe von  $n = 46$  Erwachsenen in einem Zeitabstand von knapp sechs Wochen ( $M = 40.6$  Tage,  $SD = 7.0$ ) eine zweite Erhebung der gesamten Testbatterie durchgeführt.

Die Resultate sind in Tabelle 8 zusammenfassend dargestellt. Während die Ergebnisse für Vorschule und 2. Klasse bis auf die eingesetzten RAVEN-Matrizen (Set A bei Vorschüler/-innen, Set C bei Zweitklässler/-innen) unmittelbar vergleichbar sind, ist der direkte Vergleich mit den Stichproben der Fünft- und Neuntklässler/-innen sowie Erwachsenen nicht möglich. Insgesamt liegen die Ergebnisse für alle Altersgruppen im erwarteten Bereich für derartige Kurztests.

Tabelle 8: Retest-Reliabilität des NEPS-BZT und des NEPS-MAT und weiterer Tests

$r_{tt}$	NEPS- BZT	NEPS- MAT	<u>Konvergente Tests</u>				<u>Diskriminante Tests</u>				
			KO	ZST	RAVEN A	RAVEN C	LPS	MWT	IST-V	IST-F	
Vorschüler/- innen	.68	.77	.73		.50			-	-	-	-
2. Klasse	.61	.80	.50				.77	-	-	-	-
Erwachsene	.76	.68		.83		.76	.79	.77	.67	.58	

Anm.: Alle dargestellten Koeffizienten sind  $p < .01$ . Aufgrund zu geringer Stichprobengrößen muss auf eine separate Darstellung der Retest-Reliabilitäten für die 5. + 9. Klasse verzichtet. KO = Symbole Kodieren (HAWIVA-III); ZST = Zahlen-Symbol-Test B (HAWIK-IV bzw. HAWIE); RAVEN = Standard Progressive Matrices, Set A bzw. C; MWT = MWT-B; LPS = Wortflüssigkeit (L-P-S); IST-V = Merkfähigkeit Verbal (I-S-T 2000 R); IST-F = Merkfähigkeit Figural (I-S-T 2000 R)

### 4.3 Konvergente und divergente Validität

*Vorschule und 2. Klasse.* In Tabelle 9 sind die Korrelationen der beiden neuen Leistungstests für die NEPS-Altersgruppen 1 und 2 mit konvergenten und diskriminanten Außenkriterien dargestellt. Ein konvergentes Außenkriterium des Bilder-Zeichen-Tests war der Untertest „Symbole Kodieren“ des HAWIVA-III (Petermann & Petermann, 2008). Konvergentes Außenkriterium des NEPS-MAT war das Set A der SPM-C (Raven, 1977) bei den Vorschüler/-innen, bzw. Set C der SPM-C (Raven, 1977) bei den Zweitklässler/-innen. Zwischen beiden neuen Leistungstests werden moderate Korrelationen erwartet.

Für die diskriminante Validität wurden für die Vorschüler/-innen die ersten zwölf Aufgaben der „Bildkonzepte“ des HAWIK-IV und der Untertest 3 des CFT-1 eingesetzt. Bei den Zweitklässler/-innen kam der volle Satz von 26 Items der „Bildkonzepte“ zum Einsatz sowie Subtest 3 des CFT-1 und ergänzend der Wortflüssigkeitstest aus dem L-P-S. Die jeweils erzielten Werte für konvergente und divergente Korrelationen bewegen sich insgesamt im erwarteten Bereich (siehe Tabelle 9).

Tabelle 9: Konvergente und divergente Validitäten des NEPS-BZT und des NEPS-MAT

	NEPS-BZT	NEPS-MAT	Konvergente Tests			Diskriminante Tests		
			KO	RAVEN A	RAVEN C	BK	CFT	LPS
<b>NEPS-BZT</b>								
Vorschule		.37**	.65**	.20†	-	.19†	.11	-
2. Klasse		-.08	.57**		-.04	.17	.12	.26*
<b>NEPS-MAT</b>								
Vorschule	.37**		.31**	.43**	-	.28*	.33**	-
2. Klasse	-.08		.00		.57**	.33**	.35**	-.26*

Anm.: †  $p < .05$ , \*  $p < .01$ , \*\*  $p < .001$ ; KO = Symbole Kodieren (HAWIVA-III); RAVEN A = Standard Progressive Matrices, Set A; RAVEN C = Standard Progressive Matrices, Set C; BK = Bildkonzepte (HAWIK-IV); CFT = Klassifikationen (CFT-1); LPS = Wortflüssigkeit (L-P-S)

5. Klasse bis Erwachsenenalter. Tabelle 10 gibt einen Überblick über Korrelationen der beiden neuen Leistungstests mit konvergenten und diskriminanten Konstrukten für die Gesamtstichprobe aller Probanden ab der 5. Klasse sowie getrennt für die entsprechenden Altersgruppen. Erneut gilt, dass sich konvergente und diskriminante Konstrukte je nach betrachtetem NEPS-Test unterscheiden (vgl. oben).

Konvergentes Außenkriterium des NEPS-BZT war der Zahlen-Symbol-Test B des HAWIK-IV (Petermann & Petermann, 2008) bzw. des WIE (von Aster et al., 2006). Konvergentes Außenkriterium des neu entwickelten NEPS-MAT war das Set C der SPM-C von Raven (1977). Zwischen beiden neuen Leistungstests wurden erwartungsgemäß moderate Korrelationen beobachtet.

Für beide Tests wurden schwache bis moderate Korrelationen mit den diskriminanten Außenkriterien der kristallinen kognitiven Leistung erwartet. Eingesetzt wurden hierfür der Mehrfachwahl-Wortschatz-Intelligenz-Test MWT-B (Lehrl, 1977), der Wortflüssigkeitstest des Leistungsprüfsystems (L-P-S, Horn, 1983, Untertest 6), sowie die beiden Merkfähigkeitstests des IST-2000-R (Amthauer et al., 2001). Es zeigt sich, dass die konvergenten und divergenten Validitäten des NEPS-BZT und des NEPS-MAT insgesamt stimmig und erwartungskongruent sind.

Tabelle 10: Konvergente und divergente Validitäten des NEPS-BZT und des NEPS-MAT

	NEPS- BZT	NEPS- MAT	Konvergente Tests			Diskriminante Tests		
			ZST	RAVEN C	MWT	LPS	IST-V	IST-F
<b>NEPS-BZT</b>								
5. Klasse		.46**	.43**	.36**	-.05	-.08	.11	.08
9. Klasse		.02	.65**	.09	-.17	.02	.13	.05
Erwachsene		.23†	.42**	.29*	.29*	.14	.31*	.41**
<b>NEPS-MAT</b>								
5. Klasse	.46**		.21	.44**	.22†	-.02	.05	.19
9. Klasse	.02		.12	.58**	.28**	.34**	.28**	.43**
Erwachsene	.23†		.11	.61**	.06	.25†	-.06	.18

Anm.: †  $p < .05$ , \*  $p < .01$ , \*\*  $p < .001$ ; ZST = Zahlen-Symbol-Test B (HAWIK-IV bzw. HAWIE); RAVEN C = Standard Progressive Matrices, Set C; MWT = MWT-B; LPS = Wortflüssigkeit (L-P-S); IST-V = Merkfähigkeit Verbal (I-S-T 2000 R); IST-F = Merkfähigkeit Figural (I-S-T 2000 R)

Allerdings sind substanzielle Altersunterschiede in den konvergenten und diskriminanten Außenkorrelationen beobachtbar. Die Unterschiede in den Ergebnissen der Validierung zwischen den Altersgruppen dürften zum Teil auf Besonderheiten der verglichenen Stichproben zurückzuführen sein. Im Rahmen des Projekts zur Entwicklung der beiden Kurztests war keine Erhebung mit vollständig heterogenen oder gar repräsentativen Stichproben möglich. Wie weiter unten dargestellt, spricht die durchaus befriedigende Messäquivalenz über die Altersgruppen hinweg dafür, dass die altersdifferentiellen Außenkorrelationen entweder rein stichprobenspezifisch sind (z. B. Vergleich von Haupt- und Realschüler/-innen mit Gymnasialabschluss bei den Erwachsenen) oder aber auf Probleme der Äquivalenz der Außenkriterien zurückzuführen sind.

#### 4.4 Darstellung der Messäquivalenz

Um die Äquivalenz der Testverfahren über verschiedene Altersgruppen vergleichen zu können, wurde die psychometrische Messinvarianz über die verschiedenen Altersstichproben berechnet. Hierzu wurden in einem ersten Schritt konfirmatorische Faktorenanalysen für jede Altersstichprobe berechnet, um die Validität des Messmodells in jeder Gruppe sicherzustellen. In einem zweiten Schritt wurden Multigruppenmodelle mit zunehmenden psychometrischen Restriktionen berechnet. Besonders bedeutsam ist hierbei die Frage, inwiefern Wahrnehmungsgeschwindigkeit und schlussfolgerndes Denken durch die entsprechenden Aufgaben äquivalent über verschiedene Altersgruppen erfasst werden.

##### 4.1.1 Einzelgruppenanalysen

Das zweifaktorielle Modell beinhaltet die korrelierten NEPS-BZT und NEPS-MAT Faktoren, die durch die entsprechenden Testdurchgänge auf ihren Skalen erfasst werden. Diese umfassen in Vorschule und 2. Klasse zwei Durchgänge und somit zwei Indikatoren, ab 5. Klasse aufwärts entsprechend drei Indikatoren. Die Kovarianzmatrix wurde mit Mplus 6.1 und einem robusten Maximum-Likelihood Schätzer analysiert, um für Abweichungen von der Normalverteilungsannahme zu korrigieren (Muthén & Muthén, 1998-2011). Die Metrik der Faktoren wurde durch die Fixierung der jeweilig ersten Faktorladung gesetzt. Aufgrund eines

Konvergenzprobleme war es erforderlich, die Fehlervarianz des Items BZT2 in der 2. Klasse auf Null zu setzen. Dies ist bei der Interpretation der Modellparameter zu beachten. Zur Beurteilung der Anpassungsgüte der Modelle wurden neben dem Chi-Quadrat Test ( $\chi^2$ ) folgende Indizes mit entsprechend konservativen Kriterien nach Hu und Bentler (1999) betrachtet: Standardized Root Mean Square Residual (SRMR;  $\leq .08$ ), Root mean Square Error of Approximation (RMSEA;  $\leq .08$ ), Comparative Fit Index (CFI;  $\geq .95$ ) sowie der Tucker-Lewis Index (TLI;  $\geq .95$ ).

Tabelle 11 zeigt die Fit-Indizes für die Modelle in den verschiedenen Altersgruppen. Der  $\chi^2$ -Test fällt für alle Teilstichproben nicht signifikant aus und impliziert eine exakte Modellanpassung über die verschiedenen Altersgruppen. Die approximativen Fit-Indizes entsprechen den Empfehlungen von Hu und Bentler (1999) und zeigen eine hohe Modellgüte. Eine Ausnahme bildet die 5. Klasse, in der die Cut-off-Werte für RMSEA, CFI und TLI knapp verfehlt werden, aber aufgrund der in dieser Gruppe besonders geringen Stichprobengröße dennoch als akzeptabel einzustufen sind. Die Abweichung der Freiheitsgrade ist auf die bereits erläuterte abweichende Spezifikation der Modelle zurückzuführen.

*Tabelle 11: Fit-Indizes für jede Testgruppe*

	$\chi^2$	<i>p</i>	<i>df</i>	RMSEA	SRMR	CFI	TLI
Vorschule ( <i>n</i> = 88)	.24	.623	1	.00	.004	1.00	1.05
2. Klasse ( <i>n</i> = 99)	1.09	.581	2	.00	.034	1.00	1.04
5. Klasse ( <i>n</i> = 60)	12.86	.117	8	.101	.072	.943	.893
9. Klasse ( <i>n</i> = 92)	6.61	.579	8	.00	.037	1.00	1.03
Erwachsene ( <i>n</i> = 70)	8.65	.373	8	.034	.031	.997	.994

Anm.:  $\chi^2$  = Chi-Quadrat, *p* = Signifikanz, *df* = Freiheitsgrade, RMSEA = Root Mean Square Error of Approximation, SRMR = Standardized Root mean Square Residual, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index

Die standardisierten Faktorladungen, Residuen und Faktorinterkorrelationen sind in Tabelle 12 dargestellt. Die Indikatoren zeigen in jeder Altersgruppe die erwarteten salienten Ladungen auf ihren latenten Faktoren. Es ist zu beachten, dass die bereits erläuterte Nullsetzung der Residualvarianz für BZT2 in der Gruppe der Zweitklässler/-innen zu einer perfekten Ladung von 1.00 führt. Die Faktorinterkorrelationen auf der latenten Ebene entsprechen in ihrer Richtung den in Abschnitt 4.3 dargestellten konvergenten Validitäten der Summenwerte, wobei auch hier die nicht erwarteten Korrelationen in den Gruppen der Zweit- und Neuntklässler/-innen zu beobachten sind.

Die Einzelgruppenanalysen zeigen, dass die spezifizierten Faktorenstrukturen der NEPS-BZT und NEPS-MAT Instrumente gut zu den beobachteten Daten passen und die Indikatoren ihre latenten Faktoren differenziert erfassen. In einem nächsten Schritt wird dargestellt, in welchen strukturellen Aspekten sich die einzelnen Altersstichproben unterscheiden.

Tabelle 12: Standardisierte CFA Parameter für jede Testgruppe

	BZT1	BZT2	BZT3	MAT1	MAT2	MAT3	$r(\text{BZT}, \text{MAT})$
Vorschule ( $n = 88$ )							.46
Ladung	.86	.85	–	.88	.73	–	
Residuum	.26	.28	–	.23	.47	–	
2. Klasse ( $n = 99$ )							-.20
Ladung	.52	1.00	–	.75	.83	–	
Residuum	.73	.00	–	.44	.31	–	
5. Klasse ( $n = 60$ )							.67
Ladung	.57	.79	.95	.57	.51	.62	
Residuum	.68	.38	.10	.67	.74	.62	
9. Klasse ( $n = 92$ )							.06
Ladung	.61	.65	.90	.64	.48	.68	
Residuum	.63	.57	.19	.60	.77	.54	
Erwachsene ( $n = 70$ )							.66
Ladung	.86	.93	.80	.63	.68	.73	
Residuum	.27	.13	.36	.60	.54	.47	

### 4.3.3 Multigruppenanalysen

Das in 4.4.1 spezifizierte Modell wurde in Multigruppenanalysen simultan für die dargestellten Altersgruppen geschätzt, um verschiedene Arten der Messinvarianz zu prüfen. Konfigurale Invarianz liegt vor, wenn die Modellparameter frei über verschiedene Altersgruppen variieren und bezieht sich auf die Frage, ob von einer identischen Faktorenstruktur in den verschiedenen Altersgruppen ausgegangen werden kann. Bei metrischer Invarianz müssen zusätzlich die Faktorladungen über verschiedene Altersgruppen invariant gesetzt werden. Hierbei geht es um den Nachweis, dass die Testaufgaben ihre latenten Faktoren über verschiedene Altersgruppen gleich gut erfassen. Skalare Invarianz liegt vor, wenn zusätzlich zu den Bedingungen der metrischen Invarianz die Intercepts der Indikatoren über die Gruppen invariant sind. Diese Restriktion wurde aus Gründen der Vollständigkeit aufgenommen und ist für den Nachweis der psychometrischen Qualität der Testinstrumente nicht erforderlich, da durchaus von Mittelwertunterschieden in den verschiedenen Gruppen auszugehen ist. Um die Vergleichbarkeit zwischen den Altersgruppen zu gewährleisten, wurden nur die ersten beiden Testdurchgänge in den Analysen berücksichtigt. Hierbei wurde für die Modellidentifikation die erste Faktorladung freigesetzt und die Varianz des latenten BZT Faktors auf eins gesetzt. Invarianz liegt vor, wenn sich die Modelle nicht signifikant voneinander unterscheiden. Hierzu wurden die Modelle mittels eines Satorra-Bentler Chi-Quadrat Differenztests mit Korrektur für robuste Maximum-Likelihood Schätzer verglichen (Satorra & Bentler, 2001).

Die Ergebnisse der Multigruppenanalysen sind in Tabelle 13 dargestellt. Das Basismodell der konfiguralen Invarianz weist eine hohe Modellgüte auf und zeigt, dass allen Altersgruppen eine äquivalente Faktorenstruktur zugrunde liegt. Die weitere Festsetzung der Faktorladungen über alle Altersgruppen führt mit  $\chi^2_{diff}(8) = 13.78$ , *ns* zu keiner signifikanten Verschlechterung der Modellgüte, was auf das Vorliegen metrischer Invarianz hinweist. Die noch restriktivere Annahme skalarer Invarianz führt zu einer deutlichen Verschlechterung der Indizes und weist darauf hin, dass sich die Punktwerte im Mittel in den Altersgruppen unterscheiden.

Insgesamt weisen die Ergebnisse auf die psychometrische Äquivalenz der Testinstrumente auf der Ebene der Faktorenstruktur und der Indikatoren hin. Die Instrumente sind somit sinnvoll in verschiedenen Altersbereichen einzusetzen.

*Tabelle 13: Messinvarianz über alle Altersgruppen (N = 409)*

	$\chi^2$	<i>P</i>	<i>df</i>	$\chi^2_{diff}$	$\Delta df$	<i>RMSEA</i>	<i>CFI</i>	<i>TLI</i>
Konfigurale Invarianz	9.27	.413	9	–	–	.02	1.00	1.00
Metrische Invarianz	23.75	.127	17	14.48	8	.07	.98	.96
Skalare Invarianz	145.15	.000	25	121.4	8	.24	.61	.53

Anm.:  $\chi^2$  = Chi-Quadrat, *df* = Freiheitsgrade,  $\chi^2_{diff}$  = Chi-Quadrat Differenz,  $\Delta df$  = Veränderung Freiheitsgrade, *RMSEA* = Root Mean Square Error of Approximation, *CFI* = Comparative Fit Index, *TLI* = Tucker-Lewis Index

## 4.5 Durchführungs- und Auswertungsobjektivität

Zur Objektivität der erstellten Tests lässt sich sagen, dass aufgrund umfassender Dokumentation der Instruktionen in enger Abstimmung mit den inhaltlichen und organisatorischen Ansprechpartnern des NEPS bzw. dem IEA Data Processing and Research Center (IEA) im Zusammenspiel mit den speziell für den Einsatzzweck im NEPS aufbereiteten Testheften einer standardisierten Durchführung der Tests nichts im Wege steht.

Für die Eingabe und Auswertung der vorgelegten Tests sind einfache Regeln (NEPS-BZT) bzw. Lösungsschlüssel (NEPS-MAT) gegeben, die Fehler unwahrscheinlich machen. Lediglich im NEPS-BZT für Vorschulkinder (eingeschränkt auf die 2. Klasse) können vereinzelt durch sehr undeutlich eingezeichnete Lösungen verschiedene Rater zu unterschiedlichen Ergebnissen kommen. Aufgrund der standardisierten Testleiter/-innentrainings sollte dies jedoch zu vernachlässigen sein. Aufgrund der einfachen schriftlichen Antwortformate, zumindest für die Altersgruppen ab der 5. Klasse bis zum Erwachsenenalter (direktes Eintragen von Zahlen im NEPS-BZT bzw. Ankreuzen einer Lösungsmöglichkeit im NEPS-MAT), ist ein computerisiertes Auslesen der Daten gut möglich, ohne Einschränkungen der Reliabilität und Validität zu erwarten. Über die Protokollbögen für die Vorschüler/-innen und Grundschüler/-innen, die die jeweiligen Testleiter/-innen ausfüllen, ist bei Erhebung mit Computern ein verbessertes objektives Einlesen der Daten gewährleistet. Da die Testergebnisse jeweils als metrische Skalen- bzw. Gesamtwerte ausgegeben werden, die als grober Schätzer fluider kognitiver Potenziale zu verstehen sind, ist auch die Interpretationsobjektivität gewährleistet.

## 5. Diskussion und Fazit

Die beiden neu entwickelten Tests zur Erfassung der Wahrnehmungsgeschwindigkeit (WG) und des schlussfolgernden Denkens (SD) stellen Kurztests dar, die Indikatoren der basalen fluiden Leistungskapazitäten der untersuchten Personen im Altersbereich von fünf bis 70 Jahren liefern. Bei Kurztests müssen jedoch gegenüber den in der Literatur üblichen differenzierten Erfassungen strukturell bedingte Einschränkungen in Kauf genommen werden. Vor diesem Hintergrund sind die hier dargestellten Ergebnisse als zufriedenstellend zu bezeichnen. Die entwickelten Tests erlauben eine grobe, reliable und in Grenzen valide Erfassung der basalen fluiden Leistungskapazitäten in einem heterogenen Altersbereich von fünf bis 70 Jahren und älter. Die durchgeführten Validierungsstudien belegen, dass die beiden

ultrakurzen Tests eine – wenn auch nur sehr grobe – für die korrelative Forschung jedoch zweckmäßige Einschätzung der fluiden Kognitionspotenziale ermöglichen. Im Folgenden diskutieren wir die Befunde zur Objektivität, Reliabilität und Validität der Ultrakurztests und geben ein abschließendes Fazit zu deren Einsatzfähigkeit im Rahmen von Survey- und Panelstudien der Bildungsforschung.

*Objektivität im Hinblick auf Durchführung und Auswertung.* Beide Tests sind in ihrer Anwendung für Testleiter/-innen leicht erlernbar und erweisen sich als weitgehend robust gegen Variationen der Durchführungssituation. Testleiter/-innentraining ist dennoch erforderlich, vor allem im Hinblick auf die korrekte Einhaltung der Zeitgrenzen sowie im Hinblick auf die Einhaltung einer standardisierten Testsituation. Die Auswertung der Tests erfolgt standardisiert und ist daher kaum fehleranfällig.

*Reliabilität.* Die für die verschiedenen Testversionen erzielten Werte der internen Konsistenz (im Bereich von  $.59 \leq \alpha \leq .90$ ) entsprechen den üblichen und erwartbaren Werten für vergleichbare heterogene Leistungstests mit ansteigender Schwierigkeit. Auch die Retest-Reliabilitäten bestätigen weitgehend die Erwartungen. Einschränkungen gibt es hier allerdings im Hinblick auf die ungleiche Varianzverteilung über die Altersgruppen hinweg. Allerdings führen diese stichprobenspezifischen Störungen eher zu einer Unterschätzung der tatsächlich zu erwartenden Retest-Reliabilität der Tests.

*Validität.* Im Hinblick auf konvergente und diskriminante Außenkriterien zeigen die beiden neu entwickelten Tests ein erwartetes Muster und durchgehend höhere Korrelationen mit konstruktgleichen Tests als mit divergenten Tests. Dieses Muster konnte strukturäquivalent in allen einbezogenen Altersgruppen gefunden werden.

In der Zusammenschau der Befunde bleibt festzustellen, dass die beiden neu entwickelten Testverfahren NEPS-BZT und NEPS-MAT als Ultrakurzinstrumente zur Erfassung der fluiden Leistungspotenziale im Vorschul-, Schul- und Erwachsenenalter geeignet sind. Die vorgelegten Testverfahren NEPS-BZT und NEPS-MAT erweisen sich als tauglich für die Survey- und Panelforschung, sofern nur wenig Zeit bleibt für eine gründlichere und umfassendere Erhebung kognitiver Leistungsparameter. Die hier vorgestellten ultrakurzen kognitiven Tests erlauben dabei keine Rückschlüsse auf die individuellen Intelligenz- oder Kognitionspotenziale einzelner Individuen im Sinne der Einzeldiagnostik. Allerdings ist festzuhalten, dass die ultrakurzen kognitiven Tests des NEPS-BZT und NEPS-MAT aufgrund der Kriterien der Objektivität, Sparsamkeit, Reliabilität und Validität durchaus für Forschungskontexte empfohlen werden, in denen beispielsweise kognitiv-fluide Basispotenziale der Individuen als Randvariable etwa zur Differenzierung von Bildungs- und Kontexteffekten in der kognitiven Entwicklung betrachtet werden sollen. Die ultrakurzen Tests erlauben dabei allerdings nur eine grobe Schätzung der auf individuelle Unterschiede zurückführbaren Varianz in der fluiden kognitiven Leistung.

Zusammenfassend ist festzustellen, dass die entwickelten Tests geeignet erscheinen für den Einsatz im Rahmen der vergleichenden, korrelativen Surveyforschung und Panelstudien, sofern eine grobe, ultrakurze Erfassung der fluiden kognitiven Leistungsfähigkeit gewünscht wird.

## 6. Literatur

- Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.
- Brunner, M., Lang, F. R. & Lüdtke, O. (2014). Erfassung der fluiden kognitiven Leistungsfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Expertise (NEPS Working Paper No. 42). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.
- Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the RAVEN Progressive Matrices Test. *Psychological Review*, 97(3), 404-431.
- Cattell, R. B., Weiß, R. H. & Osterland, J. (1997). *Grundintelligenztest Skala 1 (CFT 1)* (5., revidierte Auflage). Göttingen: Hogrefe.
- Horn, W. (1983). *Leistungsprüfsystem (L-P-S)*. Göttingen: Hogrefe.
- Hu, L. T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Jacobs, P. I. & Vandeventer, M. (1972). Evaluating the teaching of intelligence. *Educational and Psychological Measurement*, 32, 235-248.
- Lang, F. R. (2005). Erfassung des kognitiven Leistungspotenzials und der "Big Five" mit Computer-Assisted-Personal-Interviewing (CAPI): Zur Reliabilität und Validität zweier ultrakurzer Tests und des BFI. Deutsches Institut für Wirtschaftsforschung (DIW Berlin), Research Notes 9.
- Lang, F. R., Weiss, D., Stocker, A. & Rosenblatt, B. v. (2007). Assessing cognitive capacities in computer-assisted survey research: Two ultra-short tests of intellectual ability in the German Socio-Economic Panel (SOEP). *Schmollers Jahrbuch. Journal of Applied Social Science Studies*, 127, 183-192.
- Lehrl, S. (1977). *Mehrfachwahl-Wortschatz-Intelligenztest (MWT-B)*. Erlangen: Verlag Dr. med. D. Straube.
- Muthén, L. K. & Muthén, B. O. (1998-2011). *MPlus 6.1* [computer software]. Los Angeles, CA: Muthén & Muthén. <http://statmodel.com>
- Petermann, F. & Petermann U. (Hrsg.). (2008). *Hamburg-Wechsler-Intelligenztest für Kinder – IV (HAWIK-IV)* (2., ergänzte Auflage). Bern: Huber.
- Primi, R. (2001). Complexity of geometric inductive reasoning tasks. Contribution to the understanding of fluid intelligence. *Intelligence*, 30, 41-70.
- Raven, J. C. (1977). *Standard Progressive Matrices: Sets A, B, C, D & E*. San Antonio, TX: Harcourt.
- Ricken, G., Fritz, A., Schuck, K. D. & Preuß, U. (Hrsg.). (2007). *Hannover-Wechsler-Intelligenztest für das Vorschulalter – III (HAWIVA-III)*. Bern: Huber.
- Satorra, A. & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66, 507-514.

- Smith, A. (1995). Symbol Digit Modalities Test (SDMT) (7., unveränderte Auflage). Los Angeles, CA: Western Psychological Services.
- von Aster, M., Neubauer, A. & Horn, R. (Hrsg.). (2006). Wechsler Intelligenztest für Erwachsene (WIE). Frankfurt am Main: Pearson Assessment.