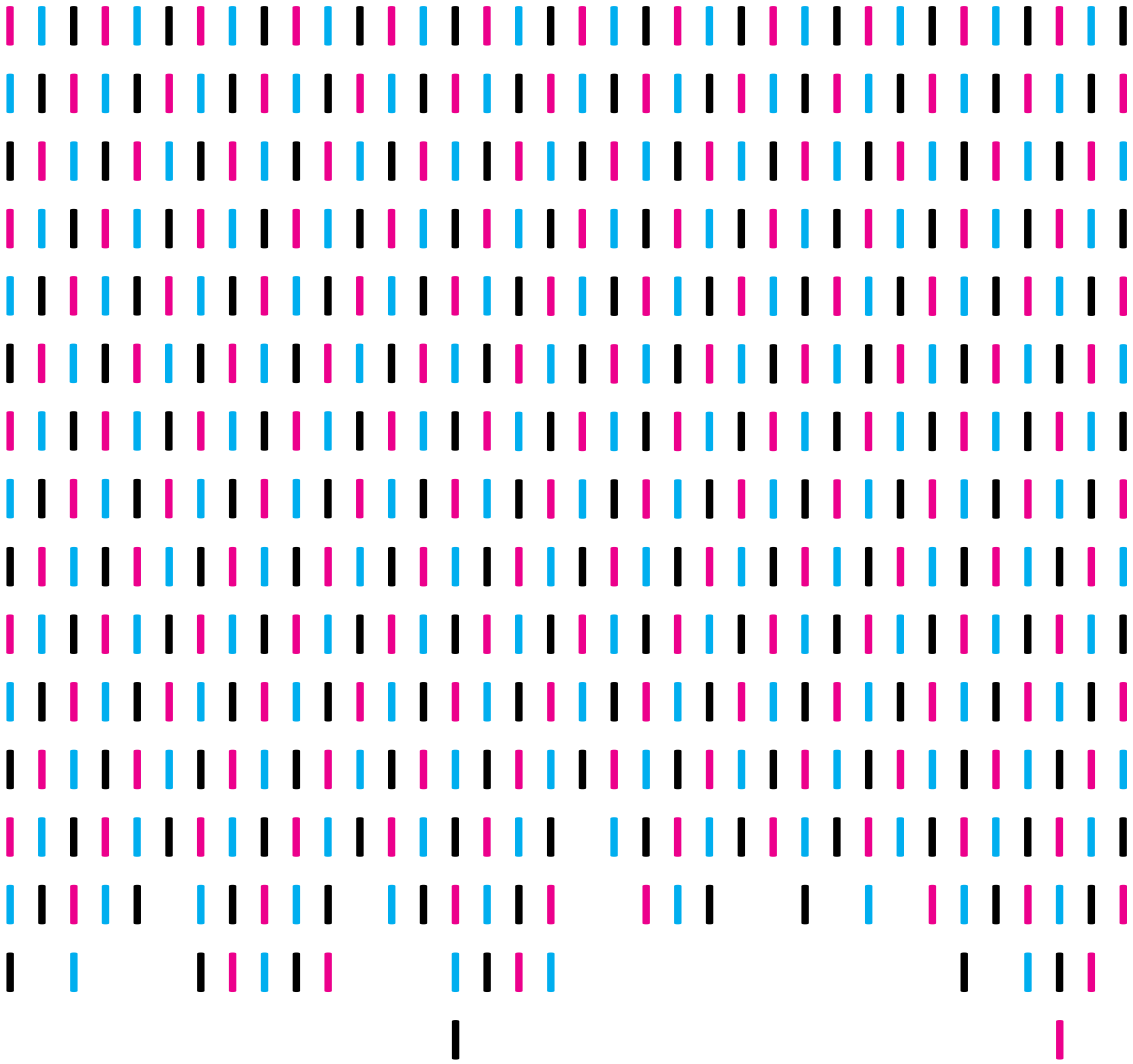


NEPS WORKING PAPERS

Martin Brunner, Frieder R. Lang und Oliver Lüdtke

ERFASSUNG DER FLUIDEN KOGNITIVEN LEISTUNGSFÄHIGKEIT ÜBER DIE LEBENSSPANNE IM RAHMEN DER NATIONAL EDUCATIONAL PANEL STUDY: EXPERTISE

NEPS Working Paper No. 42
Bamberg, Juni 2014

LIfBi
LEIBNIZ-INSTITUT FÜR BILDUNGSVERLÄUFE e.V.

NEPS
National Educational Panel Study

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

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

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

Martin Brunner, Freie Universität Berlin und Institut für Schulqualität der Länder Berlin und Brandenburg

Frieder R. Lang, Friedrich-Alexander-Universität Erlangen-Nürnberg

Oliver Lüdtke, IPN - Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik

Expertise[1]

E-Mail-Adresse des Erstautors:

Martin.Brunner@isq-bb.de

Bibliographische Angaben:

Brunner, M., Lang, F. R. & Lüdtke, O. (2014). Erfassung der fluiden kognitiven Leistungfähigkeit über die Lebensspanne im Rahmen der National Educational Panel Study: Expertise (NEPS Working Paper No. 42). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

[1] Die vorliegende Expertise wurde 2009 für das Nationale Bildungspanel (NEPS) verfasst.

Inhalt

Zusammenfassung

Die Expertise behandelt die Möglichkeiten und Grenzen einer zeitökonomischen Erfassung fluider, non-verbaler Intelligenzleistungen[2] Gf bei Menschen zwischen 5 und 67 Jahren. Als Indikatoren von Gf schlagen wir Messinstrumente zur Erfassung der Wahrnehmungsgeschwindigkeit WG und zum schlussfolgernden Denken SF vor: Diese Indikatoren liegen im konzeptionellen Zentrum von Gf und es wurde wiederholt empirisch gezeigt, dass WG und SF einen bedeutsamen Beitrag für eine erfolgreiche Entwicklung leisten. Zur Evaluation bestehender Instrumente dieser kognitiven Kompetenzen haben wir die psychometrischen Kriterien Validität, Reliabilität und Messinvarianz herangezogen. Wir haben uns dabei auf Instrumente zur Erfassung von Gf konzentriert, die am häufigsten im deutschen Sprachraum in der Individualdiagnostik eingesetzt werden oder sich in large-scale-Studien sowie bedeutsamen Längsschnittstudien bewährt haben. Auf Grundlage dieser Evaluation sprechen wir zwei Empfehlungen aus: (1) Bei Verwendung von Tests aus der Wechslertest-Familie wird ein national und international sehr bewährtes Instrument der Individualdiagnostik gewählt. Es ist hierbei keine Neuentwicklung notwendig, jedoch werden relativ hohe Lizenzgebühren für die Verwendung der Wechslertests in NEPS anfallen. (2) Die Verwendung und Weiterentwicklung bestehender, nicht-kommerzieller Tests zur Erfassung von Gf, nämlich des Zeichen-Zahlen-Tests aus dem SOEP und der Gerotest-Matrizen des IPG Erlangen. Folgt man diesem zweiten Vorschlag, fallen zwar einmalig größere Kosten für die altersgerechte Weiterentwicklung der Instrumente an, jedoch sind anschließend keine Lizenzgebühren mehr zu entrichten. Unseres Erachtens werden beide Empfehlungen darin münden, Gf bzw. SF und WG valide und reliabel über die Lebensspanne von 5 bis 67 Jahren zu erheben.

[2] Die im Kontext von large-scale-Projekten durchführbaren kognitiven Tests zur Erfassung kognitiv-fluider Basispotenziale „erlauben dabei keine Rückschlüsse auf die individuellen Intelligenz- oder Kognitionspotenziale einzelner Individuen im Sinne der Einzeldiagnostik. … Die [hier in der Regel einsetzbaren] ultrakurzen Tests erlauben … nur eine grobe Schätzung der auf individuelle Unterschiede zurückführbaren Varianz in der fluiden kognitiven Leistung“ (Lang, Kamin, Rohr, Stünkel & Williger, 2014).

1. Zielsetzung der Expertise

In neueren Survey- und Panelstudien stellt sich vermehrt die Frage, in welcher Weise in erhebungsökonomisch angemessenem Umfang zentrale und robuste Schätzer der fluiden bzw. non-verbalen Intelligenzleistungen erfasst werden können. Meist geht es in den jeweiligen Forschungszusammenhängen nicht um ein primäres Erkenntnisinteresse an der Ontogenese oder der Struktur von fluiden Intelligenzleistungen. Viel mehr geht es bei den auf ein breites Themenspektrum zielenden Großprojekten (z.B. sozio-ökonomisches Panel – SOEP, National Educational Panel Study – NEPS) darum, empirische Befunde auch gegen mögliche Einflüsse von individuellen Unterschieden der kognitiven Ausgangskapazität absichern zu können. Wenn beispielsweise die Einflüsse von Bildungskontexten auf die Kompetenzentwicklung des Individuums interessieren, so kann es von Bedeutung sein, ob sich mögliche Bildungsgewinne über die Zeit in Abhängigkeit von der jeweiligen intellektuellen Kapazität des Schulkindes abgeschwächt oder verstärkt zeigen (z.B. Snow, 1989).

Ein wesentliches Problem ist dabei die Tatsache, dass eine reliable, valide und objektive Diagnostik der intellektuellen Fähigkeiten eines Menschen in aller Regel nur von professionell geschultem Personal (vgl. DIN 33430, Westhoff, Hellfritsch, Hornke, Kubinger, Lang, Moosbrugger, Püschel & Reimann, 2005) und mit erheblichem Zeitaufwand möglich ist. Erst in jüngster Zeit wurde damit begonnen, auch in langfristigen Panelstudien (z.B. SOEP) eine grobe Erfassung von kognitiven Leistungspotenzialen der Studienteilnehmer/-innen vorzunehmen. Trotz der deutlichen Begrenzung solcher Kurzinstrumente im Hinblick auf Gültigkeit und Reliabilität (Lang, Weiss, Stocker & von Rosenbladt, 2007) ermöglicht die Berücksichtigung fluider Leistungsparameter eine grobe Abschätzung und Abgrenzung von bildungsunabhängigen und stärker bildungs- und umweltbezogenen Kompetenzen sowie deren Effekten.

Mit der vorliegenden Expertise wird die Frage behandelt, ob und inwiefern eine Erfassung der fluiden Intelligenz über die Lebensspanne im Rahmen von NEPS möglich ist. Dabei berücksichtigen wir eingehend die theoretischen und methodischen Überlegungen des ursprünglichen Forschungsantrags “Lifelong learning – A Proposal for a National Educational Panel Study (NEPS) in Germany, Part A and Part B“ (Blossfeld et al., 2008a, b)[3].

Im Vordergrund unserer Expertise steht die übergeordnete Frage, ob relativ robuste, nonverbale Indikatoren der fluiden Intelligenz bei Kindern, Jugendlichen und Erwachsenen in entwicklungs- und messäquivalenter Weise angemessen und zeitökonomisch erfasst werden können. Im Hinblick auf diese Zielsetzung geben wir empirisch belastbare Antworten zu folgenden spezifischen Einzelfragen:

- Welche Indikatoren der fluiden Intelligenz erscheinen im Rahmen einer Survey- und Panelstudie geeignet (Abschnitt 3)?
- Welche methodischen Anforderungen sollten Indikatoren der fluiden Intelligenz erfüllen (Abschnitt 4)?
- Welche Instrumente liegen vor, die geeignet erscheinen, entsprechende Indikatoren zu messen (Abschnitt 5)?

[3] Für eine Beschreibung der NEPS-Studie vgl. Blossfeld, Roßbach & von Maurice (2011).

- Welche Instrumente werden für die einzelnen Altersgruppen empfohlen (Abschnitt 6)?
- Welche Möglichkeiten bestehen für computergestütztes adaptives Messen, bzw. Testen (Abschnitt 7)?

Die Expertise mündet in Abschnitt 8 in ein Fazit und in konkrete Empfehlungen zur Erfassung kognitiver, non-verbaler Leistungen der fluiden Intelligenz in verschiedenen Lebensphasen (bzw. Etappen der NEPS-Studie).

2. Die Studienanlage von NEPS

Für eine Beantwortung der dargestellten Fragen erscheint es zunächst notwendig, die Studienanlage von NEPS einleitend darzustellen, um die sich ergebenden Anforderungen an eine Kurzerfassung der fluiden bzw. non-verbalen Intelligenz zu verdeutlichen.

NEPS erfasst Kompetenzen, die für ein erfolgreiches und verantwortungsbewusstes Leben sowie das Funktionieren einer modernen Gesellschaft relevant sind (Blossfeld, 2008a, S. 35). Daher sollen neben domänenspezifischen Kompetenzen (bspw. in den Domänen Deutsch, Mathematik und Naturwissenschaften) auch domänenübergreifende, möglichst bildungsunabhängige Indikatoren der fluiden Intelligenz *Gf* bzw. der kognitiven Mechanik bei Personen sehr unterschiedlichen Alters erhoben werden (Blossfeld, 2008a, S. 36)[4].

Als zentrale Indikatoren von *Gf* werden figurale (non-verbale) Aufgaben zum Schlussfolgernden Denken (*SF*: „*Reasoning*“) sowie Aufgaben zur Messung der Wahrnehmungsgeschwindigkeit betrachtet (*WG: „Speed“*, Blossfeld, 2008b, S. 10).

Diese Indikatoren sollen für die folgenden sechs Altersgruppen erfasst werden (vgl. Tabelle 1): (1) 5 Jahre, (2) 7 Jahre, (3) 10 Jahre, (4) 14 Jahre, (5) 20-26 Jahre und (6) 28-70 Jahre. Zur Untersuchung dieser Altersgruppen bedient sich NEPS eines Multikohorten Sequenzdesigns (Blossfeld, 2008a, Tabelle 7.8, S. 83)[5], das vier Längsschnitt-Kohorten (Kohorte 1: 7 Monate bis 13 Jahre[6]; Kohorte 2: 4 bis 17 Jahre; Kohorte 3: 10 bis 23 Jahre; Kohorte 4: 14 bis 23 Jahre) vorsieht sowie zwei Kohorten untersucht, die ein querschnittliches und längsschnittliches Design kombinieren (Kohorte 5: 18 bis 29 Jahre; Kohorte 6: 23-67 Jahre). Abbildung 1 und Tabelle 1 fassen die wichtigsten Elemente der Studienanlage von NEPS zusammen.

[4] Für eine ausführliche Darstellung vgl. Weinert, Artelt, Prenzel, Senkbeil, Ehmke & Carstensen (2011).

[5] Blossfeld, von Maurice & Schneider, 2011, S. 14.

[6] Hier sind die Planungen für 13 Jahre benannt; die genaue Laufzeit der einzelnen Kohorten ist damit nicht abschließend festgelegt.

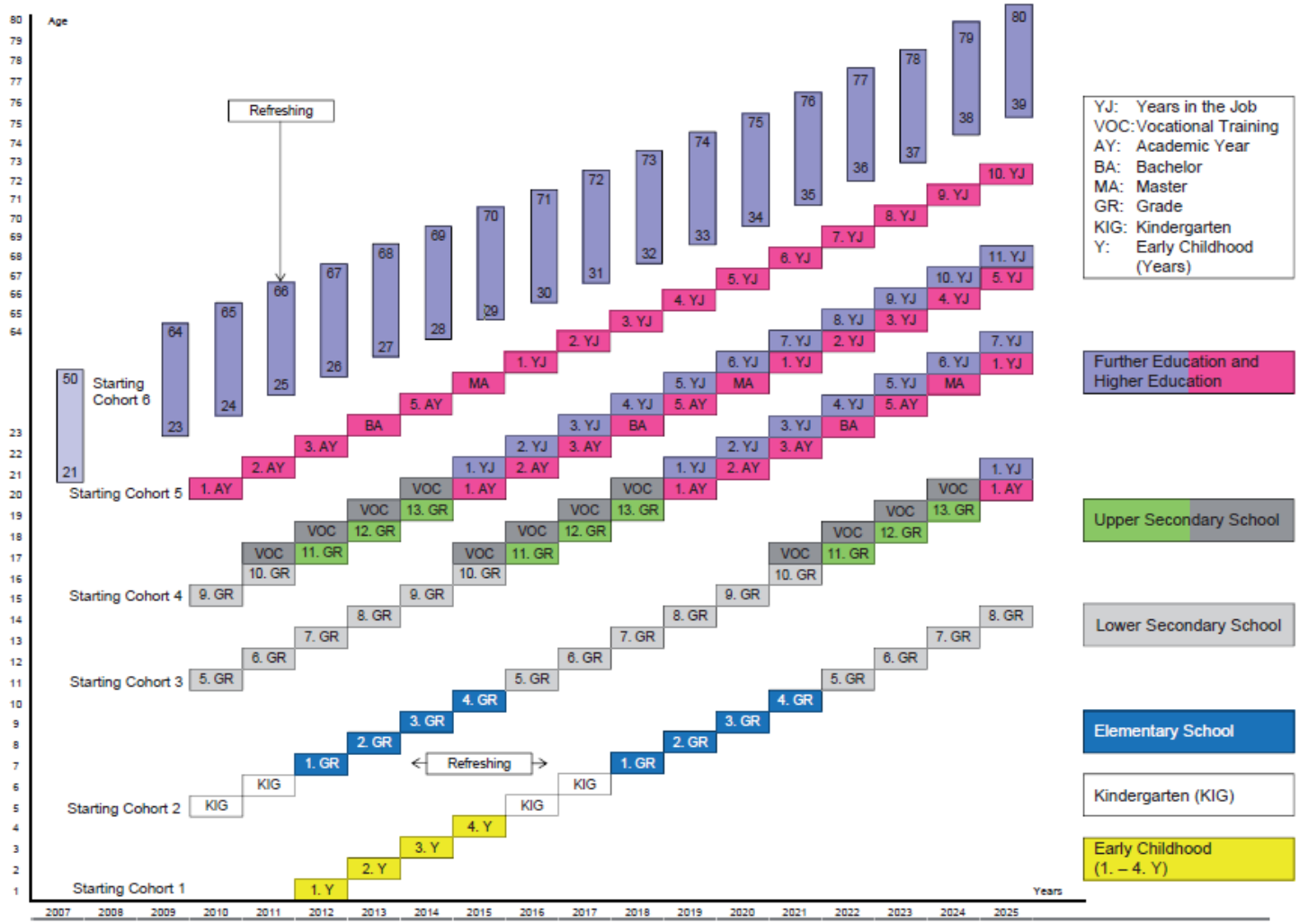


Abbildung 1. Kohortensequenzdesign des NEPS (übernommen aus Artelt, Weinert & Carstensen, 2013, S. 7; diese Abbildung wurde zur Aktualisierung der Expertise in das vorliegende NEPS Working Paper aufgenommen)

Tabelle 1. Kompetenzmessungen des NEPS in den Jahren 2010-2014 für die verschiedenen Startkohorten (über die genannten Kompetenzen hinaus wurden zudem Kompetenzen in der Erstsprache (L1) und im Englischen erhoben; übernommen aus Artelt, Weinert & Carstensen, 2013, S. 11; die Tabelle wurde zur Aktualisierung der Expertise in das vorliegende NEPS Working Paper aufgenommen)

Age	Stage	Assessment	*Domain-general competencies*	*Domain-specific competencies*				*Metacompetencies*		*Stage-specific tests*
				German language						
			DGCF	Reading	Listening	Math	Science	Metacogn.	ICT	
Starting Cohort 2 – Kindergarten										
4	Preschool	2010			x		x			
5	Preschool	2011	x			x				x
6	E: Grade 1	2012			x	x	x	x		
7	E: Grade 2	2013	x	x		x				
8	E: Grade 3	2014			x		x	x	x	
Starting Cohort 3 – Grade 5										
10	S: Grade 5	2010	x	x		x				x
11	S: Grade 6	2011			x		x	x	x	
12	S: Grade 7	2012		x		x				x
13	S: Grade 8	2013								
14	S: Grade 9	2014	x	x	x	x	x	x	x	x
Starting Cohort 4 – Grade 9										
14	S: Grade 9	2010	x	x	x	x	x	x	x	x
15	S: Grade 10/Voc	2011								
16	S: Grade 11/Voc	2012					x			
17	S: Grade 12/Voc	2013		x		x			x	
Starting Cohort 5 – First-Year Students										
18–24	U: Acad. Y. 1	2010		x		x				
19–25	U: Acad. Y. 2	2011								
20–26	U: Acad. Y. 3	2012	x				x			
21–27	U: Acad. Y. 4	2013								
22–28	U: Acad. Y. 5	2014			x					
Starting Cohort 6 – Adults										
24–66		2010		x		x				
25–67		2011								
26–68		2012					x		x	
27–69		2013								
28–70		2014	x		x					

Note. DGCF = domain-general cognitive functions; reading = reading competence; listening = listening comprehension; Math = mathematical competence; Metacogn. = metacognition; ICT = information and communication technologies literacy; E = elementary school; S = secondary school; Voc = vocational training; U = university.

Starting Cohort 1 – Early Childhood is not depicted because only stage-specific competence assessments were assessed.

Für die valide und reliable Erfassung von *Gf* im Rahmen einer large-scale-Studie stehen für die Kohorten 2, 3 und 4 jeweils 30 Minuten zur Verfügung; bei den Kohorten 5 und 6 sind jeweils 15 Minuten Testzeit vorgesehen (Blossfeld, 2008a, Tabelle 7.8, S. 83). Im Idealfall sollten die eingesetzten Instrumente der fluiden Intelligenz *SF* und *WG* bei Kindern mit 5 Jahren in gleicher Art und Weise gemessen werden wie bei Erwachsenen im Alter von 67 Jahren (vgl. Blossfeld, 2008a, S. 47).

3. Welche Indikatoren der fluiden Intelligenz erscheinen geeignet?

Im Einklang mit dem Rahmenkonzept von NEPS (Blossfeld, 2008b, S. 10) betrachten wir Tests zum Schlussfolgernden Denken mit figuralem Inhalt (SF) sowie zur Wahrnehmungsgeschwindigkeit (WG) als sehr gut geeignete Indikatoren der fluiden Inteligenz (Gf) bzw. der kognitiven Mechanik.

Unsere Schlussfolgerung basiert auf zwei Überlegungen: Erstens begründen wir, weshalb SF und WG konzeptuell wie auch empirisch im Zentrum von Gf liegen. Zweitens referieren wir Studien, die zeigen, dass Indikatoren von Gf bedeutsam für eine erfolgreiche Entwicklung sind.

3.1 Was ist fluide Intelligenz Gf?

Die Frage was Intelligenz sei, beschäftigt Wissenschaftler und Philosophen schon seit jeher. So findet sich bereits bei Aristoteles eine Unterscheidung zwischen dem schon „angelegten Vermögen“ und dem noch zu erwerbenden Wissen des Menschen (vgl. Ackrill, 1985). Eine durchaus vergleichbare Unterscheidung zwischen „absoluten“ und „relativen“ Vermögensweisen des Menschen verwendete bereits Johann Nicolaus Tetens (1777) in seinen „Versuchen über die menschliche Natur“. Tetens legte damit die Grundlage einer modernen, die gesamte Lebensspanne umfassenden anthropometrischen Intelligenzforschung, die mit Francis Galton und den ersten psychometrischen Arbeiten von Alfred Binet zu Beginn des 20. Jahrhunderts ihren eigentlichen Anfang nahm.

Allerdings ist es bis heute dennoch nicht gelungen, eine einheitliche Definition des Intelligenzbegriffs in der Psychologie zu etablieren. Obwohl viele Detailfragen noch strittig sind, kristallisiert sich aber zunehmend ein Konsens in der Definition und Konzeptualisierung von Intelligenz heraus. Zwei Forschungslinien sind hierbei bedeutsam.

Erstens, eine wichtige Unterscheidung im Kontext kognitiver Theorien über die Lebensspanne betrifft die Differenzierung von eher biologisch determinierten und eher kulturell determinierten Komponenten der Intelligenz. Diese Unterscheidung findet sich in aktuellen zwei-Komponenten-Theorien der Intelligenz wieder, wie zum Beispiel in der Theorie fluider (Gf) und kristalliner (Gc) Intelligenz von Cattell und Horn (Cattell, 1963; Horn & Noll, 1997) oder der Theorie der kognitiven Mechanik und kognitiven Pragmatik von Baltes und Kollegen (Baltes, Staudinger & Lindenberger, 1999). Geschwindigkeit, Genauigkeit und Koordination kognitiver Prozesse werden hierbei der fluiden Intelligenz zugeordnet, wohingegen erworbenes Wissen der kristallinen Intelligenz zugerechnet wird.

Zweitens besteht unter zahlreichen Experten der Kognitionswissenschaften ein Konsens darin, dass höhere kognitive Prozesse wie beispielsweise Schlussfolgerndes Denken (SF) ein Kernelement der Intelligenz darstellen (Gottfredson, 1997a; Intelligence and its measure-

ment: A symposium, 1921; Mayer, 1992; Neisser et al., 1996; Snyderman & Rothman, 1987; Sternberg & Detterman, 1986). Mit Gottfredson (1997a) bzw. mit 52 Experten der Intelligenzforschung, kann Intelligenz wie folgt definiert werden: “A very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience.” (Gottfredson, 1997a, p. 13).

Empirisch unterfüttert wird die konzeptuelle Unterscheidung von Gf und Gc vor allem durch zwei oftmals replizierte Befunde. Zum Einen kann bei Vorliegen einer heterogenen Testbatterie von kognitiven Aufgaben immer ein genereller Faktor g extrahiert werden (Lubinski, 2004; Spearman, 1904). Wichtig ist hierbei, dass viele Studien zeigen, dass im Vergleich zu anderen Aufgabengruppen, bei Tests von SF deutlich mehr Varianz durch g erklärt wird (Arendasy, Hergovich & Sommer, 2008; Carroll, 1993; Gustafsson, 1984, 1988). Dies deutet daraufhin, dass der generelle, domänenunspezifische Aspekt kognitiver Fähigkeiten am besten durch SF abgebildet wird. Diese Schlussfolgerung wird auch dadurch gestützt, dass viele Forscher Maße des SF wie beispielsweise den Raven Progressive Matrizentest (Raven, Raven & Court, 1998) als direktes Maß von Gf (bzw. Intelligenz) betrachten (Irwing & Lynn, 2005; Mackintosh, 1998).

Zum Zweiten unterscheiden sich die Altersgradienten für Tests der WG wie auch für Tests des SF deutlich von den Altersgradienten für Tests von Gc. Während mit zunehmendem Alter die Leistung bei WG und SF deutlich abnimmt, bleibt die Leistung bei Tests von Gc relativ konstant, oder nimmt erst im hohen Erwachsenenalter ab (Schaie, 1996). Dies deutet daraufhin, dass Gf (und damit auch deren Komponenten SF und WG) deutlich stärker biologisch determiniert sind als Gc.

Bis hierhin ist festzuhalten, dass im Kontext kognitiver Theorien über die Lebensspanne mit Gf und Gc zwei Komponenten der Intelligenz unterschieden werden. Weiterhin fokussieren Expertendefinitionen von Intelligenz vornehmlich auf Gf als Kernaspekt des Intelligenzbegriffs. Konzeptuell stehen also generelle, domänenübergreifende kognitive Prozesse der Geschwindigkeit, Genauigkeit, Koordination und des Schlussfolgernden Denkens im Zentrum von Gf.

Als zentrale Markiervariable von Gf werden daher häufig Tests der Wahrnehmungsgeschwindigkeit (Baltes et al., 1999; Ghisletta & Lindenberger, 2003) wie auch Tests des Schlussfolgernden Denkens (Cattell, 1987) verwendet. Die Markiervariablen von Gf zeichnen sich im Allgemeinen einheitlich dadurch aus, dass sie möglichst bildungsunbhängig, neuartig und domänen-unspezifisch sind. Aus diesen Gründen werden meist Aufgaben mit figuralen Anforderungen gewählt bzw. bevorzugt (Cattell, 1963).

Auf der anderen Seite fokussieren Markieraufgaben der kristallinen Intelligenz (Gc) auf Aufgaben, deren Lösung Wissen erfordert, das im Kontext schulischer (aber auch außerschulischer) Lernprozesse erworben wurde. Diese Markieraufgaben sind meist bildungsabhängig, relativ verbreitet bzw. bekannt und domänenspezifisch (Cattell, 1963, 1987; Ghisletta & Lindenberger, 2003; Horn & Noll, 1997). Als Indikatoren von Gc können somit auch Vokabeltests oder Schülerleistungstests in Mathematik oder Deutsch betrachtet werden.

3.2 Stellenwert fluider Intelligenz für erfolgreiche Entwicklung

Fluide Intelligenz und ihre Subkomponenten gehören zu den wichtigsten psychologischen Konstrukten, da sie prädiktiv für das Erreichen vieler Lebensziele sind und Lernprozesse determinieren.

Schulischer Erfolg. Die ersten Intelligenztests wurden von Alfred Binet und seinen Kollegen entwickelt, um schulischen Erfolg vorherzusagen (Mayer, 2000). Und in der Tat, „If there is any unquestioned fact in applied psychometrics, it is that IQ tests have a high degree of predictive validity for many educational criteria, such as [...] school and college grades“ (Jensen, 1998, S. 277). Zahlreiche Meta-Analysen und Einzelstudien bestätigen diese Schlussfolgerung (Deary, Strand, Smith & Fernandes, 2007; Gustafsson & Balke, 1993; Kuncel, Hezlett & Ones, 2004; Neisser et al., 1996; Tent, 2001). So lag beispielsweise in der Meta-Analyse von Kuncel et al. die über 70 Einzelstudien gemittelte Korrelation zwischen fluider Intelligenz und dem Graduate Grade Point Average bei r = .27 (Kuncel et al., 2004, Tabelle 2).

Beruflicher Erfolg. Fluide Intelligenz hat sich wiederholt als ein bedeutsamer Prädiktor des beruflichen Erfolgs und damit auch des erreichten sozio-ökonomischen Status herausgestellt. In zahlreichen Meta-Analysen (Bertua, Anderson & Salgado, 2005; Kuncel et al., 2004; Salgado et al., 2003; Schmidt & Hunter, 1998) fanden sich substanzielle Korrelationen zwischen Indikatoren der fluiden Intelligenz einerseits und Kriterien des beruflichen Erfolgs andererseits (z.B. Abschluss der beruflichen Ausbildung, berufsrelevantes Wissen oder Vorgesetztenbeurteilungen).

Gesundheit. In jüngerer Zeit haben insbesondere die empirischen Arbeiten von Deary und Kollegen wiederholt den Zusammenhang zwischen fluider Intelligenz und Gesundheit bzw. Mortalität nachgewiesen (Batty, Deary & Gottfredson, 2007; Deary et al., 2005; Gottfredson, 1997b; Gottfredson & Deary, 2004): Intelligentere Menschen leben gesünder und dabei wohl auch länger.

Lernprozesse. Ein wichtiger Grund für die prädiktive Kapazität fluider Intelligenz liegt unter anderem darin, dass diejenigen Personen, die hohe Werte bei Tests von *Gf* (und hier insbesondere Tests von *SF*) erzielen, tendenziell mehr deklaratives und prozedurales Wissen erwerben, um erfolgreich komplexe Informationen zu verarbeiten, die relevant für schulisches Lernen, Ausbildung, Beruf oder eben Gesundheit sind (Gottfredson, 1997b, 2004; Gottfredson & Deary, 2004; Helmke & Weinert, 1997).

Der Zusammenhang zwischen Gf und Wissenserwerb zeigt sich unter anderem auch darin, dass sowohl SF als auch WG wichtige Determinanten des Erwerbs komplexer Fertigkeiten (z.B. bei Fluglotsenaufgaben) sind (Ackerman, 1987, 1988; Ackerman & Kanfer, 1993): So geht Ackerman im Einklang mit Theorien aus der Forschung zur Informationsverarbeitung (Anderson, 1993) davon aus, dass der Fertigkeitserwerb (zur Lösung von Aufgaben mit konsistenten Anforderungen) in drei Phasen abläuft. In der ersten kognitiven Phase sind Gedächtnis- und „Reasoning“-prozesse notwendig. Die Bearbeitung von Aufgaben in dieser Phase ist langsam und fehleranfällig, die gesamte Aufmerksamkeit eines Lerners ist notwendig, um die Aufgabe zu verstehen und auszuführen. Die Leistungen in dieser Phase korrelieren sehr hoch mit SF. Sobald ein Lerner eine adäquate mentale Repräsentation der Aufgabe gebildet hat, geht er über in die assoziative Phase. Während dieser zweiten Phase fügen Lernende die Sequenzen kognitiver und psychomotorischer Prozesse, die zur Auf-

gabenbearbeitung notwendig sind, zusammen (Kompilierung). Die Leistungen in der assoziativen Phase korrelieren am höchsten mit der Wahrnehmungsgeschwindigkeit (Ackerman, 1989, S. 179). In der dritten, autonomen Phase ist die Aufgabenbearbeitung weitestgehend automatisiert. Die Leistungen in dieser Phase korrelieren am höchsten mit psychomotorischen Fähigkeiten.

Resümierend stellen wir fest, dass mit Tests zur Erfassung von *SF* und *WG* zwei wichtige Indikatoren von *Gf* im Rahmen von NEPS gewählt wurden, die einerseits im Kern der Konzeption von *Gf* liegen und andererseits empirisch nachgewiesenermaßen bedeutsame Determinanten von zentralen Lebenszielen und Lernprozessen sind.

4. Methodische Anforderungen zur Erfassung von Gf über die Lebensspanne

Eine wichtige Frage ist nun, welche methodischen Anforderungen die Instrumente zur Erfassung von *Gf* erfüllen müssen, die im Rahmen von NEPS eingesetzt werden sollen. Aus diesen methodischen Anforderungen werden wir einen Kriterienkatalog ableiten, um konkrete Empfehlungen für Messinstrumente zur Erfassung von *Gf* auszusprechen.

4.1 Klassische psychometrische Gütekriterien

Zur Bewertung von Messinstrumenten ziehen wir einerseits die Validität und Reliabilität als zentrale psychometrische Gütekriterien heran (American Educational Research Association, 1999; Testkuratorium der Föderation deutscher Psychologenverbände, 1986); die beiden Gütekriterien werden in diesem Abschnitt besprochen.

Weiterhin hat NEPS das Anliegen, „*that the measurement instruments are comparable across the various cohorts of the NEPS*“ (Blossfeld, 2008a, S. 47): Die Studienanlage von NEPS ermöglicht längsschnittliche Entwicklungsprozesse von *Gf* (z.B. für Kohorte 2; vgl. Abb. 1), bzw. altersbezogene Unterschiede in *Gf* (z.B. für Kohorte 6; vgl. Abb. 1, Tab. 1) zu untersuchen. Deshalb beurteilen wir auch die Messinvarianz der Erhebungsinstrumente über die Zeit bzw. über Altergruppen hinweg als weiteres Kriterium. In Abschnitt 4.2 gehen wir detailliert auf dieses psychometrische Konzept ein.

Validität. Zur Definition des Validitätsbegriffs greifen wir auf die einflussreiche Arbeit von Borsboom, Mellenbergh und van Heerden (2004) zurück. Sie verwenden hierbei die „klassische“ Definition von Validität nach der ein Test valide ist, wenn er das misst, was er messen soll (für eine alternative, doch deutlich komplexere Definition, s. Messick, 1989). Das besondere an der Arbeit von Borsboom und Mitarbeitern ist, dass sie den Validitätsbegriff im Kontext von latenten Variablenmodellen verankern; diese Messmodellklasse (z.B. Item Response Theorie [IRT] oder Strukturgleichungsmodelle [SEM]) wird gegenwärtig in der Bildungsforschung am häufigsten verwendet.

Entsprechend der Validitätsdefinition von Borsboom ist beispielsweise ein Test des *SF* ein *valides* Instrument zur Messung von *Gf*, falls (1) *Gf* existiert und (2) interindividuelle Unterschiede in *Gf* kausal für beobachtete interindividuelle Unterschiede in den Testwerten von *SF* verantwortlich sind. Gleichermaßen gilt, dass ein Test der *WG* valide *Gf* erfasst, falls (1) *Gf* exisitiert und (2) interindividuelle Unterschiede in *Gf* kausal für beobachtete interindividuelle Unterschiede in den Testwerten von *WG* verantwortlich sind. Abbildung 2 stellt diese Überlegungen im Rahmen eines konfirmatorischen Faktormodells für *Gf* dar. Die manifesten

Testwerte (*SF* und *WG*) sind in Abbildung 2 als Rechtecke, die latente Fähigkeit *Gf* ist als Ellipse dargestellt. *Gf* beeinflusst die Leistung bei *SF* und *WG*. Dies wird durch gerichtete Pfeile dargestellt, die von *Gf* auf *SF* bzw. *WG* zielen: Individuelle Unterschiede in *Gf* (repräsentiert durch die Varianz σ^2(Gf)) erklären somit beobachtete individuelle Unterschiede in den Tests. Die Stärke des kausalen Einflusses wird durch die Faktorladungen λ1 und λ2 repräsentiert. Weiterhin ist davon auszugehen, dass nicht die gesamte Varianz von *SF* und *WG* durch *Gf* erklärt wird: Ein Teil dieser Varianz σ^2(e1) sowie σ^2(e2) ist spezifisch für den jeweiligen Test; der restliche Varianzanteil von σ^2(e1) sowie σ^2(e2) kann auf zufällige Messfehler zurückgeführt werden. Testspezifische und Fehlervarianzen können weder getrennt voneinander analysiert werden noch direkt gemessen werden. Deshalb sind diese in Abbildung 2 jeweils in Form eines Fehlerterms als latente Variablen e1 und e2 dargestellt.

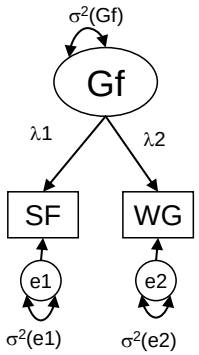


Abbildung 2. Messmodell der fluiden Intelligenz (Gf) mit Tests zum schlussfolgernden Denken und der Wahrnehmungsgeschwindigkeit (WG)

Die Frage, ob die Items zur Messung der latenten Fähigkeiten *SF* und *WG* valide sind, kann auch im Kontext von IRT-Modellen analysiert werden. Die Rationale ist hier analog zu *Gf*.

Warum sind diese Ausführungen zur Modellierung wichtig? Wir verwenden die Definition von Borsboom et al. (2004) um die Validität von Instrumenten zu beurteilen. Daher sind vor allem zwei Nachweise für die Validität eines Instruments entscheidend:

- Es werden Gründe angegeben, weshalb ein latentes Konstrukt kausal für beobachtbare Unterschiede in den manifesten Testwerten verantwortlich sein soll. Diese Begründung kann (a) auf Expertenratings oder besser (b) auf der systematischen Manipulation der Itemeigenschaften basieren (Arendasy et al., 2008; Embretson, 1983, 1998).
- Die Struktur des Messinstrumentes wurde mittels latenter Variablenmodelle (IRT oder SEM) geprüft.

Reliabilität. Die Überlegungen von Borsboom und Mitarbeitern, wie man den Validitätsbegriff in den Kontext von latenten Variablenmodellen integriert, haben wichtige Implikationen für die Reliabilität im Sinne der internen Konsistenz (McDonald, 1999). *Reliabilität* kann damit definiert werden als die Genauigkeit, mit der ein Testwert das Zielkonstrukt misst.

Reliabilität wird gemeinhin als der Anteil von wahrer Varianz zur Gesamtvarianz der Testwerte bestimmt (Lord & Novick, 1968; Schmidt, Le & Ilies, 2003). Hierbei ist zu beachten, dass mit zunehmenden Faktorladungen der Anteil wahrer Varianz in den Indikatoren ebenfalls

zunimmt (Bollen, 1989; McDonald, 1999). *Ceteris paribus* nimmt also mit Zunahme des kausalen Einflusses von *Gf* auf SF und *WG* (in Form der Faktorladungen $\lambda 1$ und $\lambda 2$) die Reliabilität der Tests SF und *WG* zu. Höhere Faktorladungen schlagen sich also in höheren Werten von Maßen der internen Konsistenz (z.B. Cronbach's Alpha) nieder.

Ein weiterer Aspekt der Reliabilität ist die *Retest-Reliabilität* der Testwerte. Diese wird allgemein als Korrelation der Testwerte derselben Personen zu zwei verschiedenen Zeitpunkten bestimmt. Je höher die Korrelation ist, desto stabiler ist beispielsweise die Rangreihe der Schüler/-innen in einem Mathematiktest: Diejenigen Schüler/-innen, die im Vergleich zu ihren Mitschüler(n)/-innen zum ersten Messzeitpunkt besser abschnitten, schneiden also auch besser beim zweiten Messzeitpunkt ab. Perfekte Stabilität impliziert (unter anderem), dass sich die Leistungsrangreihe nicht verändert, d.h. durch den Test mit einer hohen Messgenauigkeit erfasst wird.

Bei der Erfassung der *Retest-Reliablität* gilt es zu beachten, dass die Messzeitpunkte relativ nah beieinander liegen sollten (maximal 2-3 Monate Abstand). Bei größeren Zeitabständen kann nicht mehr eindeutig entschieden werden, ob eine geringe Korrelation zwischen den beiden Messzeitpunkten in einer niedrigen Reliabilität des Testinstruments oder einer differenziellen Entwicklung in dem zu erfassenden Konstrukt begründet liegt. So konnte Becker (2008) Hinweise dafür liefern, dass auch die Entwicklung der fluiden Intelligenz sensitiv für den Einfluss unterschiedlicher Entwicklungsumwelten ist. Auf Grundlage einer Teilstichprobe der BIJU-Studie (Sachsen-Anhalt und Mecklenburg-Vorpommern), in der die psychometrische Intelligenz (Markiervariable Figurenanalogien) der Schüler/-innen direkt zum Zeitpunkt des Übergangs in die gegliederte Sekundarstufe (Beginn 7.Klassenstufe) sowie nach weiteren vier Schuljahren (Ende 10. Klassenstufe) erhoben wurde, wurde ein *Propensity Score Matching* durchgeführt, bei dem Gymnasiasten/-innen und Nicht-Gymnasiasten/-innen, die eine ähnliche Wahrscheinlichkeit besaßen, dem Gymnasium zugewiesen zu werden, in ihrer Intelligenzentwicklung verglichen wurden. Es zeigten sich substanzielle Unterschiede in der fluiden Intelligenz zugunsten der Gymnasiasten/-innen am Ende der 10. Klassenstufe, die für einen differenziellen Fördereffekt der Schulform sprechen.

4.2 Messinvarianz

Ein Ziel von NEPS ist, dass die Messinstrumente über Kohorten hinweg vergleichbar sind (Blossfeld, 2008a, S. 47). Konkret bedeutet dies, dass im Idealfall *Gf* und deren Subkompetenzen *SF* und *WG* in gleicher Art und Weise über die Alterspanne von 4 bis 67 Jahren gemessen werden sollen. Um dies zu gewährleisten, muss die psychometrische *Messinvarianz* der Instrumente vorliegen. Diese kann je nach Skalenniveau der vorliegenden Daten (Ordinal vs. Intervallskalenniveau) detailliert im Rahmen von Item-Response-Modellen (Millsap & Yun-Tein, 2004) und Strukturgleichungsmodellen (Horn & McArdle, 1992) untersucht werden. Das Vorliegen unterschiedlicher Abstufungen der Messinvarianz hat entscheidende Implikationen für die Analysen zur kognitiven Entwicklung, bzw. die Vergleichbarkeit von *Gf* über Altersstufen hinweg. Diese Überlegungen wollen wir für zwei Messzeitpunkte (hier die Altersstufen 5 und 7 Jahre) von NEPS verdeutlichen. Im Beispiel wird die *Stabilität* und *mittlere Veränderung* von *Gf* im Rahmen eines Strukturgleichungsmodells analysiert.

Gf wird zu beiden Erhebungszeitpunkten als latente Variable konzipiert (Gf_5, Gf_7) und jeweils durch figurale Matrizenaufgaben zur Erfassung des Schlussfolgernden Denkens (SF_5, SF_7)

sowie durch Aufgaben mit figuralem Inhalt zur Erfassung der Wahrnehmungsgeschwindigkeit (WG_5, WG_7) gemessen. Die Korrelation $r(Gf_5, Gf_7)$ zwischen *Gf* im Alter von 5 Jahren und *Gf* im Alter von 7 Jahren ist als Doppelpfeil dargestellt; die Korrelation repräsentiert die *Stabilität* von *Gf*.

Zu beiden Erhebungszeitpunkten wird angenommen, dass *Gf* die Leistung bei SF und WG kausal beeinflusst: Individuelle Unterschiede in *Gf* (repräsentiert durch die Varianzterme $\sigma^2(Gf_5)$ und $\sigma^2(Gf_7)$) erklären somit beobachtete individuelle Unterschiede in den Testwerten. Analog zu einem Regressionsmodell ist *Gf* jeweils eine unabhängige Variable und die manifesten Testwerte sind abhängige Variablen. Daher ist bei allen Testwerten ein Intercept α vorgesehen (der in Abbildung 3 durch einen Pfeil vom Dreieck auf die manifesten Variablen dargestellt ist). Der Mittelwert der latenten Variable Gf_7 wird durch $\alpha(Gf_7)$ repräsentiert.

Um mathematisch eindeutige Schätzungen der Modellparameter zu ermöglichen (Identifikation des Modells), sind die Faktorladungen des SF auf *Gf* auf den Wert 1 fixiert (Festlegung der Metrik von *Gf*). Zudem wurde der Mittelwert der Fehlerterme sowie der Mittelwert von Gf_5 auf 0 fixiert. Aufgrund dieser Restriktion stellt der Mittelwert von Gf_7 (bei Vorliegen skalarer oder strikter Invarianz, s.u.) die *mittlere Veränderung* in *Gf* vom 5. zum 7. Lebensjahr dar.

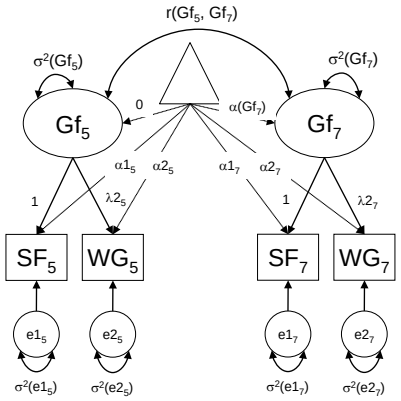


Abbildung 3. Ein Beispiel zur Messung von fluider Intelligenz (Gf) zu zwei Messzeitpunkten im Alter von 5 und 7 Jahren. SF: Aufgabenwert zum Schlussfolgernden Denken mit figuralem Inhalt. WG: Aufgabenwert der Wahrnehmungsgeschwindigkeit

In Abhängigkeit davon, welche psychometrischen Restriktionen für die Aufgabenwerte erfüllt sind, können unterschiedliche Aussagen über die Entwicklung von *Gf* sowie der Entwicklung der Subkompetenzen *Schlussfolgerndes Denken* und *mentale Geschwindigkeit* gemacht werden (Tabelle 2).

Konfigurale Invarianz. Bei dieser Form der Invarianz laden zu jedem Messzeitpunkt die manifesten Variablen auf korrespondierenden latenten Variablen, und das Muster der Null-Ladungen bleibt über die Zeit hinweg gleich. Die Faktorladungen, die Residualvarianzen und die Mittelwerte der manifesten Variablen können (innerhalb und) zwischen den Messzeitpunkten variieren, dass heißt, diese Modellparameter werden *nicht* invariant über die Zeit gesetzt. Wenn nur die Bedingung konfiguraler Invarianz erfüllt ist, ist der Vergleich manifes-

ter wie auch latenter Varianzen sowie manifester wie auch latenter Mittelwerte über die Zeit (bzw. Altersstufen) hinweg *nicht* möglich. Hierfür müssen weitere Bedingungen erfüllt sein.

Metrische Invarianz. Bei der metrischen Invarianz sind zusätzlich zu den Bedingungen der konfiguralen Invarianz die unstandardisierten Ladungen der manifesten Variablen invariant über die Messzeitpunkte. Wenn die Bedingung der metrischen Invarianz erfüllt ist, ist die Bedeutung einer latenten Variable *Gf* (zur Erklärung interindividueller Unterschiede bei den jeweiligen manifesten Aufgabenwerten) zu jedem Messzeitpunkt gleich. Somit können latente Varianzen (und Kovarianzen) von *Gf* über die Zeit hinweg verglichen werden. Manifeste und latente Mittelwerte sowie die Varianz der manifesten Variablen kann jedoch *nicht* über die Zeit hinweg verglichen werden.

Skalare Invarianz. Möchte man die latenten Mittelwerte über die Zeit hinweg vergleichen, muss zum einen die metrische Invarianzbedingung erfüllt sein. Zum anderen müssen zusätzlich die Intercepts der manifesten Variablen über die Zeit hinweg invariant sein. Bei Vorliegen der skalaren Invarianz können jedoch die Varianzen der manifesten Variablen *nicht* über die Zeit hinweg verglichen werden.

Strikte Invarianz. Der Vergleich der Varianzen manifester Variablen über die Zeit erfordert, dass noch zusätzlich zu den Restriktionen der skalaren Invarianz die Varianzen korrespondierender Fehlerterme über die Zeit hinweg invariant gesetzt werden.

Wichtig ist hierbei zu betonen, dass diese Überlegungen zur längsschnittlichen Invarianz von Messinstrumenten von *Gf* analog auf die Invarianz der Messung von *Gf* für verschiedene Altersgruppen angewendet werden können (s. z.B. Lubke, Dolan, Kelderman & Mellenbergh, 2003; Meredith, 1993).

4.3 Computergestütztes adaptives Testen

Bislang haben wir psychometrische Kriterien diskutiert. Eine interessante Frage ist in diesem Zusammenhang, welchen Einfluss die Form der Testadministration hierbei auf diese Kriterien hat. Zwei Formen sind dabei von besonderer Bedeutung: Computergestütztes adaptives Testen (*CAT*) und Papier-und-Bleistift-Tests (*PBT*).

CAT zielt darauf ab, für jeden Studienteilnehmer den optimalen Fähigkeitstest zu konstruieren (Frey, 2007; Meijer & Nering, 1999). Hierzu wird während der Testdurchführung die Personenfähigkeit geschätzt. Die zu bearbeiteten Testitems, deren psychometrische Eigenschaften auf Grundlage eines IRT-Modells vorab bestimmt wurden, werden dann aus einer Itembank nach einem vorgegebenen Algorithmus ausgewählt. Die meisten dieser Algorithmen maximieren bei einer möglichst geringen Anzahl bearbeiteter Items die Messgenauigkeit der Personenfähigkeit für jede(n) Testteilnehmer/-in: Items werden so ausgewählt, dass sie maximal informativ für das jeweilige Fähigkeitsniveau sind.

Um die Möglichkeiten für CAT zu beurteilen, ist es notwendig, sich die Vor- und Nachteile von CAT zu vergegenwärtigen. Vor- und Nachteile ergeben sich sowohl aus einer computerbasierten Testung als auch adaptiven Testung (Frey, 2007). Zu den Vorteilen der computerbasierten Testung gehören eine höhere Testsicherheit, ein standardisierter Testablauf, schnelle und fehlerfreie Testauswertung, zeitnahe Ergebnisrückmeldung sowie die Verwendung neuer Itemformate.

Die Nachteile computerbasierten Testens liegen vor allem in der organisatorischen Durchführbarkeit. Der Entwicklungsaufwand für eine solche Testung ist im Vergleich zu PBT deutlich höher. Ferner ist die Verfügbarkeit von Computern am Testort zu gewährleisten (entweder durch Nutzen bestehender Computerlabors vor Ort oder durch den Einsatz mobiler Notebooklabors). Letztlich können individuelle Unterschiede in der Erfahrung mit Computern oder Computerängstlichkeit die Leistung positiv wie auch negativ beeinflussen (Frey, 2007).

Zu den Vorteilen des adaptiven Testens zählen vor allem die deutlich höhere *Messeffizienz*: Für eine vergleichbare Messpräzision werden bei adaptiven Verfahren häufig nur 40 bis 60 Prozent der Items benötigt wie bei nicht-adaptiven Verfahren (Frey, 2007; Frey & Ehmke, 2008). Weiterhin besteht die Möglichkeit, bei einer vorgegebenen Itemzahl in den Extrembereichen des Fähigkeitsbereichs (bei Vorliegen entsprechender Items) mit hoher Präzision zu messen. Dies ist bei nicht-adaptiven Verfahren häufig nicht möglich, da bei diesen vor allem Items mittlerer Schwierigkeit eingesetzt werden. Bei gleicher Testzeit hat CAT also gegenüber von PBT den Vorteil der höheren Reliabilität.

Zu den Nachteilen adaptiven Testens zählt der Aufwand zum Aufbau der notwendigen Itembank: Hier sind große Anstrengungen (a) für die Entwicklung und (b) Kalibrierung der psychometrischen Itemeigenschaften sowie der Normierung der Items erforderlich. Darüber hinaus besteht die Gefahr, dass sich - entgegen der bislang vorherrschenden Meinung - adaptives Testen nachteilig auf die Motivation zur Testbearbeitung auswirkt (Frey, Hartig & Moosbrugger, 2009). Bei gleicher Testzeit hat CAT also gegenüber von PBT den Nachteil der höheren Kosten für die Entwicklung der Itembank und eventuell der geringeren Validität, da über das Zielkonstrukt hinaus weitere psychologische Eigenschaften die Testleistung beeinflussen können.

4.4 Zusammenfassung

Im Rahmen von NEPS sollen der altersspezifische Status, aber auch generelle und differenzielle Entwicklungsverläufe von Gf erfasst werden (Blossfeld, 2008a, S. 55-56). Als psychometrische Kriterien zur Beurteilung von Instrumenten ziehen wir daher vor allem drei Kriterien heran: (1) Validität (Misst ein Test Gf, bzw. die Subkomponenten SF und WG oder nicht?), (2) interne Konsistenz als Maß der Reliabilität (Wie genau misst ein Test Gf, bzw. die Subkomponenten SF und WG?) und (3) empirische Evidenz zu Fragen der Messinvarianz (misst ein Test Gf, bzw. die Subkomponenten SF und WG in gleicher Art und Weise über die Zeit oder Altersgruppen hinweg?). Die Möglichkeiten und Grenzen zur Erfassung von Gf mittels CAT diskutieren wir in Abschnitt 7.

5. Welche Instrumente liegen vor?

In diesem Abschnitt geben wir einen groben Überblick über bereits entwickelte Instrumente zur Erfassung der fluiden Intelligenz über die Lebensspanne. Die in diesem Abschnitt diskutierten Instrumente stellen natürlich nur einen Bruchteil der national und international verfügbaren Instrumente dar und können hier aus Platzgründen auch nicht umfassend dokumentiert werden. Im Rahmen einer Recherche mit der einschlägigen Datenbank *PSYNDEX-plus Tests* (von 11. bis 13. März 2009) haben wir insgesamt 246 Instrumente zur Messung der Intelligenz identifiziert. Bei dieser beträchtlichen Zahl ist weiter zu bedenken, dass es

sich hier nur um die Instrumente handelt, die im deutschen Beschreibungsfeld (*MJG*) *„Intelligenztests“* enthielten.

Um die Einsatzmöglichkeiten dieser Intelligenztests auszuloten, haben wir in einem zweiten Schritt die Tests danach gruppiert, für welche Altersgruppen ihre Anwendung empfohlen ist. Mit *PSYNDEXplus Tests* haben wir vier Altersgruppen unterschieden: Vorschulalter *(„Preschool-Age“: 2-5 Jahre)*, Schulalter *(„School-Age“: 6-12 Jahre)*, Adoleszenz *(„Adolescence“: 13-17 Jahre)* und Erwachsene *(„Adulthood“: 18-64 Jahre)*. Die farblich hervorgehobenen Balken in Tabelle 2 stellen dar, ob Tests für eine Altersgruppe bzw. für mehrere Altersgruppen geeignet sind. Wir haben die Anzahl an Tests in Tabelle 2 eingetragen, die für eine bestimmte Altersgruppe bzw. über mehrere Altersgruppen hinweg verwendet werden können (dadurch ist die Gesamtzahl der in Tab. 2 eingetragenen Tests größer als 246).

Tabelle 2. Anzahl an deutschsprachigen Intelligenztests, die für eine oder mehrere Altersgruppen geeignet sind

Vorschulalter (2-5 Jahre)	Schulalter (6 - 12 Jahre)	Adoleszenz (13-17 Jahre)	Erwachsene (18-64 Jahre)	Anzahl Tests
■				16
	■			48
		■		125
			■	119
■	■			2
	■	■		7
		■	■	85
■	■	■		0
	■	■	■	1
■	■	■	■	0

Aus Tabelle 2 wird deutlich, dass die überwiegende Zahl aller verfügbaren Intelligenztests auf Adoleszenz und Erwachsenalter zielen. Weiterhin ist die Anzahl verfügbarer Tests für das Vorschulalter deutlich geringer als für alle anderen Altersgruppen.

Mit Blick auf altersgruppenübergreifende Verfahren gab es nur zwei Tests, die sowohl für das Vorschulalter und das Schulalter geeignet sind: Der Hannover-Wechsler-Intelligenztest für das Vorschulalter (HAWIVA-III; Ricken, Fritz, Schuck & Preuss, 2007) und die Kaufman Assessment Battery for Children in der deutschsprachigen Fassung (K-ABC; Melchers & Preuss, 1991).

Das wichtigste Rechercheergebnis war, dass es derzeit keinen verfügbaren (kommerziellen) Test gibt, der *Gf* bzw. die Subkomponenten *WG* und *SF* über eine Lebensspanne von 5 bis 67 Jahren erfasst. Der Test von *SF*, der für die meisten Altersgruppen geeignet erscheint, sind die Standard Progressive Matrices (Heller, Kratzmeier & Lengfelder, 1998). Matrizenaufgaben scheinen also mit am besten geeignet, um altersgruppenübergreifend *SF* zu erfassen. Im nächsten Recherche-Schritt haben wir die Zahl der detailliert zu diskutierenden Tests durch folgendes mehrstufiges Vorgehen deutlich reduziert.

(1) Wir haben uns auf Instrumente konzentriert, die im deutschen Sprachraum am häufigsten zur Diagnose von *Gf* eingesetzt werden (Amelang & Schmidt-Atzert, 2006). Die fünf am häufigsten eingesetzten Intelligenztests sind nach Schorr (1995) bzw. Steck (1997):

1. der Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK-IV, Petermann & Petermann, 2007),
2. der Wechsler-Intelligenztest für Erwachsene (WIE, von Aster, Neubauer & Horn, 2006),
3. Varianten der Raven's Progressive Matrices (CPM, Raven, Raven & Court, 2002; SPM, Heller, Kratzmeier & Lengfelder, 1998; APM, Raven, Raven & Court, 1998),
4. der Intelligenz-Struktur-Test (I-S-T 2000 R, Amthauer, Brocke, Liepmann & Beauducel, 2001) und
5. der Culture Fair Test (CFT 20-R, Weiß, 2006).

Wir stellen hier jeweils die aktuellste Testfassung vor, für die detaillierte Informationen in *PSYNDEXplus Tests* verfügbar waren. Für die *Raven Progressive Matrizen* beschränken wir uns in der Darstellung auf die CPM und SPM, da die APM speziell für die Diagnose kognitiver Leistungsfähigkeit im überdurchschnittlichen Bereich entwickelt wurden.

(2) Diese Liste an Tests haben wir komplettiert durch Instrumente, die in deutschen *large-scale-assessment*-Studien oder bedeutenden deutschen Längsschnittstudien zur Untersuchung der kognitiven Entwicklung verwendet wurden. Insbesondere wurden die folgenden Tests hinzugefügt:

- Der Subtest Figuren-Analogien N2 aus dem Kognitiven Fähigkeitstest (KFT 4-12+R, Heller & Perleth, 2000). Dieser Test wurde beispielsweise in der PISA-Studie, in der PALMA Studie und in ELEMENT eingesetzt.
- Die Columbia Mental Maturity Scale (CMMS, Bondy, Cohen, Eggert & Lueer, 1975), die z.B. in der LOGIK-Studie (Schneider, Bullock & Sodian, 1998) verwendet wurde.

- Der nicht-kommerzielle Zahlen-Zeichen-Test aus dem sozio-ökonomischen Panel SOEP (Lang 2005; Lang et al., 2007).
- Die nicht-kommerzielle Gerotest-Matrizen (Lang & Matthes, in Vorb.).

(3) Schließlich haben wir noch den Hannover-Wechsler-Intelligenztest für Kinder im Vorschulalter (HAWIVA-III, Ricken, Fritz, Schuck & Preuss, 2007) der Liste zu besprechender Tests hinzugefügt, da dies einer der wenigen Tests ist, der für das Vorschulalter und das Schulalter geeignet ist (s. Tabelle 2) und der sowohl einen Test für *WG* und Matrizentests zur Messung von *SF* enthält. Der HAWIVA wurde auch bereits in der LOGIK-Studie eingesetzt (allerdings nur die verbalen Aufgaben, vgl. Schneider, Bullock & Sodian, 1998). Den K-ABC haben wir nicht weiter betrachtet, da dieser Test keine Matrizentests zur Messung von *SF* beinhaltet.

Anschließend haben wir die so ausgewählten Tests danach eingeteilt, ob sie die Erfassung von *Gf* in einer oder mehreren von drei Altersgruppen erlauben. Im Einzelnen sind dies die Altersgruppen von (1) 5 Jahren, (2) 7-10 Jahren und (3) 14-67 Jahren. Diese Altersgruppen stellen unseres Erachtens jeweils spezifische Anforderungen bei der Durchführung von Tests zur Erfassung von *SF* oder *WG* in Rahmen von NEPS bzw. der Gestaltung und Schwierigkeitsanforderung der Testmaterialien. Diese Einteilung entspricht auch weitestgehend der Altersgruppeneinteilung in *PSYNDEXplus Tests.*

Für die ausgewählten Instrumente führen wir die folgenden Informationen in Tabelle 3 auf, falls diese in der Datenbank *PSYNDEXplus Tests* bzw. technischen Berichten vorhanden waren.

- Testname und Name des Subtests; Angaben darüber, ob der Test als Einzel- (ET) und/oder Gruppentest (GT) durchgeführt werden kann;
- Fähigkeit (F.K.), die der jeweilige (Sub-)Test erfasst (WG = Wahrnehmungsgeschwindigkeit; SF = Schlussfolgerndes Denken);
- Die notwendige Testzeit in Minuten;
- Eine Beurteilung, ob er für die Altersgruppen von NEPS geeignet ist (Spalten „geeignet für Altersgruppen“);
- Informationen zur internen Konsistenz für die jeweiligen Altersgruppen. Entsprechende Koeffizienten sind beispielsweise Cronbach’s Alpha, Kruder-Richardson oder Split-Half-Koeffizienten. Falls mehrere Angaben zur internen Konsistenz vorliegen (z.B. Cronbach’s Alpha für die Altersstufen 7, 8 und 9 Jahre), wird der Mittelwert über die angegebenen Werte berechnet;
- Informationen zur Retest-Reliabilität sowie die Zeitspanne zwischen erster und zweiter Messung in Monaten (letzere Information steht in Klammern). Hierbei ist zu beachten, dass wir die Retest-Reliabilität als die Korrelation zwischen Testwerten für zwei Messzeitpunkte, die wenige Monate auseinander liegen, definieren. Wir grenzen damit temporäre Stabilität von potenziellen, tatsächlichen Veränderungsprozessen einer Fähigkeit über längere Zeiträume ab;
- Validitätsnachweise in Form latenter Variablenmodelle (s. Abschnitt 4.1 zu Validität), die die theoretisch postulierte Struktur stützen. Latente Variablenmodelle können bei-

spielsweise exploratorische Faktorenmodelle, konfirmatorische Faktorenmodelle oder Modelle der Item Response Theorie (IRT) sein. Falls solche Validitätsnachweise vorliegen, wird dies für die jeweilige Altersgruppe angegeben.

Beim Lesen von Tabelle 3 ist zu bedenken, dass es zwei Arten von fehlenden Informationen gibt. (1) Informationen sind nicht relevant (z.B. ein Test ist nur für die Altersgruppe von Kindern mit 5 Jahren geeignet; die Spalten „7-10“, bzw. „14-67“ sind also nicht relevant für diesen Test): Dies wird durch „--“ vermerkt. (2) Relevante Informationen sind nicht verfügbar (z.B. ein Test ist nur für die Altersgruppe von Kindern mit 5 Jahren geeignet, es liegen uns aber keine Informationen zur internen Konsistenz für diese Altersgruppe vor): Dies wird durch „n.v.“ (= nicht verfügbar) vermerkt. Hierbei ist zu beachten, dass wir dieses Kriterium sehr strikt angewendet haben und „n.v.“ auch dann verwendeten, wenn beispielsweise summarisch die Spannbreite der Reliabilitäten für verschiedene Subtests angegeben wurde.

Auf Nachweise zur Messinvarianz gehen wir dann in Abschnitt 6 ein, wenn wir konkrete Empfehlungen für Instrumente zur Erfassung von *Gf* aussprechen.

Tabelle 3. Übersicht über Instrumente zur Erfassung der fluiden Intelligenz (sortiert nach Altersgruppen)

Test (ET / GT)	F.K.	Testzeit	geeignet für Altersgruppe			Interne Konsistenz			Stabilität (Zeitspanne)			Validitätsnachweis		
		(Minuten)	5	7-10	14-67	5	7-10	14-67	5	7-10	14-67	5	7-10	14-67
CMMS (ET)	SF	10-30	ja	Ja	nein	.94	.86	--	.60 (12)	n.v.	--	nein	nein	--
HAWIVA-III: Symbol-Suche (ET)	WG	2	ja	nein	nein	n.v.	--	--	.75 (n.v.)	--	--	ja	--	--
HAWIVA-III: Matrizentest (ET)	SF	n.v.	ja	nein	nein	.80	--	--	n.v.	--	--	ja	--	--
Raven – CPM (ET & GT)	SF	20-90	ja	Ja	nein	n.v.	n.v.	--	n.v.	n.v.	--	n.v.	n.v.	--
HAWIK-III: Zahlen-Symbol-Test (ET)	WG	2	nein	Ja	nein	--	.85	--	--	n.v.	--	--	ja	--
HAWIK-IV: Zahlen-Symbol-Test (ET)	WG	n.v.	nein	Ja	nein	--	n.v.	--	--	n.v.	--	--	n.v.	--
HAWIK-IV: Matrizentest (ET)	SF	n.v.	nein	Ja	nein	--	n.v.	--	--	n.v.	--	--	n.v.	--
Raven – SPM (ET & GT)	SF	60	nein	Ja	ja	--	n.v.	n.v.	--	n.v.	n.v.	--	n.v.	n.v.
CFT 20-R: Matrizen (ET & GT)	SF	12	nein	Ja	ja	--	n.v.	n.v.	--	n.v.	n.v.	--	n.v.	n.v.
KFT 4-12+R: Figurenanalogien N2 (ET & GT)	SF	15[a]	nein	Ja	ja	--	n.v.	.91[a]	--	n.v.	n.v.	--	ja	ja
WIE: Zahlen Symbol Test (ET)	WG	2	nein	nein	ja	--	--	.84	--	--	n.v.	--	--	ja
WIE: Matrizentest (ET)	SF	n.v.	nein	nein	ja	--	--	.92	--	--	n.v.	--	--	ja
IST 2000 R: Matrizen (ET & GT)	SF	10[b]	nein	nein	ja	--	--	.71[a]	--	--	n.v.	--	--	ja
Gerotest Matrizen (ET)	SF	10-20	nein	nein	ja	--	--	.74	--	--	.85 (n.v.)	--	--	ja
SOEP: Zahlen-Zeichen-Test (ET)	WG	3	nein	nein	ja	--	--	.87	--	--	.55 (1,5)	--	--	ja

Anmerkungen. ET / GT = Einzel- (ET) und/oder Gruppentest (GT); F.K. = erfasste Fähigkeit; WG = Wahrnehmungsgeschwindigkeit; SF = Schlussfolgerndes Denken; Validitätsnachweis: Nachweis, dass ein latentes Variablenmodell (z.B. Faktorenanalyse, oder Item Response Modell) die Kovariation zwischen manifesten Testindikatoren erklären kann. n.v. = Information war nicht verfügbar (z.B. weil die entsprechenden Testmanuale nicht an unseren Forschungsinstituten vorlagen, bzw. nicht mehr rechtzeitig zur Ansicht bestellt werden konnten). -- = nicht relevant, da der Test nicht für diese Altersgruppe geeignet ist. Stabilität: temporäre Stabilität; die Zeitspanne zwischen erster und zweiter Messung wird in Monaten angegeben.

a: Angaben aus Brunner (2008). b: Angaben aus Amelang und Schmidt-Atzert (2006, S. 215).

Die wichtigsten Ergebnisse, die wir bei der Recherche für diesen Abschnitt unserer Expertise und bei der Erstellung von Tabelle 2 und 3 gewinnen konnten, fassen wir folgendermaßen zusammen:

- Die meisten Intelligenztests erfassen Gf in Form von SF.
- Nahezu alle Intelligenztests enthalten Matrizenaufgaben zur Messung von SF. Dies steht auch damit im Einklang, dass viele Experten Matrizentests als Standardmarker von Gf betrachten (Mackintosh, 1998). Da die vorgesehene Testzeit in NEPS zur Erfassung von Gf nur den Einsatz von ein oder maximal zwei Tests ermöglicht, haben wir deshalb in Tabelle 3 die uns verfügbaren Informationen zu den Matrizentests aufgeführt, die in der jeweiligen Testbatterie enthalten sind.
- Trotz des Vorteils, dass die SPM für ein breites Altersspektrum empfohlen werden, besteht ein Problem der SPM darin, dass dieser Test anscheinend zu viele leichte Items enthält. Dies war sicherlich auch einer der Gründe, weshalb die APM entwickelt wurden, um auch im oberen Leistungsbereich reliabel differenzieren zu können.
- Nur die Tests aus der Wechslerfamilie (HAWIVA-III, HAWIK-III/IV und WIE) enthalten Aufgaben zur Erfassung von SF und von WG.
- Häufig fehlten Angaben zu Bearbeitungszeiten pro Subtest; es lagen nur Informationen für die gesamte Testdauer vor.
- Altersgruppen- und subtestspezifische Informationen zur internen Konsistenz, zeitlichen Stabilität (Retest-Reliabilität) sowie Validitätsnachweise waren in der Datenbank PSYNDEXplus Tests nur für wenige Subtests verfügbar. Meist wurde nur summarisch die Spannbreite der Reliabilitäten für verschiedene Subtests oder die Reliabilität für aggregierte Skalen angegeben, zu denen diese Subtests gehörten. Die in dieser Form dargestellten internen Konsistenzen wie auch die Retest-Reliabilitäten lagen aber in einem für Forschungszwecke akzeptablen Bereich, d.h. interne Konsistenzen lagen meist um .70 oder höher; Retest-Stabilitäten um .50 oder höher. Prinzipiell kann somit ausgeschlossen werden, dass die Reliabilität der in Tabelle 3 aufgelisteten Instrumente nicht für Forschungszwecke ausreicht.

Auf Grundlage dieser Erkenntnisse entwickeln wir im nächsten Abschnitt zwei Empfehlungen für die Erfassung von Gf im Rahmen von NEPS.

6. Welche Instrumente werden für die einzelnen Altersgruppen empfohlen?

Im Idealfall sollten im Rahmen von NEPS Gf und deren Subkompetenzen SF und WG messinvariant, reliabel und valide über die Alterspanne von 4 bis 67 Jahren mittels von 15 bis 30 Minuten Testzeit am besten in Form von Gruppentests erfasst werden.

Wichtig ist hierbei zu betonen, dass aus konzeptioneller aber auch psychometrischer Sicht jeweils ein Indikator von SF und WG im Rahmen von NEPS erfasst werden sollte: (1) SF und WG liegen beide im konzeptionellen Zentrum von Gf. (2) SF und WG erfassen jeweils unterschiedliche Facetten von Gf (vgl. Brunner & Süß, 2005). (3) Zwei manifeste Indikatoren sind minimal erforderlich, um Gf als latente Variable zu erfassen; so werden auch die Varianz von Gf und subtestspezifische Varianzen getrennt voneinander modelliert (vgl. Bollen, 1989). Wir

entwickeln nachfolgend zwei Empfehlungen, die diesen Anforderungen weitestgehend nachkommen und diskutieren deren Vor- und Nachteile.

6.1 Empfehlung 1: Verwendung der Tests aus der „Wechsler-Familie“

In Abschnitt 5 haben wir resümiert, dass es derzeit kein Einzelinstrument gibt, dass *Gf* für 5- bis 67-Jährige in gleicher Art und Weise erfasst. Allerdings kommt diesem Ziel ein kombinierter Einsatz der Tests zur Erfassung von *WG* und *SF* aus den Intelligenztestbatterien HAWIVA-III, HAWIK-IV und dem WIE nahe: *WG* wird dann durch die Tests Symbol-Suche (aus dem HAWIVA-III) und dem Zahlen-Symbol-Test (aus dem HAWIK-IV und dem WIE) gemessen. *SF* wird in allen drei Testbatterien durch Matrizentests erfasst. Tabelle 4 enthält eine Beschreibung dieser Tests zur Erfassung von *SF* und *WG*.

Tabelle 4. Beschreibung der Aufgaben zur Erfassung von WG und SF, die in den Wechslertests enthalten sind

Testbeschreibung
HAWIVA-III
Symbol-Suche (WG). Der Test besteht aus 50 Items, die Bearbeitungszeit beträgt zwei Minuten. Bei jedem Item wird ein Symbol vorgegeben. Dieses Symbol soll unter drei nachfolgenden Symbolen erkannt und markiert werden. Fehlt es, soll das am Ende stehende Fragezeichen angekreuzt werden.
Matrizentest (SF). Der Test enthält 17 Items, die jeweils aus einer unvollständigen 2x2-Matrix bestehen, deren Inhalt in einem sachlogischen Zusammenhang steht. Es muss aus vier, bzw. fünf Möglichkeiten die richtige Lösung ausgewählt werden.
HAWIK-IV
Zahlen-Symbol-Test (WG). Einfach strukturierte geometrische Formen (Symbole A) oder Ziffern (Symbole B) sind abstrakten Symbolen zugeordnet. Das Kind soll nach dem Zuordnungsschlüssel die abstrakten Symbole in die geometrischen Formen bzw. unter die Ziffern zeichnen (Speed-Test mit Zeitbegrenzung von 120 Sekunden). Symbole A werden den 6- und 7-jährigen Kindern, Symbole B den älteren Kindern vorgegeben. Es liegen sieben Beispielaufgaben und 59 Aufgaben vor.
Matrizentest (SF). Der Matrizentest enhält 35 Items, die in vier Kategorien eingeteilt werden können: Vervollständigen von Mustern, Klassifizieren, Analoges Schlussfolgern und Fortsetzen von Reihen.
WIE
Zahlen-Symbol-Test (WG). Nach einer vorgegebenen Zahlen-Zeichen-Zuordnung sind unter einer Reihe von Zahlen unter Zeitbegrenzung die entsprechenden abstrakten Zeichen zu notieren. Es stehen 120 Sekunden Testzeit zur Verfügung.
Matrizentest (SF). Der Test enthält 26 Items. Die Testperson muss das Konstruktionsprinzip von Serien geometrischer Muster erkennen und die Serien durch Auswahl eines von fünf vorgegebenen Mustern ergänzen.

Die internen Konsistenzen (IK) der Subtests, die altersgruppenübergreifend angegeben wurden, erscheinen für Forschungszwecke mehr als ausreichend (HAWIVA-III: IK_{WG} = .75, IK_{SF} = .80; HAWIK-III: IK_{WG} = .85, IK_{SF}: noch keine Angabe, da dieser Test erst im HAWIK-IV aufgenommen wurde; WIE: IK_{WG} = .84, IK_{SF} = .92). Sofern Retest-Reliabilitäten angegeben wurden, waren diese für Subtests zur Erfassung von SF und WG mit minimalen Werten von .48 ausreichend.

Für alle Subtests der Wechslerfamilie liegen Validitätsnachweise in Form von Faktorenanalysen vor. Wichtig ist hierbei, dass auf Datengrundlage der jeweiligen Normierungsstichproben Studien aus dem anglo-amerikanischen Raum die Annahme der konfiguralen Invarianz für die amerikanische Testversion des HAWIVA-III (Blaha & Wallbrown, 1991) sowie der metrischen Invarianz für den HAWIK-III (Keith, Goldenring Fine, Taub, Reynolds & Kranzler, 2006) und den WIE (Bowden, Weiss, Holdnack & Lloyd, 2006; Taub, McGrew & Witta, 2004) über Altersgruppen hinweg stützen. Diese Ergebnisse untermauern also den Einsatz der Wechslertests zur Erfassung von Gf über die Lebensspanne.

Generell werden üblicherweise alle Intelligenztests der Wechslerfamilie als Einzeltests vorgegeben. Wir folgen dieser Empfehlung insoweit, dass Kinder mit 5 und 7 Jahren Aufgaben zur Erfassung von Gf im Rahmen von Einzeltestsitzungen bearbeiten sollten. Allerdings sollten Kinder ab 9 Jahren in der Lage sein, Tests zur Messung von Gf in Form von Gruppentests zu bearbeiten. Dieser Schluß liegt auch nahe, da beispielsweise der CFT 20-R, der ähnliche kognitive Anforderungen stellt, schon bei Kindern ab 8 Jahren als Gruppentest eingesetzt werden kann. Allerdings ist eventuell in diesem Fall eine geringe Modifikation der Testinstruktionen notwendig.

Ebenso sollte in Pilotierungsstudien die notwendige Testzeit für die Durchführung der Tests zur Erfassung von WG und SF mittels der Wechslertests untersucht werden. Während die Bearbeitungszeit der „Speed"-Tests (WG) meist recht präzise auf wenige Minuten eingeschränkt bleibt, sind die Bearbeitungszeiten von „Reasoning"- bzw. Matrizentests recht unterschiedlich und liegen erfahrungsgemäß je nach Version und je nach Umfang der verwendeten Einzelaufgaben zwischen 30 und 90 Minuten. Allerdings sind hier Kürzungen gut möglich.

Trotz der aufgezeigten Vorzüge bei einem Einsatz der „Wechslertest-Familie" ergibt sich ein gravierender Nachteil: Für eine large-scale-Studie wie NEPS werden hohe Kosten durch die zu zahlenden Lizenzgebühren fällig. Vor allem aus diesem Grund sprechen wir noch eine zweite Empfehlung aus, wie man Gf im Rahmen von NEPS messen kann.

6.2 Empfehlung 2: Weiterentwicklung nicht-kommerzieller Instrumente

Wahrnehmungsgeschwindigkeit: Der Zeichen-Zahlen-Test (SOEP). Ein verbreiteter und gut etablierter Indikator der Wahrnehmungsgeschwindigkeit (*WG*) ist der Zahlen-Zeichen-Test aus dem HAWIE-R (Tewes, 1994). Bei diesem Test geht es um eine möglichst voraussetzungsarme (z.B. geringe Gedächtniseffekte) und kontextfreie Zuordnung von unbekannten, nicht vertrauten Zeichen zu Zahlen nach einer jederzeit sichtbaren Zuordnungstabelle. Eine vereinfachte Abwandlung dieses Tests ist der Symbol-Digit-Modalities-Test (SDMT; Smith 1973/1995), bei dem nicht die Zeichen der Zahl sondern die korrekten Zahlen den Zeichen zugeordnet werden. Der Test hat sich für den Altersbereich ab 16 Jahren als robuster Schätzer der *Gf* erwiesen und ist breit anwendbar. Dabei können durch minimale Änderungen der jeweils zugeordneten Zeichen mögliche Lerneffekte weitgehend ausgeschlossen werden.

Eine Variante dieses Zeichentests wurde im Rahmen des SOEP zur Erfassung der Wahrnehmungsgeschwindigkeit erfolgreich eingesetzt (Lang, 2005; Lang et al., 2007). Eine Anpassung des Tests für Papier-Bleistift-Erhebungen in Einzel oder Gruppentesterhebungen ist im Prinzip vorstellbar, bedarf aber sicherlich eines gewissen Entwicklungsaufwandes. Ein exemplari-

sches Beispiel dieses Zeichentests, wie er im Rahmen des SOEP eingesetzt wurde, findet sich im Anhang.

Schlussfolgerndes Denken (SF). Bislang liegen noch keine Kurzinstrumente für figurales schlussfolgerndes Denken („*progressive Matrizen*“) vor, bei denen SF in einem zeitökonomischen, robusten und validen Verfahren erfasst werden kann. Vorstellbar wäre aber eine Entwicklung von altersgerechten Matrizenaufgaben, die auf einer möglichst breiten Heterogenität verschiedener Aufgabenmerkmale beruht (z.B. mentale Rotation, Erkennung von Konstanz, Regelerkennung). Dabei könnten die jeweiligen Aufgabenmerkmale altersspezifisch variiert werden, um eine Vergleichbarkeit über verschiedene Altersgruppen hinweg zu erzielen. Im Rahmen des Forschungsprojekts *Gerotest* am Instituts für Psychogerontologie der Friedrich-Alexander-Universität Erlangen-Nürnberg wurde ein ultrakurzer Matrizentest (*Gerotest-Matrizen*) entwickelt, der sich in einem Altersbereich von 20 bis 80 Jahren bereits gut bewährt hat (Lang & Matthes, in Vorb.). Die Bearbeitungszeit liegt für 20 Einzelaufgaben bei rund 15 Minuten, variiert aber altersspezifisch. Eine Beispielaufgabe der Gerotest-Matrizen ist im Anhang abgebildet.

Für den Einsatz im Rahmen einer altersheterogenen Panelstudie wäre eine daran anknüpfende Weiterentwicklung von Items möglich, bei der insbesondere für die jüngeren Altersgruppen noch entsprechende leichtere Aufgaben (z.B. zu Addition, Subtraktion, Konstanz, Regelerkennung) mit einem geringeren Anteil von Distraktor-Anforderungen entwickelt werden müssten. Dabei wäre auf eine möglichst einfache Abänderbarkeit der Aufgaben zu achten, um mögliche Lern- und Testungseffekte im Falle einer Messwiederholung ausschließen zu können.

Insgesamt kann die Entwicklung solcher schwierigkeitsadaptierter Items entscheidend durch die Berücksichtigung der Literatur zum automatisierten Erstellen von Testitems durch sogenannte Itemgeneration-Rules profitieren (Arendasy & Sommer, 2005; Arendasy, Sommer, Gittler & Hergovich, 2006; Embretson, 1998; Holzman, Pellegrino & Glaser, 1982). Eine Übersicht über bereits existierende Instrumente geben Arendasy et al. (2008, Tabelle 1).

7. Welche Möglichkeiten bestehen für computergestütztes adaptives Testen?

Die Erfassung von *WG* und *SF* ist nicht an ein bestimmtes Testmedium gebunden. Falls man im Rahmen von NEPS die Vorteile von CAT nutzen möchte, dann sollte man unseres Erachtens versuchen, in allen Kohorten und Altersstufen CAT einzusetzen. Diese Designüberlegung maximiert a priori die Vergleichbarkeit der erhobenen Daten zwischen Kohorten und Altersstufen.

Andererseits verstehen wir den derzeitigen Designvorschlag von NEPS so, dass die Teilnehmer/-innen der verschiedenen Studienkohorten die Tests in unterschiedlichen Testmodi bearbeiten werden (vgl. Blossfeld, 2008a, S. 81): Während Kinder und Schüler/-innen in Gruppensitzungen überwiegend PBT bearbeiten, werden Erwachsene an Einzeluntersuchungen

teilnehmen, bei denen der Einsatz von Computern geplant ist und sich somit der Einsatz von CAT anbietet.[7]

Diesem Design stehen wir etwas kritisch gegenüber. Warum? Erstens, bei Wechsel des Administrationsmodus (Computer vs. PBT) gehen in die beobachteten Testleistungen individuelle Unterschiede, die auf tatsächliche Fähigkeitsunterschiede zurückgehen, und individuelle Unterschiede, die auf die computerbasierte Testadministration zurückgehen (z.B. Computerängstlichkeit, Computererfahrung), konfundiert ein. Generell ist also a priori davon auszugehen, dass immer, wenn Tests in zwei verschiedenen Administrationsmodi vorgegeben werden, zumindest von kleinen Unterschieden zwischen den beiden Modi auszugehen ist (Clariana & Wallace, 2002; Kolen & Brennan, 2004; Mead & Drasgow, 1993). Es ist eine offene Frage, ob die anvisierten Vergleichbarkeitsstudien (Blossfeld, 2008a, Tabelle 7.7, S. 81) tatsächlich die Vergleichbarkeit von computerbasierten Tests und PBT für alle Kohorten und alle untersuchten Alterstufen gewährleisten können.

Zweitens, durch die gewählten Testmodi beeinflussen zusätzlich zu den Zielkonstrukten modusspezifische individuelle Unterschiede die Leistung bei den Tests. Diese Erkenntnis motivierte die Multi-Trait-Multi-Method-Forschung (MTMM) (Campbell & Fiske, 1959) und die Entwicklung entsprechender psychometrischer Modelle (z.B. Eid, Lischetzke, Nussbeck & Trierweiler, 2003). Falls also zum Beispiel *Gf* mit CAT und domänenspezifische Kompetenzen mit PBT erhoben werden, werden die korrelativen Zusammenhänge zwischen *Gf* und domänenspezifischen Kompetenzen geringer ausfallen als bei Verwendung des gleichen Testadministrationsmodus. Falls dies (a) bei der Interpretation der Ergebnisse, (b) der Kommunikation der Ergebnisse gegenüber der Politik oder Presse wie auch (c) der Rezeption der Ergebnisse durch die Politik oder Presse nicht beachtet wird, besteht die große Gefahr, dass inhaltlich falsche Schlussfolgerungen gezogen werden.

Unabhängig von diesen Überlegungen ist es angesichts der ausgesprochenen Empfehlungen jedoch fraglich, ob der Vorteil von CAT in Form der höheren Messeffizienz eingelöst werden kann. Die Anzahl der Items zur Erfassung von SF und WG in den Wechslertests ist relativ gering. Wenn die nicht-kommerziellen Tests weiterentwickelt werden, sind sehr große Anstrengungen nötig, um eine ausreichende Anzahl an Items mit befriedigender psychometrischer Qualität zu erstellen und die erforderlichen Itembanken aufzubauen. Schließlich erfordert der Einsatz von CAT insbesondere in Kindergärten und Grundschulen, dass mobile Computerlabors angeschafft und gewartet werden müssen.

Zusammenfassend ist also festzuhalten, dass CAT sicherlich aus theoretischer Sicht viele Vorteile hat. Jedoch stellen sich in der praktischen Anwendung Probleme, die nur mit großem finanziellem und zeitlichem Aufwand zu lösen sind.

[7] Die Möglichkeit und Vergleichbarkeit von computergestützten Testungen wird im Rahmen von NEPS systematisch überprüft. Die ersten Erhebungswellen erfolgten für alle Alterskohorten (abgesehen von Kohorte 1) als PBT.

8. Fazit

Die vorliegende Expertise untersuchte die übergreifende Frage, ob relative robuste, nonverbale Indikatoren der fluiden Intelligenz bei Kindern, Jugendlichen und bei Erwachsenen in entwicklungs- und messäquivalenter Weise angemessen und zeitökonomisch im Rahmen von NEPS erfasst werden können. Als Indikatoren von *Gf* schlagen wir non-verbale Messinstrumente zur Erfassung der Wahrnehmungsgeschwindigkeit und zum schlussfolgernden Denken vor: Diese Indikatoren liegen im konzeptionellen Zentrum der fluiden Intelligenz und es wurde wiederholt empirisch gezeigt, dass *WG* und *SF* einen bedeutsamen Beitrag für eine erfolgreiche Entwicklung leisten. Zur Evaluation bestehender Instrumente haben wir die psychometrischen Kriterien Validität, Reliabilität und Messinvarianz herangezogen. Wir haben uns dabei auf Instrumente zur Erfassung von *Gf* konzentriert, die am häufigsten im deutschen Sprachraum in der Individualdiagnostik eingesetzt werden oder sich in large-scale-Studien und bedeutsamen Längsschnittstudien bewährt haben.

Auf Grundlage dieser Evaluation sprachen wir zwei Empfehlungen aus, die jeweils spezifische Vor- und Nachteile haben: (1) Bei Verwendung der Wechslertest-Familie wird ein national und international vielfach bewährtes Instrument der Individualdiagnostik gewählt. Es ist hierbei keine Neuentwicklung notwendig, jedoch werden relativ hohe Lizenzgebühren für die Verwendung der Wechslertests in NEPS *anfallen*. (2) Die Verwendung und Weiterentwicklung bestehender, nicht-kommerzieller Tests zur Erfassung von *Gf*, nämlich *des Zeichen-Zahlen-Tests* aus dem SOEP und des *Gerotests.* Folgt man diesem Vorschlag, fallen zwar einmalig größere Kosten für die Weiterentwicklung dieser Instrumente an, jedoch sind anschließend keine Lizenzgebühren mehr zu entrichten. Unseres Erachtens werden beide Empfehlungen darin münden, *Gf* bzw. *SF* und *WG* valide und reliabel über die Lebensspanne von 5 bis 67 Jahren zu erheben.

Gleich welcher Empfehlung gefolgt wird, wollen wir unterstreichen, dass wir die Erfassung von Gf im Rahmen von NEPS ausdrücklich befürworten und begrüßen: Dadurch wird das Spektrum der erhobenen kognitiven Kompetenzen stark bereichert. Gleichzeitig haben wir aber mit Blick auf das Untersuchungsdesign von Gf im Rahmen von NEPS auch gewisse Bedenken.

In der empirischen Bildungsforschung und pädagogischen Psychologie wird häufig die Intelligenz, die sich weitgehend situationsunabhängig und bereichsübergreifend manifestiert, von schulischen Kompetenzen abgegrenzt, unter denen in Anlehnung an F.E. Weinert kontextspezifische Leistungsdispositionen verstanden werden, die das Resultat von spezifischen Lern- und Erwerbsprozessen darstellen (Baumert, Lüdtke, Trautwein & Brunner, 2009). Während in NEPS eine regelmäßige Erhebung der schulischen Kompetenzen geplant ist, geht aus dem Untersuchungsdesign hervor, dass für die meisten Kohorten nur eine einmalige Erfassung der Intelligenz vorgesehen ist. Dabei gilt es zu bedenken, dass auch die fluide Intelligenz differenziellen Fördereffekten unterliegt und somit nicht als eine zeitlich stabile „Kovariate" über längere Erhebungszeiträume angesehen werden darf. So ist z.B. gut bestätigt, dass die *Quantität* der Beschulung einen positiven Effekt auf die Intelligenzentwicklung besitzt (siehe z.B. Ceci, 1991). Es liegen inzwischen aber auch Befunde vor, die für einen positiven Effekt der *Qualität* der Beschulung sprechen (Becker, 2008). Allerdings gilt zu berücksichtigen, dass für den deutschsprachigen Raum so gut wie keine längsschnittlichen Studien vorliegen, in denen eine mehrfache Erfassung der Intelligenz vorgenommen wurde (Ausnahmen: LOGIK

und BIJU; siehe Übersicht in Becker, 2008, S. 43). Das NEPS könnte hier noch einen wichtigen Beitrag zum besseren Verständnis der kognitiven Entwicklung über die Lebensspanne in Abhängigkeit von institutionellen Lernumwelten liefern. Aus diesem Grund finden wir eine Ausweitung der geplanten Erhebungszeitpunkte von *Gf* wünschenswert.

9. Literatur

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. Journal of Experimental Psychology: General, 117, 288-318.

Ackerman, P. L. (1989). Individual differences and skill acquisition. In P. L. Ackerman, R. J. Sternberg & R. Glaser (Eds.), Learning and individual differences. Advances in theory and research (pp. 165-217). New York, NY: W. H. Freeman and Company.

Ackerman, P. L. & Kanfer, R. (1993). Integrating laboratory and field study for improving selection: Development of a battery for predicting air traffic controller success. Journal of Applied Psychology, 78, 413-432.

Ackrill, J. L. (1985). Aristoteles: Eine Einführung in sein Philosophieren. Berlin. Walter de Gruyter.

Amelang, M. & Schmidt-Atzert, L. (2006). Psychologische Diagnostik und Intervention (4., vollständig überarbeitete und erweiterte Auflage). Heidelberg: Springer.

American Educational Research Association, A. P. A., National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: Author.

Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R). Göttingen: Hogrefe

Anderson, J. R. (1993). Problem solving and learning. American Psychologist, 48, 35-44.

Arendasy, M., Hergovich, A. & Sommer, M. (2008). Investigating the 'g'-saturation of various stratum-two factors using automatic item generation. Intelligence, 36, 574-583.

Arendasy, M. & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. Intelligence, 33, 307-324.

Arendasy, M., Sommer, M., Gittler, G. & Hergovich, A. (2006). Automatic generation of quantitative reasoning items. A pilot study. Journal of Individual Differences, 27, 2-14.

Artelt, C., Weinert, S. & Carstensen, C. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. Journal for Educational Research Online, 5(2), 5-14.

Baltes, P. B., Staudinger, U. M. & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. Annual Review of Psychology, 50, 471-507.

Batty, G. D., Deary, I. J. & Gottfredson, L. S. (2007). Premorbid (early life) IQ and later mortality risk: Systematic review. Annals of Epidemiology, 17, 278-288.

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,* 165-176.

Becker, M. (2008). Kognitive Leistungsentwicklung in differenziellen Lernumwelten: Effekte des gegliederten Sekundarschulsystems in Deutschland. Dissertation, Freie Universität Berlin.

Bertua, C., Anderson, N. & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. Journal of Occupational and Organizational Psychology, 78, 387-409.

Blaha, J. & Wallbrown, F. H. (1991). Hierarchical factor structure of the Wechsler Preschool and Primary Scale of Intelligence - Revised. Psychological Assessment, 3, 455-463.

Blossfeld, H.-P. (2008a). A proposal for a national educational panel study (NEPS) in Germany. Part A: Overview. Bamberg: Bamberg University.

Blossfeld, H.-P. (2008b). A proposal for a national educational panel study (NEPS) in Germany. Part B: Theories, operationalizations and piloting strategies for the proposed measurements. Bamberg: Bamberg University.

Blossfeld, H.-P., Roßbach, H.-G. & von Maurice, J. (Hrsg.). (2011). Education as a lifelong process. The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, Sonderheft 14.

Blossfeld, H.-P., von Maurice, J. & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. Zeitschrift für Erziehungswissenschaft,Sonderheft 14, 5-17.

Bollen, K. A. (1989). Structural equations with latent variables. New York, NY: John Wiley & Sons.

Bondy, C., Cohen, R., Eggert, D. & Lueer, G. (1975). Columbia Mental Maturity Scale (CMMS). Weinheim: Beltz.

Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061-1071.

Bowden, S. C., Weiss, L. G., Holdnack, J. A. & Lloyd, D. (2006). Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale-III. Psychological Assessment, 18, 334-339.

Brunner, M. (2008). No g in education? Learning and Individual Differences, 18, 152-165.

Brunner, M. & Süß, H.-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. Educational and Psychological Measurement, 65, 227-240.

Campbell, D. T. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York, NY: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. Journal of Educational Psychology, 54, 1-22.

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

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. Developmental Psychology, 27, 703-722.

Clariana, R. & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. British Journal of Educational Technology, 33, 593-602.

Deary, I. J., Strand, S., Smith, P. & Fernandes, C. (2007). Intelligence and educational achievement. Intelligence, 35, 13-21.

Deary, I. J., Taylor, M. D., Hart, C. L., Wilson, V., Smith, G. D., Blane, D. et al. (2005). Intergenerational social mobility and mid-life status attainment: Influences of childhood intelligence, childhood social factors and education. Intelligence, 33, 455-472.

Eid, M., Lischetzke, T., Nussbeck, F. W. & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. Psychological Methods, 8, 38-60.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. Psychological Methods, 3(3), 380-396.

Frey, A. (2007). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (S. 261-278). Heidelberg: Springer.

Frey, A. & Ehmke, T. (2008). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. Zeitschrift für Erziehungswissenschaft, Sonderheft 8, 169-184.

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. Diagnostica, 55, 20-28.

Ghisletta, P. & Lindenberger, U. (2003). Age-based structural dynamics between perceptual speed and knowledge in the Berlin Aging Study: Direct evidence for ability dedifferentiation in old age. Psychology and Aging, 18, 696-713.

Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history and bibliography. Intelligence, 24, 13-23.

Gottfredson, L. S. (1997b). Why g matters: The complexity of everyday life. Intelligence, 24, 79-132.

Gottfredson, L. S. (2004). Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? Journal of Personality and Social Psychology, 86, 174-199.

Gottfredson, L. S. & Deary, I. J. (2004). Intelligence predicts health and longevity, but why? Current Directions in Psychological Science, 13, 1-4.

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Gustafsson, J. E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 4, pp. 35-72). Hillsdale, NJ: Erlbaum.

Gustafsson, J. E. & Balke, G. (1993). General and specific abilities as predictors of school achievement. Multivariate Behavioral Research, 28, 407-434.

Heller, K. A., Kratzmeier, H. & Lengfelder, A. (1998). Matrizen-Test-Manual, Band 1. Ein Handbuch mit deutschen Normen zu den Standard Progressive Matrices von J.C. Raven. Göttingen: Beltz.

Heller, K. A. & Perleth, C. (2000). Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision (KFT 4-12+R). Manual. Göttingen: Hogrefe.

Helmke, A. & Weinert, F. E. (1997). Bedingungsfaktoren schulischer Leistungen. In F. E. Weinert (Hrsg.), Psychologie des Unterrichts und der Schule (Enzyklopädie der Psychologie D/I/3, S. 71-176). Göttingen: Hogrefe.

Holzman, T. G., Pellegrino, J. W. & Glaser, R. (1982). Cognitive dimensions of numerical rule induction. Journal of Educational Psychology, 74, 360-373.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. Experimental Aging Research, 18, 117-144.

Horn, J. L. & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (pp. 53-91). New York, NY: The Guilford Press.

“Intelligence and its measurement”: A symposium. (1921). Journal of Educational Psychology, 12, 123–147, 195–216, 271–275.

Irwing, P. & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. British Journal of Psychology, 96, 505-524.

Jensen, A. R. (1998). The g factor. The science of mental ability. Westport: Praeger.

Keith, T. Z., Goldenring Fine, J., Taub, G. E., Reynolds, M. R. & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children - Fourth Edition: What does it measure? School Psychology Review, 35, 108-127.

Kolen, M. J. & Brennan, R. L. (2004). Test equating, scaling, and linking. Methods and practices (second edition). New York, NY: Springer.

Kuncel, N. R., Hezlett, S. A. & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? Journal of Personality and Social Psychology, 86, 148-161.

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-S. Berlin: Deutsches Institut für Wirtschaftsforschung.

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 (NEPS Working Paper No. XX). Bamberg: Leibniz-Institut für Bildungsverläufe, Nationales Bildungspanel.

Lang, F. R. & Matthes, B. (2009, in Vorb.). Ultrakurztests zur Erfassung des figuralen schlussfolgernden Denkens: Technischer Bericht zu den Gerotest-Matrizen. Erlangen: IPG research notes.

Lang, F. R., Weiss, D., Stocker, A. & von Rosenbladt, B. (2007). Assessing cognitive capacities in computer-assisted survey research: Two ultra-short tests of intellectual ability in the Germany Socio-Economic Panel (SOEP). Schmollers Jahrbuch. Journal of Applied Social Science Studies, 127, 183-192.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman´s (1904) "'General intelligence', objectively determined and measured". Journal of Personality and Social Psychology, 86, 96-111.

Lubke, G. H., Dolan, C. V., Kelderman, H. & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. Intelligence, 31, 543-566.

Mackintosh, N. J. (1998). IQ and human intelligence. Oxford: Oxford University Press.

Mayer, R. E. (1992). Thinking, problem solving, cognition. New York, NY: W. H. Freeman and Company.

Mayer, R. E. (2000). Intelligence and education. In R. J. Sternberg (Ed.), Handbook of intelligence (pp. 519-533). Cambridge, UK: Cambridge University Press.

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449-458.

Meijer, R. R. & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. Applied Psychological Measurement, 23(3), 187-194.

Melchers, P. & Preuss, U. (1991). Kaufman Assessment Battery for Children (K-ABC) von Alan S. Kaufman und Nadeen L. Kaufman. Deutschsprachige Fassung. Amsterdam: Swets & Zeitlinger.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58, 525-543.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York, NY: American Council on Education and Macmillan Publishing Company.

Millsap, R. E. & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. Multivariate Behavioral Research, 39, 479-515.

Neisser, U., Boodoo, G., Bouchard Jr., T. J., Boykin, A. W., Brody, N., Ceci, S. J. et al. (1996). Intelligence: Knowns and unknowns. American Psychologist, 51, 77-101.

Petermann, F. & Petermann, U. (2007). Hamburg-Wechsler-Intelligenztest für Kinder - IV (HAWIK-IV). Übersetzung und Adaptation der WISC-IV von David Wechsler. Bern: Huber.

Raven, J. C., Raven, J. & Court, J. H. (1998). Raven's Progressive Matrices and Vocabulary Scales (herausgegeben von Hartmut Haecker und Stephan Bulheller). Advanced progressive matrices (APM). Frankfurt: Swets.

Raven, J. C., Raven, J. & Court, J. H. (2002). Raven's Progressive Matrices und Vocabulary Scales. Coloured Progressive Matrices mit der Parallelform des Tests und der Puzzle-Form. Deutsche Bearbeitung und Normierung Stephan Bulheller und Hartmut Haecker. Frankfurt: Swets.

Ricken, G., Fritz, A., Schuck, K. D. & Preuss, U. (2007). Hannover-Wechsler-Intelligenztest für Kinder im Vorschulalter - III (HAWIVA III). Bern: Huber.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F. & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. Journal of Applied Psychology, 88, 1068-1081.

Schaie, K. W. (1996). Intellectual development in adulthood. The Seattle longitudinal study. Cambridge, UK: Cambridge University Press.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. Psychological Bulletin, 124, 262-274.

Schmidt, F. L., Le, H. & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. Psychological Methods, 8, 206-224.

Schneider, W., Bullock, M. & Sodian, B. (1998). Die Entwicklung des Denkens und der Intelligenzunterschiede zwischen Kindern. In F. E. Weinert (Hrsg.), Entwicklung im Kindesalter (S. 53-74). Weinheim: Psychologie Verlags Union.

Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. Diagnostica, 41, 3-20.

Smith, A. (1973/1995). Symbol Digit Modalities Test (SDMT). Los Angeles, CA: Western Psychological Services.

Snow, R. E. (1989). Aptitude-treatment interaction as a framework of research in individual differences in learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), Learning and individual differences (pp. 13-59). New York, NY: Freeman.

Snyderman, M. & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. American Psychologist, 42, 137-144.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. American Journal of Psychology, 15, 201-293.

Steck, P. (1997). Psychologische Testverfahren in der Praxis: Ergebnisse einer Umfrage unter Testanwendern. Diagnostica, 43, 267-284.

Sternberg, R. J. & Detterman, D. K. (Eds.). (1986). What is intelligence? Contemporary viewpoints on its nature and definition. Norwood, NJ: Ablex.

Taub, G. E., McGrew, K. S. & Witta, E. L. (2004). A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale - Third Edition. Psychological Assessment, 16, 85-89.

Tent, L. (2001). Zensuren. In D. H. Rost (Hrsg.), Handwörterbuch Pädagogische Psychologie (2. Aufl., S. 805-811). Weinheim: Psychologie Verlags Union.

Testkuratorium der Föderation deutscher Psychologenverbände. (1986). Beschreibung der einzelnen Kriterien für die Testbeurteilung. Diagnostica, 32, 358-360.

Tetens, J. N. (1777). Philosophische Versuche über die menschliche Natur und ihre Entwicklung (Band 2). Leipzig: Bey M.G. Weidmanns erben und Reich.

Tewes, U. (1994). Hamburg-Wechsler-Intelligenztest für Erwachsene, Revision 1991 (HAWIE-R). Bern: Huber.

von Aster, M., Neubauer, A. & Horn, R. (2006). Wechsler Intelligenztest für Erwachsene *(WIE). Deutschsprachige Bearbeitung und Adaptation des WAIS-III von David Wechsler*. Frankfurt: Harcourt.

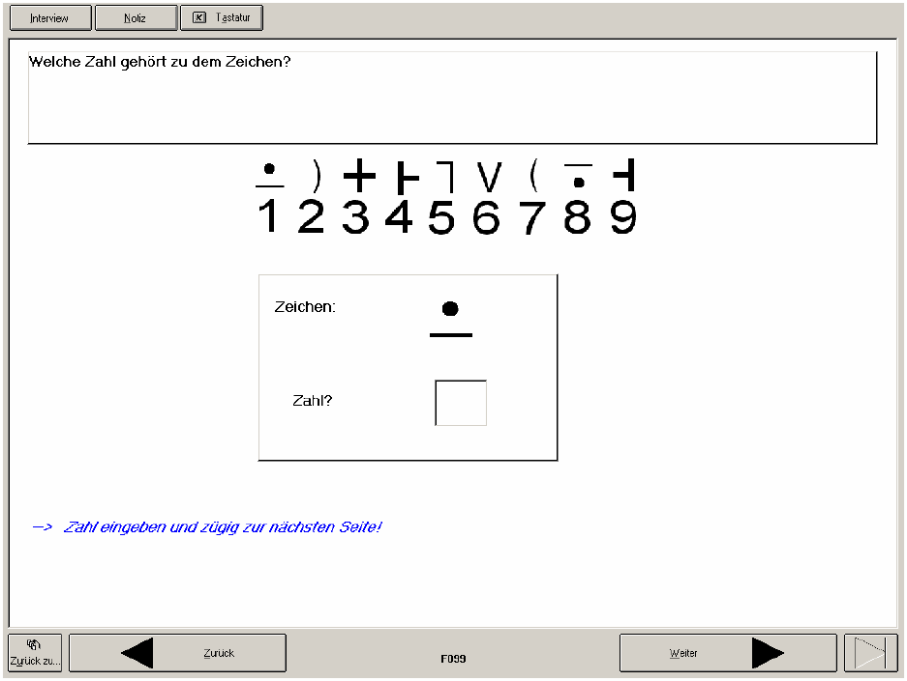
Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T. & Carstensen, C. H. (2011). Development of competencies across the life span. *Zeitschrift für Erziehungswissenschaft, Sonderheft 14*, 67-86.

Weiß, R. H. (2006). *Grundintelligenztest Skala 2 – Revision (CFT 20-R).* Göttingen: Hogrefe.

Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (Hrsg.). (2005). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2., überarbeite Auflage). Lengerich: Pabst.

10. Anhang

Beispielitem aus dem SOEP-Test (CAPI-Version)



Beispielitem aus den Gerotest Matrizen