

Torsten Joél

Die Anwendung von Intelligenztests im sonderpädagogischen Kontext

Eine empirische Untersuchung
unter besonderer Berücksichtigung
der Durchführungs- und
Auswertungsobjektivität



Mit Online-Materialien

BELTZ JUVENTA

Torsten Joél

Die Anwendung von Intelligenztests im sonderpädagogischen Kontext

Torsten Joél

Die Anwendung von Intelligenztests im sonderpädagogischen Kontext

Eine empirische Untersuchung unter
besonderer Berücksichtigung der
Durchführungs- und Auswertungsobjektivität

Mit Online-Materialien

BELTZ JUVENTA

Der Autor

Dipl.-Psych. Torsten Joél lehrt die Anwendung und Interpretation von Intelligenz- und Persönlichkeitstests im Rahmen von bundesweit durchgeführten Seminaren und an der Europa-Universität Flensburg (Institut für Sonderpädagogik). Weitere Arbeitsschwerpunkte sind Verhaltensauffälligkeiten, Intelligenzminderungen und Lernbeeinträchtigungen im Kindes- und Jugendalter.

Dissertation zur Erlangung des Doktorgrades, Europa-Universität Flensburg, Institut für Sonderpädagogik, im November 2019.

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Der Text dieser Publikation wird unter der Lizenz **Creative Commons Namensnennung-Nicht kommerziell-Keine Bearbeitungen 4.0 International (CC BY-NC-ND 4.0)** veröffentlicht. Den vollständigen Lizenztext finden Sie unter: <https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>
Verwertung, die den Rahmen der **CC BY-NC-ND 4.0 Lizenz** überschreitet, ist ohne Zustimmung des Verlags unzulässig. Das gilt insbesondere für die Bearbeitung und Übersetzungen des Werkes.



Dieses Buch ist erhältlich als:
ISBN 978-3-7799-6399-8 Print
ISBN 978-3-7799-5706-5 E-Book (PDF)

1. Auflage 2021

© 2021 Beltz Juventa
in der Verlagsgruppe Beltz · Weinheim Basel
Werderstraße 10, 69469 Weinheim
Einige Rechte vorbehalten

Herstellung: Ulrike Poppel
Satz: text plus form, Dresden
Druck und Bindung: Beltz Grafische Betriebe, Bad Langensalza
Printed in Germany

Weitere Informationen zu unseren Autor_innen und Titeln finden Sie unter: www.beltz.de

Danksagung

Ich bedanke mich herzlich bei meinem Erstgutachter Herrn Prof. Dr. Armin Castello und bei meiner Zweitgutachterin Frau Prof. Dr. Anna-Maria Hintz.

Insbesondere Herr Prof. Dr. Armin Castello als mein Doktorvater hat mich auf seine kollegiale und unkomplizierte Art und Weise ausnahmslos konstruktiv unterstützt. Aus jedem Gespräch ging ich gestärkt hervor. Ich freue mich für die DoktorandInnen, die in Zukunft von ihm betreut werden.

Ich möchte mich ausdrücklich bedanken bei den vielen engagierten ForscherInnen und Fachleuten, die ihr Wissen im Internet teilen und mir an vielen Stellen zügig weiterhalfen.

Karin danke ich für die Unterstützung, ebenso wie meiner Liebsten Christina, die mich auch emotional gestärkt hat. Es ist ein gutes Gefühl, nach der Arbeit an Statistiken und nach der Suche nach der besten Formulierung zu wissen, dass mich meine meist gutgelaunte Ehefrau erwartet mit ihrer optimistischen Grundhaltung. Danke!

Zusammenfassung

Die vorliegende Arbeit untersucht, unter welchen Bedingungen im Rahmen einer Begutachtung zur Feststellung sonderpädagogischen Unterstützungsbedarfs Intelligenztests durchgeführt werden und wie sich die unterschiedlichen Bedingungen auf Schwierigkeiten bei der Anwendung auswirken. Die Analyse der Schwierigkeiten ist hilfreich, um Empfehlungen für den Umgang mit den Tests zu entwickeln.

Intelligenztests – im Besonderen mehrdimensionale Verfahren – sind durch eine Vielzahl von zu beachtenden Regeln in der Anwendung schwierig und benötigen Routine und Erfahrungswissen, da ansonsten die Durchführungs- als auch die Auswertungsobjektivität gefährdet sein kann.

Unter anderem wird mit einem erstellten Schwierigkeiten-Index untersucht, ob Arbeitsbedingungen auf Problematiken bei der Anwendung Einfluss nehmen können, wie sich Unterschiede in der universitären Ausbildung auswirken, oder ob es von Bedeutung ist, in welchem Bundesland getestet wird.

Mit Hilfe eines selbst konstruierten Fragebogens resultieren die Berechnungen aus der Befragung von 1 077 SonderpädagogInnen, die Fragen konzentrierten sich auf elf häufig angewendete Intelligenztests.

Ergänzt wird die Studie durch die Analyse von 248 Testformularen, welche auf Auswertungsfehler untersucht wurden – angefertigt während einer Gutachtererstellung. Während die Antworten aus den Fragebögen Einschätzungen darstellen, resultieren aus der Analyse der Testformulare objektiv feststellbare Hinweise auf typische Anwendungsfehler.

Zusammengenommen belegen die Ergebnisse vielfältige Schwierigkeiten bei der Anwendung von Intelligenztests, signifikante Unterschiede in Abhängigkeit vom Bundesland, in dem getestet wird, und belegen die Bedeutung einer umfassenden universitären Ausbildung zur Testdiagnostik.

Abschließend werden aus den Studienergebnissen resultierende Vorschläge zur Verbesserung der Durchführungsbedingungen vorgestellt, z. B. eine Spezialisierung weniger SonderpädagogInnen bei der Anwendung der komplexeren Tests, aber auch Hinweise zur Konstruktion der Verfahren, die angemessener den Rahmen berücksichtigen, in denen in der Sonderpädagogik Intelligenztests genutzt werden.

Inhalt

Abkürzungsverzeichnis	11
1 Einleitung	13
2 Theoretischer Hintergrund	16
2.1 Intelligenz	17
2.2 Kritik an der Intelligenzmessung	19
2.2.1 Methodische Schwierigkeiten	19
2.2.2 Benachteiligung von Randgruppen durch unfaire Testbedingungen	21
2.2.3 Exklusion, Selektion und Separation als mögliche Folge der Statusdiagnostik	22
2.3 Intelligenzmodelle	23
2.3.1 Was ist Intelligenz: eine Übersicht	24
2.3.2 Intelligenzmodelle im sonderpädagogischen Kontext	31
2.3.2.1 Lurija-Modell	31
2.3.2.2 Kramer-Modell in Anlehnung an den Binet-Simon-Test	34
2.3.2.3 CHC-Modell	37
2.3.2.4 Exkurs: Eugenik und Intelligenzforschung	37
2.3.2.5 Das CHC-Modell als integrierendes Intelligenzmodell	42
2.4 Anwendungen von Intelligenztests durch SonderpädagogInnen	47
2.4.1 Untersuchte Schwierigkeiten bei der Testanwendung in Deutschland	48
2.4.2 Testanwendungen durch SonderpädagogInnen außerhalb Deutschlands	62
2.4.3 Überblick über die Anwendung von Intelligenztests durch <i>special education teachers</i> außerhalb Deutschlands	65
2.5 Intelligenztests	68
2.5.1 Testgütekriterien	69
2.5.1.1 Hauptgütekriterien	69
2.5.1.2 Kritische Werte der Hauptgütekriterien	74
2.5.1.3 Nebengütekriterien	76
2.5.2 Beschreibung der Testverfahren	78
2.5.2.1 K-ABC (Kaufman Assessment Battery for Children)	79

2.5.2.2	KABC-II (Kaufman Assessment Battery for Children – II)	80
2.5.2.3	CFT1-R (Grundintelligenztest Skala 1)	82
2.5.2.4	CFT20-R (Grundintelligenztest Skala 2 – Revision mit Wortschatztest und Zahlenfolgentest)	82
2.5.2.5	WISC-IV (Wechsler Intelligence Scale for Children (Deutsche Ausgabe) – fourth Edition, ehemals HAWIK-IV)	83
2.5.2.6	WPPSI-III (Wechsler Preschool and Primary Scale of Intelligence – III Deutsche Version)	84
2.5.2.7	WNV (Wechsler Nonverbal Scale of Ability)	85
2.5.2.8	SON-R 2½–7 (Non-verbaler Intelligenztest)	86
2.5.2.9	SON-R 5½–17 (Non-verbaler Intelligenztest)	87
2.5.2.10	SON-R 6–40 (Non-verbaler Intelligenztest)	88
2.5.2.11	IDS (Intelligence and Development Scales)	89
2.5.3	Zusammenfassende Übersicht der Testgütekriterien	90
2.5.4	Bedeutungsvolle Aspekte bei der Testanwendung in der Sonderpädagogik	91
2.5.5	Rahmenbedingungen im Umgang mit Intelligenztests auf der Ebene der Bundesländer	95
2.5.6	Antwort- und Verzerrungstendenzen, Beobachtungsfehler und TestleiterInneneffekte	104
3	Forschungsfragen	110
4	Methoden	116
4.1	Fragebogenkonstruktion	117
4.2	Vorannahmen für die Auswertung der Fragebögen	125
4.2.1	Ein- versus mehrdimensionale Intelligenztests	125
4.2.2	Komplexe vs. weniger komplexe Intelligenztests	127
4.2.3	Konstruktion eines Schwierigkeiten-Index	129
4.3	Analyse von ausgewerteten Intelligenztestformularen	134
4.4	Beschreibung der Stichprobe: Fragebogen	137
4.5	Beschreibung der Stichprobe: Formularprüfung	138
5	Ergebnisse	139
5.1	Gewichtungen	140
5.2	Gesamt-, Versuchs- und Kontrollgruppe	141
5.3	Deskriptivstatistische Auswertung	142
5.3.1	Auswertung Fragebögen	142
5.3.2	Auswertung Testformulare	155
5.4	Inferenzstatistische Auswertungen	166

5.4.1	Empfundene Aussagekraft der Tests	168
5.4.2	Unterschiede zwischen <i>Komplexität</i> und Anwendungshäufigkeit	170
5.4.3	Unterschiede zwischen Verfügbarkeit und Vorlieben für Tests	173
5.4.4	Unterschiede in der Anwendung der Tests abhängig vom Bundesland	183
5.4.5	Zusammenhänge zwischen Alter, empfundenen Schwierigkeiten und Anwendung der Tests	209
5.4.6	Unterschiede zwischen Geschlecht und empfundenen Schwierigkeiten bei der Anwendung der Tests	212
5.4.7	Zusammenhänge zwischen Schwierigkeiten bei der Anwendung der Tests und der universitären Ausbildung	212
5.4.8	Zusammenhänge zwischen Schwierigkeiten bei der Anwendung der Tests und der außeruniversitären Fortbildung	224
5.4.9	Unterschiede zwischen Auswertungsfehlern und der Anwendung von Auswertungsprogrammen	225
5.4.10	Zusammenhänge zwischen Durchführungs- und Auswertungsfehlern und der Komplexität der Tests	226
6	Interpretation und Diskussion	228
6.1	Fragebogen	228
6.1.1	Anwendung	228
6.1.1.1	Interpretation der Ergebnisse zur Anwendung von Intelligenztests	229
6.1.1.2	Bedeutung der Ergebnisse für die Sonderpädagogik	234
6.1.2	Vergleiche zwischen den Bundesländern	235
6.1.2.1	Interpretation der Ergebnisse zu den Bundesländervergleichen	236
6.1.2.2	Bedeutung der Ergebnisse für die Sonderpädagogik	241
6.1.3	Alter und Geschlecht	243
6.1.3.1	Problematiken im Zusammenhang mit Alter und Geschlecht	243
6.1.3.2	Bedeutung der Ergebnisse für die Sonderpädagogik	244
6.1.4	Ausbildung	244
6.1.4.1	Auswirkungen der Ausbildung auf Problematiken	245
6.1.4.2	Bedeutung der Ergebnisse für die Sonderpädagogik	247
6.2	Formularanalyse	248
6.2.1	Analyse von Intelligenztestformularen	249
6.2.2	Zusammenfassung und Bedeutung der Ergebnisse	253

6.3 Methodenkritik und Einschränkungen der Untersuchung	255
6.4 Fazit und Ausblick	259
Literatur	263

Hinweis zu den Online-Materialien

Die Online-Materialien sind unter der Homepage des Verlages (www.beltz.de) zu finden. Wenn Sie sich dort das vorliegende Buch anzeigen lassen (Sie können dort nach dem Titel suchen lassen), finden Sie einen Link zum Herunterladen der Online-Materialien.

Die Online-Materialien enthalten Folgendes:

A1 Fragebogen¹

A2: Anschreiben/Datenschutzerklärung an Schulämter

Tabelle B1: Ergänzende Tabelle zu Hypothese 2

Tabelle B2: Ergänzende Tabelle zu Hypothese 2

Tabelle B3: Ergänzende Tabelle zu Hypothese 4.5

Tabelle B4: Ergänzende Tabelle zu Hypothese 4.6

Tabelle B5: Ergänzende Tabelle zu Hypothese 7.2

Tabelle B6: Ergänzende Tabelle zu Hypothese 7.2

Tabelle B7: Ergänzende Tabelle zu Hypothese 7.2

Tabelle B8: Ergänzende Tabelle zu Hypothese 7.2

Tabelle B9: Ergänzende Tabelle zu Hypothese 7.2

Abbildung C: Verteilung der Anzahl gemachter Gesamtfehler

1 Anmerkung: Abgebildet sind alle für diese Studie relevanten Fragen. Fragen zu anderen Themen, die aus Gründen der Ökonomie mitgestellt worden sind, werden hier nicht aufgeführt. Es sind Fragen zur Anwendung von Intelligenztests mit geflüchteten Kindern und Jugendlichen (Joël, 2018) und Fragen, die in eine Studie zu Einstellungen von SonderpädagogInnen zur Intelligenzdiagnostik münden werden.

Abkürzungsverzeichnis

AFI	Allgemeiner-Fähigkeiten-Index
AG	Arbeitsgedächtnis
AID	Adaptives Intelligenz Diagnostikum
AO-SF	Ausbildungsordnung zum sonderpädagogischen Förderbedarf
Ba.-Wü.	Baden-Württemberg
BayEUG	Bayerisches Gesetz über das Erziehungs- und Unterrichtswesen
BUEVA-II	Basisdiagnostik umschriebener Entwicklungsstörungen im Vorschulalter
BVFG	Gesetz über die Angelegenheiten der Vertriebenen und Flüchtlinge
CFT1-R	Grundintelligenztest Skala 1
CFT20-R	Grundintelligenztest Skala 2
CHC-Modell	Cattell-Horn-Carroll-Modell
COTAN	Committee On Test Affairs Netherlands
CPM	Coloured Progressive Matrices
EEG	Elektroenzephalografie
EVuP	Erste Verordnung für unterstützende Pädagogik
Gc	Kristalline Intelligenz
Gf	Fluides Denken und Problemlösen
Gs	Verarbeitungsgeschwindigkeit
Gsm	Kurzzeitgedächtnis
Gv	Visuelle Verarbeitung
HAWIK	Hamburg-Wechsler-Intelligenztest für Kinder
HAWIVA	Hannover-Wechsler-Intelligenztest für das Vorschulalter
IDS	Intelligence and Development Scales
IQ SH	Institut für Qualitätsentwicklung an Schulen Schleswig-Holstein
IQ	Intelligenzquotient
IVI	Individueller Verarbeitungsindex
K-ABC	Kaufman Assessment Battery for Children
KABC-II	Kaufman Assessment Battery for Children – II
KFT 4–12	Kognitiver Fähigkeitstest für 4. bis 12. Klassen
KMK	Kultusministerkonferenz
KTT	Klassische Testtheorie
LAG	Landesarbeitsgemeinschaft Baden-Württemberg – Gemeinsam leben – gemeinsam lernen e.V.
LISUM	Landesinstitut für Schule und Medien Berlin-Brandenburg
MBJS	Ministerium für Bildung, Jugend und Sport

Nds. GVBl	Verordnung zum Bedarf an sonderpädagogischer Unterstützung
NIO	Nederlandse Intelligentietest voor Onderwijsniveau
ReBBZ	Regionales Bildungs- und Beratungszentrum
ReBUZ	Regionales Beratungs- und Unterstützungszentrum
SBA-VO	Verordnung des Kultusministeriums über die Feststellung und Erfüllung des Anspruchs auf ein sonderpädagogisches Bildungsangebot
SchpflG	Schulpflichtgesetz
SD	Standardabweichung
SFI	Sprachfrei-Index
SHP	Schulische HeilpädagogInnen
SIBUZ	Schulpsychologisches Beratungszentrum
Sig	Signifikanz
SOFS	Verordnung des Sächsischen Staatsministeriums für Kultus über Förderschulen
SoFVO	Landesverordnung über sonderpädagogische Förderung
SON-R	Non-verbaler Intelligenztest
SoPädFV	Verordnung über die Förderung von Schülerinnen und Schülern mit sonderpädagogischem Bildungs-, Beratungs- und Unterstützungsbedarf
SoSchulO RP	Schulordnung für die öffentlichen Sonderschulen
TBS-TK	Testbeurteilungssystem des Diagnostik- und Testkuratoriums
ThürSoFöV	Thüringer Verordnung zur sonderpädagogischen Förderung
TLS	Teilleistungsstörungen
VOSB	Verordnung über Unterricht, Erziehung und sonderpädagogische Förderung von Schülerinnen und Schülern mit Beeinträchtigungen
WET	Wiener Entwicklungstest
WISC	Wechsler Intelligence Scale for Children
WLD	Wahrnehmungsgebundenes Logisches Denken
WNV	Wechsler Nonverbal Scale of Ability
WP	Wertpunkt
WPPSI	Wechsler Preschool and Primary Scale of Intelligence
ZUP	Zentrum für unterstützende Pädagogik

1 Einleitung

Die Durchführung von Intelligenztests (z.B. WISC-IV, IDS, SON R 6–40, K-ABC/K-ABC II) setzt eine genaue Kenntnis der Durchführungsregeln der jeweiligen Verfahren voraus. Darüber hinaus postuliert Bundschuh (2010), dass Aussagen über die Berechnung eines Intelligenzquotienten hinaus es „einer guten Kenntnis dessen [bedarf], was der jeweilige Test beinhaltet, seiner Konstruktion, seiner Implikationen, vor allem der ihm zugrunde liegenden Theorie“ (ebd., S. 184). Helmke (2007, S. 85) beschreibt diagnostisches Wissen als zwingende Voraussetzung für eine individuelle Förderung.

Es gehört zur Stellenbeschreibung von SonderpädagogInnen, standardisierte Verfahren durchführen, auswerten und interpretieren zu können. Nach durchgeführten Schulungen, an denen zwischen 2009 und 2016 bisher ca. 9 000 Personen teilnahmen² – überwiegend mit SonderpädagogInnen als TeilnehmerInnen –, kann behauptet werden, dass die Durchführungsregeln nicht hinreichend gewürdigt werden und die Bedeutung der Regeln für die objektive Begutachtung unterschätzt wird.

Dies gilt im Übrigen auch für die an den Seminaren teilnehmenden PsychologInnen. Hinzu kommt, dass vorliegende Testergebnisse unterschiedlich interpretiert werden, die aus den Interpretationen abgeleiteten (sonderpädagogischen) Empfehlungen also nicht einheitlich ausfallen. Beispiele:

Bei der Durchführung der Diagnostikseminare konnte beobachtet werden³,

- dass das Testalter häufig falsch berechnet wird, so dass die Rohwerte in der falschen Alterstabelle mit den standardisierten Werten verglichen werden,
- es wird häufig ein Gesamt-IQ bei Intelligenzverfahren errechnet, ohne das für die Bestimmung der kognitiven Leistungsfähigkeit aussagekräftigere Vertrauensintervall anzugeben,
- es werden Umkehrregeln häufig falsch angewendet, so dass der Ablauf eines Subtests falsch, die Reihenfolge der durchzuführenden Items fehlerhaft ist,
- Durchführungsregeln werden wohlwollend ausgelegt, z. B. die Bearbeitungszeit eines Items weggelassen, ohne zu berücksichtigen, dass das Ergebnis dann keine Vergleichbarkeit mehr mit den Daten der Normstichprobe liefert.

2 Schulungen zu normierten standardisierten Testverfahren, durchgeführt ab 2009.

3 Z.B. durch Sichtung von Durchführungsformularen, die Teilnehmende mir zur Begutachtung oder für besondere Fragestellungen vorlegten; es war allerdings nicht Ziel der Vorlage, diese auf Fehler zu untersuchen. Diese sind eher zufällig entdeckt worden.

Nach der Durchführung von mehr als 250 Diagnostikseminaren bis 2016 kann folgendes angenommen werden:

- Die Durchführungsregeln der Testverfahren sind nicht hinreichend bekannt, die Bedeutung der richtigen Anwendung und somit die Bedeutung einer ausführlichen Einarbeitung für eine korrekte Interpretation wird unterschätzt (bzw. wird für die Vorbereitung nicht genügend Zeit zur Verfügung gestellt).
- Durch die angenommene falsche Anwendung der Intelligenztests kann vermutet werden, dass falsche Ergebnisse Einfluss auf die abschließende Beurteilung der zu testenden Kinder hatten.
- Es besteht eine Diskrepanz zwischen dem Anspruch einer objektiven Testdurchführung und strukturellen Vorgaben der Schulen, die eine Verwirklichung des objektiven Anspruchs erschweren, aber auch unverhältnismäßig komplizierten Durchführungsregeln der Testverfahren, die in einigen Fällen nachweislich widersprüchlich und nicht eindeutig sind, zuweilen sogar versierten Fachleuten unverständlich erscheinen.

Es könnte eingewendet werden, dass unerfahrene Personen an einer Fortbildung teilnehmen, um eben diese Fehler zu vermeiden. Zu Beginn eines Seminars wird allerdings die Testerfahrung erfragt und es gab praktisch keine Seminare ohne TeilnehmerInnen, die überwiegend bereits häufig Tests durchgeführt hatten. Als Motivation an der Teilnahme der Seminare wurde meist das Kennenlernen neuerer Testverfahren genannt.

Hauptsächlich soll die Dissertation untersuchen, welche Schwierigkeiten bei der Durchführung von Intelligenztests auftreten und warum diese Schwierigkeiten auftreten.

Abgeleitet werden könnten Hinweise für die Konstruktion zukünftiger Intelligenztests. Denn ein valider Test mit ausgezeichneten Testgütekriterien verliert an Aussagekraft, wenn er die Anforderungen des sonderpädagogischen Schulalltags nicht berücksichtigt und deshalb möglicherweise falsch durchgeführt wird.

Abgeleitet werden könnte zudem eine Auswahl von aktuell häufig durchgeführten Testverfahren, welche im sonderpädagogischen Kontext gut anwendbar und aus Sicht der sonderpädagogischen Lehrkräfte von hohem Nutzen sind.

Die bis hierhin beschriebenen Schwierigkeiten beruhen auf nicht validierten Erfahrungen und Beobachtungen, die weder in einem Setting festgestellt worden sind, welches wissenschaftlichen Standards entspricht, noch als Quelle für eine wissenschaftliche Studie dienen dürfen, da sie einem Eindruck entsprechen. Ziel dieser Dissertation ist es, zu überprüfen, ob die subjektiv beobachteten Schwierigkeiten objektiv vorliegen. Dies ist nur im Rahmen einer wissenschaftlich fundierten Studie möglich.

Die Dissertation soll nicht die Sinnhaftigkeit von Intelligenztests diskutieren und/oder untersuchen, nicht das Konstrukt Intelligenz generell in Frage stellen und auch nicht die Durchführung von psychologischen Tätigkeiten der SonderpädagogInnen (z. B. die Durchführung von Tests) in Frage stellen.

Im theoretischen Kapitel wird nach einer Auseinandersetzung mit dem Konstrukt Intelligenz und der Beschreibung, wie Intelligenztests von SonderpädagogInnen im In- und Ausland eingesetzt werden, der Aufbau und die Güte von Intelligenztests im Mittelpunkt stehen. In der Sonderpädagogik gebräuchliche Intelligenztests werden beschrieben inklusive der jeweils dazugehörigen theoretischen Grundlage bzw. des Intelligenzmodells⁴, welches die AutorInnen der jeweiligen Testverfahren als gültig postulieren.

Im methodischen Kapitel wird begründet, warum die Fragestellungen nicht nur mit Hilfe eines Fragebogens (ausgefüllt von SonderpädagogInnen) beantwortet werden sollen, sondern ergänzend Analysen von Durchführungsprotokollen zur genaueren Beantwortung der Forschungsfragen beitragen sollen. Es wird begründet, warum und wie ein Schwierigkeiten-Index erstellt wird und warum und wie die Intelligenztests nach Dimensionalität und Komplexität unterteilt werden.

Im Ergebnisteil werden die Daten deskriptiv- und inferenzstatisch ausgewertet. Nach der Analyse von Testformularen, die im Zuge einer Begutachtung angefertigt worden sind und im Rahmen dieser Studie auf Fehler untersucht werden, wird zudem qualitativ dargestellt, welche Auswirkungen aus Auswertungsmängeln resultieren.

Abschließend werden im Kapitel 6 (Interpretation und Fazit) Ableitungen diskutiert für die Konstruktion von Intelligenztests, für die (schulischen) Rahmenbedingungen, für die Durchführung von Intelligenztests durch SonderpädagogInnen und für die Qualifikation von SonderpädagogInnen bezüglich der Durchführung von Intelligenztests. Da Intelligenztests die komplexesten standardisierten normierten Testverfahren darstellen, lassen sich somit generell Ableitungen für die Durchführung von standardisierten normierten Testverfahren durch SonderpädagogInnen als Teilbereich innerhalb der Diagnostik im sonderpädagogischen Kontext vornehmen.

4 Die KABC-II basiert sogar auf zwei Intelligenzmodellen, von denen eines wahlweise vor der Testung gewählt werden muss.

2 Theoretischer Hintergrund

Im Folgenden sollen die Kapitel des theoretischen Teils näher begründet und die Bedeutung für diese Dissertation herausgestellt werden:

Intelligenz (Kapitel 2.1): Bei der Anwendung eines Intelligenztests sind Kenntnisse über das Konstrukt Intelligenz unerlässlich, um die aus den Tests gewonnenen Ergebnisse interpretieren und in pädagogische Handlungen umsetzen zu können. Das erste Kapitel beschreibt die Schwierigkeiten bei der Definition, aber auch die Bedeutung des Konstrukts Intelligenz.

Kritik an der Intelligenzmessung (Kapitel 2.2): Intelligenztests sind in der Sonderpädagogik nicht unumstritten. Kritische Einwände gegenüber dem Generalfaktor der Intelligenz, der meist mit einem Gesamt-IQ (Gesamt-Intelligenzquotient) dargestellt wird, sind berechtigt. Es werden nicht nur methodische, sondern auch ethische Aspekte bei der Berücksichtigung von Testergebnissen vorgestellt. Kenntnisse über kritische Einwände sind für Intelligenztests anwendende SonderpädagogInnen nützlich, um die Relevanz von Testergebnissen ermessen und in die Gewichtung innerhalb anderer Bausteine der sonderpädagogischen Diagnostik einordnen zu können. Die Problematik, ob bei der Interpretation von Testergebnissen die dargestellten kritischen Einwände gegenüber der Intelligenztestung berücksichtigt werden, oder Testergebnisse als per se gültig akzeptiert werden, soll im methodischen Teil geklärt werden.

Intelligenzmodelle (Kapitel 2.3): Die heute angewendeten Intelligenztests stehen teils in einer jahrzehntealten (teils jahrhundertealten) Tradition von Intelligenztests und -theorien, so dass ein kurzer Blick auf Forschungszweige zur Intelligenz nützlich ist, die in die heute angewendeten Intelligenztests münden. Ein ausführlicherer Blick gilt Intelligenztheorien, auf die sich explizit aktuell angewendete Intelligenztests berufen (Lurija-Modell; Kramer-Modell in Anlehnung an den Binet-Simon-Test und vor allem das derzeit wichtige CHC-Modell). Ausgespart wird nicht ein Blick auf eugenische Gedanken, die im Rahmen der Intelligenzforschung keine Randerscheinung darstellen. Die Optimierung und Förderung der Menschen unter Verhinderung weniger nützlicher Menschen (Eugenik) steht im Gegensatz zur Akzeptanz und Förderung des einzelnen Menschen (Sonderpädagogik). Kenntnisse über die Motivation eugenisch denkender (durchaus bedeutender) IntelligenzforscherInnen ist nützlich, um den Stellenwert von Intelligenztests und vor allem den Stellenwert hierarchischer Intelligenzmodelle angemessen beurteilen zu können.

Anwendungen von Intelligenztests durch SonderpädagogInnen (Kapitel 2.4): Um im methodischen Teil die Schwierigkeiten im Umgang mit Intelligenztests besser erforschen zu können, werden bereits durchgeführte Untersuchungen

zum Thema vorgestellt. Dies ist notwendig, um beurteilen zu können, woran angeknüpft werden kann, welche Befunde vertieft erforscht werden sollten und für welche Themen Neuland betreten wird. Unterschieden wird zwischen Untersuchungen in Deutschland und im Ausland, die sich mit Schwierigkeiten im Umgang mit Intelligenztests beschäftigen. Dies macht Sinn, da sich die Stellenbeschreibungen deutscher SonderpädagogInnen von denen der SonderpädagogInnen im Ausland unterscheiden können. Unterschiede sind möglich im Umfang psychologischer Tätigkeiten, wie z.B. die Anwendung von Intelligenztests. Dies ist zu klären, denn führten SonderpädagogInnen im Ausland kaum Intelligenztests durch, werden entsprechend keine Forschungsergebnisse im Umgang mit Intelligenztests durch SonderpädagogInnen vorliegen.

Intelligenztests (Kapitel 2.5): Im letzten Kapitel des theoretischen Teils werden testtheoretische Aspekte vorgestellt, die im Zusammenhang mit der Fragestellung von Bedeutung sind. Es gibt z.B. bei der Anwendung von Intelligenztests bekannte Effekte (Mildefehler, Härtefehler, Beurteilungsfehler, Rosenthal-Effekt etc.) auf Seiten der AnwenderInnen, die die Beurteilung der Testergebnisse erschweren und verzerren könnten. Bei der Anwendung von Intelligenztests kommt hinzu, dass sich immer auch die Frage stellt, wie gut die Testverfahren konstruiert und validiert sind. Testgütekriterien belegen die Qualität der Intelligenztests. Wird die Qualität eines Intelligenztests nicht hinreichend durch die Testgütekriterien belegt, führt die Beurteilung von Testergebnissen, erzielt mit einem Test mit schwachen Testgütekriterien, zu weiteren Schwierigkeiten auf Seiten der AnwenderInnen. Begründet wird eine Unterscheidung der Tests in ein- bzw. mehrdimensionale Tests und die Unterscheidung nach Komplexität.

2.1 Intelligenz

Das Vorliegen der Persönlichkeitsdimension *Intelligenz* beruht auf einer Annahme. Ist ein Konstrukt nicht beweisbar und basiert somit auf der Annahme über dessen Existenz, ist dieses Konstrukt in der Regel nicht nur umstritten, sondern es existieren verschiedene theoretische Überlegungen dazu. So ist z.B. umstritten, ob Intelligenz mit Hilfe hierarchischer Modelle beschrieben werden kann, an dessen Spitze ein übergeordneter Intelligenzfaktor steht (vgl. Spearman, 1904; Schneider & McGrew, 2012) oder ob es mehrere unabhängige Intelligenzen nebeneinander gibt (vgl. Guilford, 1977; Jäger, 1982). Dass Intelligenz lediglich das ist, was ein Intelligenztest misst, (Boring, 1923, S. 35) negiert konsequent zu Ende gedacht nicht die Existenz dieser Persönlichkeitsdimension, sondern beschreibt die methodische Schwierigkeit, einer Person eine bestimmte und somit definierte Intelligenz zuzuschreiben. In der Tat bleibt es fraglich, wie Intelligenz valide gemessen werden kann, wo weder dessen Existenz eindeutig nachgewiesen ist noch ein Konsens über eine Definition vorliegt. Ein grund-

sätzliches Problem auch in anderen Bereichen der Psychologie, die oft mit Annahmen arbeitet. Intelligenz ist aber kein neutraler Begriff, sondern ein gewerteter. Eine hohe Intelligenz wird positiv belegt, eine niedrige Intelligenz negativ. Würde man in einem Hochschulseminar fragen, ob jemand mit einem IQ von 88 zufrieden wäre, denn immerhin sei man ja noch im Normbereich, so wäre die Bejahung dieser Frage unwahrscheinlich⁵. Man möchte intelligent sein, denn mit einer hohen Intelligenz wird beruflicher und sozialer Erfolg assoziiert, wenn nicht ein besseres Leben als mit einer niedrigen Intelligenz. Wer möchte nicht im Hochschulseminar sitzen, um später zur Intelligenz zu gehören, dem Sammelbegriff „als Name einer ganzen gesellschaftlichen Gruppe, der intellektuellen und kulturproduzierenden Elite“ (Holert, 2004, S. 126). AbiturientInnen demonstrieren gelegentlich stolz ihre mit dem Abschluss des Abiturs assoziierte höhere Intelligenz und die damit verbundene Zugehörigkeit zu einer Elite in einem Schriftzug im Fond ihrer Autos, z. B. „ABI 2015“. Im Gegensatz zu anderen durchaus auch als bedeutsam postulierten Persönlichkeitsdimensionen wie Verträglichkeit, Extraversion, Gewissenhaftigkeit (vgl. McCrae & Costa, 1989) usw. ragt Intelligenz als Schwergewicht unter den Persönlichkeitsdimensionen heraus, es ist ein hoch „gelobtes Gut“ (Zimbardo, 1992, S. 444). Somit ist die Zuschreibung einer bestimmten Intelligenz für eine Person eine verantwortungsvolle Angelegenheit, da diese laufbahn- und schullaufbahnentscheidend sein kann. Die Belege dafür, dass die Zuschreibung einer bestimmten Intelligenz für eine Person sogar lebensentscheidend sein kann, sind eindeutig. Im Gerichtsverfahren Atkins gegen Virginia beschreibt Chwallek (2005) den Fall von Daryl Atkins, der nach einer Verurteilung wegen Entführung, Raubes und Mord nicht zum Tode verurteilt wurde, da er mit einem IQ von 59 als geistig behindert galt. Eine Todesstrafe ist nach der amerikanischen Verfassung bei Vorliegen einer geistigen Behinderung nicht möglich. Nach Durchführung von Testwiederholungen erzielte er zu einem späteren Zeitpunkt einen IQ von 76 bzw. 74 und galt somit nicht mehr als geistig behindert. Nach 13 Stunden Beratung wurde ein Hinrichtungsdatum festgelegt, da die Kriterien für die geistige Behinderung nicht mehr vorgelegen haben⁶.

Der von R.M. Yerkes 1917 entwickelte Army-Alpha-Test sah die Testung von Rekruten nach Eintritt der USA in den Ersten Weltkrieg vor (Funke & Vatterodt, 2009). Mit diesem Test und dem ähnlichen Army-Beta-Test wurden über 1 700 000 Rekruten auf Intelligenz getestet. Ziel war die Einordnung in einen militärischen Rang. Die Einordnung erfolgte in Form eines Buchstabens von A bis E. Für die Offizierslaufbahn kamen lediglich Rekruten von mindes-

5 Manchem läge vermutlich sogar das Bonmot auf der Zunge, dass ein Studierender mit einem IQ von 88 gar nicht Studierender wäre.

6 Die Todesstrafe wurde später in eine lebenslange Haftstrafe umgewandelt, da Verfahrensfehler vorlagen.

tens C (eher A oder B) in Betracht. Auf die Spitze getrieben muss davon ausgegangen werden, dass eine niedrige Intelligenzeinstufung ein Soldatendasein zur Folge hatte, welches gelegentlich umgangssprachlich als *Kanonenfutter* bezeichnet wird.

Der Soziologe Pierre Bourdieu bringt die Intelligenztestung mit Rassismus in Zusammenhang. Unter anderem begründet er eine ablehnende Haltung so:

Die Klassifizierung durch die Schule ist eine legitimierte und wissenschaftlich ausgewiesene soziale Diskriminierung. Das Auftauchen von Intelligenztests (...) hängt damit zusammen, dass dank der Schulpflicht Schüler in das Schulsystem kamen, mit denen dieses Schulsystem nichts anzufangen wusste, weil sie nicht „prädisponiert“ waren, „nicht begabt“, das heißt, nicht von ihrem familiären Milieu her mit jenen Prädispositionen ausgestattet, die die Voraussetzung für das normale Funktionieren des Schulsystems sind: Kulturelles Kapital und guter Wille in Bezug auf die Schulabschlüsse. Diese Tests, die die von der Schule verlangten sozialen Prädispositionen messen, sind genau dazu da, jene schulischen Verdikte im Voraus zu legitimieren, durch die sie legitimiert werden; daher auch ihre Aussagekraft in Bezug auf den Schulerfolg. (Bourdieu, Beister & Schwibs, 1993, S. 254)

Es wird nicht bezweifelt, dass es so etwas wie Intelligenz gibt. Es wird bezweifelt, dass man den Grad der individuellen Intelligenz einer Person erfassen kann. Nicht nur, weil das Konstrukt *Intelligenz* unterschiedlich definiert wird, auch weil angezweifelt wird, der individuellen Intelligenz einer Person methodisch auf die Spur kommen zu können und dies dann (radikal zu Ende gedacht) in einer Zahl zu manifestieren: dem Gesamt-IQ.

2.2 Kritik an der Intelligenzmessung

Bevor die unterschiedlichen Intelligenzmodelle vorgestellt werden, sollen die wichtigsten kritischen Einwände zur Intelligenzmessung zusammengefasst werden.

2.2.1 Methodische Schwierigkeiten

Unabhängig von der Definition wird Intelligenz als Intelligenz bezeichnet. Wenn aber je nach Definition außer der Überschrift die Annahmen dessen, was Intelligenz darstellen soll, sehr variieren, wie will man dann Intelligenz messen? Intelligenz ist nicht gleich Intelligenz, dies wird aber unterstellt, wenn das Resultat eines Intelligenztests meist ein Intelligenzquotient ist. Dann wäre Intelligenzquotient nicht gleich Intelligenzquotient, was der Wahrnehmung des IQ

sowohl in der Wissenschaft als auch in der Allgemeinheit widerspricht. Steht in einer Sportzeitschrift, dass eine französische Tennisspielerin hochbegabt sei und einen IQ von 175 hat, dann wird angenommen, dass der IQ 175 beträgt und es würde kaum hinterfragt werden, welchem theoretischen Modell der durchgeführte Intelligenztest zu Grunde liegt⁷. Ein weiteres methodisches Problem ist die Abhängigkeit des Testergebnisses von Umgebungsvariablen während der Testsituation, z. B. Lichtverhältnisse, Klima, Ablenkung, ist das Kind ausgeruht, ist das Kind belastet, ist das Kind motiviert, ist die TesterIn nett, ist die TesterIn wohlwollend, ist die TesterIn kompetent usw. Da eine mögliche mangelhafte Durchführungsobjektivität ein Hauptbestandteil dieser Untersuchung ist, wird an dieser Stelle nicht genauer darauf eingegangen.

Erwähnt sei in diesem Zusammenhang die Annahme, dass die Intelligenz, wie auch andere Konstrukte in einer Population, normalverteilt sein soll und mit der Glockenkurve nach Gauß (nach Carl Friedrich Gauß) dargestellt wird. Mienert & Pitcher (2011, S. 111) beklagen, dass diese Annahme weder bewiesen noch widerlegt werden kann und dass Intelligenztests so konstruiert werden, bis die Ergebnisse zum Konstrukt passen.

Ob sich ein Intelligenztest als solcher bewährt oder nicht wird mit Hilfe von Testgütekriterien geprüft. Dazu wird unter anderem verglichen, ob ein neuer Test mit einem älteren anerkannten Test korreliert, ob also die Testergebnisse sich decken. Es wurde z. B. geprüft, ob die Testergebnisse der neuen KABC-II (Melchers & Melchers, 2014) mit den Testergebnissen der IDS (Grob, Meyer & Arx, 2009) korrelieren. Decken sich die Ergebnisse, wird dies als Beleg gewertet, dass der neue Intelligenztest ebenfalls Intelligenz misst. Wenn nun aber die bewährten Tests gar nicht Intelligenz testen, sondern Irgendetwas und es liegt eine Korrelation zu einem neuen Test vor, dann ist dies nicht ein Beleg dafür, dass der neue Intelligenztest Intelligenz misst, sondern Irgendetwas. Es bleibt zu hoffen, dass die ersten Testverfahren in dieser Reihe von fortlaufenden Korrelationsstudien auch wirklich das gewünschte Konstrukt Intelligenz getestet haben. Ansonsten müsste man davon ausgehen, dass die Korrelationsstudien eine Art Hermann-Teig⁸ darstellen. Im Gedanken ähnlich formulieren es Amelang und Zielinski (1994, S. 146), die den Intelligenztest von Binet & Simon (1905) zwar als ersten ernstzunehmenden Test bezeichnen, aus dem sich aber

7 Es würde auch kaum hinterfragt werden, welcher Test denn auf den Punkt genau so gut misst und welcher Test überhaupt Gesamtwerte von IQ 175 in seinen Tabellen enthält. Eine Nachfrage über Facebook bei Marion Bartoli, welcher Test angewendet wurde, blieb unbeantwortet.

8 Der Hermann-Teig ist ein Sauerteigansatz. Vor dem Backen wird ein Teil des Ansatzes entnommen, dieser vermehrt sich durch Hefepilzreaktionen und kann nach einiger Zeit für ein weiteres Brot verwendet werden – nachdem wiederum ein Teil vorher entnommen wurde (Dr. Oetker Homepage, 2015).

keine differenzierten Diagnosen ableiten lassen, da die Aufgaben zu heterogen seien. Die Güte der ersten Tests aus der Wechsler-Reihe belegte Wechsler durch eine hohe Korrelation mit dem aus Sicht von Amelang und Zielinski fragwürdigen Test von Binet & Simon und sicherte den Wechsler-Tests die „historische Kontinuität“ (Amelang & Zielinski 1994, S. 146).

2.2.2 Benachteiligung von Randgruppen durch unfaire Testbedingungen

Häufig wird gegen den Einsatz von Intelligenztests eingewendet, dass diese besondere Bevölkerungsgruppen benachteiligen, z.B. Kinder mit Migrationshintergrund, Kinder aus sozial benachteiligten Familien. Typische Intelligenztests repräsentieren eher den kulturellen Hintergrund des meist westlich geprägten Landes, in dem der Test angewendet wird (Joél, 2018, S. 204). Verzerrungen können auftreten, wenn das getestete Kind einen anderen „kulturellen, sozialen und sprachlichen Hintergrund“ hat (Zimbardo, 1992, S. 454).

Angenommen, ein syrisches Flüchtlingskind wird drei Wochen nach seiner Ankunft in Deutschland mit einem Test getestet, bei dem die Anweisungen in Deutsch vorgegeben werden müssen, z.B. Text-Rechenaufgaben wie im Subtest *Rechnerisches Denken* im HAWIK-IV (Petermann & Petermann, 2007). Das Kind würde eine Textaufgabe nicht verstehen und das Item würde mit falsch bewertet werden. Auch wenn das Kind gut rechnen könnte, würde die rechnerische Kompetenz nicht erfasst werden können. In diesem Fall würde kaum dieser sprachlastige Subtest durchgeführt werden. Denkbar wäre jedoch die Testdurchführung, nachdem das Kind bereits vier Jahre in Deutschland lebt und gut deutsch spricht. Es macht aber nach wie vor einen Unterschied, ob ein Kind in seiner Muttersprache oder in einer erlernten Sprache getestet wird. Durch die Konstruktion sprach- und kulturfairer Intelligenztests sollte dieses Problem behoben werden. Doch erscheint fraglich, ob ein Test losgelöst von kulturellen Einflüssen konstruiert werden kann. So muss z.B. das Kind gewohnt sein, am Tisch Denkaufgaben lösen zu können und logisch planvoll vorzugehen. Es sind z.B. Verzerrungen im Testergebnis denkbar für aus Bürgerkriegsregionen geflüchtete Kinder durch die ungewohnten Rahmenbedingungen der Testsituation (strukturiert planvolles Vorgehen am Tisch in einer Eins-zu-Eins Situation mit teils abstrakten Items), sofern diese Kinder niemals eine Schule besucht haben und das Arbeiten am Tisch unbekannt ist.

Serpell (1979) konnte kulturelle Unterschiede bei der Bewältigung von Aufgaben nachweisen, indem er englische und sambische Kinder aufforderte, Muster mit unterschiedlichen Materialien nachzuvollziehen. Während sambischen Kindern dies mit Draht besser gelang, konnten englische Kinder die Aufgabe besser mit Stift und Papier bewältigen (Geißler, 2008, S. 37).

Der Mitte der 70er Jahre ermittelte Unterschied zwischen weißen und schwarzen Kindern in den USA (Loehlin, Lindzey & Spuhler, 1975) (schwarze Kinder lagen knapp eine Standardabweichung unter dem Durchschnittswert weißer Kinder) wird mit ungleichen Bildungschancen und unfairen Testbedingungen erklärt werden müssen, würde man nicht eine intellektuelle Überlegenheit der weißen Rasse gegenüber der schwarzen Rasse postulieren, wie es im Anschluss an diesen und ähnlichen Studien rassistisch motivierte PolitikerInnen taten.

2.2.3 Exklusion, Selektion und Separation als mögliche Folge der Statusdiagnostik

Insbesondere in der Sonderpädagogik wurde in teilweise heftig geführten Disputen eine Diagnostik in Frage gestellt, die einen Ist-Zustand mit Hilfe standardisierter normierter Testverfahren ermittelt. Dies führe zu einer Diagnostik, die sich an pathologisch-medizinischen Modellen orientiere und der Selektion diene (Eberwein, 1996), stigmatisierend sei und mit dem Recht auf inklusive Bildung „unvereinbar“ ist (Schumann, 2013, ohne Seitenangabe). Das aus der Kritik an der Selektionsdiagnostik und Einweisungsdiagnostik (Kobi, 1977) resultierende Modell der Förderdiagnostik fand in der Sonderpädagogik starke Beachtung und verbreitete sich so stark, dass von einem Paradigmenwechsel gesprochen wurde, sofern die Testung mit normierten Testverfahren als Paradigma und die Förderdiagnostik als das neue Paradigma bezeichnet werden kann. Das Konzept der Förderdiagnostik wurde von renommierten WissenschaftlerInnen aufgegriffen und gelehrt (Bundschuh, 1985, 2007; Eberwein & Knauer, 1998). Das mit der Durchführung normierter Testverfahren assoziierte Menschenbild stehe im Widerspruch zu einer Sonderpädagogik, die sich an humanitären Grundsätzen orientiert, während die herkömmliche Diagnostik mit Hilfe von normierten Verfahren als Einweisungsdiagnostik beschrieben wird (Eggert, 1997). Eggert beschreibt diesen Zustand mit einem „Unbehagen an der Diagnostik“ (Eggert, 1997, S. 71). Überraschend moderat resümiert Bundschuh, dass die Berechnung eines Intelligenzquotienten „ein gewisses unsicheres Moment darstelle“ (2008, S. 184), doch ist der Konflikt Statusdiagnostik versus Förderdiagnostik eher von Schärfe geprägt. Letztlich wurde dieser Art von (Status-)Diagnostik vorgeworfen, sie orientiere sich an der Fragestellung, ob ein Kind vom bisherigen Schulsystem separiert werden soll und in Folge sonderpädagogisch in Förderschulen beschult wird und damit ausgesondert (separiert) wird. Dies sei an einer gesellschaftlich festgelegten Norm gekoppelt und nicht an den individuellen Bedürfnissen, Zielen und Fortschritten des Kinds, für das die Förderdiagnostik bzw. Lernprozessdiagnostik stehe. So werden zuweilen Testverfahren wie Intelligenztests als „harte“ Verfahren bezeichnet, während förderdiagnosti-

sche Verfahren als „weiche“ Verfahren bezeichnet werden (Kottmann, 2006, S. 125). Da mit dem Gebrauch von Worten Vorstellungen und Gedanken gekoppelt sind, spiegelt der Gebrauch der Worte *hart* und *weich* gut die Schärfe wider, die die Diskussion bestimmte, denn wer wollte mit einem Kind *harte* Maßnahmen durchführen⁹. Nicht minder scharf fielen die Kritiken an die Kritiker bisheriger sonderpädagogischer Diagnostik aus. In einer Bilanz zu 30 Jahre „Förderdiagnostik“ [sic, Förderdiagnostik wird in Anführungszeichen gesetzt, Anmerkung T.J.] (Schlee, 2008) wird den VertreterInnen dieses Ansatzes vorgeworfen, dass das Konzept der Förderdiagnostik auf falschen Annahmen beruhe. So ist eine Beschulung nach einer Statusdiagnostik (die z.B. den sonderpädagogischen Förderbedarf *Geistige Entwicklung* attestiert) in einer Sonderschule nicht verwerflich, sondern das Nicht-Einlösen der in der Sonderpädagogik „proklamierten Ansprüche“ (Schlee, 2008, S. 124). Indem mit einer Beschulung in einer Sonderschule schlechtere Zukunftschancen angenommen werden, werden die Methoden kritisiert (z.B. Intelligenztests), die zu dieser Beschulung führen, obwohl die Diagnostik gar nichts mit den schlechteren Zukunftschancen zu tun hat, sondern die pädagogische Umsetzung in der Sonderschule. Es wird kritisiert, dass die Förderdiagnostik zwar einer guten Absicht entspringt, aber weder empirisch belegt wurde noch die Nützlichkeit nachgewiesen ist (Schlee, 2008, S. 122). Häufig wird in der Pädagogik und Psychologie die ganzheitliche Betrachtung des Menschen gefordert, so auch von VertreterInnen der Förderdiagnostik. Doch darf dies als nebulöse Metapher betrachtet werden, deren Umsetzung angesichts der Komplexität des Menschen unmöglich erscheint. Einem wichtigen Vertreter der Förderdiagnostik, Prof. Bundschuh, wirft Schlee vor, sich als „ein einsamer Rufer für die Menschlichkeit (...)“ darzustellen, der sich am „Beginn des dritten Jahrtausends (...) an vielen Fronten in ganzheitlicher Sicht engagiert (...)“ (Schlee, 2008, S. 129). Schlee schlussfolgert, dass man angesichts der Vielzahl von proklamierten hehren Zielen der Förderdiagnostik sich kaum traue, „nach Begründungen oder Konkretisierungen zu fragen (...)“ (Schlee, 2008, S. 130).

2.3 Intelligenzmodelle

Bei vorherrschender Uneinigkeit dessen, was Intelligenz sein oder auch nicht sein soll, müssen die verschiedenen Intelligenztheorien beschrieben werden. Da ein Intelligenztest auf einer Intelligenztheorie beruht, wird an späterer Stelle auf diese Theorien zurückgegriffen werden müssen.

⁹ Auch der Begriff Selektion ist negativ konnotiert und erinnert an die Selektion von JüdInnen und anderen Personengruppen während des Nationalsozialismus.

Die Intelligenzforschung, die daraus resultierenden Konzeptionen und die sich daraus abgeleiteten Versuche, die Konzeptionen psychometrisch zu erfassen, lassen sich nur schwer in einer kurzen Zusammenfassung darstellen, beispielsweise in einem Kapitel mit den fünf wichtigsten Intelligenztheorien. Es gibt Theorien, die einen übergeordneten Intelligenzfaktor erkennen, Theorien, die mehrere Intelligenzfaktoren gleichbedeutsam nebeneinander annehmen, Theorien, die eher die Art und Weise, wie ein Individuum zu einer Problemlösung kommt als Kern der Intelligenz annehmen und nicht unveränderliche und eher zeitstabile Persönlichkeitsmerkmale dahinter vermuten, Theorien, die negieren, dass die vorher genannten Theorien gültig sein können, da *die* Intelligenz kulturabhängig sei und also eine Intelligenztheorie an die kulturellen Gegebenheiten adaptiert sein müssen und Theorien, die Intelligenz eher genetisch bedingt vermuten und nicht umweltbedingt sowie umgekehrt. Intelligenzforschung und Intelligenzmessung werden in der Regel als gleichbedeutsam beschrieben. Allerdings wäre es auch möglich, Intelligenz unabhängig von der Messung zu untersuchen. Es ist denkbar, dass eine Definition von Intelligenz richtig ist, diese aber nicht mit Intelligenztests belegt werden kann. Es gibt viele methodische Schwierigkeiten bei der Bestimmung eines Intelligenzquotienten, die vielfach im Rahmen dieser Arbeit beschrieben werden. Es wäre also auch möglich, Intelligenz unabhängig von der Intelligenzmessung zu untersuchen. Da Intelligenz und Intelligenzmessung in der Regel gemeinsam diskutiert, zuweilen sogar synonym verwendet werden, soll im Folgenden auch Intelligenz und Intelligenzmessung im Kontext betrachtet werden. Dies sei auch damit begründet, dass Intelligenztheorien faktorenanalytisch belegt werden. Die Faktorenanalysen wiederum resultieren aus der Durchführung von psychometrischen Verfahren zur Bestimmung von Intelligenzfaktoren. Letztlich würde eine Intelligenztheorie nicht belegt werden können ohne die Durchführung von Testverfahren. Sollten diese Testverfahren an sich methodisch fragwürdig sein, müsste bei objektiver Betrachtung festgestellt werden, dass Intelligenztheorien auf wackeligen Beinen stehen.

2.3.1 Was ist Intelligenz: eine Übersicht

„There seem to be almost as many definitions of intelligence as there were experts asked to define it.“ (Sternberg, 1987, S. 376).

Eine weite Definition von Intelligenz umfasst die akademische, praktische, soziale Intelligenz, Lernfähigkeit, Kreativität und komplexes Problemlösen (Brocke & Beauducel, 2001). Hinzu kommt die *Emotionale Intelligenz* (Goleman, 2012) und die *Künstliche Intelligenz* (Legg & Hutter, 2007). In dieser Arbeit soll sich auf die engeren Definitionen von Intelligenz bezogen werden, die kognitive

Prozesse beinhalten. Eine Darstellung von allem, was im Zusammenhang mit der Intelligenzforschung diskutiert wurde und wird, sprengt den Rahmen dieser Untersuchung. Übersichtsarbeiten zur Thematik liegen vor von Funke und Vatterodt (2009) und Lamberti (2006).

Alfred Binet gilt als Erfinder des ersten Intelligenztests (Holling, Preckel & Vock, 2004). Dieser sollte die Frage klären, wie geistig behinderte Kinder zu unterrichten seien (Zimbardo, 1992, S. 441)¹⁰. Ziel der Diagnostik mit Hilfe dieses ersten Intelligenztests war es, objektiv die Gruppe kognitiv sehr schwacher Kinder zu identifizieren und diese Identifizierung nicht dem subjektiven Urteil der Lehrkräfte zu überlassen. Somit wurde angenommen, dass ein Test den kognitiven Stand eines Kinds besser erfassen kann als das LehrerInnenurteil. Belege, dass diese Annahme begründet ist, liegen vielfach vor. Berühmt ist z.B. das Experiment von Robert Ulshöfer, der Ende der 40er Jahre einen einzigen Abituraufsatz 42 Lehrkräften zur Benotung vorlegte. Die Urteile in Form von Schulnoten verteilten sich von der Note 1–6 (Kahl, 2006). Dieses Experiment wurde 1981 von Gottfried Schröter (1981) repliziert. Die Untersuchung der Notenvergabe durch 11 000 Lehrkräfte ergab ein ähnliches Ergebnis. Eine Übersicht über die Subjektivität von LehrerInnenurteilen liefern z.B. Brügelmann (2006), Dalbert (2013) oder Zaborowski, Meier und Breidenstein (2011).

Das kurze Eingehen auf die Subjektivität von LehrerInnenurteilen sei damit begründet, dass Intelligenztests gerade auch im sonderpädagogischen Kontext eingesetzt und legitimiert werden, um der Subjektivität entgegenzuwirken.

In dem Test von Binet, den er zusammen mit seinem Kollegen Theophile Simon entwickelte, werden Kindern Testaufgaben vorgelegt, die im Schwierigkeitsgrad ansteigen, ein typisches Merkmal auch heutiger Intelligenztests. Es wurde für die Normierung das durchschnittliche Ergebnis von Kindern verschiedener Altersstufen erfasst. Das Ergebnis des getesteten Kinds wurde dann verglichen mit diesen Durchschnittswerten und dementsprechend erhielt das getestete Kind ein Intelligenzalter, je nachdem, welcher Altersgruppe das Testergebnis entsprach. Das Berechnen eines Intelligenzalters verlor durch das Berechnen eines Intelligenzquotienten (Stern, 1914) an Bedeutung, welches das Verhältnis des Intelligenzalters zum Lebensalter darstellt ($IQ = \text{Intelligenzalter} / \text{Lebensalter} \times 100$). Doch gerade bei der Präsentation von Testergebnissen gegenüber Sorgeberechtigten wird wieder häufig das Intelligenzalter als Beschreibung für Testergebnisse genutzt, welches nun *Referenzalter* oder *Äquivalenzalter* genannt wird. Viele der neu aufgelegten Testverfahren (z.B. WISC-IV,

10 Bereits an dieser Stelle würden KritikerInnen einwenden, dass Diagnostik neutral ist und die daraus resultierenden pädagogischen Ableitungen erst einmal nichts mit den diagnostischen Befunden zu tun haben und sich mit dieser auch nicht begründen lassen (Schlee, 2008), sondern mit (pädagogischen) Idealen, Normvorstellungen und Werten, z.B. über eine gute Beschulung, eine gute Pädagogik etc.

KABC-II) bieten im Anhang der Manuale Tabellen zum schnellen Errechnen des Referenz- bzw. Äquivalenzalters. In den Verfahren der SON-Reihe wird das Referenzalter gar in der Computerauswertung immer mit angegeben.

Theoretische Grundlage des Binet-Simon-Tests war die Annahme, Intelligenz zeige sich in komplexen Denkvorgängen wie z. B. Urteilsfähigkeit und Anpassungsfähigkeit. Es wurden verbale, numerische und räumliche Aufgaben vorgegeben (Kramer, 2009, S. 20). Der Test beruhte auf den Kriterien, dass er in der Anwendung standardisiert und die Schwierigkeitsanordnung der Aufgaben empirisch geprüft ist, die Validität und Reliabilität als Gütekriterien vorausgesetzt sind und die Stichprobe von Verhaltensweisen nicht in separate Teilfähigkeiten zergliedert, sondern zur allgemeinen Intelligenz vereinigt werden (Speckmeier, 2011, S. 93). Das Testverfahren dieser „Urväter der Testpsychologie“ (ebd., S. 94) wurde in revidierten Fassungen noch Jahrzehnte angewendet.

Spearman (1904) beschrieb die Allgemeine Intelligenz mit dem Generalfaktor g . Demnach ist ein allgemeiner Intelligenzfaktor g maßgeblich beteiligt an allen kognitiven Fähigkeiten. Daraus resultiert, dass eine Person mit einem hohen Intelligenzfaktor g praktisch in allen kognitiven Teilbereichen hohe Werte erzielen würde. Dies widerspricht konsequent zu Ende gedacht dem Stereotyp vom zerstreuten Professor, der in seinem Fach zwar nobelpreisverdächtig agiert, aber ansonsten völlig schusselig nicht in der Lage wäre, sich ein Spiegelei zu braten. Tatsächlich müsste bei Vorliegen eines allgemeinen Intelligenzfaktors angenommen werden, dass der Professor auch gute Strategien beim Kochen entwickeln müsste (Anpassungsfähigkeit im Umgang mit den räumlichen Gegebenheiten, beim sinnvollen Arbeiten mit den Kochwerkzeugen, beim divergenten Umgang mit den Zutaten, beim Erstellen eines Zeitplans, beim Lernen aus vorher gemachten Erfahrungen, z. B. Essen falsch gewürzt usw.)¹¹. Wer also über eine hohe Grundintelligenz verfügt, dargestellt durch den Generalfaktor g , kann auf sein gutes intellektuelles Potential in allen Bereichen des Lebens zurückgreifen, z. B. beim Lernen einer Sprache, beim Rechnen, beim logischen Denken usw. Die allgemeine Intelligenz g kann besser als IQ dargestellt werden als das Postulieren verschiedener unabhängiger Intelligenzen, die gleichberechtigt nebeneinander existieren. Diese müssten eigentlich mit mehreren IQs dargestellt werden, da es sich nicht um ein hierarchisches Modell wie bei Spearman handelt. Thurstone (1938) postulierte ein nicht-hierarchisches Modell der Intelligenz in Form von gleich bedeutsamen Primärfaktoren, die nicht abhängig von einem Generalfaktor sind (Holling, Preckel & Vock, 2004; Kramer, 2009): *verbales Verständnis, Wortflüssigkeit, schlussfolgerndes Denken, räumliches Vorstel-*

11 Tatsächlich entspricht das Stereotyp vom zerstreuten Professor wohl eher dem Konzept von Savants: Teilleistungsbegabte oder Inselbegabte, die teils Symptome aus dem autistischen Spektrum zeigen.

lungsvermögen, Merkfähigkeit, Rechenfähigkeit und Wahrnehmungsgeschwindigkeit. Bei Betrachtung aktueller Testverfahren, die im sonderpädagogischen Bereich eingesetzt werden, kommen die Primärfaktoren bekannt vor, denn diverse Subtests der Testbatterien testen diese Primärfaktoren, auch wenn sich nicht ausdrücklich auf Thurstone berufen wird. *Schlussfolgerndes Denken* kommt in Form von Matrizen-tests in fast jedem Test vor, *Rechenfähigkeit* in der IDS und im WISC-IV, *Wahrnehmungsgeschwindigkeit* in Form von Speed-Tests in fast jeder Testbatterie, Rotationstests (*räumliches Vorstellungsvermögen*) z.B. in der KABC-II und im WISC-IV. Obwohl er zunächst die Primärfaktoren als unkorreliert ansah (Kramer, 2009, S. 21), ging er später doch von einer Korrelation aus und nahm an, dass ein Generalfaktor höherer Ordnung extrahiert werden könne. Der Wilde-Intelligenztest (Jäger & Althoff, 1983) und ältere Formen des Intelligenz-Struktur-Tests (Amthauer, 1953, 1973) berufen sich auf Thurstones Annahmen über die Intelligenz (Kramer, 2009, S. 21).

Mit Stern zeigten sich erste Ansätze, die das Zusammenwirken von Anlage und Umwelt thematisieren, z. B. die Anpassungsfähigkeit des Individuums auch unter verschiedensten Bedingungen und auf verschiedensten Gebieten (Stern, 1912, S. 4). Stern definiert die Intelligenz als „die allgemeine Fähigkeit eines Individuums, sein Denken bewusst auf neue Forderungen einzustellen; sie ist allgemeine geistige Anpassungsfähigkeit an neue Aufgaben und Bedingungen des Lebens“ (1912, S. 3).

Der von Stern eingeführte Begriff Intelligenzquotient erhielt seinen auch heute noch verwendeten Namen durch Wechsler, der den Intelligenzquotient (IQ) als Abweichungsquotienten einführte (Wechsler, 1958). Wechsler betrachtete die Hypothese von Stern und Binet bei der Berechnung der Testergebnisse in Form von Intelligenzalter und Intelligenzquotient als problematisch, da die Beziehung zwischen Intelligenz und Alter nicht linear sei (Speckemeier, 2011, S. 95). Er schlug den IQ als Abweichungsquotienten vor, der auch Vergleiche im Alter zulässt (ebd., S. 95; Amelang & Bartussek, 1990, S. 181). Die Bezeichnung IQ als Größe für die Intelligenz einer Person erhält damit auch heute noch eine große Bedeutung, weil er sich unabhängig vom biologischen Alter einer Person als feste Größe darstellt. Somit sind 70-Jährige mit 20-Jährigen vergleichbar. Die Testverfahren aus der Wechsler-Reihe (z. B. Wechsler-Bellevue; HAWIK-R; HAWIK-III; HAWIK/WISC-IV; WNV)¹² berufen sich in den Be-

12 Wechsler-Bellevue: Wechsler Bellevue Intelligence Scale (Wechsler, 1939).

HAWIK-R: Hamburg-Wechsler-Intelligenztest für Kinder-Revision 1983 (Tewes, 1983).

HAWIK-III: Hamburg-Wechsler-Intelligenztest für Kinder (Tewes, Rossmann & Schallberger, 1999).

HAWIK/WISC-IV: Hamburg-Wechsler-Intelligenztest für Kinder-IV (Petermann & Petermann, 2007).

WNV: Wechsler Nonverbal Scale of Ability (Petermann, 2014).

gründungen auf die theoretischen Überlegungen zur Intelligenz von Raymond Cattell. Cattell (ein Schüler Spearman) postulierte, dass Intelligenz aus zwei allgemeinen Faktoren bestehe, der *fluiden* und der *kristallinen* Intelligenz (Cattell, 1957). *Fluide* Intelligenz beschreibt die kognitiven Fähigkeiten, deren Vorliegen eher genetisch bedingt angenommen werden und mit dem intellektuellen Potential einer Person umschrieben werden kann. Während die Ausprägung der *fluiden* Intelligenz überwiegend genetisch erklärt wird, soll die eher umweltbedingte *kristalline* Intelligenz die Summe von Lernerfahrungen sein. Dabei wird angenommen, dass Lernerfahrungen sich eher kumulieren und herauskristallisieren, je mehr man das intellektuelle Potential in Form der *fluiden* Intelligenz nutzt und dieses in die *kristalline* Intelligenz investiert, weshalb dieser Ansatz in der Intelligenzforschung auch Investmenttheorie (Cattell, 1963) genannt wird. Rindermann, Flores-Mendoza und Mansur-Alves (2010) setzen sich kritisch mit dieser Theorie auseinander.

Entsprechend dieses Ansatzes erklärt sich auch, dass z. B. im Intelligenztest HAWIK-IV reine Wissensaufgaben vorkommen wie folgende: *Woraus bestehen Diamanten?* (richtige Antwort z. B.: *Kohlenstoff/Karbon*; falsche Antwort z. B.: *Kohle*) (Petermann & Petermann, 2007, S. 327). Nach der Theorie Cattells und der Anwendung dieser Theorie im HAWIK-IV müsste also angenommen werden, dass ein Kind sein intellektuelles Potential (*fluide* Intelligenz) investiert haben sollte, um sich damit zu beschäftigen, woraus Diamanten bestehen (Folge der Kenntnis: *kristalline* Intelligenz). Obwohl die Theorie umstritten ist und die Grundannahmen bezweifelt werden (Horn, 1998, zitiert nach Johnson & Boucard, 2005; Holling et al., 2004), dient die Erklärung der Intelligenz nach Cattell, erweitert und modifiziert von Horn und Carroll vielen Intelligenztests, die in der Sonderpädagogik eingesetzt werden, als Grundlage und wird mit dem CHC-Modell beschrieben: Cattell-Horn-Carroll Modell.

Guilford (1967) beschreibt mit dem *Strukturmodell der Intelligenz* 120 intellektuelle Fähigkeiten. Drei Faktoren der Intelligenz (Inhalt, Produkt/Form, Operation) bestimmen dabei die jeweilige intellektuelle Fähigkeit. Dieses Modell als Analogie zur Tafel der Elemente (Zimbardo, 1992, S. 447) diene als Grundlage für die Forschung verschiedenster kognitiver Fähigkeiten, ist aber zumindest in Teilaspekten schlecht empirisch belegt (Kramer, 2009, S. 23) und sei an dieser Stelle der Vollständigkeit halber erwähnt.

Als theoretische Grundlage für die Konstruktion von Intelligenztests, die in der deutschen Sonderpädagogik eingesetzt werden, finden weder Strukturmodelle wie das nach Guilford, noch Intelligenzmodelle mit starker Berücksichtigung der kulturellen Hintergründe, noch theoriegeleitete Modelle von Intelligenz Verwendung. Letztere, weil hier der Beleg über psychometrische Verfahren abgelehnt wird. Bedeutsame Intelligenzmodelle nach diesen Ansätzen sollen an dieser Stelle erwähnt werden, für eine tiefergehende Beschäftigung wird angesichts der Fragestellung auf weiterführende Literatur verwiesen, z. B.

von Sternberg und Detterman (1986), Funke und Vaterrodt (2009) und Lamberti (2006).

Einen wichtigen Beitrag in der Intelligenzforschung wird den Ansätzen von Gardner, Sternberg und Jäger attestiert, so dass diese hier kurz beschrieben werden.

Gardner (1983) betrachtet Intelligenz im kulturellen Kontext. In westlichen Gesellschaften wird z. B. mehr Wert auf *linguistisch* und *logisch-mathematische Fähigkeiten* gelegt. Gardner beschreibt sieben Arten von Intelligenz. Auf Bali wäre dementsprechend die *körperlich-kinästhetische Fähigkeit* (Fertigkeiten der motorischen Bewegung und Koordination) von Bedeutung (Zimbardo, 1992, S. 448). Eine Testung über ein psychometrisches Verfahren wird abgelehnt, um eine Einschätzung über eines oder mehrere der sieben Intelligenzen zu erhalten. Gardner postuliert neben den oben erwähnten Intelligenzen *räumliches Vorstellungsvermögen*, *musikalische Fähigkeit*, *interpersonale Fähigkeit* (Verstehen anderer) und die *intrapersonale Fähigkeit* (Verstehen des Selbst). Auch wenn im Kontext dieser Arbeit dieser Ansatz zur Definition von Intelligenz wenig bedeutsam ist, stellt sich die Frage, ob nicht in der Tat der kulturelle Hintergrund bei der Testung im sonderpädagogischen Kontext mehr beachtet werden sollte. Die Schaffung kultur- und sprachfairer Testverfahren ist ein berechtigtes Anliegen, welches vielfach versucht wurde zu lösen. So sind die aus überwiegend abstrakten Symbolen bestehenden Aufgaben aus der CFT-Reihe der Versuch, kulturelle Hintergründe nicht in ein Testergebnis einfließen zu lassen. Der CFT heißt ausgesprochen *Culture Fair Intelligence Test*. Die Testverfahren aus der SON-Reihe (Snijders-Oomen non-verbaler Intelligenztest) sind der Versuch einer Intelligenztestung gänzlich ohne Worte (auch die TestleiterIn muss nicht sprechen, die Anweisungen können pantomimisch-gestisch erläutert werden). Angesichts der steigenden Zahl von in Deutschland aufgenommenen geflüchteten Kindern bleibt die Frage nach einer Berücksichtigung des kulturellen Hintergrunds des zu testenden Kinds berechtigt. Es kann also diskutiert werden, ob Ansätze nach Berücksichtigung des jeweils kulturellen Hintergrunds bei der Bestimmung von Intelligenz wie der nach Gardner an Aktualität gewonnen haben.

Die meisten Intelligenzmodelle gehen davon aus, dass Intelligenz ein vorhandenes Maß an kognitiven Fähigkeiten (und je nach Modell auch Fertigkeiten) darstellt, welches abgerufen werden kann. Von dieser Vorstellung ausgehend müssen also Verfahren entwickelt werden, die dieses individuelle Maß abrufen können, z. B. mit einem Intelligenztest. Sternberg (1985, 1986) versucht, die Intelligenz als einen Weg zu einem Ziel zu beschreiben, nicht als die feststehende kognitive Kompetenz, die zu einem Ziel führen muss, wenn man intelligent genug sei.

Intelligenz besteht nach Sternberg aus drei fundamentalen Aspekten: *analytische*, *kreative* und *praktische* Intelligenz (Sternberg, 1985). Differenzierter be-

schreibt Zimbardo (1992) die Intelligenz-Triade mit *komponentenbezogener*, *erfahrungsbezogener* und *kontextabhängiger* Intelligenz (ebd., S. 448). Die *komponentenbezogene* Intelligenz ist psychometrisch erfassbar und untersucht kognitive Prozesse auf dem Weg zu einer Lösung, z.B. eines Items in einem IQ-Test. Nicht erfassbar mit psychometrischen Verfahren ist die *erfahrungsbezogene* Intelligenz, die das innere Erleben einer Person im Zusammenhang mit der Umwelt beschreibt. In diesem Zusammenhang ergibt sich die Frage danach, wie Intelligenz Erfahrungen beeinflusst.

Ebenfalls nicht psychometrisch erfassbar ist die *kontextabhängige* Intelligenz, die beschreibt, wie eine Person die Umwelt beeinflusst, z.B. ob vorhandene Gegebenheiten optimal und klug genutzt werden, ob eine Person also den Kontext der Umwelt erfasst und sinnvoll nutzt. Nach Sternbergs Ansatz ist es durchaus möglich, dass eine Person in einem Intelligenztest ein schwaches oder gar sehr schwaches Resultat erzielt, aber durch eine sinnvolle Integration von Umwelterfahrungen und eine sinnvolle Beeinflussung der Umwelt durch Berücksichtigung des Umwelt-Kontextes sehr gut zurecht kommt¹³. Ausgehend von diesem Intelligenzmodell könnte sich die gelegentlich von SonderpädagogInnen empfundene Diskrepanz erklären zwischen den Testergebnissen aus Intelligenztests und dem guten Zurechtkommen im Schulalltag, in den sozialen Beziehungen, bei der Lösung von Problemen und auch beim kreativen Umgang mit Etwas, denn das Konstrukt Kreativität beinhaltet Sternbergs Intelligenz-Triade.

Das Berliner Intelligenzstruktur Modell nach Jäger (1982, 1984) resultiert aus der Auseinandersetzung mit 2000 Items aus Intelligenztests (Jäger, Süß & Beauducel 1997a). Diese konnten zu 191 Blöcken mit 98 Aufgabentypen extrahiert werden. Aus der anschließenden Analyse der Struktur erkannte Jäger vier operative Fähigkeiten: *Bearbeitungsgeschwindigkeit*, *Merkfähigkeit*, *Einfallstiefe* (auch *Kreativität*) und *Verarbeitungskapazität*. Durch eine Kreuzklassifikation (Speckemeier, 2011, S. 107) konnten die *inhaltsgebundenen* Fähigkeiten *sprachgebundenes*, *zahlengebundenes* und *anschauungsgebundenes (figural-bildhaftes)* Denken nachgewiesen werden. Es wird auch in dieser Intelligenztheorie ein übergeordneter Generalfaktor der Intelligenz angenommen.

Die Leistung Jägers besteht darin, ein Aufgabenpool zu erstellen, welches für die bis dahin eingesetzten Intelligenzaufgaben in der Intelligenzforschung repräsentativ ist. Aus diesem Aufgabenpool resultierte der Berliner Intelligenzstruktur-Test (Jäger, Süß & Beauducel, 1997b). Insgesamt gilt das Berliner Intelligenzstruktur-Modell als empirisch gut bewährt (Brocke & Beauducel, 2001).

13 Umgangssprachlich (und auch diskriminierend gegenüber LandwirtInnen) wird diese Kombination (niedriger IQ, dennoch lebenspraktisch agierend) manchmal abwertend mit Bauernschläue umschrieben.

2.3.2 Intelligenzmodelle im sonderpädagogischen Kontext

Bei der Konstruktion eines Intelligenztests ist das Erfinden von Testaufgaben sicherlich der einfachste und erfreulichste Teilschritt. Jede/Jeder hat eine Vorstellung davon, mit welchen Items Intelligenz getestet werden könnte. Aufwendig wird im Anschluss die Prüfung, ob diese Items auch tatsächlich Intelligenz testen oder nicht, erschwert zudem durch den Mangel, dass es kein *State of the Art* bezüglich der Definition von Intelligenz gibt. Noch aufwändiger wird die Durchführung statistischer Verfahren, mit deren Hilfe Gütekriterien die Qualität des Tests insgesamt belegen. Am aufwändigsten ist aber sicherlich die anschließende Normierung des Tests mit Hilfe einer Normstichprobe, die aus einer ProbandInnenzahl mindestens im vierstelligen Bereich (bzw. je Altersgruppe mindestens 30 ProbandInnen) bestehen sollte und die auch den oft erstaunlich hohen Kaufpreis der Tests begründet. Es kann ausgeschlossen werden, dass dies alles ohne eine theoretische Vorstellung von dem gelingt, was Intelligenz sein soll. Ideal ist natürlich, auf ein selbst erstelltes theoretisches und am besten auch gut belegtes Konzept über Intelligenz zugreifen zu können bei der Konstruktion eines Tests wie z.B. bei Cattell mit dem daraus entwickelten CFT oder wie bei Jäger mit dem daraus entwickelten Berliner-Intelligenzstruktur-Test. Doch ist es untypisch bei den in der Sonderpädagogik eingesetzten Intelligenztests, dass die AutorInnen auf eigene Konzeptionen bezüglich dessen, was Intelligenz sein soll, zugreifen. Während im vorherigen Kapitel die Meilensteine der Intelligenzforschung vorgestellt worden sind, ohne weitestgehende konkrete Bezüge zu aktuell durchgeführten Intelligenztests herzustellen, die in der Sonderpädagogik gebräuchlich sind, sollen nun die konkreten Intelligenztheorien dargestellt werden, auf die sich die in der Sonderpädagogik verwendeten Intelligenztests beziehen.

Die Definition für einen psychologischen Test orientiert sich an Moosbrugger & Kelava (2007, S. 2), die einen Test als ein wissenschaftliches Routineverfahren zur Erfassung eines oder mehrerer empirisch abgrenzbarer psychologischer Merkmale mit dem Ziel einer möglichst genauen quantitativen Aussage über den Grad der individuellen Merkmalsausprägung definieren.

2.3.2.1 Lurija-Modell

Neben der Idee, Intelligenz mit Hilfe von auf Faktorenanalysen begründeten psychometrischen Verfahren testen zu wollen, gibt es auch den kognitionspsychologischen Ansatz, Intelligenz zu testen (Maltby, Day & Macaskill, 2011). Biologische und physiologische Unterschiede werden hier in Verbindung mit Intelligenz gebracht. Eine Idee ist z.B. die Annahme über eine positive Korrelation zwischen Gehirngröße und Intelligenz(testergebnissen). Durch entspre-

chende Korrelationsstudien konnte diese ursprünglich von Tiedemann (1836) postulierte Annahme später bestätigt werden (Willerman et al., 1991; McDaniel, 2005)¹⁴. Einen ebenfalls biologisch-physiologischen Ansatz verfolgte Jensen (1998). Die Bewältigung kognitiver Aufgaben benötigt unterschiedliche Bearbeitungszeiten, deren Länge mit Intelligenz in Verbindung gebracht wird. So nimmt er z. B. an, dass die über ein evoziertes Potenzial gemessene Verarbeitungszeit (erkennbar über ein Spitzenpotenzial im EEG) mit Intelligenz in Verbindung steht, eine kürzere Verarbeitungszeit bedeutete eine höhere Intelligenz. Jensen vermutet generell eine bessere Messbarkeit der Intelligenz über kurze und einfache kognitive Aufgaben (z. B. die Geschwindigkeit, mit der auditive oder visuelle Reize erkannt werden), bei denen Wissen, Schlussfolgern und Problemlösungstechniken nicht im Vordergrund stehen wie bei herkömmlichen Intelligenztests (Maltby et al., 2011, S. 553).

Maltby et al. (2011) beschreibt einen wichtigen kognitionspsychologischen Ansatz mit den Arbeiten von Lurija (1902–1977). Alexander Romanowitsch Lurija¹⁵ (1970) entwickelte eine „Karte der Systeme und Funktionen des Gehirns, die für komplexe Verhaltensprozesse verantwortlich sind“ (Melchers & Melchers, 2015, S. 43). Diese Einteilung des Gehirns in drei Blöcke, die für unterschiedliche Prozesse verantwortlich gemacht werden im Zusammenhang mit dem Abruf von intellektuellen Fähigkeiten, dienten dem Ehepaar Nadeen und Alan Kaufman u. a. als theoretische Grundlage bei der Konstruktion der K-ABC und der KABC-II.

Lurija war Psychologe, der sein Psychologie- und Medizinstudium sowie sein Studium der Gesellschaftswissenschaften bereits mit 16 Jahren an der 800 Kilometer östlich von Moskau gelegenen Universität von Kazan begann. Seinen Universitätsabschluss erzielte er mit 19 Jahren, somit als Teenager. Zusammen mit Lew Wygotskij und Alexej Leont'ev begründete er die *Kulturhistorische Schule*, die ein Zusammenwirken von physischer Entwicklung und sensorischen Mechanismen mit kulturellen Faktoren zur Hervorbringung psychologischer Prozesse und Funktionen (einschließlich der Intelligenz) bei Erwachsenen (Maltby et al., 2011, S. 554) untersucht. Im Gegensatz zu Tieren sind Menschen in der Lage, sich nicht nur an die Umgebung anzupassen, sondern sich verschiedene Fertigkeiten anzueignen, die dann zu Verinnerlichungen führen und Kognitionen steuern können. Im Zusammenhang dieser Studie sind seine Untersuchungen zur Messung von Denkprozessen interessant. Lurija fand z. B. heraus, dass die Bearbeitungszeit eines Items bei inneren emotionalen Konflik-

14 Konsequent – und unter Ausblendung von Assoziationen zu der Rassenlehre im 3. Reich – zu Ende gedacht resultiert daraus ein höheres intellektuelles Potenzial für größere Menschen, auf den Punkt gebracht von McDaniel (2005) mit folgender Artikelüberschrift: *Big brained people are smarter*.

15 Gelegentlich auch Luria geschrieben.

ten verlängert ist und entwickelte zur Messung die gekoppelte motorische Methode (Lurija, 1932). Auf Grund dieser Beobachtungen wurde später der Polygraf (Lügendetektor) entwickelt (Maltby et al., 2011, S. 553).

Die Ideen von Lurija, Vygotskij und Leont'ev sowie die Aufsätze zu dem Wirken dieser Forschergruppe sind sehr differenziert, auch durch die Ablehnung von Vereinfachungen komplexer Sachverhalte. Veranschaulichungen über Schematisierungen werden als unzulässig reduktionistisch verworfen, gezielt anti-reduktionistische Positionen werden vertreten, z.B. ein Objekt niemals in seiner Statik, sondern seiner Entwicklung zu untersuchen (Jantzen, 2004). In diesem Zusammenhang ist interessant, dass Lurija nie einen standardisierten Test entwickelte, da die Einzigartigkeit des Menschen eine individuelle Anpassung erfordere (Maltby et al., 2011). Eine umfassendere Betrachtung von Lurijas, Vygotskijs und Leont'evs Wirken und zu der *Kulturhistorischen Schule* bietet Jantzen (2003, 2011), eingehendere Betrachtungen von Lurijas Theorien bietet er in seiner Veröffentlichung *Human brain and psychological processes* (1966).

Aus der oben beschriebenen von Lurija angestrebten Entwicklung einer Karte der Systeme und Funktionen des Gehirns, welche für komplexe Verhaltensprozesse verantwortlich sind, „insbesondere der auf hohem Funktionsniveau ablaufenden Prozesse, die mit der Aufnahme und der Integration von Informationen sowie mit Problemlösefähigkeiten assoziiert sind“ (Melchers & Melchers, 2014, S. 43), entstand die Idee eines funktionalen Systems, welches durch drei Blöcke veranschaulicht wird, die die basalen Funktionen des Gehirns zusammenfassen (ebd., S. 45):

Der 1. Block, der mit dieser Terminologie wie auch die anderen Blöcke als veranschaulichende Bezeichnung für ein funktionales System betrachtet werden sollte, ordnet Lurija dem retikulären Aktivierungssystem (Medulla Oblangata) zu und ist im Wesentlichen für Wachheit und Aufmerksamkeit verantwortlich und eng verbunden mit dem 3. Block, da beide Blöcke sich mit der Gesamteffizienz der Hirnfunktionen beschäftigen (Melchers & Melchers, 2015, S. 43). Dieser 3. Block lokalisiert sich in präfrontalen Anteilen des Frontallappens, in dem sich Handlungskonzepte und geplante Verhaltensweisen entwickeln.

Hauptsächlich mit der Speicherung, Kodierung und Analyse von Informationen wird der nahe der Rolandofurche befindliche 2. Block assoziiert. Viele Subtests der K-ABC und K-ABC-II beziehen sich auf die diesem 2. Block zugeordnete Verarbeitung von visuellen, auditiven, haptischen und kinästhetischen Stimuli.

Im Sinne Lurijas soll auf eine schematische und somit reduktionistische Darstellung der drei Blöcke z.B. mit Hilfe eines Schaubilds mit definierten Aufgabenstellungen definierter Hirnteile verzichtet werden, denn letztlich geht es um die komplexe Interaktion zwischen den von Lurija angenommenen Eigenschaften der drei Blöcke, die erst menschliches Verhalten und das Abrufen von Fähigkeiten erklären. Dementsprechend wird mit der Konstruktion der

KABC-II (und der ehemaligen K-ABC) versucht, die Art und Weise zu ermitteln, wie die den drei Blöcken zugeschriebenen Aspekte integriert sind und interagierend genutzt werden können. Eine weitere wichtige Annahme in Lurijas Modell ist die Annahme über zwei verschiedene Formen der Verarbeitung, der sukzessiven und der simultanen Verarbeitung (Lurija, 1966, S. 74). Diese Arten der Verarbeitung werden vor allem mit dem 2. Block in Verbindung gebracht, also mit der Integration, Speicherung und Kodierung von Sinneswahrnehmungen. So werden seriell präsentierte Stimuli wie Zahlenfolgen (z. B. 5-7-2-9) vom Kind mit Hilfe des Kurzzeitspeichers gemerkt und wiederholt. Es folgt ein Item nach dem anderen, welches also nach und nach (sukzessiv) bzw. sequentiell (Terminologie der KABC-II) bearbeitet wird. Die Skala *Simultan* der KABC-II hingegen präsentiert dem Kind Items, die integrativ und unter Berücksichtigung komplexer Wahrnehmungen und unter Einbezug mehrerer Aspekte simultan bearbeitet werden müssen. In der der KABC-II vorangegangenen K-ABC wurden diese beiden Skalen noch *einzelheitliches* und *ganzheitliches* Denken genannt¹⁶.

2.3.2.2 Kramer-Modell in Anlehnung an den Binet-Simon-Test

Obwohl in der Differentiellen Psychologie bereits Galton (1869) Versuche unternahm, mit Hilfe von systematischen Prüfverfahren Unterschiede zwischen Menschen zu belegen und mit ihm der Testbegriff Einzug in die Psychologie hielt, und obwohl er auch viele aktuell noch gebräuchliche statistische Verfahren entwickelte, wird im Zusammenhang mit Intelligenztests die Binet-Simon-Skala (Binet & Simon, 1905) als erster wirklicher Intelligenztest im heutigen Sinne betrachtet. Daran ändert auch nicht die von James McKeen Cattell publizierte erste Testbatterie 1890 mit sensorischen, motorischen und teils kognitiven Aufgaben (Lamberti, 2006, S. 13).

Alfred Binet (1857–1911) studierte nach einem Jurastudium Medizin und Biologie. Seine Promotion beschäftigte sich mit dem Nervensystem von Insekten (Funke, 2006). Er gründete ein psycho-physiologisches Laboratorium sowie eine psychologische Fachzeitschrift, die er bis zu seinem Tod durch einen Hirntumor 1911 leitete (Funke, 2006, S. 25). Er ist Verfasser von rund 300 Fachartikeln. Zusammen mit Théodore Simon, den Binet ursprünglich als studentischen Mitarbeiter aufnahm, entwickelte er 1905 einen Intelligenztest im Auftrag des französischen Erziehungsministeriums¹⁷. Intelligenz als wichtigstes Kriterium

16 Man nahm an, dass die Gedanken Lurijas nicht sehr verbreitet sind und hatte sich dann für diese eingedeutschten Begriffe entschieden. Inzwischen wird davon ausgegangen, dass die Begriffe *sukzessive* bzw. *sequentielle* und *simultane* Verarbeitung bekannter sind.

17 Eine von Piaget später entwickelte Affinität für das Erfassen von Leistungen soll übrigens auf die Beschäftigung Piagets mit Simon zurückgehen (Funke, 2006, S. 26).

für Prognosen bezüglich zukünftiger Schulerfolge wurden zunächst angenommen, bekanntlich in späteren Jahren bestätigt (Gottfredson, 2002). Folgende Kriterien waren wichtig für die Entwicklung des Tests (Groffmann, 1983):

- Standardisierte Durchführung und Auswertung,
- empirische Itemprüfung nach Schwierigkeitsgrad,
- Voraussetzung von Validität und Reliabilität,
- keine Aufteilung in Teilfähigkeiten, sondern die Postulierung einer allgemeinen Intelligenz.

Binet nahm bereits hier einen Generalfaktor der Intelligenz an, der später mit *g* bezeichnet werden wird. Er nahm weiter an, dass die Intelligenz hierarchisch in einer Stufenleiter der Intelligenz aufgebaut ist (*échelle métrique de l'intelligence*, Speckemeier, 2011, S. 93). Dies entspräche den späteren speziellen Faktoren, die dem Generalfaktor untergeordnet sind. Die Konstruktion des Fragebogenkatalogs beinhaltete ein Ansteigen des Schwierigkeitsgrads der Fragen. Nach Auswertung der Fragen kamen Binet und Simon der Antwort näher, ob das Kind „gut urteilen, gut verstehen und gut denken könne“ (Amelang & Bartussek, 1994, S. 177), welches die Grundannahme über die Definition von Intelligenz neben dem Vorliegen eines allgemeinen Intelligenzfaktors nach Binet und Simon ist. In modifizierten Skalen wurde später versucht, kritische Einwände zu berücksichtigen, z. B. den großen Zeitaufwand bei der Durchführung, Mängel in der Durchführungsobjektivität und die Vernachlässigung von Intelligenzstrukturen (Speckemeier, 2011, S. 94). Der von Stern eingeführte Intelligenzquotient, der damals noch kein Abweichungsquotient war (dies wurde erst später von Wechsler in der heute gebräuchlichen Form vorgeschlagen), sondern der Quotient von Intelligenzalter und Lebensalter, wurde übernommen.

Der Test von Binet und Simon fand starke Beachtung und verbreitete sich weltweit in den unterschiedlichen Staaten oder in den besonderen Bedürfnissen der Kinder entsprechenden revidierten Fassungen, wie z. B. das Binetarium von Norden (1956), der Binet-Test für Blinde von Strehle (1961) oder der Kramer Intelligenz-Test (Kramer, 1972).

Josefine Kramer ist eine in der Bodenseeregion geborene Heilpädagogin, die später die Schweizer Staatsbürgerschaft annahm. Sie gilt als eine der „hochgeschätzten Heilpädagoginnen“ (Berger, 2014, S. 24), die aus einer kinderreichen Familie stammte und in eher ärmlichen Verhältnissen aufwuchs. Sie beschäftigte sich mit sprachauffälligen Kindern (insbesondere mit dem Sigmatismus) und leitete mehrere Jahre eine Erziehungsberatungsstelle. Unter anderem auch in Würdigung ihrer langjährigen psychodiagnostischen Tätigkeit erhielt sie 1963 als erste Frau von der Philosophischen Fakultät der Universität Fribourg die Ehrendoktorwürde. Der von ihr entwickelte Kramer-Test (Kramer, 1972) basierte auf dem Binet-Simon-Test. Zwischen der ersten Version, die noch Binet-

Simon-Kramer-Test¹⁸ hieß und der zuletzt erschienenen Version 1972 (ebd., 1972) entwickelte sich der Kramer-Test zu einem der am häufigsten angewendeten Testverfahren im deutschsprachigen Raum.

In der vierten Auflage besteht der Test aus verbalen Items, einigen Handlungs- und einigen Zeichenaufgaben (Seidler-Brandner, 2002) und einem optionalen Labyrinth-Test nach Porteus (1965). Geeicht wurde der Test für Kinder von 3 bis 15 Jahren ($N = 2719$), Resultat ist ein Intelligenzquotient nach Stern (Quotient aus Intelligenzalter und Lebensalter). Dieser Stufentest beinhaltet für die Kleinkind-Altersstufen auch nichtsprachliche Aufgaben, Kinder der oberen Altersstufen führen vermehrt Denkaufgaben durch. Es gibt eine Kurzform, aber keine Gruppenform. Die Instruktionen werden wortwörtlich vorgelesen; evtl. dürfen sie ergänzt werden. Für diesen Fall gibt es genaue Hinweise, in welcher Form ergänzend die Subtests erläutert werden dürfen. In der Regel beginnt ein Kind in einer Altersstufe ein Jahr unter dem eigenen biologischen Alter. Es gibt Regeln, die den heute üblichen Abbruch- und Umkehrregeln entsprechen. Eine Faktorenanalyse ergab sechs Faktoren, von denen fünf wie folgt interpretiert worden sind: *Klassifikation, Sprachverständnis, Reasoning, Umgang mit sprachlich bezeichneten Inhalten, Erfassen der sprachlichen Struktur* (Seidler-Brandner, 2002, S. 182). Der Kramer-Test gilt somit als sprachlastig. Die angegebenen Gütekriterien sind zufriedenstellend.

Auch nach seiner letzten Revision war der Kramer-Test noch Jahrzehnte im Einsatz (Castello & Nestler, 2003), obwohl „das darin enthaltene Stufenkonzept als überholt betrachtet werden muss, sprachliche Formulierungen und Abbildungen der Aufgabenstellungen nicht mehr dem Zeitgeist entsprechen und das gesamte Erscheinungsbild veraltet ist“ (Grob, Meyer & Arx, 2009). In der Tradition des Kramer-Test wollte eine Gruppe von Schweizer PsychologInnen diesen überarbeiten und neu normieren. Den Anspruch der Gruppe, den überarbeiteten Kramer-Test nicht mehr als Stufentest zu konstruieren, ergänzend psychomotorische, sprachliche, mathematische, motivationale und sozial-emotionale Aspekte zu erfassen, und das Arbeits-, Bild- und Spielmaterial attraktiver zu gestalten, stieß an seine Grenzen, so dass eine vollständige Überarbeitung nahelag (Grob et al., 2009, S. 148). Eine Überarbeitung des Kramer-Test wurde verworfen, eine in der Tradition des Kramer-Test entwickelte Neukonzeptionierung wurde angestrebt und mit der Intelligence and Development Scales 2009 veröffentlicht (Grob et al., 2009).

18 In einer ganz frühen Fassung Binet-Simon-Bobettag-Kramer-Test, da der Kramer-Test nicht nur auf dem Test von Binet und Simon basierte, sondern auch auf Weiterentwicklungen der im deutschsprachigen Raum adaptierten Fassungen des Binet-Simon-Tests, maßgeblich vorangetrieben von Bobettag (1928) und Norden (1953).

2.3.2.3 CHC-Modell

Die in Deutschland von SonderpädagogInnen zur Erkennung sonderpädagogischen Förderbedarfs und in Facheinrichtungen wie z. B. Sozialpädiatrischen Zentren durchgeführten Intelligenztests basieren überwiegend auf dem CHC-Modell, so dass dieses Modell im Zusammenhang mit dieser Arbeit das bedeutendste Intelligenzmodell darstellt. Mickley und Renner (2010) fordern gar auf Grundlage dieser Feststellung, dass auch im Sinne einer besseren Vergleichbarkeit zwischen den Testergebnissen die im deutschen Sprachraum angewendeten Testverfahren sich an dem CHC-Modell orientieren sollten.

Aus der Zusammenführung verschiedener Intelligenzmodelle zu einem integrierenden Modell resultiert das CHC-Modell, wobei CHC die Initialen der maßgeblich an diesem integrierenden Modell beteiligten Forscher sind: Raymond B. Cattell, John L. Horn und John B. Carroll.

Raymond B. Cattell (1905–1998) war ein ursprünglich britischer, später US-amerikanischer Psychologe, der zunächst in Großbritannien arbeitete und lehrte. Nach seiner Immigration in die USA 1937 lehrte er als Professor Psychologie und entwickelte auf Grundlage von Faktorenanalysen das Kristallin-Fluid Modell der Intelligenz. Cattell sollte 1997 die Goldmedaille der American Psychological Association (APA) für sein Lebenswerk erhalten (Knebel & Marquardt, 2012, S. 97), die er ablehnte, um einer Beurteilung einer Untersuchungskommission der APA vorzubeugen, die sich mit Cattells eugenischen Ansichten beschäftigte. Eugenische Ideen im Zusammenhang mit der Intelligenzforschung waren bedeutsam, denn wichtige IntelligenzforscherInnen waren Anhänger der Eugenik. Deshalb soll die Bedeutung der Eugenik im Zusammenhang mit der Intelligenzforschung genauer betrachtet werden.

2.3.2.4 Exkurs: Eugenik und Intelligenzforschung

Es ist nicht nachvollziehbar, dass die ethisch und politisch umstrittene eugenische Bewegung nur am Rande – wenn überhaupt – im Zusammenhang mit der Geschichte der Intelligenzforschung Erwähnung findet. Die Befürchtung liegt nahe, dass nicht sein soll, was nicht erwähnt wird, ansonsten muss Ignoranz für die Vernachlässigung dieses Themas angenommen werden. Eine reflektierte Position zur Eugenik darf von TestanwenderInnen erwartet werden, denn die Anwendung von Intelligenztheorien und -tests, die auch unter Einbezug eugenischer Gedanken und motiviert von eugenischen Vorstellungen erstellt worden sind, würde unreflektiert zur Übernahme entsprechender Positionen führen können. So ist insbesondere der Vergleich von Gesamtwerten (Gesamt-IQs) zwischen Ethnien und/oder kulturellen Gruppen nur mit großer Vorsicht zu interpretieren.

Mit Absicht soll dieser Exkurs eingebettet werden in die Beschäftigung mit der CHC-Intelligenztheorie. Das CHC-Modell ist das derzeit bedeutsamste Intelligenzmodell und ist untrennbar mit R.B. Cattell verbunden, dessen Forschungen mit Eugenik in Verbindung gebracht werden (vgl. Haller & Niggelschmidt, 2012; Kühl, 2014).

Nach Maltby et al. (2011) hat Eugenik das Ziel, im Rahmen eines Selektionsprozesses in der menschlichen Fortpflanzung, Menschen mit erwünschten Eigenschaften hervorzubringen (z.B. mit einer höheren Intelligenz). Positive Eugenik ist hier die Förderung erwünschter Merkmale, z.B. die staatlich gelenkte Förderung der Geburtenzahl aus AkademikerInnenfamilien, da bei diesen eine höhere Intelligenz angenommen wird. Negative Eugenik ist das kontrollierte Senken der Fortpflanzungsrate bei Personen, die mit nicht erwünschten Merkmalen assoziiert sind, z.B. die Sterilisation von geistig behinderten Menschen. In der Zeit des Dritten Reichs nahm die Eugenik und die damit im Zusammenhang zu nennende Euthanasie eine zentrale Rolle ein. Vor dieser Zeit formulierte Alfred Ploetz (1895) in Deutschland in seiner Schrift *Die Tüchtigkeit unserer Rasse und der Schutz der Schwachen* die konsequente Anwendung auf Grundlage der Eugenik:

Stellt es sich heraus (...), daß das Neugeborene ein schwächliches oder mißgestaltetes Kind ist, so wird ihm von dem Ärzte-Collegium, das über den Bürgerbrief der Gesellschaft entscheidet, ein sanfter Tod bereitet, sagen wir, durch eine kleine Dosis Morphium. Die Eltern, erzogen in strenger Achtung vor dem Wohle der Rasse, überlassen sich nicht lange rebellischen Gefühlen, sondern versuchen es frisch und fröhlich ein zweites Mal, wenn ihnen dies nach ihrem Zeugnis über Fortpflanzungsfähigkeit erlaubt ist. (Ploetz, 1895, S. 144f.)

Eugenische Ansichten spiegelten sich in der ersten Hälfte des zwanzigsten Jahrhunderts in vielen Gesetzen wider, z.B. in den Gesetzen zur Zwangssterilisation in den USA, Kanada, Schweden, Australien, Norwegen, Finnland, Dänemark und der Schweiz, insbesondere aber in Deutschland während der Zeit des Nationalsozialismus, z.B. im *Gesetz zur Verhütung erbkranken Nachwuchses*, abgeleitet aus einem 1922 erlassenen Gesetz in den USA, dem *Model Eugenic Sterilisation Law* (Maltby et al., 2011).

Cattell war lediglich ein Anhänger der Eugenik in einer Reihe eugenisch denkender Intelligenzforscher. Aufsätze zur Intelligenzforschung beginnen in der Regel mit einem historischen Überblick und beginnen hier häufig mit den Forschungen Sir Francis Galtons. Genau bei Galton hat auch die Eugenik seinen Ursprung, denn der Begriff wurde von ihm geprägt. Er ging von einer Vererbung der Intelligenz aus und nahm an, dass weniger intelligente Menschen sich stärker vermehren würden und somit die menschliche Rasse schwächen würden. Das Galton Institute in London hieß ursprünglich *Eugenics Education*

Society. Lewis Terman, zeitweilig Präsident der *American Psychological Association*, war Anhänger von Galtons Theorien und unterstellte hispano-indianischen, mexikanischen und afroamerikanischen Menschen per Zugehörigkeit zu der Ethnie eine geringere Intelligenz (Maltby et al., 2011, S. 651).

Charles Spearman ist bekannt geworden durch seine Theorie des übergeordneten Intelligenzfaktors *g*, auch er ein Anhänger eugenischer Ideen. Im Zusammenhang mit Studien zur Rassenforschung half er bei der Modifizierung von Intelligenztests zur Erforschung von „nichteuropäischen und primitiven Völkern“ (Kühl, 2014, S. 104). Cattell war ein Schüler Spearmans und entwickelte Spearmans Theorien weiter in der Grundannahme eines übergeordneten Intelligenzfaktors. Cattell als Anhänger der Eugenik befürchtete, dass weniger intelligente Menschen überproportional viele Kinder bekämen (Cattell, 1936, S. 181), sein *fight for our national intelligence* begleitete er mit einer ausgesprochenen Anerkennung der in Deutschland von den NationalsozialistInnen durchgeführten Rassenverbesserungsprogrammen (Cattell, 1936, S. 141), diese Anerkennung revidierte er nach 1945 kaum (Knebel & Marquardt, 2012). Als Cattell von sich die Goldmedaille der APA ablehnte, um einer möglichen negativen Beurteilung einer Untersuchungskommission der APA vorzubeugen, wehrte er sich allerdings in einem Brief an die APA gegen Rassismuskorrekturen (Cattell, 1997, o.S.) heftig und führte unter anderem an, dass er mit der Entwicklung kultur- und sprachfairer (*fluid*) Tests gerade versucht hat, Chancengleichheit herzustellen. Vorwürfe auf Grundlage jahrzehntealter Aussagen („from material that is sixty years old“) (ebd., o.S.), würden nicht im Kontext der damaligen Zeit betrachtet, seine Ansichten seien sinntest und verdreht wiedergegeben: „Their presentation reeks with all the little tricks that journalists use. They have quoted loaded terms I have used and then surrounded them with ‚paraphrased‘ statements of my position.“ (ebd., o.S.) Im Gegensatz zu Knebel und Marquardt, die keine wirkliche Distanzierung Cattells erkennen (2012), distanziert sich Cattell deutlich in dem offenen Brief von den Verbrechen Hitlers, indem er schreibt: „We must long remember the evil actions of Hitler lest we repeat the mistake of the German people who followed his utter lunacy to violate the most fundamental of human rights.“ (ebd., o.S.). Eine Distanzierung gegen die Ideen der Eugenik nimmt er nicht in dem offenen Brief vor.

Die Reihe eugenisch denkender und im Rahmen der Intelligenzforschung als renommiert zu nennenden WissenschaftlerInnen¹⁹ wird u.a. ergänzt durch Cyril Burt (dem Fälschung von empirischen Daten, die die Erblichkeit von

19 In der Regel handelt es sich um männliche Intelligenzforscher, die eugenische Ideen propagieren, so dass von eugenisch denkenden Intelligenzforschern gesprochen werden könnte. Doch zumindest muss an dieser Stelle Audrey Shuey genannt werden, deren Studien (1966) zur intellektuellen Überlegenheit der US-Weiß-AmerikanerInnen gegenüber den US Schwarz-AmerikanerInnen oft zitiert werden.

Intelligenz belegen sollten, nach seinem Tod nachgewiesen werden konnte), Arthur Jensen oder einem der berühmtesten Psychologen überhaupt, Hans Jürgen Eysenck (Knebel & Marquardt, 2012, S. 98; Kühl, 2014).

Die Eugenik wird zuweilen als unrühmlicher Abschnitt in der Psychologie dargestellt, z. B. von Maltby et al. (2011, S. 650). Dies unterstellt aber ein abgeschlossenes Kapitel, so wie die unrühmliche Zeit des Nationalsozialismus oder die unrühmliche Anwendung der Lobotomie. Es ist jedoch zweifelhaft, ob eugenische Ideen tatsächlich lediglich ein Kapitel darstellen. Das Wort Eugenik findet praktisch keine Verwendung mehr, dennoch muss dies nicht bedeuten, dass die hinter einem stigmatisierten Begriff stehenden Inhalte ebenfalls keine Verwendung finden. Eine Kontinuität im Ausagieren eugenischer Ideen darf attestiert werden angesichts der Diskussion um die *Verdummung der Deutschen* (Sarrazin, 2010), der Präimplantationsdiagnostik oder der Forderung des Politikers Bernd Lucke, dass Akademikerinnen mehr Kinder bekommen sollen. Speck (2003) schreibt in Anspielung an die nach kapitalistischen Maßstäben angestrebte Ökonomisierung vieler Lebensbereiche, die Heilpädagogik sei herausgefordert, wenn zunehmend „Lebensbereich und Handlungsmuster von ökonomischen Wertmaßstäben bestimmt werden, und sich eine neue Eugenik anmeldet, die einen Menschen nach Maß anstrebt.“ (2003, S. 133 ff.)

Auch in der Psychologie sind im Besonderen in der Intelligenzforschung heftige Diskussionen geführt worden, bei denen eugenische Ideen WissenschaftlerInnen unterstellt worden sind. Beispielhaft angeführt sei hier die Diskussion um das Buch *The Bell Curve* von Herrnstein und Murray (1994). Sie ziehen aus ihrem Vergleich zu Intelligenztestergebnissen verschiedener ethnischer Gruppen u. a. den Schluss, dass weiße US-AmerikanerInnen ca. eine Standardabweichung (also ca. 15 IQ Punkte) intelligenter seien als schwarze US-AmerikanerInnen, attestierten aber auch anderen ethnischen Gruppen eine geringere Intelligenz. Ein Kernpunkt vieler kritischer Auseinandersetzungen war Herrnsteins und Murrays Schlussfolgerung, die mit einer geringeren Intelligenz versehenen Angehörigen der Ethnien wirkt sich nachteilig auf die Durchschnittsintelligenz in den USA aus, seien aber auch maßgeblich verantwortlich für Armut, Arbeitslosigkeit oder Kriminalität (Maltby et al., 2011, S. 639). Somit wird Angehörigen einer bestimmten Ethnie eine geringere Intelligenz (im Durchschnitt zur weißen Kontrollgruppe) unterstellt und die Zugehörigkeit zu dieser Ethnie auch als Risikofaktor für Armut und Kriminalität. Ausgeblendet würde bei dieser Sichtweise die Ausgrenzung verschiedener Bevölkerungsgruppen am Wohlstand der herrschenden (weißen) Schichten und die mit der Ausgrenzung verbundene Arbeitslosigkeit oder höhere Kriminalität. Finzsch (1999) fasst die seiner Meinung nach wichtigsten Aussagen des Buches so zusammen:

Man nehme eine allgemeine Intelligenz G (General Intelligence), definiert als ‚a person’s capacity for complex mental work‘ als breites Maß der Intelligenz. Dieses Maß

G kann auch mit einem Intelligenzquotienten IQ gleichgesetzt werden, der genau und ohne Verzerrung (Cultural Bias) festgelegt werden kann. Dieser IQ sei zu 40 % bis 80 % vererbt und relativ stabil, unabhängig vom Lebensalter der Versuchspersonen. Die Testergebnisse von IQs bei African Americans seien [sic] signifikant niedriger als die von Weißen. Niedrige IQ-Werte seien unter anderem die Folge von gesellschaftlichen Problemen wie Armut, Verbrechen, Arbeitslosigkeit, unehelichen Geburten und der Abhängigkeit von der Sozialhilfe. Hohe IQ-Werte korrelierten mit gesellschaftlichem Erfolg. Die Welt differenziere sich rasch in eine kognitive Elite und eine kognitiv defizitäre Unterklasse. Es sei unmöglich, IQ-Werte anzuheben und Sozialprogramme (...) seien sinnlos, kontraproduktiv oder beides. Da sich also an den IQ-Werten verschiedener Populationen nichts ändern lasse, sei es geboten, das Prinzip des Laissez-faire walten zu lassen und Menschen dort ihren gesellschaftlichen Platz finden zu lassen, wo sie hingehörten. (ebd., S. 86)

Ein bewährtes Argumentationsmuster eugenisch denkender IntelligenzforscherInnen auf Kritik ist die Berufung auf die Wahrheit, das Beklagen eines angeblichen Denkverbotes und die Berufung auf objektives Forschen nach den Kriterien wissenschaftlicher Standards, welches auch Ergebnisse nach sich zieht, die nicht en vogue sind. Doch dieses vordergründige Berufen auf die Wahrheit und auf wissenschaftliche Standards unabhängig vorgefertigter Meinungen wird konterkariert durch den Nachweis vielfältiger Querverbindungen zwischen EugenikerInnen und faschistisch-rassistischen Publikationen und Denkmustern (Velden, 2013; Billig, 1979; Knebel & Marquardt, 2012; Finzsch, 1999; Kaupen-Haas & Saller, 1999; Kühl, 1999, 2014; Haller & Niggeschmidt, 2012; Mecklenburg, 2002; Sesín, 2012).

Letztlich stehen eugenische Ansichten diametral sonderpädagogisch-heilpädagogischen Grundwerten entgegen, die Unterschiede zwischen den Menschen akzeptiert, sich nicht auf die Suche nach der Optimierung der menschlichen Rasse, sondern auf die Suche nach der optimalen Förderung jedes Menschen macht. Dabei werden intellektuelle Abweichungen nicht als für die Gesellschaft schädlich betrachtet, sondern akzeptiert. Geistig behinderten Menschen wird weder das Recht auf Sexualität noch das Recht auf Fortpflanzung und schon gar nicht das Recht auf Leben verwehrt. Die sich am Humanismus orientierenden Ideale der Sonderpädagogik sind ein Ende eines Pols, an deren anderem Ende die Eugenik steht. Die Gustav Heinemann zugeschriebene Aussage, man erkenne den Wert einer Gesellschaft daran, wie sie mit den schwächsten Gliedern verfährt, steht im Widerspruch zu eugenischen Ideen, die schwache Glieder einer Gesellschaft gar nicht erst zulassen möchte.

KritikerInnen der Eugenik werfen bei der Konstruktion von Intelligenztests ForscherInnen vor, die Intelligenz mit einer die Intelligenz repräsentierenden Zahl kennzeichnen zu wollen, denn nur dann wären die eugenischen Denkmuster und Schlussfolgerungen möglich, die intellektuelle Unterschiede zwi-

schen Ethnien postulieren. Die Kennzeichnung der Intelligenz auf einen Nennwert, dem Gesamt-IQ, ist nur möglich, orientiert man sich an einer entsprechenden Theorie, die von einem Generalfaktor der Intelligenz ausgeht (Knebel & Marquardt, 2012; Finzsch, 1999). Würde man von multifaktoriellen Intelligenztheorien ausgehen wie bei Gardner (1983) oder Sternberg (1985), wäre dies nicht möglich. Es stellt sich also die Frage, ob die Postulierung des Generalfaktors der Intelligenz, aktuell repräsentiert durch das CHC-Modell, nicht in der Tradition eugenischer Denkmuster zur Instrumentalisierung entsprechender Vorstellungen entstanden ist. Dies würde konsequent zu Ende gedacht bedeuten, dass SonderpädagogInnen mit der Durchführung eines Intelligenztests eine Handlung im Geiste der Eugenik durchführten.

Dass eugenisch denkende WissenschaftlerInnen maßgeblich an der Entwicklung des Generalfaktors der Intelligenz beteiligt waren, steht nicht in Frage. Auch wenn Horn und Carroll, deren Initialen neben Cattells Initialen für das CHC-Modell stehen, nicht mit Eugenik in Verbindung gebracht werden, so darf angenommen werden, dass Cattells Wirken als Eugeniker Horn und Carroll bekannt gewesen war. Es ist irritierend, dass dies nicht dazu geführt hat, nicht als Bestandteil eines absehbar in der Wissenschaft bedeutsamen Intelligenzmodells in einem Zug mit einem Wissenschaftler genannt werden zu wollen, der sich deutlich in der Tradition der Eugenik positionierte, dem Cattell-Horn-Carroll-Modell.

Angesichts der Tatsache, dass die Eugenik innerhalb der Intelligenzforschung eine bedeutsame Rolle spielte, stellt sich weiterhin die Frage nach der Motivation für das Forschungsinteresse an dem Konstrukt Intelligenz und es bleibt zu hoffen, dass dies nicht zu einem großen Teil mit diesem Zitat beantwortet ist, welches Myers (2008) Lewis Terman zuschreibt: Intelligenztests dienen nach Terman dazu „letztendlich die Fortpflanzung von Schwachsinn deutlich einzuschränken und dadurch zur Beseitigung eines hohen Maßes an Kriminalität, Massenarmut und Ineffizienz in der Industrie beitragen zu können“ (Terman, zitiert nach Myers, 2008, S. 411).

2.3.2.5 Das CHC-Modell als integrierendes Intelligenzmodell

Maßgeblich an diesem Modell beteiligt waren die Wissenschaftler Cattell, Horn, Carroll, aber auch W. Woodcock. John Leonard Horn (1929–2006) lehrte bzw. arbeitete an Universitäten in den USA, England und Schweden und studierte u. a. in Australien (Intelltheory, 2013a). Sein ursprüngliches Ziel, Mathematik- und Chemielehrer zu werden, gab er zugunsten eines Psychologiestudiums auf. Als Student lernte er R. B. Cattell kennen und wurde sein Schüler, zusammen arbeiteten sie langjährig zusammen. Wie Cattell postulierte Horn (1965) eine *fluide* und *kristalline* Intelligenz, verwarf aber später den Glauben

an die Existenz eines Generalfaktors der Intelligenz (Maltby et al., 2011, S. 530). Obwohl er die Bezeichnung Gf-Gc-Theorie weiterverwendete, entdeckte er neben der *fluiden* und der *kristallinen* Intelligenz sieben weitere Intelligenzdimensionen, z. B. *auditive Verarbeitung*, *Verarbeitungsgeschwindigkeit* oder *visuelle Verarbeitung* (ebd., S. 530). Zeitlebens stand Horn der Existenz eines Generalfaktors der Intelligenz skeptisch gegenüber. Dies würde bedeuten, dass eine Extrahierung eines Gesamt-IQ auf Grundlage des CHC-Modells, auf dem die meisten Tests basieren, die von SonderpädagogInnen angewendet werden, eigentlich nicht möglich ist. Maltby et al. (2011, S. 531 f.) beschreiben das CHC-Modell sogar als ein Modell ohne übergeordneten Generalfaktor.

John B. Carroll (1916–2003) war ein pädagogischer Psychologe und Militärpsychologe und Schüler von B.F. Skinner und der Intelligenzforscherin F. Goodenough [Anm. T.J.: Goodenough entwickelte unter anderem den Mann-Zeichen-Test (*Draw-a-Man*) und stand der Darstellung der Intelligenz in Form eines IQ ebenso wie Horn skeptisch gegenüber] (Intelltheory, 2013b). Er lehrte an US-amerikanischen Universitäten. Unter anderem beschäftigte sich Carroll mit der Linguistik und dem Fremdsprachenunterricht (ebd.), bekannt wurde er allerdings mit einer aufwendigen Metaanalyse von Daten, deren Resultat die Postulierung eines Intelligenzmodells nach sich zog, welches als Drei-Schichten-Modell (*Three-Stratum Theory of Human Cognitive Abilities*) bekannt wurde (Carroll, 1993). Mit 461 zwischen 1927 und 1987 erhobenen Datensätzen, deren Inhalte die Untersuchung von statistischen Zusammenhängen zwischen Aspekten der Intelligenz (Baudson, 2012) waren, führte er Faktorenanalysen durch, deren Resultat das Drei-Schichten-Modell darstellt, welches drei Ebenen der Intelligenz beschreibt (Carroll, 1993). Auf *Stratum III* (Ebene 3) dieses hierarchischen Modells befindet sich der Generalfaktor der Intelligenz, ähnlich dem Generalfaktor, wie er auch von Spearman postuliert wurde. Auf *Stratum II* befinden sich acht übergeordnete Faktoren, die in gängigen Intelligenztests wie KABC-II oder WISC-IV *Indice* genannt werden, in diesen Tests allerdings unter Berufung auf die *Stratum-II*-Faktoren des auch aus dem Carroll Modell resultierenden CHC-Modells. Die 8 Faktoren nach Carroll sind:

1. *Fluide Intelligenz*,
2. *kristalline Intelligenz*,
3. *allgemeine Gedächtnisfähigkeit*,
4. *visuelle Wahrnehmung*,
5. *auditive Wahrnehmung*,
6. *Abruffähigkeit*,
7. *kognitive Geschwindigkeit*,
8. *Verarbeitungsgeschwindigkeit*.

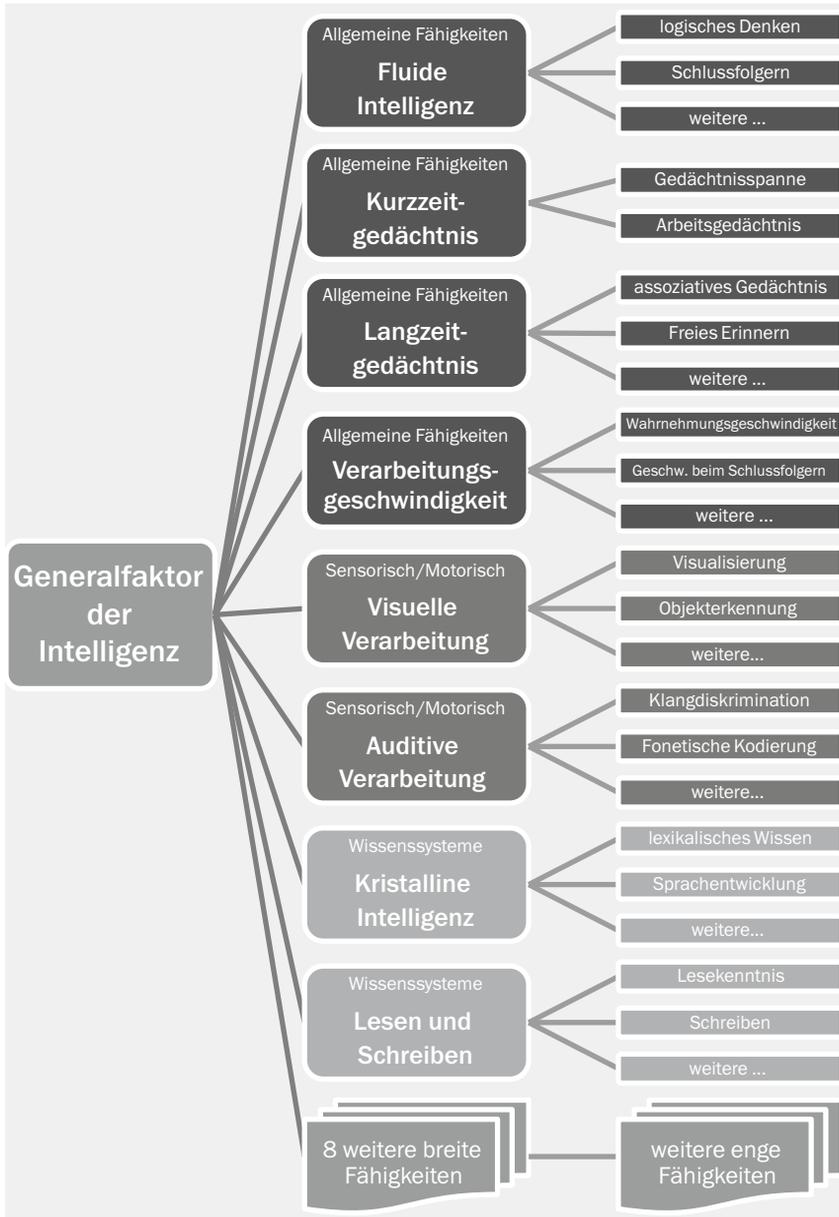
Auf *Stratum I* wurden 69 (Maltby et al., 2011, S. 530) spezifische Fähigkeiten erkannt, auf deren Darstellung zu Gunsten einer ausführlicheren Darstellung der *Stratum-I*-Ebene des auch aus dem Drei-Schichten-Modells resultierenden CHC-Modells an dieser Stelle verzichtet wird. Carroll war Mitunterzeichner des Aufsatzes *Mainstream Science on Intelligence* (Gottfredson, 1994), einer Solidaritätsnote für die Autoren des heftig u. a. als rassistisch kritisierten Buches *The Bell Curve* von Herrnstein und Murray (1994).

Auf einem 1999 stattfindenden Treffen versuchte der Psychologe und Testautor Richard W. Woodcock ein integrierendes Modell von Intelligenz zu erarbeiten und traf sich zu diesem Zweck mit Horn und Carroll (Cattell war verstorben). Woodcock (geboren 1928) hat u. a. den in den USA renommierten Test *Woodcock-Johnson Psychoeducational Battery – Revised: Tests of Cognitive Ability* (Woodcock, McGrew & Mather, 2001) entwickelt und unternahm den Versuch einer Integration sich ähnelnder Intelligenzmodelle. Erstaunlich problemlos und zügig gelang dies, so dass Budson (2012, S. 10) feststellte, dass sich manchmal Streitigkeiten und vermeintliche Widersprüche auf kreative Art lösen lassen. Im Wesentlichen wird mit dem CHC-Modell das Cattell-Horn-Modell der *fluiden* und *kristallinen* Intelligenz sowie das Drei-Schichten-Modell nach Carroll vereinigt, aber auch Ansätze von Thurstones Primärfaktoren-Theorie, deren ursprüngliche Negierung eines Generalfaktors von Thurstone modifiziert wurde zugunsten der Annahme eines Generalfaktors. Erstmals beschrieben wurde das CHC-Modell von Flanagan, McGrew und Ortiz (2000).

Ähnlich dem Drei-Schichten-Modell nach Carroll werden auf drei Ebenen (*Stratum I* bis *Stratum III*) hierarchisch von oben nach unten der Generalfaktor der Intelligenz (*Stratum III*) beschrieben, darunter auf *Stratum II* breite Fähigkeitsbereiche (*broad abilities*) wie z. B. *fluide* und *kristalline* Intelligenz, *auditive Verarbeitung* oder *Langzeitspeicherung*. Auf der untersten Ebene *Stratum I* befinden sich spezifische oder enge Fähigkeiten (*narrow abilities*), die sich in Gruppen den breiten Fähigkeitsbereichen zuordnen lassen. In ersten Erläuterungen (*first Generation*) zu dem CHC-Modell wird von 10 Faktoren auf zweiter Ebene und 73 Faktoren auf der ersten Ebene ausgegangen (Mickley & Renner, 2010, S. 449). Eine aktuellere Beschreibung des CHC-Modells wurde von Flanagan, Ortiz und Alfonso (2013) und Schneider und McGrew (2012) vorgenommen und beinhaltet auf *Stratum-II*-Ebene 16 Bereiche, untergliedert in erworbene Wissenssysteme, allgemeine Fähigkeiten und sensorisch/motorische Fähigkeiten. Auf *Stratum I* befinden sich ca. 80 enge Fähigkeiten (siehe Abbildung 1).

Es ist ein grundlegendes Problem bei der Entwicklung eines Intelligenztests, die Erklärung dessen, was Intelligenz darstellen soll, durch ein Testverfahren abrufen zu können. Es ist möglich, dass eine theoretische Erklärung über das Konstrukt Intelligenz richtig ist, aber nicht durch einen Test validiert werden kann. Wird Intelligenz definiert als ein Zusammenspiel von kognitiven Fähigkeiten, die im deutlichen Zusammenhang mit der augenblicklichen (Test-)Situ-

Abbildung 1. Das CHC-Modell. Die 16 breiten Fähigkeiten sind in drei inhaltlich ähnliche Bereiche unterteilt und farblich gekennzeichnet. Wenig testrelevante breite Fähigkeiten wie *Olfaktorische Fähigkeiten* werden vernachlässigt dargestellt und unter weitere breite Fähigkeiten subsummiert (vgl. Flanagan et al., 2013; Schneider & McGrew, 2012; Renner & Mickley, 2015b).



tion betrachtet werden muss, also situativ variabel ist – auch durch möglicherweise noch nicht verstandene Abläufe im Gehirn während der Entwicklung von Problemlösungsstrategien während einer Testsituation – und zudem noch stark kulturabhängig (auch innerhalb einer Kultur wie der deutschen Kultur, die es auf Grund vielfacher Subkulturen so sicherlich nicht gibt), dann ist eine Testkonstruktion methodisch schwer möglich. Wird zudem noch postuliert, dass es verschiedene gleichbedeutende Intelligenzen gibt, die nebeneinander Bestand haben und nicht in einen Gesamtwert münden, der der Tendenz von Menschen nach Vereinfachung und Kategorisierung entgegenkommt; wenn dann noch nicht testbare Konstrukte wie Kreativität (wer wollte festlegen, was kreativ ist und was nicht?) Bestandteile der Intelligenz sein sollen, dann wird die Beschreibung von Intelligenz bei einer Beschreibung bleiben, nicht mündend in einer möglichen Überprüfung durch einen Intelligenztest.

Das CHC-Modell der Intelligenz ist verlockend für die Entwicklung eines Intelligenztests. Auch wenn es sich um eine Beschreibung der Intelligenz und nicht um eine Definition handelt (Mickley & Renner, 2010; Renner & Mickley, 2015b), liegt die mögliche Konzeptualisierung der im Modell beschriebenen Aspekte auf der Hand. Sogar der Aufbau eines Tests wird durch die drei Schichten bereits in Bahnen gelenkt. Die im *Stratum I* beschriebenen engen Fähigkeiten geben Hinweise auf die Gestaltung der Subtests, die im *Stratum II* beschriebenen breiten Fähigkeiten geben Hinweise auf die Gestaltung von *Indices* (siehe KABC-II, WISC-IV), übergeordneten und gut interpretierbaren Bereichen und alles mündet in einen Gesamtwert, der das intellektuelle Potential der getesteten Person widerspiegelt und Prognosen z. B. über den Schulerfolg zulässt. Es bleibt zweifelhaft, ob Bereiche auf *Stratum-II*-Ebene wie die auf den Geruchssinn bezogenen *Olfaktorischen Fähigkeiten* als Teilbereich der Intelligenz überprüfbar werden. In der aktuellen Beschreibung der CHC-Theorie werden ebenso *Taktile Fähigkeiten* und *Kinästhetische Fähigkeiten* als Bestandteile der Intelligenz beschrieben (Flanagan et al., 2013; Schneider & McGrew, 2012). Allerdings sind diese noch wenig untersuchten Bestandteile des aktuellen CHC-Modells auch nicht Gegenstand eines Intelligenztests, welcher in Deutschland von SonderpädagogInnen durchgeführt wird.

Mit Hilfe des CHC-Modells ist es sogar möglich, sich einer Fragestellung im diagnostischen Prozess mit Hilfe mehrerer Subtests aus verschiedenen Testverfahren zu nähern, sofern die Testverfahren auf Grundlage des CHC-Modells konstruiert worden sind. Dies ist bei den meisten gebräuchlichen Intelligenztests der Fall, welche in Deutschland von SonderpädagogInnen angewendet werden. Fällt z. B. bei einem Kind im Unterricht auf, dass dieses sich nicht gut merken kann, was die Lehrkraft auditiv vorträgt (*Stratum-II: Auditive Verarbeitung*; auch *Kurzzeitgedächtnis*), besteht also bei einem Kind der Verdacht auf ein Defizit in der *akustischen Merkfähigkeit* (als Teilbereich der *auditiven Verarbeitung*), so könnten die Subtests *Zahlennachsprechen* und *Wortreihe* der

KABC-II mit *Zahlennachsprechen* der IDS und *Zahlennachsprechen vorwärts*²⁰ des WISC-IV kombiniert werden. Interessante Ansätze zu einem *Cross-battery assessment* liefern hierzu Renner und Mickley (2015b).

2.4 Anwendungen von Intelligenztests durch SonderpädagogInnen

Die vorherigen Kapitel haben Intelligenzmodelle, die Intelligenzmessung und Widersprüche in diesem Zusammenhang beschrieben, um das Konstrukt bei der Anwendung von Intelligenztests verstehen zu können: Intelligenz bzw. die zur Ermittlung von Intelligenz verwendeten Tests. Darauf aufbauend soll sich nun der Fragestellung dieser Arbeit konkreter genähert werden: Rahmenbedingungen und Schwierigkeiten im Umgang mit Intelligenztests durch SonderpädagogInnen.

Ziel dieses Kapitels ist die Darstellung bereits in der Forschung beschriebener Schwierigkeiten bei der Anwendung von Intelligenztests. Im ersten Teil sollen Forschungsergebnisse bei der Anwendung von Intelligenztests in Deutschland beschrieben werden. Bereits vorhandene Forschungsergebnisse tragen dazu bei, den Fokus auf Fragestellungen im methodischen Teil richtig auszurichten, tragen aber auch dazu bei, bereits gut beforschte Teilbereiche nicht wiederholt zum Gegenstand von Fragestellungen werden zu lassen.

Im zweiten Teil dieses Kapitels soll geprüft werden, ob bekannte Schwierigkeiten im Sinne der Fragestellung auch im Ausland bekannt sind. Dies würde allerdings unterstellen, dass SonderpädagogInnen außerhalb Deutschlands vergleichbar häufig und institutionalisiert Intelligenztests durchführen. Weltweit unterscheiden sich Schulsysteme und die Tätigkeiten von SonderpädagogInnen. Bevor nach Belegen zu Schwierigkeiten bei der Anwendung von Intelligenztests durch *special education teachers* (dieser Begriff soll im Folgenden die Gruppe der nicht deutschen SonderpädagogInnen beschreiben, auch wenn je nach Staat und Sprache andere Übersetzungen vorliegen) gesucht wird, muss geklärt werden, ob *special education teachers* ähnlich wie SonderpädagogInnen Intelligenztests anwenden. Wäre die häufige Anwendung von Intelligenztests durch SonderpädagogInnen in Deutschland eine weltweite Ausnahme in den jeweiligen Stellenbeschreibungen der *special education teachers*, erübrigte sich die Recherche nach vorhandenen Befunden: führen *special education teachers* außerhalb Deutschlands keine Intelligenztests durch, können Belege zu Schwierigkeiten bei der Anwendung von Intelligenztests nicht vorliegen.

20 *Zahlennachsprechen rückwärts* bedarf bei der Fragestellung *auditive Merkfähigkeit* einer besonderen Interpretation.

2.4.1 Untersuchte Schwierigkeiten bei der Testanwendung in Deutschland

Neben der Sichtung entsprechender Fachbücher zur (sonderpädagogischen) Diagnostik und der Sichtung der Indexe sonderpädagogischer Fachzeitschriften, im Besonderen der Zeitschrift für Heilpädagogik, wurden Datenbanken genutzt.

PubPsych ist eine Metadatenbank mit über 930 000 Datensätzen, welche sich vor allem aus Aufsätzen aus psychologischen Fachzeitschriften, aber auch den gängigen sonderpädagogischen und pädagogischen Zeitschriften zusammensetzen. Renommierete Datenbanken wie PSYINDEX, PsychOpen oder PsychData gehören zu PubPsych. Da das Thema dieser Arbeit der psychologischen Sonderpädagogik zugeordnet werden kann, erscheint eine Recherche in PubPsych angemessen.

Bewusst wurden auch allgemeine Suchworte wie *Intelligenztest* verwendet, auch wenn die Trefferlisten groß waren. Bei der Thematik dieser Arbeit kann vermutet werden, dass nur wenige Veröffentlichungen einen eindeutigen Hinweis auf den Forschungsstand geben würden. Wie erwartet wiederholten sich die Datensätze, z. B. bei Eingabe der Suchworte *Intelligenztest* und *Intelligenztests*, doch hatte dies im Rahmen der Sichtung gleichzeitig eine Kontrollfunktion, so dass es unwahrscheinlicher war, interessante Artikel zu übersehen. Es wurde damit gerechnet, dass es nur wenige Hinweise auf Schwierigkeiten bei der Anwendung von Intelligenztests im sonderpädagogischen Kontext geben würde, deshalb wurden Hinweise auf Schwierigkeiten bei der Anwendung von Intelligenztests auch in anderen Kontexten (Erziehungsberatungsstellen, Schulpsychologie etc.) nicht ausgeschlossen, sofern diese Hinweise interessant im Sinne der Fragestellung schienen. Die Suchworte in Tabelle 1 ergaben folgende Trefferlisten, die gesichtet worden sind:

Tabelle 1. Trefferlisten von Suchmaschinen.

Stichwort	Anzahl Treffer
„Intelligenztest“	796
„Intelligenztests“ (ab 1995)	597
„Durchführungsobjektivität“	17
„IQ“ (nur Publikationen ab 1995)	320
„Testleiter“	57
„Flynn“	19

Anmerkung. Stand 23. 1. 16. Sofern nicht anders angegeben alle Jahre und nur in dt. Sprache.

In die folgende Darstellung von Untersuchungsergebnissen werden auch Schwierigkeiten bei der Anwendung von Intelligenztests durch andere Berufsgruppen einbezogen, z.B. PsychologInnen. Dies honoriert die schwache Befundlage, kann aber auch inhaltlich begründet werden. Beschriebene Schwierigkeiten bei der Anwendung von Intelligenztests in der Praxis ähneln sich sowohl in sonderpädagogischen als auch in psychologischen Kontexten, da die Anwendung standardisiert ist und die Durchführungs- und Auswertungsregeln unabhängig vom beruflichen Hintergrund der AnwenderInnen identisch sind. Ein Einbezug von Untersuchungsergebnissen im Sinne der Fragestellung auch in psychologischen Kontexten ist gewinnbringend, da übertragbar auf den sonderpädagogischen Bereich. Generell kann angenommen werden, dass Intelligenztests in psychologischen Beratungsstellen, durch SchulpsychologInnen, in sozialpädiatrischen Zentren oder psychiatrischen Einrichtungen häufiger angewendet werden als durch SonderpädagogInnen²¹. Die daraus resultierende größere Routine bei der Anwendung von Intelligenztests durch PsychologInnen hat zur Folge, dass die beschriebenen Schwierigkeiten in psychologischen Kontexten potenziert angenommen werden können bei der (vermuteten) selteneren Anwendung durch SonderpädagogInnen, da die Routine tendenziell geringer ist. Beziehen sich Untersuchungsergebnisse ausschließlich auf den sonderpädagogischen Kontext, wird dies kenntlich gemacht.

Huber (2000) beschreibt die sonderpädagogische Diagnostik im *Spannungsfeld traditioneller und gegenwärtiger Sichtweisen*, wobei traditionelle Diagnostik in der Sonderpädagogik wertfrei mit normorientierter und objektiv-quantitativer Diagnostik; gegenwärtige Diagnostik als individuumorientiert und subjektiv-qualitativ beschrieben wird (ebd., S. 411). Untersucht wurde die Anwendung von Intelligenztests innerhalb dieses *Spannungsfelds*. Während vor 1995 in Nordrhein-Westfalen die Verwendung eines Intelligenzquotienten maßgeblich zum Erkennen sonderpädagogischen Förderbedarfes²² beitragen sollte, verlor die Bestimmung eines Gesamt-IQ an Bedeutung. Obwohl also die Anwendung eines Intelligenztests nicht mehr zwingend war, konnte er durchgeführt werden. Basierend auf den Ergebnissen einer empirischen Untersuchung (Huber, 2000) – es wurden 313 LehrerInnen aus 14 Sonderschulen für Körperbehinderte in NRW befragt – konnte festgestellt werden, dass die Anwendung der Intelligenztests häufig nicht den Anforderungen der Durchführungsobjek-

21 Die Häufigkeiten der Anwendung von Intelligenztests durch SonderpädagogInnen wird im methodischen Teil näher untersucht werden.

22 Vor 1995 nannte sich das sonderpädagogische Gutachten in NRW „Sonderschulnahmeverfahren“ (SAV), danach „Verordnung zur Feststellung des sonderpädagogischen Förderbedarfes“ (VO-SF), heute nennt es sich „Ausbildungsordnung zum sonderpädagogischen Förderbedarf“ (AO-SF).

tivität entsprachen. Nur 2,4 Prozent der LehrerInnen gaben eine Durchführung unter den vorgeschriebenen standardisierten Bedingungen an (ebd., S. 412), am häufigsten (78 % der Befragten) wurden die Zeitgrenzen (zusätzliche Zeitgaben, begründet mit der Körperbehinderung der getesteten Kinder) missachtet. Aber auch zusätzliche Hilfestellungen (35 %) bis hin zu „leichten Hinweisen zum Lösungsweg“ (ebd., S. 412) sowie andere Verletzungen der Durchführungsregeln scheinen eher die „Regel als die Ausnahme“ (ebd., S. 412) zu sein. Huber befürchtet resümierend, dass „zumindest ein Teil der quantitativen Testergebnisse (...) durch einen fragwürdigen Umgang mit dem Testmaterial aussagelos sind“, dieser Trend sei problematisch, da auf Grundlage dieser Ergebnisse Schul- und Lebenswege beeinflusst sind (ebd., S. 412).

Auch die Testverfahren selbst sind problematisch, da häufig veraltet. So wurde am dritthäufigsten eine Version des HAWIK von 1956 angewendet. Bezogen auf die Debatte Statusdiagnostik vs. Förderdiagnostik beschreibt Huber die Anwendung von Intelligenztests als nützlichen Beitrag zu objektiven Ergebnissen, welche die Gefahr subjektiver Fehleinschätzungen verringert. Intelligenztests werden nicht als gefährlich beschrieben, „sondern nur Personen, die den verantwortungsvollen Umgang mit Tests und Testergebnissen nicht beherrschen“ (ebd., S. 415).

Es liegt auf der Hand, dass die Durchführung eines komplexen Intelligenztests mit all seinen Regeln einer gründlichen Vorbereitung bedarf. Ebenfalls auf der Hand liegt, dass eine gelegentliche Durchführung, z. B. alle vier Monate, dazu führen muss, sich erneut ausführlich mit den Durchführungsregeln zu beschäftigen. Würde in einem Kollegium die Anfertigung der Gutachten weitestgehend auf alle Lehrkräfte verteilt werden, resultierte daraus ein gelegentliches Testen, welches zur Folge hätte, dass immer wieder erneut eine Beschäftigung mit den Durchführungsregeln stattfinden müsste, da die Anwendung der Durchführungsregeln nur mit Hilfe von Routine ohne besondere Vorbereitung gewährleistet wäre. Eine logische Konsequenz dieser Schwierigkeiten könnte die Spezialisierung weniger SonderpädagogInnen bei der Anwendung von Intelligenztests sein. Obwohl zu Recht eingewendet wird, dass die Anwendung standardisierter Testverfahren zu den Kompetenzen von SonderpädagogInnen gehört, stellt sich die Frage, ob dies nicht an der Realität vorbeigeht. Ohne Berücksichtigung standesrechtlicher Belange, besonderer Vergütungen auf Grund psychologischer Tätigkeiten und Stellenbeschreibungen von SonderpädagogInnen wäre zu überlegen, ob nicht wenige spezialisierte SonderpädagogInnen oft standardisierte Tests routiniert durchführen anstatt alle SonderpädagogInnen unroutiniert selten.

Bekannt sind entsprechende Bestrebungen in Hamburg, Berlin und Brandenburg. Dort werden tendenziell die sonderpädagogischen Gutachten und komplexeren standardisierten Testverfahren von MitarbeiterInnen von Diagnos-

tikteams durchgeführt, speziell und kontinuierlich in Testverfahren ausgebildete SonderpädagogInnen (Land Brandenburg, 2013; Senat Berlin, 2012)²³.

Müller (2009) beschreibt die Schwierigkeiten einer gelegentlichen Anwendung von Intelligenztests und anderer standardisierter Verfahren, um zu dem Schluss zu kommen, dass schulinterne Diagnostikansprechpartner Abhilfe schaffen könnten. Die beschriebenen Schwierigkeiten bei der seltenen Anwendung von u. a. Intelligenztests sind (ebd., S. 180):

- eine schwierige Integration in den Unterrichtsablauf,
- oft nur in der zeitaufwändigen Einzelsituation durchführbar,
- wenn Gruppentestung möglich, erschweren mögliche Verhaltensauffälligkeiten der Kinder die Anwendung,
- begrenzte personelle Ressourcen für die Anwendung vorhanden.

Aus dem erkannten Widerspruch „zwischen der Anforderung einer ausführlichen Diagnostik als Voraussetzung für individualisierte Förderung und der nur sehr eingeschränkten Umsetzung in die Praxis“ (ebd., S. 180) könnte die Implementierung von *Diagnostikansprechpartner* resultieren. Diese führen dann auch aufwändige formelle Verfahren durch, um objektivere und grundsätzlichere Einblicke zu gewähren im Gegensatz zu den schneller und einfacher durchgeführten informellen Verfahren, die oft einer subjektiven Färbung unterliegen (ebd., S. 182). Aufgaben der Diagnostikansprechpartner sind neben der Einarbeitung in Fragen der Diagnostik, der Durchführung und Auswertung diagnostischer Verfahren inkl. Ableitungen von Fördermaßnahmen aus den Testergebnissen auch die Beratung in Fragen der Diagnostik und die Vorstellung von Testverfahren (ebd., S. 182). In einem Modellversuch an einer Düsseldorfer Förderschule (Förderschwerpunkte *Lernen* und *Emotional-Sozial*) wurde die Diagnostik auf eine Lehrkraft gebündelt [Anm.: Müller selbst], die dafür 2 Stunden wöchentlich Ausgleich erhielt. Unter Einbezug der anderen Lehrkräfte, die Wünsche für durchzuführende Testverfahren und Fragestellungen nannten, führte die Umsetzung des Konzepts dazu, dass die anderen Lehrkräfte sich in ihren Anliegen nach diagnostischer Abklärung ernst genommen fühlten, verwertbare diagnostische Hinweise erhielten ohne selbst diagnostisch tätig werden zu müssen, und Schulproblematiken der Kinder besser erklärt werden konnten (ebd., S. 185). Die kontinuierliche Anwendung der Diagnostik konnte durch den Diagnostikansprechpartner eher verwirklicht werden, da die schuli-

23 Die Diagnostik-SpezialistInnen in Hamburg arbeiten ohne veröffentlichtes Konzept in Absprache. Auskunft erteilt die Referatsleitung Landesinstitut für Lehrerbildung und Schulentwicklung (LI); Referat Sonderpädagogik & Individuelle Förderung; Felix-Dahn-Str. 3, 20357 Hamburg.

schen Rahmenbedingungen (ebd., S. 185) eine systematische Diagnostik nicht zuließen.

Kritisch darf bei dem von Müller beschriebenen Konzept eingewendet werden, dass die Einverständniserklärung der Sorgeberechtigten als Voraussetzung für die Anwendung eines standardisierten normierten Testverfahrens (Avenarius, 1990) nicht in dem ansonsten detailliert beschriebenen Ablauf des diagnostischen Prozesses erwähnt wird.

Als Standardwerk der sonderpädagogischen Diagnostik kann die *Einführung in die sonderpädagogische Diagnostik* (Bundschuh, 2010) angeführt werden. Bereits im dritten Satz des Kapitels Durchführungsobjektivität merkt Bundschuh an, „es sei sehr fraglich, ob diese Forderung [Anm. T.J.: nach Durchführungsobjektivität] bei einem Teil der Kinder, mit denen wir es zu tun haben, in vollem Umfang eingehalten werden kann“ (ebd., S. 83). Die Bedeutung von der Vorgabe wortwörtlicher Instruktionen und die negativen Auswirkungen bei Nichteinhaltung dieser Vorgaben werden beschrieben. Es wird angemerkt, dass Kinder mit sonderpädagogischem Förderbedarf über die in den Handbüchern beschriebenen Durchführungsregeln besondere Erklärungen und häufigere Pausen etc. benötigen (ebd., S. 83).

Nachdem der Sonderpädagoge mit diesem Sachverhalt rechnen muss, ihn kennt, besteht bei ihm die Neigung zu besonderen Erklärungen, Wiederholungen der Testinstruktion mit jeweils anderen Worten und neuer Akzentuierung. Es besteht die Gefahr, dass die Testleistung – weil eben die Instruktion in ihrer ursprünglichen Form, also standardisiert, gegeben wurde – eine zu gute Bewertung erfährt (...). (ebd., S. 84)

Die von Bundschuh beschriebenen Problematiken bei der korrekten Anwendung der Durchführungsobjektivität in der Sonderpädagogik resultieren bei ihm in vier Aspekten:

1. Bei (...) psychometrischen Verfahren muss man die vorgegebenen Instruktionen und Testbedingungen einhalten.
2. Es gibt Verfahren, bei denen zusätzliche Erklärungen in einem bestimmten Rahmen abgegeben werden dürfen bzw. müssen.
3. Bei besonders schwierigen Kindern werden manchmal eine Abänderung der zeitlichen Abfolge von Testaufgaben, das Einlegen von Pausen, eine zusätzliche Ermutigung oder Lob nötig sein. Solche Maßnahmen sollten jedoch grundsätzlich im Gutachten vermerkt werden.
4. Es gibt Testverfahren, bei denen bei jüngeren, bei stark gehemmten und bei auf sozialen Kontakt angewiesenen Kindern zusätzliche Motivation oder Ermutigung empfohlen und nahegelegt wird. (ebd., S. 84).

Angemerkt sei, dass Punkt 1 im Widerspruch zu den anderen Punkten steht und keine Hinweise erfolgen, wie bei den empfohlenen Änderungen der Durchführungsobjektivität die Testergebnisse zu bewerten seien.²⁴

Schmidt-Atzert und Amelang (2012) beschreiben die Annahmen der Klassischen Testtheorie und merken an, „dass Testwerte, also die Ergebnisse, die uns Persönlichkeitsverfahren, Intelligenztests (...) liefern, fehlerbehaftet sind“ (ebd., S. 41). Diese als Messfehler beschriebenen Fehler könnten z.B. auch durch Erinnerungs- und Übungseffekte auftreten.

Übertragen auf den sonderpädagogischen Bereich könnte es z.B. vorkommen, dass ein Kind regelmäßig mit dem CFT 1 getestet wird, weil vielleicht der CFT 1 der einzig verfügbare Test im Testschrank ist.

Messfehler entstehen durch Fehler „bei der Testkonstruktion, bei der Durchführung und bei der Auswertung des Tests“ (ebd., S. 43). Bei der Testsituation könnten die Bedingungen variieren, z.B. „Lichtverhältnisse, Geräusche, Luftqualität, Raumtemperatur, Sitzkomfort, Art und Anzahl der Testteilnehmer“ (ebd., S. 44).

Werning und Lichtblau (2012) bezweifeln, ob eine nicht nach den Regeln der Handbücher durchgeführte Testung überhaupt zu verwertbaren Ergebnissen führt. Sie legen sich fest, indem sie attestieren, dass „eine fehlerhafte Darbietung der Testaufgaben zu Verzerrungen in den Ergebnissen führt und somit zu Testwerten ohne jede Aussagekraft“ (ebd., S. 235). Die Autoren stehen normierten Testverfahren grundsätzlich skeptisch gegenüber und fordern bei Anwendung eines normierten Intelligenztests, dass dieser wenigstens Ableitungen von Stärken und Schwächen ermöglicht im Gegensatz zu Testverfahren, die über der Ermittlung eines Gesamt-IQ hinaus kaum weitere Interpretationsideen zulassen. Als gutes Beispiel erwähnen die Autoren die K-ABC. Zu Recht wird auf die Aktualität der Normdaten auf Grund des Flynn-Effekts hingewiesen, was wiederum die K-ABC nicht gewährleistet, dafür der WISC-IV. Dieser allerdings scheint für viele Kinder mit sonderpädagogischem Bedarf im sprachlich-kommunikativen Bereich ungeeignet (ebd., S. 235).

Staud und Staud (2011) bezweifeln ebenfalls grundsätzlich die Aussagekraft von Intelligenztests, im Besonderen bei der Anwendung bei körperbehinderten Kindern. Sie sind skeptisch, dass Kinder mit körperlichen Behinderungen mit standardisierten Tests überprüft werden können, da es einerseits den Besonderheiten der Kinder nicht gerecht wird (z.B. Kinder mit einer Spastik können schwerlich ein Puzzle unter Zeitdruck nachbauen), andererseits ein sachliches,

24 Es sei erwähnt, dass bei den meisten Tests allerdings berücksichtigt ist, dass besondere Kindergruppen besondere Rahmenbedingungen oder Erklärungen benötigen. In der Regel muss die Durchführungsobjektivität erst bei bewerteten Items rigide erfolgen, aber nicht bei der nichtbewerteten Erklärung von Aufgabenstellungen.

objektives Verhalten gerade für diese Kinder hinderlich sein kann, da sie besonders viel Zuspruch benötigen.

Lipsius, Petermann und Daseking (2008) beschreiben anschaulich mögliche Fehldiagnosen, die aus Durchführungsfehlern bei der Anwendung des HAWIK-IV resultieren können. Intelligenztests werden generell als fehleranfällig beschrieben, was Einfluss auf die Validität nehmen kann. Der HAWIK-IV wird in diesem Zusammenhang als umfangreiches und komplexes Testverfahren mit mindestens 200 Anwendungshinweisen (ebd., S. 107 f.) beschrieben. Der HAWIK-IV (Petermann & Petermann, 2007) ist eine Adaption des US-amerikanischen WISC-IV (Wechsler, 2003), dem damals aktuellen Wechsler-Verfahren. Dieses, wie auch andere Wechsler-Verfahren, wurden hinsichtlich der Durchführungsobjektivität in Studien außerhalb Deutschlands bereits auf typische und häufige Durchführungsfehler untersucht, welche im nächsten Kapitel (Forschungsstand Ausland) näher erläutert werden. Zumindest die Studie von Alfonso, Johnson, Patinella und Rader (1998) sei hier bereits erwähnt: auch nach intensiver Einarbeitung in die Durchführung und Auswertung des damals aktuellen Wechsler-Tests WISC-III (Wechsler, 1991) betrug die Fehlerquote von 60 Testprotokollen 100 Prozent, insgesamt wurden 468 Fehler gefunden. Es stellte sich also nicht die Frage, ob die Anwendung fehlerhaft war, sondern wie viele Fehler je Anwendung gemacht worden sind. Alfonso et al. (1998) wendeten übrigens eine Methode an, die auch für diese Studie geplant ist. Während Alfonso et al. die Testprotokolle von speziell geschulten Studierenden untersuchten, sollen in dieser Studie die Testprotokolle von SonderpädagogInnen auf Fehler untersucht werden, sofern die Testungen im Rahmen eines sonderpädagogischen Gutachtens angefertigt worden sind.

Die Vermutung liegt nahe, dass trotz spezieller Schulung möglicherweise Studierende eher anfällig sind für Anwendungsfehler. Lipsius et al. (2008) merken hierzu jedoch an, dass besonders erfahrene und geschulte Testleiter anfällig für Anwendungsfehler seien. In einem Vergleich zweier Studien von Slate und Jones (1990) und Slate, Jones, Coulter und Covert (1992) stellen Lipsius et al. (Lipsius et al., 2008, S. 115) fest, dass praktizierende PsychologInnen mehr Fehler machten als Studierende.

Bei der Anwendung des HAWIK-IV erkannten Lipsius et al. (2008) vor allem Fehler beim Abbruchkriterium (z. B. zu früh abgebrochen; zu spät abgebrochen), beim Nachfragen (z. B. nicht nachgefragt, obwohl dies zwingend vorgeschrieben war), bei zusätzlichen Vorgaben (z. B. Wiederholung der Zahlenfolgen bei dem Subtest *Zahlen nachsprechen*), bei der falschen Anwendung der Umkehrregel, bei zusätzlichen oder ausgelassenen Zeitnahmen (z. B. statt 120 Sekunden Zeitgabe lediglich 1.20 Minuten Zeitgabe) oder bei der Vernachlässigung von Items vor der altersbedingten Anfangsaufgabe (fängt ein Kind altersbedingt z. B. bei Item 5 eines Subtests an, fließen die Items 1–4 als richtig gelöst in den Rohwert mit ein, sofern die Umkehrregel nicht angewendet werden

musste). Anhand von Fallbeispielen zeigten Lipsius et al., welche gravierenden Auswirkungen dies auf das Testergebnis und die daraus resultierende Diagnose haben könnte.

Alle Subtests des Bereichs *Sprachverständnis* sind in einer Anwendung bei einem zwölfjährigen Kind falsch durchgeführt worden. Dies führte zu einem Unterschied von immerhin 33 IQ Punkten und beeinflusste auch den Gesamt-IQ um elf Punkte. Z. B. hatte das Kind im Subtest *Gemeinsamkeiten finden* auf die Frage *was ist das Gemeinsame von Schmetterling und Biene* eine Ein-Punkte Antwort genannt (*Tiere*), auf Nachfrage ebenfalls eine Ein-Punkte Antwort (*können fliegen*). Dies hat der Testleiter dann addiert zu zwei Punkten. Es blieb aber bei einem Punkt, da beide Antworten zu der 1-Punkte Kategorie gehörten (2-Punkte z. B. bei *Insekt*). Das falsch errechnete Gesamtergebnis basierte auf zu viel gegebenen Rohwertpunkten und hätte ein Ergebnis zur Folge gehabt, welches die Diagnose Hochbegabung zu Unrecht nach sich ziehen könnte. Eine daraus resultierende Erhöhung der Anforderungen an das Kind könnte eine Überforderung bedeuten.

In einem weiteren Beispiel errechnete ein Testleiter ein fehlerhaftes Gesamtergebnis für ein 13-jähriges Kind von IQ 76. Sechs von zehn Subtests sind fehlerhaft durchgeführt worden. Bei der Auswertung sind vor allem die Items vor dem altersspezifischen Startpunkt unberücksichtigt geblieben. Tatsächlich lag der Gesamtwert nach der Korrektur bei 91. Wäre dieses Kind getestet worden, um Hinweise für den sonderpädagogischen Förderbedarf *Lernen* zu erhalten, hätte der falsch ermittelte Wert von IQ 76 einen deutlichen Hinweis auf diesen Förderbedarf ergeben im Gegensatz zu IQ 91, die möglichen fatalen Folgen dieser falschen Durchführung liegen auf der Hand, zumal es zum Zeitpunkt der Testung noch deutlich mehr Förder- bzw. Sonderschulen gab und die Inklusion erst später umgesetzt wurde. Eine aus dem falschen Testergebnis resultierende Beschulung in einer Förderschule Schwerpunkt *Lernen* hätte eine mögliche Unterforderung des Kinds bedeuten können.

Lipsius et al. (2008, S. 115) resümieren die Notwendigkeit einer genauen und akribischen Vorbereitung vor einer Anwendung mit dem HAWIK-IV. Insbesondere eine übermäßige Hilfestellung kritisieren die AutorInnen und zu Recht wird angemerkt, dass es „ein Fehler [ist] zu glauben, dass ein höherer Testwert auch besser für das Kind ist“ (ebd., S. 115). Neben speziellen Hinweisen für eine angemessene Durchführung des HAWIK-IV (z. B. die besonderen Umkehr- und Abbruchregeln beachten) werden auch allgemeine Hinweise für Testdurchführungen gegeben, z. B. die intensive Beschäftigung mit dem Testhandbuch, die Teilnahme an Schulungen, das Üben mit Hilfe von Probe-ProbandInnen und das Auswerten von Probe-Testungen (ebd., S. 116).

Obwohl Eser (2007) wie Lipsius et al. (2008) unzulässige Hilfen während der Testdurchführung ablehnen, sollte die Testdurchführung z. B. einer Testbatterie, aus der ein *g*-Maß resultiert, einen den Prüfungssituationen analogen Nach-

teilsausgleich bieten. Auch wenn Eser (2007, S. 10) die Durchführung einer Testbatterie unter Anwendung der ganzen Kunst und berufsständischen Ethik der Leistungs- und PsychodiagnostikerInnen fordert, widerspricht sich dieser Wunsch nach einem Nachteilsausgleich während der Testdurchführung mit der berufsständischen Ethik der PsychodiagnostikerInnen (Leibniz-Zentrum für Psychologische Information und Dokumentation, 2017), denn die Kunst derjenigen, die normierte standardisierte Tests anwenden besteht darin, die Durchführungsregeln nach den Vorgaben anzuwenden und nicht darin zu entscheiden, was ein Nachteilsausgleich während einer Testdurchführung darstellt. Hier deutet sich eine weitere Schwierigkeit für Tests anwendende SonderpädagogInnen an: einerseits haben sie es mit Kindern mit besonderen Bedürfnissen zu tun, die evtl. eine dem Kind angepasste Testdurchführung notwendig machen, andererseits sind aus einem normierten Test resultierende Testergebnisse nur interpretierbar, sofern die Durchführungsobjektivität gewahrt bleibt. Das Dilemma besteht nun darin, dass Autoren wie Eser (2007) und andere (Bundschuh, 2010) für Kinder mit sonderpädagogischem Bedarf eine angepasste und besondere Testdurchführung attestieren, aber nicht beschreiben, wie diese auszusehen hat. Auch der von Eser (2007, S. 10) geforderte Nachteilsausgleich während der Testdurchführung wird nicht konkretisiert. So bleibt es bei dem Dilemma, dass einerseits für besondere Kinder besondere Behandlungen gefordert werden bei der Anwendung von normierten Tests, die humanistisch formuliert wenig Spielraum für Widerspruch bieten, andererseits es zu den Regeln der Kunst gehört, die Durchführungsobjektivität zu wahren. Auch wenn Eser einen Nachteilsausgleich in Testsituationen nicht näher erläutert, verweist er zumindest auf Konkretisierungen für Nachteilsausgleiche in Prüfungssituationen (ebd., S. 10), wie sie von Keune & Frohnenberg (2004) beschrieben werden. Doch die von den Autorinnen beschriebenen Nachteilsausgleiche beziehen sich auf den beruflichen Prüfungskontext (z. B. modifizierte Berufstests auf Grund einer Blindheit oder Hörbeeinträchtigung), nicht auf mögliche Modifikationen bei der Durchführung beispielsweise des HAWIK-IV, der K-ABC oder ähnlicher standardisierter Intelligenztests.

Dabei stellt sich die Frage, ob es wirklich so schwierig ist, die an sich berechtigte Forderung nach einer angepassten Berücksichtigung von besonderen Kindern zu verwirklichen. Zwei kurze Beispiele sollen verdeutlichen, wie zulässig die Testsituation angepasst werden könnte, ohne die Durchführungsobjektivität zu verletzen:

- Ein vor dem Krieg geflüchtetes und niemals beschultes syrisches Kind mit Verdacht auf durchschnittliche bis überdurchschnittliche Intelligenz soll mit der KABC-II getestet werden. Im Subtest *Wort- und Sachwissen* kann das Kind auf die Frage *Wo begann die Renaissance* nicht auf eines der sechs Auswahl-Landkarten zeigen, welches die Gegend um Rom zeigt. Ein unzu-

lässiger Nachteilsausgleich könnte sein, dass die in Deutsch gestellte Frage auf z. B. hocharabisch übersetzt wird und das Wort Renaissance²⁵ erläutert wird, da das Kind noch nie in einer Schule war. Jedes darauffolgende Ergebnis könnte nicht verwertet werden, da die Kinder aus der Normstichprobe nicht die Frage auf hocharabisch gestellt und auch nicht das Wort Renaissance erklärt bekommen haben. Ein zulässiger Nachteilsausgleich wäre die ausschließliche Anwendung des *SFI-Index* (*Sprachfrei-Index* innerhalb der KABC-II) oder die Anwendung von nonverbalen Verfahren wie WNV oder SON-R 6–40.

- Ein Kind mit Spastizität hat Schwierigkeiten im Umgang mit den Würfeln des *Mosaik Tests* des WISC-IV. Unter Zeitdruck soll das Kind nach einer Vorlage mit Hilfe von Würfeln die Muster nachbauen, was dem Kind schwerfällt, da es die Hände nicht so gut verwenden kann wie die Kinder aus der Normstichprobe ohne Spastizität. Ein unzulässiger Nachteilsausgleich könnte in der Verdoppelung oder gar im Weglassen der Zeitvorgaben sein. Doch diese Fairness ist nur scheinbar, denn die Testergebnisse sind ohne Wert, da das Kind unter zu stark veränderten Bedingungen getestet wurde im Vergleich zu den Kindern aus der Normstichprobe. Ein zulässiger Nachteilsausgleich könnte darin bestehen, dass lediglich Tests oder Subtests ohne Zeitvorgaben durchgeführt und zusammengestellt werden, so dass das Kind mit der Spastizität sich so lange Zeit nehmen kann, wie es benötigt oder möchte. Die Durchführungsobjektivität wäre gewahrt ebenso wie eine faire Testbedingung.

Aus dem zweiten Beispiel wird deutlich, dass ein gut sortierter Testschrank mit einer größeren Auswahl an Testverfahren für eine heterogene Kindergruppe wie Kinder mit sonderpädagogischem Förderbedarf oder vermutetem Förderbedarf den Kindern eher gerecht wird als die Testung mit einem einzigen Intelligenztest. In diesem Fall bestünde eine weitere Schwierigkeit bei der Anwendung von Intelligenztests im sonderpädagogischen Kontext darin, dass angemessene Tests oft nicht zur Verfügung stünden. Im Zuge der Umsetzung der Inklusion kündigt sich hier ein Folgeproblem an. Da vermehrt Förderschulen aufgelöst und SonderpädagogInnen auf Regelschulen²⁶ verteilt werden, um dann dort die Kinder mit sonderpädagogischem Bedarf zu betreuen, ist kaum anzunehmen, dass die Regelschulen sich jeweils einen gut sortierten Testschrank anschaffen, da die Tests oft sehr teuer sind. Die neue KABC-II kostet inkl. Com-

25 Es läge sogar die Vermutung nahe, dass eine Frage zur Renaissance nicht kulturfair wäre in diesem Fall.

26 Ob Sonder- bzw. Förderschulen auch zu den Regelschulen gehören, soll hier nicht diskutiert werden; Regelschulen bezeichnen im Kontext dieser Arbeit Schulen ohne Sonder- bzw. Förderschulen.

puterauswertung 1 529 Euro (plus Steuern)²⁷, die WISC-IV inklusive Software 1 513,50 Euro (plus Steuern), der differentialdiagnostisch kaum aussagekräftige SON-R 6–40 komplett 2 076 Euro, obwohl als Ergebnis lediglich der Generalfaktor damit bestimmt werden kann und selbst der wegen veralteter Normen und veralteter Stimuli nutzlose SON-R 5½–17 ist noch zu kaufen für 2 087 Euro²⁸. Relativ günstig sind hingegen der CFT 1-R (118 Euro) und der CFT 20-R (232 Euro). Die letztgenannten Tests sind in der Sonderpädagogik recht beliebt und es ist zu befürchten, dass in Zukunft ausschließlich mit diesen kurzen und ein-dimensionalen Tests getestet wird, denn diese sind im Gegensatz zu den komplexeren Testverfahren für die Regelschulen erschwinglicher. Können also an Regelschulen tätige SonderpädagogInnen nicht auf komplexe Intelligenztests zurückgreifen, z. B. durch Schaffung von Testleihen, und wird die Anfertigung von sonderpädagogischen Gutachten nicht auf das ganze Jahr verteilt, damit nicht alle zur gleichen Zeit die Tests leihen müssen, könnte eine Folge bei der Umsetzung der Inklusion eine reduzierte Intelligenzdiagnostik mit kurzen Tests wie denen der CFT-Reihe sein mit dem Ergebnis einer reinen Statusdiagnostik ohne die Möglichkeit der Ableitung von Stärken und Schwächen aus Intelligenztests.

Renner und Mickley (2015a) gehen unter anderem der Frage nach, ob die Testverfahren überhaupt die Möglichkeit anbieten, bei Kindern mit sensorischen, körperlichen und/oder geistigen Beeinträchtigungen die Durchführungsobjektivität zu gewährleisten. Denn dazu müssten die Testmanuale detaillierte Hinweise geben, wie die Subtests unter Berücksichtigung von Beeinträchtigungen und Behinderungen objektiv durchzuführen sind. 23 Manuale deutschsprachiger Intelligenztests wurden analysiert, ob die Tests auch dann valide durchgeführt werden könnten, sollten die Kinder über Zugangsfertigkeiten (ebd., S. 90) wie z. B. Hör- und Sehfähigkeit, Motorik und Sprache nur eingeschränkt verfügen. Es wurde also die Frage gestellt, ob ein in der sensorischen, motorischen oder sprachlichen Fertigkeit eingeschränktes Kind noch auf Intelligenz getestet werden würde mit dem Intelligenztest, oder ob auf Grund der Beeinträchtigungen die Bedingungen zur Ermittlung einer Aussage über die Intelligenz fehlen. Dazu müsste es Hinweise in den Testmanualen geben, ob die Tests für Kinder mit Beeinträchtigungen geeignet oder ungeeignet sind, ob in der Normierung Kinder mit Beeinträchtigungen berücksichtigt worden sind, ob bestimmte Instruktionen abgestimmt auf entsprechende Behinderungen vorliegen oder ob eine mangelnde Testfairness in der Interpretation beachtet werden muss (ebd., S. 91). Die deutliche Mehrheit der Intelligenztests haben weder bei der Konstruktion noch bei der Normierung noch bei den Vorgaben für die

27 Alle Preise Stand 26. 1. 2016.

28 Anmerkung: inzwischen nicht mehr erhältlich (28. 7. 19).

Instruktionen beeinträchtigte Kinder berücksichtigt, obwohl 52 Prozent der Testmanuale die Anwendung bei Kindern mit Beeinträchtigungen nicht ausgeschlossen haben (ebd., S. 92) und sogar 30 Prozent der Verfahren ausdrücklich auf die Anwendung mit behinderten Kindern hinweisen. Die Berücksichtigung von Kindern mit Behinderungen stellt die Ausnahme in den Testverfahren dar. So gebe es z.B. Hinweise für Kinder mit dem Asperger-Syndrom in der IDS (Grob, Meyer & Hagmann-von Arx, 2009), einige Hinweise für Kinder mit einer Intelligenzminderung in mehreren Verfahren, Hinweise für Kinder mit Trisomie 21 in der IDS-P (Grob, Reimann, Gut & Frischknecht, 2013) und im WET (Kastner-Koller & Deimann, 2012). Besondere Hinweise zur Sicherung der Durchführungsobjektivität bei behinderten Kindern liegen ebenfalls nur in Ausnahmen vor: Für Kinder mit Sprachbehinderungen werden detaillierte Vorgaben im SON-R 2½-7 (Tellegen, Laros & Petermann, 2007) und im SON-R 6-40 (Tellegen, Laros & Petermann, 2012) vorgegeben. In der Regel liegen Angaben zu den Instruktionen für Kinder mit Behinderungen nicht, in wenigen Manualen eher vage vor (Renner & Mickley, 2015a, S. 96). Besonders kritisch merken Renner & Mickley an, dass „bemerkenswerterweise“ einige Testverfahren reklamieren, für behinderte Kinder geeignet zu sein, diese aber in der Normstichprobe ausgeschlossen waren (ebd., S. 98).

Selbst bei gutem Willen, angeeigneter Routine, intensiver Auseinandersetzung und Vorbereitung würde die Anwendung von Intelligenztests durch SonderpädagogInnen zusätzlich erschwert durch das weitest gehende Ignorieren von Kindergruppen bei der Normierung und Konstruktion der Tests, bei der Vorgabe von Instruktionen und bei der Ermittlung der Testgütekriterien, mit denen im sonderpädagogischen Kontext überproportional häufig gearbeitet wird: Kinder mit Sinnes-, Sprach-, kognitiven und motorischen Beeinträchtigungen und Kinder mit Intelligenzminderungen.

Mickley (2013) beschreibt die hohen Anforderungen an die TestanwenderInnen, sollten Kinder im Vorschulalter getestet werden. Dies kommt im sonderpädagogischen Kontext häufig vor. Für diese Kinder ist ein besonders sensibles Vorgehen notwendig, so dass u. a. Kenntnisse in der psychologischen Gesprächsführung und eine routinierte Testpraxis Voraussetzung sind, um falsch negative oder falsch positive Aussagen aus Testergebnissen zu vermeiden (ebd., S. 2). Notwendig sind dazu auch Kenntnisse über die Sensitivitäts- und Spezifitätsraten der Tests, um die Aussagekraft aus Testergebnissen zu relativieren. Eine niedrige Spezifitätsrate würde z. B. bedeuten, dass Testergebnisse zu falsch positiven Zuordnungen führen können. Mickley beschreibt mögliche Schäden, die durch medizinisch Tätige (z. B. KinderärztInnen) aus der Anwendung von Entwicklungs- und Intelligenztests führen können, doch lassen sich die beschriebenen Problematiken auf die Anwendung entsprechender Tests durch SonderpädagogInnen übertragen. Erläutert sei dies anhand des BUEVA-II (Esser & Wyschkon, 2012): Dieser Test soll Entwicklungsstörungen im Lern- und

Leistungsbereich erkennen, typische Fragestellungen, denen SonderpädagogInnen nachgehen. Die angegebene Spezifitätsrate von 81,4 Prozent würde bedeuten, dass bei auffälligen Testergebnissen und einer angenommenen Prävalenzrate von 2,5–5 Prozent für Lernstörungen mit 80–90 prozentiger Wahrscheinlichkeit dies zu falsch positiven Zuordnungen führen würde (Mickley, 2013, S. 4). Die Berücksichtigung schwacher Spezifitäts- bzw. Sensitivitätsraten führt zur logischen Konsequenz, dass normierte Testverfahren lediglich einen Teilbereich der Diagnostik darstellen können, da die Testverfahren teilweise nur im Kontext aller psychosozialen Umstände und unter Einbezug von Beobachtungssituationen und anderen Bausteinen der Diagnostik interpretiert werden können. Der Frage, ob die Berücksichtigung vieler Variablen in der sonderpädagogischen Diagnostik bei der Interpretation der Testergebnisse aus Intelligenztests tatsächlich eine Schwierigkeit für SonderpädagogInnen darstellt, soll im Rahmen dieser Arbeit nachgegangen werden. So ist z.B. anzunehmen, dass ein begrenzter Zeitrahmen bei der Erstellung eines sonderpädagogischen Gutachtens der notwendigen komplexen diagnostischen Vorgehensweise gegenüberstehen könnte.

Die vielfältigen Schwierigkeiten bei der Anwendung von Intelligenztests werden im Rahmen dieser Arbeit dargestellt und untersucht. Hebenstreit (2000) ging der Frage nach, ob die banalste aller Tätigkeiten bei der Anwendung von Intelligenztests bereits eine Schwierigkeit darstellen könnte: dem Zusammenzählen der Rohwertpunkte. Am Ende einer Gutachten-Ausbildung im Rahmen des Psychologiestudiums versicherten die Verfasser zwar, dass die Gutachten, in dessen Rahmen auch Intelligenztests durchgeführt worden sind, nach besten Wissen und Gewissen und unter Wahrung der berufsethisch festgeschriebenen Richtlinien angefertigt worden sind, doch sind nach Prüfung der Formulare bemerkenswert viele Rechenfehler festgestellt worden. Nach Auswertung von 184 AID-Protokollen waren lediglich nur zwei Protokolle fehlerfrei. 12,5 Prozent machten Rechenfehler, 40,8 Prozent machten Fehler beim tabellarischen Abgleich. Im Mittel sind 11,3 Auswertungsfehler beim AID (Kubinger, 2009a) festgestellt worden. Kubinger (2009b) regt unter Berücksichtigung von Hebenstreits Untersuchung eine grundsätzliche Nutzung der computerisierten Auswertung an, sofern dies möglich ist (ebd., S. 46). Angemerkt sei jedoch, dass in den meisten Auswertungsprogrammen (z.B. IDS; WISC-IV; KABC-II) die Rohwertpunkte manuell gezählt werden müssen, bevor sie in das Auswertungsprogramm eingetragen werden. Bei der Programmierung der Auswertungsprogramme wäre zu überlegen, ob nicht die Möglichkeit sinnvoller wäre, Item für Item (wie z.B. in der ehemaligen Computerauswertung der K-ABC) einzutragen, um die Verrechnungssicherheit zu erhöhen.

Hebenstreits Befunde stellen keinen Einzelfall dar und decken sich mit den weiter oben beschriebenen Ergebnissen von Alfonso et al. (1998) und Lipsius et al. (2008).

Abschließend sei eine Rezension zum HAWIK-III (Tewes, Rossmann & Schallberger, 1999) von Renner und Fricke (2001) erwähnt. Diese Rezension ist interessant, weil sie die Schwierigkeiten bei der Bewertung von Items exemplarisch und anschaulich verdeutlicht. Selbst bei sorgfältigster Vorbereitung stellt sich die Frage, ob die Testverfahren durch entsprechende Anweisungen die Voraussetzung für eine sichere Auswertung vorstellen. Anhand des HAWIK-III werden Beispiele verdeutlicht, die daran zweifeln lassen müssen. Insbesondere bei den Subtests aus dem *Indice Sprachverständnis* wird kritisiert, dass der Bewertungsspielraum verbaler Antworten der Kinder zu groß ist und manche Vorgaben zu schwammig oder gar falsch sind. Im Untertest *Allgemeines Verständnis* wird nach dem Vorteil von Zeitungen gegenüber dem Fernsehen gefragt. Antwortet das Kind *aus Zeitungen erfährt man mehr* erhalte es einen Punkt, antwortete es *aus der Zeitung erfährt man mehr* null Punkte. Im *Wortschatztest* würde die Definition von *anstrengend* mit zwei Punkten bewertet werden, würde das Kind *aufreibend* oder *zermürend* nennen, aber lediglich einen Punkt bei *erschöpfend* oder *ermüdend*. Falsche Antworten werden als richtig bewertet wie z.B. *Bundestagsabgeordnete schützen uns davor, in Kriege verwickelt zu werden* oder *wenn man das ABC kennt, weiß man, wie man Wörter schreibt* (ebd., S. 464). Bei einigen Items wird das Wort *warum* umgangssprachlich genutzt (fragt nach dem Grund), obwohl *wozu* (fragt nach dem Zweck) gemeint war.

Kinder mit einer hohen Sprach- und Bildungskompetenz wären an einigen Stellen benachteiligt, würden sie dem elaborierten Sprachcode den Vorzug geben gegenüber den an einigen Stellen umgangssprachlich anmutenden Sprachcode der Manuale, nicht nur des Manuals des HAWIK-III. So wird z.B. im schwersten Item des Subtests *Rechnen* der K-ABC folgendes im Testordner 3 (o.S.) gefragt: „Vor dem Zoo verkauft diese Dame Tierplaketten. Wenn sie 600 dieser Plaketten zum Preis von 40 Pfennig verkauft, wie viel Geld nimmt sie dann ein?“ Obwohl *240 DM* die richtige Antwort sein soll, wäre als Antwort *40 Pfennig* ebenfalls richtig.

Im WISC-IV lautet eine Frage: „Was sollst du tun, wenn ein Junge/Mädchen dich schlägt oder haut, der/das kleiner ist als du?“ (Petermann & Petermann, 2007, S. 263). Eine richtige Antwort mit einer vollen Punktzahl soll sein: „Gewalt löst überhaupt kein Problem.“ (ebd., S. 263). Wird nach einer Handlung gefragt (*was tust du (...)*) dürfte lediglich ein Slogan als Antwort nicht als völlig richtig gelten dürfen. Einer der derzeit aktuellsten mehrdimensionalen Tests, die KABC-II, verwirrt AnwenderInnen bei den Erläuterungen des Subtests *Rover* u. a. mit den Worten: „Der Zweck besteht darin, der Tp [Testperson] bei Ausführung der Aufgabe zu helfen und dem Testleiter, die Antwort der Tp zu kontrollieren“ (Melchers & Melchers, 2014, Testordner 2, o.S.).

Renner und Fricke (2001) beschreiben also Schwierigkeiten bei der Anwendung von Intelligenztests, die durchaus aktuell sind. Vielfältig können auch an-

dere Fehler in den Manualen attestiert werden. Beispielfhaft sei hier angefügt, dass im Handbuch des HAWIK-IV (Petermann & Petermann, 2007) ein zur Erklärung des Tests abgebildetes Testprofil eines Kinds falsche Angaben zum Vertrauensintervall enthält (ebd., S. 87) oder in der KABC-II Subtestbezeichnungen falsch genannt werden oder in der Erstauflage der KABC-II in den Tabellen ein T-Wert in einen Statine falsch transformiert wurde. SonderpädagogInnen und andere TestanwenderInnen müssen sich aber sicher sein, dass alle Angaben fehlerfrei sind, denn das sich verlassen können auf die Angaben der Testmanuale ist notwendig, um sich zweifelsfrei vorbereiten zu können.

2.4.2 Testanwendungen durch SonderpädagogInnen außerhalb Deutschlands

Belege für Schwierigkeiten bei der Anwendung von Intelligenztests durch *special education teachers*, z. B. bezüglich der Durchführungsobjektivität, scheinen nicht vorhanden. Eine Suche über die Metadatenbank PubPsych²⁹ ergab zwar Treffer bei den Stichworten *special education teacher intelligence* (58), *special education teacher intelligence test* (18), *teacher intelligence* (532), *teacher intelligence test* (174) und *special education teacher assessment* (296), doch ergab die Analyse der abstracts keine Hinweise im Sinne der Fragestellung.

Eine Suche nach vorhandener Literatur wäre müßig, würden *special education teacher* gar keine Intelligenztests durchführen, dem zur Folge gäbe es auch keine Literatur über Schwierigkeiten bei der Anwendung von Intelligenztests durch *special education teachers*.

Es stellte sich also die Frage, ob *special education teachers* ähnlich institutionalisiert wie in Deutschland Intelligenztests durchführen (dürfen).

Dazu wurden in ausgewählten Staaten ExpertInnen schriftlich per E-Mail befragt. Auch wenn sich eine Tendenz bei der Beantwortung der Frage abzeichnet, ob Intelligenztests durch *special education teachers* angewendet werden, wird nicht der Anspruch erhoben, dies für jeden der 194–207 Staaten der Erde untersucht zu haben, da dies den Rahmen der eigentlichen Untersuchung übersteigt.

Eine Anfrage an einen englischen Experten soll hier beispielhaft vorgestellt werden, Anfragen an ExpertInnen anderer Staaten wurden entsprechend geändert (z. B. statt *english danish* etc.). Auf Englisch wurden ExpertInnen in Belgien, Dänemark, Großbritannien, Kanada, Niederlanden, USA und Schweden, auf Deutsch in Österreich und der Schweiz befragt:

29 Abfrage am 1.9.16.

I examine difficulties in dealing with intelligence tests by *special education teachers* (teachers who teach mentally handicapped, mentally retarded, behavioral problems, speech-impaired and physically disabled children).

In this study, I hope for your help in answering these questions:

1. Do english *special education teachers* also perform intelligence tests (for example WISC; CFT; SON etc.)?
2. If so, what difficulties are well known in the procedure of intelligence tests, which are performed by *special education teachers*?
3. Are there any studies on these difficulties?
4. Is the performance of intelligence tests by *special education teachers* regionally differently regulated in England/Wales/Scotland?
5. When *special education teachers* do not perform intelligence tests, then who is leading intelligence tests for children with special needs?

Ein Beispiel für auf Deutsch gestellte Fragen (hier an ein Sonderpädagogisches Zentrum in Österreich):

Im Rahmen einer wissenschaftlichen Studie untersuche ich die Schwierigkeiten von deutschen SonderpädagogInnen bei der Durchführung von standardisierten Testverfahren (z. B. Intelligenztests, Persönlichkeitstests etc.).

Gerne möchte ich im Rahmen dieser Studie klären, wie die Durchführung von Testverfahren außerhalb Deutschlands durch SonderpädagogInnen geregelt ist.

1. Ist das Berufsbild der österreichischen SonderpädagogInnen in etwa mit dem der deutschen vergleichbar?
2. Führen österreichische SonderpädagogInnen auch Intelligenz- und Persönlichkeitstests durch (z. B. WISC-IV; SON; CFT; K-ABC (...))
3. Wenn Ja, welche z. B.?
4. Wenn Ja, sind Schwierigkeiten bei der Durchführung der Testverfahren bereits untersucht worden?
5. Wenn Nein, wer führt Intelligenz- und Persönlichkeitstests durch, sollten diese hilfreich eingesetzt werden bei der Feststellung eines sonderpädagogischen Förderbedarfes?

Angemerkt sei, dass in einem frühen Stadium dieser Arbeit nicht nur die Anwendung von Intelligenztests, sondern allgemein die Anwendung normierter standardisierter Testverfahren untersucht werden sollte. Aus diesem Grund wurde auch nach der Anwendung von Persönlichkeitstests gefragt. Dies beeinträchtigt jedoch nicht die Antworten bezüglich der Anwendung von Intelligenztests, es wurden lediglich alle Aussagen zu den Persönlichkeitstests außen vorgelassen.

Die Gründe für die Reduzierung in dieser Forschungsarbeit auf Intelligenztests ohne Einbezug anderer normierter standardisierter Tests liegen zum einen in der (kontroversen) Bedeutung von Intelligenztests in der Sonderpädagogik, zum anderen in der Heterogenität von normierten standardisierten Testverfahren. So spielt die Durchführungsobjektivität eine sehr viel geringere Rolle bei den Persönlichkeitstests.

Zunächst gibt Tabelle 2 eine Übersicht über die Anfragen in den verschiedenen Staaten (die Zahlen beziehen sich jeweils auf eine E-Mail an eine MitarbeiterIn, die laut Beschreibungen auf den entsprechenden Homepages geeignet zur Beantwortung der Fragen erschienen³⁰):

Im Fokus stand die Frage, ob *special education teachers* Intelligenztests anwenden und wenn ja, ob mit der Anwendung verbundene Schwierigkeiten bekannt sind.

Tabelle 2. Anfragen und Rücklauf zur Anwendung von Intelligenztests durch special education teachers.

	Niederlande	Belgien	Dänemark	Schweden	Großbritannien	USA	Kanada	Schweiz	Österreich	Total
UniversitätsmitarbeiterInnen	60	8	34	60	20	6	6	8		202
Verbände	2				1	8	11	3		25
Schulen mit sonderpäd. Ausrichtung					35	37				72
Sonderpäd. Zentren (Österreich)									25	25
sonstige								1		1
Total	62	8	34	60	56	51	17	12	25	325
Rücklauf (Antworten in Form einer E-Mail)	19 31 %	3 38 %	15 44 %	15 25 %	10 18 %	7 14 %	7 41 %	11 92 %	2 8 %	89 27 %

30 Die angeschriebenen Personen wurden personalisiert angeschrieben, also nicht allgemein z. B. mit „Dear Sir“, sondern mit Titel und Namen. Eine Kopie der Betreuungsbescheinigung durch die Uni Flensburg war als Anhang beigelegt. Der Zweck der Studie wurde kurz erläutert.

2.4.3 Überblick über die Anwendung von Intelligenztests durch *special education teachers* außerhalb Deutschlands

In den USA wird die Anwendung von Intelligenztests durch *special education teachers* übereinstimmend verneint. Interessant ist der mehrfach genannte Hinweis darauf³¹, dass komplexe Intelligenztests, deren adaptierte Versionen ins Deutsche auch von SonderpädagogInnen angewendet werden (z. B. WISC-IV; KABC-II), in den USA lediglich von besonders qualifizierten Fachleuten durchgeführt werden dürfen. Diese Intelligenztests werden als Level-C Tests bezeichnet und es ist eine *doctoral level psychologists* Qualifikation notwendig. In den von Pearson festgelegten Anforderungen (Pearsonassessment, 2019) ist unter anderem festgelegt, dass eine besondere Befähigung (*high level*) vorliegen muss, um die Anwendung der zu dieser Kategorie gehörenden Tests durchführen zu dürfen. Dies kann ein Dokortitel in *psychology* sein, aber auch in *education* oder fortlaufend wahrgenommene Fortbildungen unter der Aufsicht entsprechender Fachverbände oder eine Lizenzierung zur Anwendung. Diesen Richtlinien ist nicht zu entnehmen, ob *special education teachers* ebenfalls berechtigt sind, sofern sie sich qualifiziert haben oder über einen Dokortitel verfügen.

Allen Antworten auf meine Anfrage ist gemein, dass die Anwendung von Intelligenztests durch PsychologInnen durchgeführt wird und die Voraussetzungen für die Anwendung von komplexen Intelligenztests rigide geregelt ist.

In *Kanada* fielen die Antworten kurz und eindeutig aus, Intelligenztests werden dort von PsychologInnen durchgeführt, nicht von *special education teachers*.

Ebenso in *Großbritannien*³², wo für die Anwendung von komplexen Intelligenztests *Educational Psychologists* angeführt werden. Jedoch führen z. B. *learning support teachers* Übersichtstests wie die Raven's Progressive Matrices (dt. Version: Raven, Raven & Court, 2010) durch, wenn es um den Übergang von Schulformen oder Lernstörungen geht. Studien zu Schwierigkeiten im Umgang mit diesen eindimensionalen Tests, welche einen Hinweis auf den *g*-Faktor geben, können nicht genannt werden.

Alle Antworten von Fachleuten aus *Schweden* und *Österreich* sind ebenfalls eindeutig und kurz. Einem *special education teacher* (in Österreich ebenfalls SonderpädagogIn genannt) ist es nicht erlaubt, Intelligenztests anzuwenden, diese werden ausschließlich von PsychologInnen durchgeführt.

31 Anmerkung T. J.: Die Ergebnisse der Befragungen beruhen auf privaten Kommunikationen und werden deshalb im Folgenden nicht als zitierfähige Quelle benannt; auch da von einem stillschweigenden Einverständnis zur Veröffentlichung der Antworten auf die gestellten Fragen nicht ausgegangen werden kann.

32 Es wurden Fachleute in England, Wales und Schottland angeschrieben, nicht in Nordirland.

Nach Auswertung aller Rückmeldungen aus *Dänemark* ist die Antwort eindeutig auf die Frage, ob *special education teachers* Intelligenztests anwenden. Dies ist den PsychologInnen vorbehalten.

Den *Niederlanden* und *Belgien* ist gemeinsam, dass für die Anwendung von Intelligenztests neben PsychologInnen auch OrthopädagogInnen (orthopedagogists) zuständig sind, deshalb werden die Rückmeldungen aus diesen beiden Staaten gemeinsam dargestellt. Obwohl OrthopädagogInnen sich nicht als Lehrkräfte verstehen, studieren sie z. B. an der Genter Universität an der Faculty of Psychology and Educational Sciences.

Ein vergleichbares Studium in Deutschland ist nicht bekannt. Obwohl sich die Tätigkeitsfelder von OrthopädagogInnen und deutschen SonderpädagogInnen ähneln, wird großen Wert daraufgelegt, dass OrthopädagogInnen keine Lehrkräfte sind. In einer privaten E-Mail wird erklärt, dass geschichtlich der Begriff Orthopädagogik nach dem zweiten Weltkrieg aus dem Begriff Heilpädagogik entstanden ist, da das Wort Heil negativ konnotiert war. Trotz der Abgrenzung zu der Tätigkeit einer Lehrkraft liegen Parallelen zwischen deutschen SonderpädagogInnen und OrthopädagogInnen vor. Die meisten Schulen *for special education* haben *their own orthopedagoog or psychologist that can perform intelligence tests* (private Kommunikation, 2015). Übereinstimmend wird die Anwendung von Intelligenztests durch *special education teachers* verneint (damit sind nicht die OrthopädagogInnen gemeint, sondern im sonderpädagogischen Kontext arbeitende Lehrkräfte). Eine Ausnahme stellt der NIO Test dar (Van Dijk & Tellegen, 2004), ein kurzer Übersichtstest, der gelegentlich vor Schulformwechseln auch von Lehrkräften durchgeführt werden soll.

Der Fragenkatalog an Fachleute aus der *Schweiz* wurde um die Frage erweitert, worin der Unterschied zwischen schweizer Sonderpädagogen und schweizer Heilpädagogen besteht. Fragen dieser Art sind wichtig, um eine falsche Vergleichbarkeit auf Grund identischer Begriffe mit unterschiedlichen Bedeutungen zu vermeiden. Sonderpädagogik in der Schweiz kann lediglich an der Universität Zürich studiert werden. Diese wissenschaftliche Ausbildung ist nicht vergleichbar mit dem deutschen Studiengang Sonderpädagogik. SonderpädagogInnen in der Schweiz sind WissenschaftlerInnen und nicht Regelschullehrkräfte. Mehrfach wird jedoch eine Vergleichbarkeit zwischen deutschen SonderpädagogInnen und schweizer HeilpädagogInnen genannt sowie eine synonyme Verwendung der Berufsbezeichnungen SonderpädagogIn (eher reformierte Gebiete; ein aus Deutschland übernommener Begriff) und HeilpädagogIn (eher katholische Gebiete). Kinder mit sonderpädagogischem Förderbedarf werden unterrichtet von schulischen Heilpädagogen (SHP). Das sonderschulpädagogische System in der Schweiz wird durch das föderalistische System und die Vielsprachigkeit als sehr heterogen beschrieben, was die Ableitung einer Tendenz aus den Fragen erschwert. Es wird von einigen Fachleuten nicht ausgeschlossen, dass auch Sonder-/HeilpädagogInnen Intelligenztests durchführen, doch wird

dies nicht als Bestandteil der Stellenbeschreibung von Sonder-/HeilpädagogInnen beschrieben. Hier werden übereinstimmend (Schul-)PsychologInnen genannt. Die Anwendung von Intelligenztests durch Sonder-/HeilpädagogInnen wird z.B. im Zusammenhang mit Ausbildungs- und Forschungszwecken genannt, aber für den beruflichen Alltag unüblich.

Folgendes Fazit kann gezogen werden: In einer nicht repräsentativen Anfrage an 325 Fachleuten aus 9 Staaten wurde nach der Anwendung von Intelligenztests durch die deutschen SonderpädagogInnen vergleichbare Berufsgruppen gefragt. Diese werden zusammengefasst hier als *special education teachers* bezeichnet. Im Fokus stand die Frage, ob *special education teachers* Intelligenztests durchführen und wenn ja, ob damit verbundene Schwierigkeiten und entsprechende Untersuchungen zu evtl. Schwierigkeiten bekannt sind und benannt werden können.

Obwohl diese Fragestellung nicht im Zentrum des Interesses der eigentlichen Untersuchung dieser Arbeit steht, ist der Umweg über diese Befragung hilfreich, um zu erklären, warum es keine Befunde zu Schwierigkeiten in der Anwendung von Intelligenztests durch *special education teachers* zu geben scheint. Entsprechende Befunde gibt es z.B. zu Schwierigkeiten in der Anwendung von Intelligenztests durch PsychologInnen (Slate & Jones, 1990; Slate et al., 1992; Alfonso et al., 1998).

Es gibt keine Befunde zu Schwierigkeiten bei der Anwendung von Intelligenztests durch *special education teachers* außerhalb Deutschlands, weil *special education teachers* keine Intelligenztests ähnlich institutionalisiert wie in Deutschland durch SonderpädagogInnen durchführen. Im sehr moderaten Umfang werden die Anwendungen von Intelligenztests erwähnt, teils eher zu Schulungszwecken. Dreimal wird die Anwendung von kurzen und eindimensionalen Intelligenztests durch *special education teachers* bzw. *teachers* beim Übergang von Schulformen erwähnt. In Belgien und den Niederlanden führen neben den PsychologInnen auch OrthopädagogInnen Intelligenztests durch, doch ist dieses Berufsbild nicht mit deutschen SonderpädagogInnen vergleichbar. Durch die nicht repräsentative Anfrage an Fachleute und die Beschränkung auf einige Staaten wird nicht der Anspruch erhoben, den Nachweis dafür zu erbringen, dass *special education teachers* keine Intelligenztests wie in Deutschland durchführen und dementsprechend keine Schwierigkeiten über die Anwendung in Form von Untersuchungen vorliegen können. Es gibt jedoch deutliche Hinweise als Schlussfolgerung aus dieser Befragung, die für diese Annahme sprechen und keine Hinweise, die dagegen sprechen.

2.5 Intelligenztests

In diesem Kapitel sollen Intelligenztests vorgestellt werden, die von SonderpädagogInnen angewendet werden. Die Intelligenztests basieren auf hierarchischen Intelligenzmodellen und können unterteilt werden in ein- und mehrdimensionale Testverfahren. Bei den eindimensionalen Testverfahren wird in der Regel ein Gesamtwert ermittelt, welcher auf das intellektuelle Potential des Kinds im Vergleich mit gleichaltrigen Kindern hinweist. Bei mehrdimensionalen Verfahren können zusätzlich auf individueller Ebene Stärken und Schwächen des Kinds ermittelt werden, gerade diese Stärken-Schwächen Analysen sind interessant zur Ableitung pädagogischer Maßnahmen.

Neben der Einteilung von Intelligenztests in ein- und mehrdimensionale Verfahren können Unterschiede in der Art der Durchführung beschrieben werden. Es gibt Verfahren, die ausschließlich verbal durchgeführt werden müssen und Verfahren, die darüber hinaus auch nonverbal durchgeführt werden können. Nonverbale Verfahren sind interessant bei der Testung von Kindern mit Hörbeeinträchtigungen und Kindern, die Deutsch nicht oder nicht gut sprechen (z.B. geflüchtete Kinder). Nonverbale Intelligenztests testen allerdings nicht die nonverbale Intelligenz, sondern ermöglichen die Erfassung von Intelligenz(teilen) ohne Anwendung der Sprache.

Einige tabellarisch aufgeführte Basisinformationen (Angaben über Testgütekriterien, Preise, Dauer, Altersbereich, siehe Tabelle 4, Tabelle 5 und Tabelle 6) und Empfehlungen zur Nützlichkeit für häufig auftretende Fragestellungen (siehe Tabelle 8) beenden die Vorstellung ausgewählter Testverfahren.

Die Qualität psychometrischer Testverfahren wird über dessen Testgütekriterien ermittelt. Sind nach Prüfung der Testgütekriterien die Werte nicht akzeptabel, sollte der Test nicht angewendet werden, da evtl. der Test weder exakt (reliabel) misst noch bestimmt werden kann, was der Test eigentlich misst (Validität). Zentral bei der Darstellung von Testverfahren ist also die Beschäftigung mit den Testgütekriterien, die zu Beginn aufgeführt werden.

Die Vorstellung von TestleiterInneneffekten schließt das Kapitel ab und ist als Darstellung möglicher Fehlerquellen sowohl interessant bei der Anwendung von Intelligenztests als auch bei der Beantwortung des Fragebogens, auf den sich diese Untersuchung maßgeblich bezieht. Somit ist die Darstellung von TestleiterInneneffekten das Bindeglied zwischen dem theoretischen und dem methodischen Teil.

2.5.1 Testgütekriterien

Testgütekriterien werden unterschieden in Haupt- und Nebengütekriterien. Meist werden in den Manualen der Intelligenztests die Werte der Hauptgütekriterien beschrieben.

2.5.1.1 Hauptgütekriterien

Von den Hauptgütekriterien Objektivität, Reliabilität und Validität ist die Objektivität von besonderer Bedeutung, da insbesondere die Durchführungs- und Auswertungsobjektivität starke Bezüge zu dieser Untersuchung haben.

„Ein Test ist dann objektiv, wenn er dasjenige Merkmal, das er misst, unabhängig von Testleiter, Testauswerter und von der Ergebnisinterpretation misst“ (Moosbrugger & Kelava, 2007, S. 8). Ist ein Test nicht objektiv durchführbar, ist die Prüfung der anderen Testgütekriterien obsolet, denn es kann weder exakt noch gültig getestet werden, wenn die ermittelte Leistung der ProbandIn von der Person abhängig ist, die das Kind testet. Das Testgütekriterium Objektivität wird unterteilt in Durchführungs-, Auswertungs- und Interpretationsobjektivität.

Die Durchführungsobjektivität ist gegeben, wenn der Test dermaßen standardisiert vorliegt, dass es für den Testenden keine Zweifel über die Durchführung gibt. Sind z.B. die Durchführungsregeln nicht eindeutig oder gar nicht beschrieben, würde die Durchführung im Ermessen des Testenden liegen. So ist z.B. bei einigen Subtests der KABC-II nicht eindeutig beschrieben, ob nach Ablauf der Zeitgrenze ein Hinweis über den bevorstehenden Ablauf der Zeit gegeben werden darf oder nicht. Kinder, die einen Hinweis über das nahende Ende der Zeitbegrenzung erhielten, hätten einen Vorteil vor den Kindern, die keinen Hinweis erhielten. Je mehr sich die Art und Weise, wie der Test durchgeführt wird von Testdurchführung zu Testdurchführung unterscheidet, umso mehr ist die Durchführungsobjektivität gefährdet. Aufgabe der TestautorInnen ist es, die Durchführung eindeutig zu beschreiben, Aufgabe des Testenden ist es jedoch, die Durchführungsobjektivität durch exaktes Einhalten der Regeln zu wahren. Auf beiden Seiten kann die Durchführungsobjektivität gefährdet werden. Beide Seiten stehen in einer Wechselwirkung, denn ist die Durchführung eines Tests nicht eindeutig standardisiert beschrieben, wird es den Testenden schwerfallen, die Durchführungsobjektivität zu wahren. Im Idealfall ist ein Test so standardisiert beschrieben, dass die TesterIn vollkommen austauschbar wäre. Eine relativ vollkommene Durchführungsobjektivität könnte vorliegen, wenn der Test von einem Computer durchgeführt werden würde.

Die Auswertungsobjektivität beschreibt die Ermittlung von Testergebnissen unabhängig davon, welche Person ein Kind testet. Auf die Frage, was Wut und

Freude gemeinsam haben aus dem WISC-IV (Subtest *Gemeinsamkeiten Finden*) könnte ein Kind mit *Tränen* antworten. Es ist möglich, dass diese Antwort des Kinds mit einem Punkt oder mit null Punkten bewertet wird. Beides wäre korrekt, denn die Antwort *Tränen* ist im Antwortkatalog möglicher Antworten im Handbuch nicht enthalten, daraus könnte die Bewertung mit null Punkten resultieren. Allerdings gibt es auch übergeordnete Antwortrichtlinien, die einen Spielraum bei der Bewertung lassen. Da es sowohl Wuttränen als auch Freudentränen gibt, wäre die Vergabe von einem Punkt zulässig.

Die Bewertung in Abhängigkeit vom Testenden wird noch deutlicher bei dem Subtest *Wortschatz* der KABC-II. Kinder sollen Gegenstände benennen, die auf Bildern zu sehen sind. Spricht ein Kind nicht Deutsch, benennt die Gegenstände aber in einer anderen Sprache (z.B. arabisch), kann der Testende dies als richtig bewerten, sofern der Testende ebenfalls die Sprache spricht und ermessen kann, ob das vom Kind benannte Wort (z.B. in Arabisch) korrekt das Abgebildete wiedergibt. Überspitzt formuliert bedeutet dies, dass die Intelligenz des Kinds von den Sprachkenntnissen des Testenden abhängt. Während bei der Wahrung der Durchführungsobjektivität eindeutige und standardisierte Hinweise die Durchführung regeln, wird zur Wahrung der Auswertungsobjektivität die Auswertung der ermittelten Ergebnisse durch eindeutige Vorgaben gefordert. Eine hohe Auswertungsobjektivität würde vorliegen, wenn mehrere Personen unabhängig voneinander eine Testleistung auswerten müssten, das Maß der Übereinstimmung könnte mit dem *Konkordanzkoeffizienten W* nach Kendall (1962) dargestellt werden.

Interpretationsobjektivität liegt vor, wenn Testergebnisse zu gleichen Interpretationen führen, unabhängig davon, wer die Ergebnisse interpretiert. Dies ist auf *Stratum-III*-Ebene (Generalfaktor der Intelligenz, siehe CHC-Modell) noch einfach zu gewährleisten, da im europäischen Raum entsprechend der Gaußschen Kurve der Normbereich eines Gesamtwerts mit den mittleren 2/3 (der Ergebnisse einer Population) angegeben wird (z.B. bei der Skalierung IQ 85–115: 2/3 einer Population haben einen IQ von 85–115, bei der Skalierung T-Wert 40–60 usw.). Ein Gesamtwert wird also (unter Einbezug des Vertrauens- bzw. Konfidenzintervalls) mit dem Normbereich abgeglichen und das Gesamtergebnis befindet sich im Normbereich oder nicht. Deshalb sind eindimensionale Testverfahren wie die der CFT-Reihe eindeutig zu interpretieren. Bei mehrdimensionalen Verfahren mit Ergebnissen auf *Stratum-II*-Ebene (sog. *Indice*) ist dies weniger eindeutig, da die Ergebnisse der *Indice* im Kontext betrachtet werden müssen. Interpretationsobjektivität läge für diesen Fall vor, wenn die Manuale eindeutige Hinweise auf die Inhalte der *Indices* geben, so dass Ergebnisse auf *Stratum-II*-Ebene beurteilt werden können.

„Die Reliabilität gibt den Grad der Messgenauigkeit eines Messwerts an“ (Bühner, 2011, S. 60). Mit der Reliabilität wird gemessen, wie genau ein Test ein Konstrukt misst. Dabei ist zunächst ohne Interesse, um was für ein Konstrukt

es sich handelt. Ein Test, der vorgibt die Intelligenz zu messen, tatsächlich aber auf Grund von Konstruktionsfehlern eher die Konzentration misst, kann dennoch reliabel sein. Jedoch kann ein nicht reliabler Test weder die Konzentration noch die Intelligenz noch sonst irgendetwas messen. Der Grad der Messgenauigkeit wird mit dem Korrelationskoeffizienten geschätzt. Liegt dieser bei 1, so lägen keinerlei Messfehler vor, der Test wäre also ideal reliabel, liegt dieser bei 0, so wäre der Test ein einziger Messfehler. Korrelationskoeffizienten sollten $.7$ nicht unterschreiten. (Moosbrugger & Kelava, 2007, S. 11).

Reliabilitäten werden üblicherweise in Retest-, Paralleltest-, Testhalbierungs-Reliabilität und Innere Konsistenz angegeben.

Bei der Retest-Reliabilität (auch Stabilität) wird ein Test nach einer gewissen Zeit erneut durchgeführt und die Korrelation zwischen den Ergebnissen gemessen. Übungseffekte beeinflussen allerdings die Aussagekraft der Korrelation. Diese werden durch die Paralleltest-Reliabilität verringert, bei der die Korrelation zwischen den Ergebnissen aus zwei parallelen Testformen eines Tests ermittelt wird. Dieses Verfahren „wird oftmals als Königsweg der Reliabilitätsbestimmung bezeichnet“ (Moosbrugger & Kelava, 2007, S. 12). Es muss allerdings sichergestellt sein, dass die zwei parallelen Formen eines Tests auch tatsächlich gleich sind.

Wird ein Test in zwei Hälften geteilt und die Korrelation der Teile berechnet, würde der Korrelationskoeffizient die Testhalbierungs-Reliabilität (auch Splithalf-Reliabilität) beschreiben. Hier ist zu beachten, dass ein Test gültiger werden würde, je länger er ist. Bei der Halbierung eines Tests wird er hypothetisch auf die ursprüngliche Länge gerechnet, damit der aus der Kürzung des Tests resultierende ungünstigere Wert ausgeglichen wäre. Wird jedes Item hingegen als eigenständiger Testteil betrachtet und die Korrelation zwischen den Items errechnet, würden diese Korrelationen die innere bzw. interne Konsistenz messen. Je höher die Testteile (bestehend also aus einzelnen Items) miteinander positiv korrelieren, desto höher ist die interne Konsistenz (ebd., S. 12). Die meist mit dem Cronbach- α -Koeffizienten (Cronbach, 1951) dargestellten Korrelationen sind nicht unumstritten, unter anderem, da seine Höhe stark abhängig ist von der Anzahl der Items und zu Verzerrungen führen könnte (Bortz & Döring, 2006). Dennoch wird die Reliabilität von Testverfahren häufig mit der inneren Konsistenz belegt, da der Aufwand überschaubar ist (Schermelleh-Engel & Werner, 2007).

„Die Validität gibt an, ob der Test auch wirklich misst, was er zu messen beansprucht“ (Bühner, 2011, S. 61). Dieses Kriterium gilt als das wichtigste Kriterium. Während eine hohe Objektivität und eine hohe Reliabilität die Voraussetzungen für einen guten Test sind, legitimiert die Validität letztlich die Gültigkeit des Tests. Bei einem nicht validen Test sind die Berechnungen aller anderen Testgütekriterien vergeudete Zeit. Die Validität erlaubt Rückschlüsse über ein Testergebnis mit Verhalten bzw. Fähigkeiten oder Fertigkeiten außer-

halb der Testsituation (Moosburger & Kelava, 2007). Beansprucht ein Test die Erfassung der Intelligenz, so sollte das Intelligenztestergebnis das intellektuelle Potential abbilden, über das die Testperson in allen Bereichen des Lebens verfügt, z. B. in der Schule.

Die Validität wird meist mit der Inhalts-, der Konstrukt- und/oder der kriterienbezogenen Validität angegeben.

Im engeren Sinne gibt lediglich die Inhaltsvalidität an, ob ein Test misst, was er vorgibt zu messen (Murphy & Davidshofer, 2001), während die anderen Arten der Validitätsprüfung weniger messen, ob ein Test misst, was er vorgibt zu messen, sondern ob das gemessen wird, was mit Hilfe der Testkennwerte an abgeleiteten Aussagen postuliert wurde (Bühner, 2011, S. 61).

Inhaltsvalidität liegt vor, wenn jedes Item repräsentativ das zu messende Konstrukt abbildet. Dies kann empirisch nicht bestimmt werden, deshalb wird die Inhaltsvalidität argumentativ begründet (z. B. durch Expertisen von Fachleuten) oder die Inhaltsvalidität wird vernachlässigt, denn die argumentative Begründung der Inhaltsvalidität ist umstritten. Es ist methodisch schwierig und aufwändig, die Inhaltsvalidität zu belegen. Hartig, Frey und Jude (2007) kritisieren die fehlenden Versuche, die Validität eines Tests mit Hinweisen zur Inhaltsvalidität belegen zu wollen und aus pragmatischen Gründen gleich auf die Konstrukt- und Kriteriumsvalidität auszuweichen.

Im Zusammenhang mit der Inhaltsvalidität kann die wissenschaftlich umstrittene Augenscheinvalidität genannt werden, bei der auch Laien auf den ersten Blick der Zusammenhang zwischen Test und zu testendes Merkmal nachvollziehbar scheint.

Die Konstruktvalidität misst das theoretische Konstrukt des Tests, ob z. B. *Sprachverständnis* des WISC-IV tatsächlich *Sprachverständnis*, das *Indice Gf* der KABC-II tatsächlich die *fluide* Intelligenz misst usw. Allerdings sollten sich Angaben zur Konstruktvalidität auf alle Konstrukte eines Tests beziehen. Die Methode der Wahl ist die Multitrait-Multimethod-Methode nach Campbell und Fiske (1959), infolgedessen die konvergente Validität (Korrelation verschiedener Tests, die etwas Identisches messen sollen, z. B. wurde die neue KABC-II mit verschiedenen bewährten Tests verglichen) und die Diskriminante Validität (Korrelation mit verschiedenen Tests, die etwas anderes messen sollen. In diesem Fall sollte die Korrelation niedrig oder nicht vorhanden sein) ermittelt werden. Angaben zur konvergenten Validität über das Vergleichen von neuen mit bewährten Tests wäre dann zweifelhaft, wenn die Testgütekriterien der bewährten Tests nicht ausreichend wären. Würde z. B. die konvergente Validität im Rahmen der Konstruktvalidität über einen Vergleich eines neuen Tests mit dem vielfach und seit Jahrzehnten auch in der Sonderpädagogik eingesetzten Coloured Progressive Matrices (CPM; Raven, Raven & Court, 2010) verglichen werden, so würde mit einem Test verglichen werden, der durch die reduzierte Testung auf einen so kleinen Teilbereich keine gültigen Rückschlüsse auf die

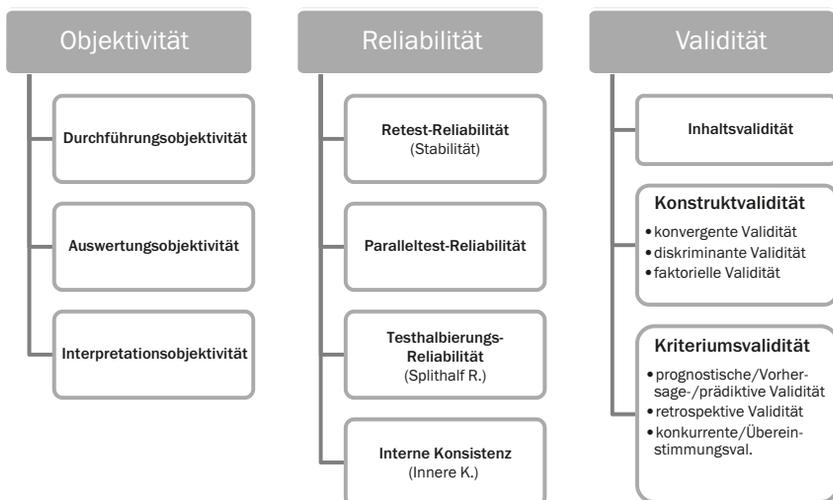
Intelligenz zuliebe (Renner & Mickley, 2015b, S. 72). Hinweise auf die konvergente Validität durch einen Vergleich mit den CPM wären in diesem Fall ohne Nutzen.

Eine weitere Möglichkeit im Rahmen der Konstruktvalidität ist die Anwendung konfirmatorischer Faktorenanalysen. Die damit ermittelte faktorielle Validität wird mit Hilfe von Faktorenanalysen ermittelt, die Zusammenhänge zwischen verschiedenen Tests untersucht. Einerseits können so homogene konstruktnahe Inhalte, andererseits konstruktferne Bereiche ermittelt werden (Moosbrugger & Kelava, 2007).

Beim Vergleich des Testergebnisses mit Außenkriterien wird die Kriteriumsvalidität bestimmt. Wird das Testergebnis eines Intelligenztests mit dem Außenkriterium Studienerfolg korreliert, läge im Rahmen der Bestimmung der Kriteriumsvalidität eine Vorhersagevalidität (auch prognostische Validität, prädiktive Validität) vor: das Testergebnis wird verglichen mit später erhobenen Kriterien. Würde mit zurückliegenden Kriterien verglichen werden (z. B. Intelligenztestergebnisse mit zurückliegenden Schulnoten), wäre die retrospektive Validität bestimmt.

Würde hingegen die Testleistung weder mit zukünftigen noch mit zurückliegenden Kriterien verglichen werden, läge die konkurrente Validität (auch Übereinstimmungsvalidität) vor, z. B. beim Vergleich der Intelligenztestleistung mit aktuellen Schultestleistungen. Einen Überblick über die Hauptgütekriterien ist in Abbildung 2 zusammengefasst.

Abbildung 2. Hauptgütekriterien im Überblick.



2.5.1.2 Kritische Werte der Hauptgütekriterien

Intelligenztests werden an ihren Testgütekriterien gemessen, diese sollten nach einer entsprechenden Prüfung akzeptabel bis gut oder sehr gut ausfallen. Es stellt sich also die Frage, was akzeptable bzw. nicht akzeptable Testgütekriterien sind. Obwohl Verlage und TestrezensentInnen ein Urteil über die in den Manualen angegebenen Gütekriterien fällen, wird kaum der zu Grunde liegende Maßstab für diese Bewertung erläutert.

Im Online Lehrbuch Medizinische Psychologie der Universität Freiburg (Schaefer, Goos & Goepfert, 2017, ohne Seitenangabe), muss ein guter Test folgende Testgütekriterien erfüllen:

Objektivität: $r =$ annähernd 1

Reliabilität: $r = .70-.95$

Validität: $r = .30-.65$

Diese Angaben stehen stellvertretend für ähnliche Angaben anderer AutorInnen und verlocken dazu, diese Richtwerte als bindend zu betrachten. Doch die statische Betrachtung dieser Richtwerte berücksichtigt nicht die komplexen Bedingungen, die zu den Angaben der Testgütekriterien führen. Es ist möglich, dass ein Test mit einer beobachteten Validität von $r = .30$ valider sein kann als ein Test mit einer Validität von $r = .50$ (Schmidt-Atzert & Amelang, 2012, S. 153).

Grundlage für die Konstruktion der meisten Tests und aller Tests, die in dieser Arbeit genannt werden, sind die Annahmen der Klassischen Testtheorie (KTT), welche von Lord und Novick (1968) zusammenfassend beschrieben worden sind. Grundsätzlich wird angenommen, dass Messfehler die Reliabilität beeinträchtigen und dass das Testergebnis – in diesem Zusammenhang beobachteter Wert genannt – nicht dem wahren Wert entspricht, oder genauer ausgedrückt, entsprechen kann. Da keine perfekte Reliabilität von 1 für einen Test angenommen wird, ist also das Verhältnis zwischen wahren und beobachtetem Wert bedeutungsvoll und somit die Berechnungen, die die notwendige Unterscheidung zwischen wahren und beobachtetem Wert berücksichtigen.

Die weiter oben beschriebenen Richtwerte vermitteln den Eindruck, dass höhere Werte prinzipiell günstiger sind. Da es jedoch unterschiedliche Messmethoden zur Bestimmung der Reliabilitäten gibt, sind die Bewertungen der Reliabilitätskoeffizienten von den Methoden zur Schätzung der Reliabilitäten abhängig (Schmidt-Atzert & Amelang, 2012, S. 137). Bei der Retest-Reliabilität hängt z.B. der Koeffizient von vermuteten Lerneffekten bei einer Testwiederholung ab, aber auch von der Zeit, die zwischen Test und Retest liegt. Dies ist besonders von Bedeutung für *kristalline* Intelligenztests, da die *kristalline* Intelligenz umweltabhängiger als genetisch bedingt ist, wie z.B. bei der *fluiden* Intel-

ligenz. In einer Metaanalyse von Charter (2003) sind die Reliabilitäten für verschiedene Testarten untersucht worden. Für die Gruppe der Intelligenztests wurde ein arithmetisches Mittel von $r = .80$ bestimmt; die meisten Reliabilitätsangaben³³ lagen zwischen $r = .71$ – $.90$. Auch diese Angaben dürfen nicht als Richtwerte interpretiert werden. Es sind lediglich die Retest-Koeffizienten, die aus Sicht der TestautorInnen zur Veröffentlichung der jeweiligen Tests ausreichten. Schuerger und Witt (1989) stellten fest, dass die Retest-Reliabilität auch vom Alter des Kinds abhängt, je älter das Kind, desto höher fallen die Reliabilitäten aus.

Koeffizienten zur Bestimmung der internen Konsistenz können ebenfalls bei der Suche nach dem höchsten Wert in die Irre führen, z. B. würden sich bei einer Berechnung nach Cronbachs Alpha (Cronbach, 1951) die Reliabilitäten mit der Anzahl der Items erhöhen. Selbst wenn die Korrelationen zwischen den Items niedrig sind, wäre bei einer großen Anzahl von Items möglicherweise die Reliabilität akzeptabel.

Auch die Validität ist von mehreren Faktoren abhängig. Schmidt-Atzert & Amelang (2012, S. 153) postulieren, dass die Höhe des Validitätskoeffizienten nur angemessen beurteilt werden kann, wenn die Bedingungen zur Ermittlung der Koeffizienten bekannt sind. Es besteht z. B. ein Zusammenhang zwischen der Reliabilität des Tests und dessen Validität, auch kann die Reliabilität des Kriteriums (an dem der Test validiert wird) die Validität beeinflussen (ebd., S. 154).

Zumindest als Referenz dienen nach Schmidt-Atzert & Amelang (ebd., S. 164) die Angaben einer Metaanalyse, welche den Zusammenhang zwischen Intelligenztests und Validitätskriterien untersuchten (Schmidt & Hunter, 1998; Salgado, Anderson, Moscoso, Bertua, de Fruyt & Rolland, 2003; Kramer, 2009; Deary, Strand, Smith & Fernandes, 2007). Die korrigierten³⁴ Werte sind für den Berufserfolg angegeben mit $.51$ – $.62$., für den Ausbildungserfolg $.53$ – $.59$, für das Bildungsniveau $.56$ und für den Schulerfolg $.69$.

In den Niederlanden gibt es bereits seit der Jahrtausendwende das Testbeurteilungssystem COTAN (Committee On Test Affairs Netherlands) (Evers, 2001a), welche u. a. Richtlinien zur Bewertung der Reliabilität eines Tests vorschlagen. Für Tests für wichtige Entscheidungen auf der individuellen Ebene – dazu dürften die Intelligenztests gehören – wird eine Reliabilität von $.80$ – $.90$ mit ausreichend, von über $.90$ mit gut, Reliabilitäten von unter $.80$ werden als unzureichend beschrieben (Evers, 2001b). Evers (ebd.) sieht nach Einführung des COTAN-Testbeurteilungssystems eine kontinuierliche Verbesserung bei

33 25 Prozent lagen unter, 25 Prozent über diesen Angaben.

34 Werte sind für Varianzeinschränkungen sowie für die Reliabilität von Test und Kriterium bzw. für Reliabilität von Prädiktor und Kriterium korrigiert (Schmidt-Atzert & Amelang, 2007, S. 164, Tabelle 2.24).

der Testkonstruktion, neuralgische Punkte der Testkonstruktionen sind die Bereiche Normen und Kriteriumsvalidität.

Problematisch an solch starren Beurteilungssystemen kann das Bemühen der TestautorInnen angenommen werden, nicht unter die kritischen Grenzen zu rutschen, z. B. eine Reliabilität von unter .80 zu berechnen. Obwohl die AutorInnen der Testbeurteilungssysteme einräumen, dass es für diese Grenzen vergleichbar mit Cut-Off Werten keine schlüssigen Begründungen gibt (Kersting, 2006), führt ein derart starres System dazu, dass „der Koeffizient in ‚gut‘ und ‚böse‘ ohne Verstand, aber mit dem Taschenrechner gehorsam abgehakt, befolgt und verfolgt werde“ (ebd., S. 248). Die Fixierung auf die Einhaltung der Kennwerte könnte dazu führen, dass wider besseren Wissens Untersuchungspläne so gestaltet werden, dass nicht die Wahrscheinlichkeit für Erkenntnisse, sondern die Wahrscheinlichkeit für hohe Koeffizienten im Vordergrund stehen (ebd., S. 248).

Das Testbeurteilungssystem (TBS-TK) des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologinnenvereinigungen (ZPID) erhebt den Anspruch, „das Beste aus verschiedenen Welten“ zu vereinen (Kersting, 2006, S. 250). Dazu gehört die Rezension eines Tests nach einem vorgegebenen Raster, die Testbeurteilung durch zwei unabhängige RezensentInnen und eine Orientierung nach dem Deutschen Institut für Normung (DIN) 33430 (Deutsches Institut für Normung, 2002), welche „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen“³⁵ formuliert (Kersting, 2006, S. 249). Eine Beurteilung ausschließlich nach nominellen Vorgaben, z. B. über die Höhe der Koeffizienten, ist in dem Testbeurteilungssystem nicht vorgesehen. Dabei werden die Testgütekriterien qualitativ bewertet, in *voll*, *weitgehend*, *teilweise* und *nicht erfüllt*. Die nach den Vorgaben des Testkuratoriums erstellten und frei zugänglichen Rezensionen werden bei der folgenden Beschreibung der Testverfahren bevorzugt zitiert, da eine nachvollziehbare und objektive Bewertung angenommen wird. Falls Rezensionen nach dem TBS-TK nicht vorliegen, werden freie Rezensionen und Bewertungen der Manuale und Testverlage nachrangig zitiert.

2.5.1.3 Nebengütekriterien

Zur Beurteilung psychologischer Tests werden weitere Gütekriterien beschrieben:

35 DIN 33430 ist generell geeignet, normierte standardisierte Tests zu beurteilen.

- *Skalierung*: Die Skalierung wird laut Bühner (2011, S. 67) eher zu den Hauptgütekriterien gezählt. Die verwendeten Testwerte müssen geeignet sein, unterschiedliche Testergebnisse miteinander gültig vergleichen zu können. Bezogen auf Intelligenztests sollten Abstufungen in den Testergebnissen auch den Abstufungen in den intellektuellen Potentialen der Testpersonen entsprechen.
- *Normierung*: Die Ergebnisse des Kinds müssen mit einem repräsentativen Durchschnitt altersgleicher Kinder verglichen werden können. Eine repräsentative und umfangreiche Normstichprobe vermindert die Gefahr von Verzerrungen beim Abgleich der vom Kind erzielten Rohwerte mit den Normdaten. Die absolute Größe der Normstichprobe ist dabei nicht ausschlaggebend, denn sie muss im Zusammenhang mit der Anzahl der Normtabellen gesehen werden, die meist in Vierteljahresschritten unterteilt sind. Vor allem auf Grund der kontinuierlichen Zunahme der Intelligenz (Flynn-Effekt: Flynn, 1987) entsteht die Notwendigkeit, Intelligenztests mit aktuellen Normdaten zu verwenden (Joél, 2017, S. 16). Moosbrugger und Höfling (2007) schlagen alle acht Jahre eine Überprüfung der Normdaten vor. Schmidt-Atzert & Amelang (2012, S. 164ff.) beschreiben die Normierung als Hauptgütekriterium.
- *Testfairness*: Ein Test sollte keine Kindergruppen benachteiligen, unabhängig vom sprachlichen und kulturellen Hintergrund, also das tatsächliche Potential ermitteln können. Würde z. B. ein geflüchtetes Kind aus Afghanistan nur wenige Wochen nach der Ankunft in Deutschland mit *kristallinen* (und meist sprachgebundenen) Subtests getestet werden, wäre die Benachteiligung wahrscheinlich und die Testergebnisse nicht gültig ermittelt. So könnten Aufgabengruppen und einzelne Items des WISC-IV als ungeeignet für nicht deutschstämmige Kindergruppen diskutiert werden.
- *Ökonomie*: Sowohl die in das Testen investierte Zeit, die mit der Testung für den Testenden verbundenen Belastungen als auch die monetären Kosten (z. B. für den Test und die Testmaterialien) sollten in einem vertretbaren Verhältnis zum Nutzen stehen. Würden SonderpädagogInnen selten Kinder testen oder diagnostisch tätig sein, wäre es sinnvoll, wären die Testverfahren schnell und einfach zu lernen oder wiederzuerlernen. Zu dem Nebengütekriterium der Ökonomie gehört also auch die Praktikabilität der Verfahren. Zeitökonomisch (und auch im Sinne der Auswertungsobjektivität) wäre zudem die Auswertungsmöglichkeit mit einer Software. Der Blick auf eine schnelle Durchführung ist allerdings auch mit Gefahren verbunden. Die Anwendung günstig anzuschaffender und schnell durchzuführender Intelligenztests wie der CFT1-R (118 Euro; Stand: 21.7.17) sind verlockend. Doch sollte die Tragweite einer durchgeführten Intelligenzdiagnostik und der damit auch verbundenen Gefahr einer Stigmatisierung bei der Ermittlung einer schnellen und somit ökonomischen Testung bedacht werden.

- *Nützlichkeit*: Die Ergebnisse eines Tests sollten von Nutzen sein, z.B. eine Fragestellung beantworten können oder eine sinnvolle Ergänzung bei der Erkennung sonderpädagogischen Förderbedarfs darstellen. Gerade im pädagogischen Bereich wäre eine mögliche Ableitung von individuellen Stärken und Schwächen neben dem Vergleich mit einer Altersgruppe wünschenswert. Eher eindimensionale Verfahren sind zwar ökonomischer, erlauben aber kaum Hinweise über die Ermittlung eines Gesamtwerts hinaus (Joél, 2017, S. 18).
- *Zumutbarkeit*: Die Testsituation sollte eine zumutbare Belastung des zu Testenden sein und sollte die aus dem Test resultierende Nützlichkeit nicht übertreffen. Die Belastung könnte z.B. reduziert werden, wenn die Items kindgerecht und spannend gestaltet sind. Kinder mit sonderpädagogischem Unterstützungsbedarf hatten oft Gefühle des Versagens in Leistungssituationen. Kinder könnten bei der Anwendung eines Intelligenztests die Testsituation mit vorherigen Leistungssituationen assoziieren und negativ gestimmt sein, dies würde mit ansprechenden Stimuli vermieden.
- *Vergleichbarkeit*: Vergleichbarkeit ist gegeben, wenn zwei parallele Formen eines Tests oder inhaltlich ähnliche Tests vorliegen. Dies könnte ein Vorteil bei einer Retestung sein.
- *Unverfälschbarkeit*: Testpersonen sollten nicht die Möglichkeit der Manipulation von Testergebnissen haben, jedoch kann dies nicht ausgeschlossen werden, z.B. bei sozial erwünschtem Verhalten bei Persönlichkeitstests. Doch auch bei Leistungstests konnten Ziegler, Schmidt-Atzert, Bühner und Krumm (2007) Manipulationsmöglichkeiten während der Testung nachweisen. Dennoch kann angenommen werden, dass bei standardisierten Intelligenztests die Möglichkeiten der Verfälschbarkeit generell gering sind.
- *Transparenz*: Die Testanweisungen sollten für das Kind verständlich sein. Häufig sind Kinder im sonderpädagogischen Kontext kognitiv schwach und benötigen einfache und intuitiv zu verstehende Einführungsaufgaben. Günstig wäre die Möglichkeit, dem Kind die Anforderungen an die Subtests frei erläutern zu dürfen, im günstigsten Fall mit nicht bewerteten Anfangsitems (Joél, 2017, S. 17). Umständliche Anweisungen, verpackt in komplizierte Schachtelsätze wären ungünstig. Testverfahren, die eine wortwörtliche Instruktionvorgabe durch Ablesen der Anweisungen vorschreiben, sind weniger geeignet bei kognitiv schwachen Kindern.

2.5.2 Beschreibung der Testverfahren

Eine Auswahl häufig im sonderpädagogischen Kontext eingesetzter Intelligenztests wird skizziert. Dabei sollen neben einer kurzen Vorstellung des jeweiligen Tests der theoretische Hintergrund, die *Dimensionalität* sowie kritische Ein-

wände erwähnt sein. Eine Übersicht stellt im Anschluss die Testgütekriterien vor (Tabellen 4 und 5), eine weitere Übersicht Erwägungen zur praktischen Anwendung der Tests im sonderpädagogischen Kontext (Tabelle 6). Eine abschließende tabellarische Auflistung nennt Preise, Altersbereich und Hinweise für die Nützlichkeit ausgesuchter sonderpädagogischer Fragestellungen (Tabellen 7 und 8).

Die Auswahl der Intelligenztests entspricht einer subjektiven Einschätzung über die Häufigkeit des Einsatzes der Verfahren in der Sonderpädagogik, entspricht aber auch den persönlichen Kenntnissen des Autoren. Eine Wertung ist mit dieser Auswahl nicht verbunden. Da sich der Fragebogen der Untersuchung ebenfalls maßgeblich auf diese Auswahl stützt und andere Testverfahren unter *Sonstige* subsummiert werden (mit der Möglichkeit, sonstige Tests von den ProbandInnen auch zu benennen), sei an dieser Stelle darauf hingewiesen, dass Verfahren wie AID-3 (Kubinger & Holocher-Ertl, 2014) oder CPM (Raven, Raven & Court, 2010) und weitere weder in ihrer Bedeutung noch Gültigkeit bewertet werden, auch wenn sie nicht ausdrücklich aufgeführt sind. So listet allein der Hogrefe Verlag 32 Intelligenztests für den Kinder- und Jugendlichenbereich auf (Hogrefe, 2017). Der Anspruch, alle Intelligenztests umfassend zu würdigen, wird nicht erhoben. Die kurze Übersicht der Testverfahren ersetzt nicht ausführliche Rezensionen, sondern soll zum besseren Verständnis einen Überblick über die Tests vermitteln.

2.5.2.1 K-ABC (Kaufman Assessment Battery for Children)

Überblick: 1983 erschien die von dem Ehepaar Kaufman konstruierte erste Fassung der mehrdimensionalen K-ABC (Kaufman & Kaufman, 2004), 1991 erschien die für den deutschsprachigen Raum adaptierte Fassung von Melchers und Preuß (2009). Die Gesamtskala *intellektueller Fähigkeiten* (eher *fluide* bzw. Grundintelligenz) unterscheidet zwischen *einzelheitlichem Denken* bzw. *sequentieller Informationsverarbeitung* und *ganzheitlichem Denken* bzw. *simultaner Informationsverarbeitung* (Lurija, 1970) und versteht Intelligenz als einen Prozess der Verarbeitung von Reizen. Lern- und Faktenwissen hingegen wird mit der *Fertigkeitenskala* erfasst und könnte mit der Testung der *kristallinen* Intelligenz verglichen werden. Im Gegensatz zum WISC-IV wird das Ergebnis der *kristallinen* Intelligenz nicht in einen Gesamtwert integriert, sondern gesondert berechnet. Somit war es möglich, die Ergebnisse der eher umgebungsabhängigen *kristallinen* Intelligenz von den Ergebnissen der eher genetisch bedingten *fluiden* Intelligenz getrennt zu interpretieren. Darüber hinaus war es differentialdiagnostisch möglich, Stärken- und Schwächenanalysen vorzunehmen.

Innovativ und vorteilhaft beim Erscheinen der K-ABC war das anwendungsfreundliche Arbeiten mit Stellordnern. In der Sonderpädagogik war die

K-ABC ein Mittel der Wahl³⁶, da es viele Items pro Subtest gab und es kognitiv schwachen Kindern durch das freie Erklären über Lernaufgaben ermöglichte, die Aufgabenstellungen besser zu verstehen. Inzwischen ist die K-ABC durch die KABC-II abgelöst und sollte sowohl auf Grund der veralteten Normstichprobe und dem damit verbundenen Flynn-Effekt, aber auch auf Grund schlechter und veralteter Bildmaterialien nicht mehr angewendet werden.

Testgütekriterien und Kritik: In einer TBS-TK-Testrezension von Rollett und Preckel (2011) wird der K-ABC Objektivität *voll*, Zuverlässigkeit und Validität jeweils *weitgehend* attestiert. Hervorgehoben wird die diagnostisch wertvolle Möglichkeit der Profilinterpretation, Hauptproblem sei die „Nichtberücksichtigung aktueller Untersuchungen“ (ebd., S. 140).

Anhand einer nicht repräsentativen Stichprobenuntersuchung untersuchten Renner, Schmid, Irblich und Krampen (2012) die psychometrischen Eigenschaften bei fünf- und sechsjährigen Kindern. Die Reliabilitätskoeffizienten der Skalen lagen zwischen .86 bis .92, (ebd., S. 196), die Stabilität der Skalen lag bei .60 bis .79. Für den Subtest *Bildhaftes Ergänzen* ist die niedrigste Retest-Stabilität und die niedrigste interne Konsistenz ermittelt worden, vor allem für diesen Subtest (aber auch für andere) sollten Befunde durch ergänzende Verfahren abgesichert werden (ebd., S. 203).

2.5.2.2 KABC-II (Kaufman Assessment Battery for Children – II)

Überblick: Die überarbeitete Version der K-ABC unterscheidet sich grundlegend von der vorherigen Version sowohl in der Durchführung, im theoretischen Modell als auch in der Gestaltung der Subtests. Acht der Subtests sind bekannt, zehn neu hinzugekommen. Von möglichen 18 Subtests wird eine Auswahl (in der Regel 8–10 Subtests, je nach Alter, gewähltem Intelligenzmodell, Fragestellung und zeitökonomischen Vorgaben) von Subtests durchgeführt, dessen Ergebnisse in 3–5 übergeordneten *Indices* münden. Die Analyse der Testergebnisse der *Indices* ermöglicht eine Interpretation von individuellen Stärken und Schwächen des Kinds. Innovativ ist die Möglichkeit, aus zwei theoretischen Modellen zu wählen. Eines orientiert sich wie in der K-ABC an dem Lurija-Modell, das andere an dem CHC-Modell der Intelligenz. Nach wie vor wird bei dem Modell nach Lurija die umgebungsabhängige *kristalline* Intelligenz bei der Berechnung eines Generalfaktors nicht einbezogen. Dies ist häufig in der Sonderpädagogik sinnvoll, z.B. bei Kindern mit Sprachproblematiken, mit Migrationshintergrund oder bei geflüchteten Kindern. Da in einer leis-

36 Die K-ABC ist deshalb auch Gegenstand dieser Untersuchung, auch wenn sie aktuell kaum noch durchgeführt wird.

tungsorientierten Gesellschaft wie Deutschland die Chance auf Bildung ungleich verteilt ist und zuweilen vom finanziellen und Bildungshintergrund der Familiensysteme abhängt, kann der Einbezug der *kristallinen* Intelligenz auch politisch kontrovers diskutiert werden. Obwohl die Qualität der 2015 erschienenen KABC-II (Melchers & Melchers, 2015) gelobt wird, erweist sich der Test auf Grund der vielen Regeln (ca. 580 Regeln allein in der Testsituation; Joél, 2017) bei einer seltenen Anwendung als unpraktikabel. Andererseits ist der Test gut geeignet, Kinder mit sonderpädagogischem Unterstützungsbedarf zu prüfen, denn die Items sind sehr kindgerecht gestaltet und es besteht die Möglichkeit des freien Erläuterns der Aufgabenstellungen und Bodeneffekte sind kaum vorhanden. Mit Hilfe ausgewählter Subtests kann über die nonverbale Darbietung der Stimuli der *Sprachfrei-Index* ermittelt werden.

Testgütekriterien und Kritik: In einer TBS-TK Rezension von Kuschel, Kamp-Becker und Ständer (2017) wird der KABC-II lediglich *weitgehend* Objektivität, Zuverlässigkeit und Validität bescheinigt. Eine Gefährdung der Objektivität wird zudem durch die *Komplexität* des Materials und der Verwendung von zwei theoretischen Modellen angenommen (ebd., S. 211).

In ausführlichen Testrezensionen äußern sich Renner (2015) und Irblich (2015) überwiegend positiv. Die wichtigsten Kritikpunkte sind zusammengefasst (Joél, 2017):

- Keine Angaben zur nonverbalen Anleitung der Subtests des *Sprachfrei-Index* (Renner/Irblich),
- Übungseffekte bei zu rascher Testwiederholung (Renner/Irblich),
- keine Hinweise zum Normierungsvorgehen (Renner/Irblich),
- Kinder aus Familien mit niedrigem Bildungsabschluss sind unterrepräsentiert in der Normierungsstichprobe (Renner/Irblich),
- leichte Bodeneffekte in einigen Subtests, vor allem bei jungen Kindern (Renner/Irblich),
- leichte Deckeneffekte in Ergänzungs-subtests für Kinder ab 13 Jahren (Irblich),
- nicht geeignet für den Förderschwerpunkt *Sehen* (Renner/Irblich),
- in Einzelfällen nicht nachvollziehbare Lösungsvorgaben (Renner),
- trotz Anweisungen, sich nicht zu beeilen, kann schnelles Arbeiten das Ergebnis erheblich beeinflussen (Renner/Irblich),
- Konstruktionsmängel beim Subtest *Symbole* (Renner/Irblich),
- deutlich mangelhafte Materialqualität (Renner/Irblich),
- Reliabilitätsberechnung auf Untertestebene unklar beschrieben (Irblich),
- *Indice Planung/Gf* eher aus statistischen als aus inhaltlichen Gründen zusammengestellt (Irblich),
- mangelnde Erläuterungen von Faktorenanalysen, dadurch sind Rückschlüsse auf Ladungen schwer möglich (Irblich),

- fehlende Angaben zur Kriteriumsvalidität und prognostischen Validität (Irblich),
- überflüssige Hintergrundinformationen auf den Instruktionseiten (Irblich),
- *Indice Wissen/Gc* bildet *kristalline* Intelligenz nur bedingt ab (Irblich),
- postulierte Kulturfairness für den Gesamttest fragwürdig (Irblich).

2.5.2.3 CFT1-R (Grundintelligenztest Skala 1)

Überblick: Der CFT1-R (Weiss & Osterland, 2013) ist ein kurzer, auch in der Gruppe durchführbarer Intelligenztest. Drei Subtests (1. Teil) messen die *wahrnehmungsgesundene Leistung*, drei weitere Subtests (2. Teil) das *figurale* und *regelmäßige Denken*. Der ermittelte Gesamtwert misst eindimensional die *fluide* Intelligenz (Renner & Mickley, 2015b, S. 72). Der Test ist kindgerecht, zügig zu erlernen und günstig in der Anschaffung, ist aber zu kurz, um einen aussagekräftigen Hinweis auf alle Bereiche der Intelligenz zu ermitteln. Der CFT1-R steht in der Tradition der culture-fair Testung, der Testung von Kindern ohne starke kulturelle und sprachliche Bezüge. Negativ fallen die versteckten und verwirrend beschriebenen Hinweise zu den Standardabweichungen auf.

Testgütekriterien und Kritik: Renner (2014) empfiehlt entsprechend den Empfehlungen des Manuals ebenfalls nicht die Testung von intelligenzgeminderten Kindern und hebt positiv eine eigene Normstichprobe für die Kinder mit dem Unterstützungsbedarf *Lernen* hervor. Die Durchführungsobjektivität sieht Renner gefährdet durch unzureichende Instruktionen, die Auswertungs- und Interpretationsobjektivität hingegen als gesichert. Die Reliabilitäten für den Gesamtwert fallen *gut* bis *sehr gut*, für die Testteile *befriedigend* bis *gut* aus. Die inhaltliche Validität wird als *gesichert* eingeschätzt, das Konstrukt Intelligenz wird nach Renner (2014) jedoch nur eingeschränkt repräsentiert.

2.5.2.4 CFT20-R (Grundintelligenztest Skala 2 – Revision mit Wortschatztest und Zahlenfolgentest)

Überblick: Ebenfalls weitgehend kulturfair wird der CFT20-R (Weiss, 2006) für ältere Kinder bzw. Jugendliche und Erwachsene durchgeführt; das Konzept ähnelt dem CFT1-R. Die postulierte Sprachfairness wird eingeschränkt durch die Vorgabe, Instruktionen wörtlich vorzutragen. Diese Vorgabe lässt wenig Spielraum bei der Erläuterung der Subtests, sollte ein Kind nicht gut oder nicht Deutsch verstehen. Die Durchführung der Subtests verläuft hingegen ohne Anwendung der Sprache. Zwei Ergänzungstests ermöglichen neben der *fluiden* auch die Ermittlung der *kristallinen* Intelligenz. Der Test besteht aus zwei Durchgängen mit jeweils vier sich ähnelnden Subtests. Besonders Subtest 4

bzw. 8 sind kognitiv schwachen Kindern schwer zu vermitteln (die Position eines kleinen schwarzen Punkts soll einer Auswahl unter bestimmten Vorgaben zugeordnet werden).

Testgütekriterien und Kritik: In einer TBS-TK-Rezension von Gruber und Tausch (2015) bewerten diese Objektivität, Zuverlässigkeit und Validität mit *weitgehend* erfüllt und kritisieren das Fehlen von gültigen Normen für Erwachsene. Positiv hervorgehoben wird die einfache Handhabbarkeit, die genauen Instruktionshinweise und die Ökonomie des Tests.

2.5.2.5 WISC-IV (Wechsler Intelligence Scale for Children (Deutsche Ausgabe) – fourth Edition, ehemals HAWIK-IV)

Überblick: Die Tests aus der Wechsler-Reihe hatten bis in die 90er Jahre eine große Bedeutung in der Sonderpädagogik. Entsprechende Suchanfragen im Register der Zeitschrift für Heilpädagogik erzielten bei den Stichworten HAWIK bzw. WISC bis 1998 21 Treffer, danach 1 Treffer. Die aktuelle Version WISC-IV besteht aus 15 Subtests, von denen mindestens 10 (in Ausnahmefällen 8) Subtests zur Erzielung eines Generalfaktors durchgeführt werden müssen. Wechsler beschreibt Intelligenz als die zusammengesetzte oder globale Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umwelt wirkungsvoll auseinanderzusetzen (Wechsler, 1956, S. 13). Diese Definition sei hier aufgeführt, da sie Grundlage der Wechsler-Testreihe ist, zu der auch die weiter unten beschriebenen Verfahren WNV und WPPSI-III gehören.

Vier *Indices* ermitteln differenziert das *Arbeitsgedächtnis* (AG), das *Sprachverständnis* (SV), *Verarbeitungsgeschwindigkeit* und das *Wahrnehmungsgebundene Logische Denken* (WLD). Die Abgrenzung der *Indices* zueinander ist ungenau. So korreliert das *Wahrnehmungsgebundene Logische Denken* (in Anlehnung an das CHC-Modell) sowohl mit der *fluiden*, der *kristallinen* und der *visuellen* Intelligenz. Daseking, Petermann und Waldmann (2008) schlagen als Ergänzung zur Interpretation eines Gesamtwerts die Betrachtung eines *allgemeinen Fähigkeitsindex* (AFI) vor, der sich aus den *Indices* SV und WLD zusammensetzt.

Die drei Subtests des *Indice Sprachverständnis* testen überwiegend akademisches Wissen und sind an die deutsche Sprache gebunden. Dieser eher *kristalline* Bereich ist problematisch bei Kindern mit Sprachschwierigkeiten, Kindern aus einem bildungsfernen Milieu und Kindern mit Migrationshintergrund. Ca. 30 Prozent des Gesamtwerts werden also aus der umgebungsabhängigen *kristallinen* Intelligenz gebildet.

Testgütekriterien und Kritik: Deimann und Kastner-Koller (2008) kritisieren die unzureichende Adaption des ursprünglich US-amerikanischen Verfahrens in deutsche Verhältnisse, unklare Erläuterungen (z. B. zur Substitution von Subtests) und eine zu kleine Normstichprobe.

In einer TBS-TK-Rezension von Schmukle & Schulze (2016), wird die Reliabilität mit *voll*, die Objektivität hingegen mit *weitgehend* und die Validität nur mit *teilweise* als erfüllt eingeschätzt. Die *eingeschränkte Validitätsevidenz* (ebd., S. 160) lässt die Frage offen, warum die WISC-IV über eine so hohe Akzeptanz verfügt und so häufig eingesetzt wird. Die Ende 2017 erschienene WISC-V bestätigt nachträglich Konstruktionsschwächen, denn nun werden entsprechend dem CHC-Modell gleich konstruierte Subtests anderen *Indices* zugeschrieben.

2.5.2.6 WPPSI-III (Wechsler Preschool and Primary Scale of Intelligence – III Deutsche Version)

Überblick: Der WPPSI-III (Petermann, 2009) für Kinder bis 7;2 Jahre ist dem WISC-IV (für Kinder von 6 bis 16 Jahren) ähnlich und resultiert aus dem HAWIVA-III (Ricken, Fritz, Schuck & Preuß, 2007). Neben dem Gesamtwert können vier weitere Werte berechnet werden, so dass der Test als mehrdimensional gewertet werden kann: *Verbalteil*, *Handlungsteil*, *Verarbeitungsgeschwindigkeit (VG)* und *Allgemeine Sprachskala*. Der *Verbalteil* misst wie das *Sprachverständnis* des WISC-IV eher *kristalline* Intelligenz und ist für einige Kindergruppen von Nachteil, die ungünstige Bildungsbedingungen haben. Die Bildmaterialien sind ansprechend und motivieren die Kinder zur Mitarbeit. Je nach Alter werden zur Ermittlung eines Gesamtwerts mindestens bis zu sieben Subtests durchgeführt, insgesamt stehen jedoch 14 Subtests zur Verfügung, die teils optional entsprechend der Fragestellungen gewählt werden können.

Testgütekriterien und Kritik: Renner (2010) kritisiert ergonomisch ungünstige Materialien, Abbruchkriterien, die eine mangelnde Compliance nach sich ziehen können und teils schwer nachvollziehbare Bewertungsrichtlinien, bescheinigt jedoch eine reliable Intelligenzmessung. Leider beziehen sich die Angaben zur Reliabilität teils auf die (weitgehend identische) Vorgängerversion HAWIVA-III. Die Durchführungsobjektivität wird als *gesichert* bescheinigt, die Auswertungsobjektivität scheint Renner fraglich. Eine Bewertung der Validität wird nicht vorgenommen, es wird allerdings empfohlen, diese durch weitere Daten zu belegen (ebd., S. 182). Irblich (2010) attestiert ebenfalls Mängel in der Auswertungsobjektivität auf Grund fehlender Angaben zur Bewertung, attestiert *befriedigende* bis *sehr gute* Reliabilitätskennwerte und bewertet die Inhaltsvalidität als *gesichert*, die Angaben zur Konstruktvalidität werden als *weitgehend zufriedenstellend* beschrieben (ebd., S. 324).

Sattler und Dumont (2004) empfehlen den WPPSI-III bei der Testung von dreijährigen Kindern eher als Screeningverfahren³⁷.

37 Bezugnehmend auf die Originalversion.

2.5.2.7 WNV (Wechsler Nonverbal Scale of Ability)

Übersicht: Es gibt nur wenige Verfahren, die tatsächlich Kinder nonverbal testen können. Im Grunde sind dies neben dem SON-R 2½–7 bzw. SON-R 6–40 die WNV (Wechsler & Naglieri, 2006; Petermann, 2014) und der *Sprachfrei-Index* der KABC-II. Ein wirklich nonverbaler Intelligenztest hat Praxisrelevanz, da kulturelle und sprachliche Hintergründe der Kinder keine wesentliche Rolle mehr spielen und die tatsächliche Intelligenz erfasst wird.

Je nach Alter werden vier der insgesamt sechs Subtests für die Berechnung des Gesamtwerts (in IQ) durchgeführt. Neben dem Gesamtwert und den vier Einzelergebnissen gibt es keine übergeordneten *Indices*, dennoch wird beansprucht, mit der WNV mehrdimensional testen zu können. Eine sinnvolle Ableitung von Stärken und Schwächen ist jedoch kaum möglich, dementsprechend gibt es auch keine Hinweise auf evtl. abzuleitende Fördermaßnahmen. Bei Bedarf können auf Seite zwei des Testformulars zwar Stärken und Schwächen berechnet werden, die sich allerdings lediglich auf Stärken und Schwächen im Vergleich zwischen den Subtests beziehen. Eine übergeordnete Ableitung (z. B. *auditive* oder *visuelle Merkschwächen*, *Gedächtnisschwierigkeiten*, *visuelle Verarbeitungsschwierigkeiten*, *Schwierigkeiten im Langzeitgedächtnis* usw.) sollten von einer Testdurchführung, die aus vier Subtests besteht, nicht erwartet werden. Bei der Annahme, ein CHC Faktor auf *Stratum-II*-Ebene wäre angemessen repräsentiert, sollten zwei dazugehörige enge Fähigkeiten auf *Stratum-III*-Ebene vorhanden sein (Renner & Mickley, 2015b, S. 72), wäre lediglich die breite Fähigkeit *Gv* (*visuelle Verarbeitung*) angemessen repräsentiert (siehe Tabelle 3).

Tabelle 3. WNV: Zuordnung enger Fähigkeiten zu den breiten (CHC-)Fähigkeiten.

	Gf (fluide Intelligenz)	Gs (Verarbeitungs- geschwindigkeit)	Gv (visuelle Verarbeitung)	Gsm (Kurzzeit- gedächtnis)	sonstige
Langversion (4 Subtests) jüngere Kinder:	o	o	+	–	–
Kurzversion (2 Subtests) ältere Kinder:	o	o	o	o	–
Langversion jüngere Kinder:	o	o	–	–	–
Kurzversion ältere Kinder:	o	–	o	–	–

Anmerkungen. Erfassung breiter CHC-Faktoren mit der WNV, angelehnt an Renner & Mickley (2015b): „–“ = nicht repräsentiert, „o“ = mit einem Subtest repräsentiert, „+“ = mit mindestens zwei Subtests angemessen repräsentiert.

Es ist auch möglich, die bereits kurze Testdurchführung noch einmal zu kürzen und lediglich zwei Subtests durchzuführen, dennoch aber einen Gesamtwert (IQ) zu ermitteln. Dies kann sinnvoll sein, wenn unter Zeitdruck getestet wird, das Kind evtl. nur schwer zu motivieren ist oder die WNV eine Zweitmeinung darstellt neben der Durchführung einer komplexeren Testbatterie.

Neben der rein nonverbalen Testung kann die WNV auch verbal durchgeführt werden. Es liegen Übersetzungen der Anweisungen in anderen Sprachen vor (Türkisch, Russisch, Spanisch, Arabisch). Es ist möglich, in der nonverbalen Durchführung neben dem Einsetzen von Gesten dem Kind ein Comic vorzulegen, auf dem die Arbeitsanweisungen in Bildern zu sehen sind. Für jeden Subtest gibt es eine Comicvorlage.

Für drei der sechs Subtests gibt es keine Zeitbegrenzungen für die Kinder, dies ist z. B. von Vorteil für Kinder mit körperlichen Einschränkungen.

Testgütekriterien und Kritik: Schroth (2015) beurteilt die Durchführungs-, Auswertungs- und Interpretationsobjektivität als *gesichert*, ebenso die Inhaltsvalidität. Die Konstruktvalidität wartet teils mit relativ hohen Werten auf. Die Reliabilitätsangaben werden mit *gut* bis *sehr gut* bewertet. Die Mehrdimensionalität wird nicht in Frage gestellt.

Mickley (2015) kritisiert unpräzise nonverbale Instruktionvorgaben und fehlende Hinweise zur Interpretation von möglichen Zusatzanalysen. Bis auf einige fehlende Instruktionsangaben wird die Durchführungs-, Auswertungs- und Interpretationsobjektivität als *ausreichend* beschrieben. Reliabilitätskoeffizienten werden mit *befriedigend* bis *gut* klassifiziert und münden in der Warnung, prognostische Entscheidungen auf der Basis des WNV mit Vorsicht zu formulieren (ebd., S. 111), da die Retest-Reliabilitäten auf kleinere Untersuchungsgruppen basieren. Die Inhaltsvalidität wird attestiert.

Die Mehrdimensionalität wird durch die begrenzte Zahl der Subtests nur eingeschränkt erfasst (ebd., S. 111), Daten zur faktoriellen Validität des Technischen Manuals deuten eher auf ein einfaktorielles Modell.

2.5.2.8 SON-R 2½-7 (Non-verbaler Intelligenztest)

Überblick: Die Snijders-Oomen Testverfahren zeichnen sich durch die nonverbale Durchführungsmöglichkeit aus, so auch der SON-R 2½-7³⁸. Ursprünglich

38 Da der Autor im Gegensatz zu den anderen Verfahren über wenig Kenntnisse zum SON-R 2½-7 verfügt, wird der SON-R 2½-7 nicht in jede Auswertung bzw. Bewertung mit einbezogen. Aus Gründen der Vollständigkeit wird der SON-R 2½-7 allerdings im Fragebogen gewürdigt, da er durchaus im sonderpädagogischen Kontext genutzt wird. Hiermit sei also begründet, dass in den folgenden Kapiteln der SON-R 2½-7 teilweise weniger Erwähnung findet.

sollte eine Intelligenzdiagnostik auch mit gehörlosen Kindern ermöglicht werden. Aus der ersten Version 1943 entwickelten sich Verfahren, die sowohl verbal als auch nonverbal Intelligenz messen.

Der zu testende Altersbereich ist jeweils im Testnamen enthalten. Der SON-R 2½–7 testet die allgemeine Intelligenz mit Hilfe von 6 Subtests. Die 14 bis 17 Items je Subtest werden adaptiv durchgeführt, d. h., die Leistung des Kinds bestimmt den weiteren Verlauf der Testung und somit den Schweregrad der Items. Das adaptive System hat den Vorteil, dass auf komplizierte Regeln wie die Umkehrregel verzichtet werden kann, deshalb ist das Erlernen der SON-Tests recht einfach (und somit das Wiedererlernen bei einer seltenen Anwendung). Je zwei Tests geben Hinweise auf das *abstrakte Denken*, das *konkrete Denken* und das *räumliche Vorstellungsvermögen*, so dass moderate differentialdiagnostische Aussagen möglich sind. Im Gegensatz zu den anderen SON-Versionen wird bei diesem Verfahren neben einer Rückmeldung auch eine Korrektur gegeben, dies ermöglicht die Beobachtung von Lernprozessen während der Testung. Je drei Untertests werden dem Bereich *Denkskala*, drei Untertests dem Bereich *Handlungsskala* zugeordnet.

Testgütekriterien und Kritik: Naescher (2009) beschreibt die Durchführungs-, Auswertungs- und Interpretationsobjektivität mit *weitgehend* vorhanden, die interne Konsistenz auf Subtestebene wird sinngemäß mit *mittel* eingestuft. Insgesamt stimmt die Autorin aber dem niederländischen Bewertungssystem COTAN (siehe Evers, 2001a) zu, der den SON-R 2½–7 mit der Bestnote *gut* auszeichnet.

Das Karg Fachportal Hochbegabung (2017) kritisiert eine demnächst veraltete Normstichprobe und verwirrende oder fehlende Angaben zu den Testgütekriterien.

2.5.2.9 SON-R 5½–17 (Non-verbaler Intelligenztest)

Überblick: Obwohl der Test veraltete und lediglich im Ausland erhobene Normdaten erhält, zählt er noch zu den wichtigeren Tests in der Sonderpädagogik. Dies ist weniger mit der Qualität des Tests begründet, sondern mehr mit der Häufigkeit seiner Anwendung in der Sonderpädagogik. Bestehend aus 7 Subtests (in einer Kurzversion 4 Subtests) kann ein Hinweis auf das intellektuelle Potential über einen Gesamtwert ermittelt werden. Moderate Hinweise auf Schwächen und Stärken können über die Auswertung übergeordneter Gruppen vorgenommen werden: *konkretes Denken*, *räumliches Vorstellungsvermögen*, *abstraktes Denken* und *Perzeption*.

Nach jedem Item erhält das Kind eine Rückmeldung, ob das Item richtig oder falsch gelöst wurde, damit gegebenenfalls die Problemlösungsstrategie überdacht werden kann. Die Anwendung ist einfach zu lernen und wird adaptiv durchgeführt.

Testgütekriterien und Kritik: In einer Rezension von Wolf (1999) wird die sorgfältige Normierung honoriert, jedoch darauf hingewiesen, dass die Normierung lediglich für die Niederlande mangels einer deutschen Normstichprobe repräsentativ ist. Die Reliabilität gilt als *gegeben*, „zahlreiche Studien zur Validität“ (ebd., o. S.) belegen den Zusammenhang mit Schulvariablen. Die Anwendung des Verfahrens wird von Wolf empfohlen.

In einer Studie zur Testung von geflüchteten Kindern in der Sonderpädagogik von 2017 wurde festgestellt, dass dieser veraltete und vor allem auf Grund des Flynn-Effekts ungeeignete Test am vierthäufigsten durchgeführt wurde³⁹ (Joel, 2018, S. 197).

2.5.2.10 SON-R 6–40 (Non-verbaler Intelligenztest)

Überblick: Relativ identisch (z. T. mit gleichen Materialien) sind vier der sieben Subtests des SON-R 5½–17 übernommen worden. Die verbliebenen vier Subtests enthalten jetzt noch weniger *kristalline* Anteile, z. B. *soziale Situationen*, auch wenn diese vorher nur wenig vorhanden waren. Die Umstellung dürfte den AnwenderInnen leicht fallen, da sich so gut wie keine Regeln geändert haben. Lediglich das adaptive Vorgehen hat sich geändert. Neben einem Gesamtwert können Stärken- und Schwächenanalysen so gut wie nicht vorgenommen werden, zwei Subtests werden dem *räumlichen Vorstellungsvermögen*, zwei dem *abstrakten Denken*⁴⁰ zugeordnet, der Generalfaktor ermittelt die *fluide* Intelligenz. Wie bei allen Tests der SON-Reihe können Kinder auch nonverbal getestet werden, zudem Erwachsene bis zum Alter von 40 Jahren.

Testgütekriterien und Kritik: Schroth (2013) bezweifelt die Kulturfreiheit des Subtests *Kategorien* und eine Fehleranfälligkeit bei einer manuellen Auswertung, betont aber die *sehr guten* Testgütekriterien. Durchführungs- und Auswertungsobjektivität sowie die Interpretationsobjektivität werden attestiert, besonders durch die Nutzung der Computerauswertung. Die Kennwerte zur Reliabilität sind hoch, die Validität wird vor allem mit hohen Korrelationen zu anderen Intelligenztests begründet.

Trotz der guten Testgütekriterien darf kritisch angemerkt werden, dass ein letztlich aus vier Subtests bestehender eindimensionaler Intelligenztest, aus dem sich lediglich ein Hinweis auf das intellektuelle Potential ergibt, mit einem Komplettpreis von 2 184 Euro⁴¹ erstaunlich hoch ist.

39 Von SonderpädagogInnen, die einmal ein geflüchtetes Kind getestet haben.

40 Hinweise zur Interpretation oder zum theoretischen Aufbau des *räumlichen Vorstellungsvermögens* und zum abstrakten Denken fehlen allerdings.

41 Stand 26.5.18.

2.5.2.11 IDS (Intelligence and Development Scales)

Überblick: Grundlage war der Versuch einer Weiterentwicklung des Kramer-Intelligenztests (Kramer, 1972), welcher wiederum in der Tradition des Binet-Tests stand, im Grunde ist allerdings ein völlig neu konzeptionierter Test konstruiert worden, der nicht nur die Intelligenz, sondern auch verschiedene Entwicklungsbereiche erfassen soll: *Psychomotorik, Sozial-emotionale Kompetenz, Mathematik, Sprache* und *Leistungsmotivation*. Jeder dieser Bereiche kann entsprechend der Fragestellungen als eigenständiges Modul durchgeführt und kombiniert werden. Der Test ist sehr spielerisch gestaltet und bei den Kindern beliebt.

Die sieben Subtests des Intelligenzteils sollen die Bereiche *Wahrnehmung, Aufmerksamkeit, Gedächtnis* und *Denken* erfassen und im Schwierigkeitsgrad aufeinander aufbauen, insgesamt die *Fluide Mechanik* testen, an dessen Ende ein Gesamtwert steht. Teilweise stehen nur wenige Items zur Verfügung, im Subtest *Wahrnehmung Visuell* z.B. sieben Items. Dies hat zur Folge, dass die Aufgaben für kognitiv schwache Kinder relativ schnell zu schwer werden und wenig differenzierende Aussagen möglich sind. Für intelligenzgeminderte Kinder ist dieser Test deshalb nicht geeignet. Für die vier postulierten Bereiche besteht nicht die Möglichkeit, diese gesondert auszuwerten. Deshalb ist über die Bestimmung eines Generalfaktors in Form eines IQ keine Differentialdiagnose möglich. Die IDS ist sehr kindgerecht gestaltet und bringt den Kindern häufig mehr Spaß als andere Intelligenztests. Dies ist von Bedeutung bei Kindern mit häufig erlebten Frustrationsmomenten in Leistungssituationen. Kindgerecht gestaltete Tests wie die IDS erhöhen gerade bei diesen Kindergruppen die *Compliance*.

Testgütekriterien und Kritik: Naescher (2010) bewertet die IDS als klar strukturiertes, reliables und valides Verfahren und äußert sich insgesamt positiv. In einer ungewöhnlich scharfen Kritik bezweifeln Koch, Kastner-Koller und Deimann (2011) die Möglichkeit, das gesamte Spektrum der Entwicklungs- und Leistungsdiagnostik im pädagogischen und klinischen Bereich mit einem Test abdecken zu können, benennen Probleme bei der Durchführungs- und Auswertungsobjektivität und raten gar von Subtests (aus dem Bereich *Psychomotorik*) ab, da die Reliabilität *mangelhaft* ist und kritisieren, dass „eine bloße Zitierung neuerer empirischer Ergebnisse (...) noch keine Entwicklungstheorie“ begründet (ebd., S. 112). Den kritischen Einwänden widersprechen Grob und Hagmann-von Arx (2011) in einer Replik, räumen aber geplante Revisionen ein.

2.5.3 Zusammenfassende Übersicht der Testgütekriterien

In den Tabellen 4 und 5 werden zum Zwecke der besseren Übersicht die Hauptgütekriterien für die überwiegend interessierenden Tests zusammengefasst. Bei mehreren Angaben zu den Gütekriterien werden die Angaben in den Testrezensionen bevorzugt dargestellt, da ein objektiverer Blick auf die Tests unterstellt wird. So ist z. B. denkbar, dass ein weniger günstiger Wert weniger prominent in einem Manual platziert bzw. vom Testverlag vorgestellt wird, während bei Testrezensionen eine kritische Grundhaltung anzunehmen ist. Die drei Hauptgütekriterien sind farblich markiert und basieren auf Angaben von Gruber und Tausch, 2015; Deimann und Kastner-Koller, 2008; Renner, 2010; Mickley, 2015; Naescher, 2009; Testzentrale, 2017; Schroth, 2015 sowie Naescher, 2010. Weitere konkrete Quellenangaben zu den jeweiligen Tests sind kenntlich gemacht.

Tabelle 4. Übersicht der Hauptgütekriterien für ausgewählte Intelligenztests.

	K-ABC ¹	KABC-II ²	CFT1(-R) ³	CFT20-R ⁴	WISC-IV ⁵
Durchführungsobjektivität	ja	ja, aber beeinträchtigt durch Komplexität	weitgehend	ja	weitgehend
Auswertungsobjektivität	ja	ja	ja	ja	weitgehend
Interpretationsobjektivität	ja	weitgehend	ja	weitgehend	ja
Retest-Reliabilität	Subtests: .57 – .95 Skalen: .84 – .97		.63 – .91 Gesamttest: .90	>.80 –>.90 WS/ZF-R: .83 – .92	Subtests: .76 – .91 Index/Gesamt: .87 – .97
Paralleltest-Reliabilität					
Testhalbierungs-Reliabilität	.69 – .93	>.70 –>.90		>.80 –>.90	
Interne Konsistenz	.59 – .90		.75 – .95 Gesamt: .97	.86 – .96	
Inhaltsvalidität			gesichert		
Konstruktvalidität	Vergleich andere Tests: .50 – .80	Vergleich andere Tests: um .70 – .80	Validität wird mit Validität des CFT 1 begründet	belegt, teils nur mit CFT20 Validität begründet	Vergleich HAWIK-III: .63 – .73
Kriteriumsvalidität					

Anmerkung. 1: Rollett & Preckel, 2011. 2: Kuschel, Kamp-Becker & Ständer, 2017; Irlich, 2015; Renner, 2014. 3: Renner, 2014; Weiss & Osterland, 2013. 4: Gruber & Tausch, 2015. 5: Schmukle & Schulze, 2016; Deimann & Kastner-Koller, 2008.

Tabelle 5. Übersicht der Hauptgütekriterien für ausgewählte Intelligenztests.

	WPPSI-III ¹	WNV ²	SON-R ³ 2 ½–7	SON-R ⁴ 5 ½–17	SON-R ⁵ 6–40	IDS ⁶
Durchführungsobjektivität	ja	weitgehend	„weitgehend gewährleistet“	nicht vollständig	ja	ja
Auswertungsobjektivität	teilweise	ja	„weitgehend gewährleistet“	ja	ja	zufriedenstellend
Interpretationsobjektivität	ja	ja	„weitgehend gewährleistet“ ⁷	ja	ja	ja
Retest-Reliabilität	.61–.75 (für zwei Subtests)	4–7;11 Jahre: .70–.95 8–21;11 Jahre: .68–.85	.79 (drei Monate)		Subtests: Ø .79 Gesamt: .92	Subtests: .34–.88 Gesamt: .83
Paralleltest-Reliabilität						
Testhalbierungs-Reliabilität	belegt; Gesamtwert: .95					
Interne Konsistenz		.72–.90	Ø .70	Gesamt: Ø .93	Gesamt: Ø .95	.68–.96
Inhaltsvalidität	belegt	Begründet mit Expertenratings			durchgeführt	
Konstruktvalidität	Vergleich HAWIK-IV: Ges.: .91 Subtests: .61–.78	begründet mit US-amerikanischen kanadischen Daten	.46 (Lehrkräfte) Ø .65 andere Tests	Cito-Test: .66	Bewertung Lehrer: .42 Andere Tests: Ø .80	Vergleich mit Tests: .21–.69
Kriteriumsvalidität				.54–.63 (Schul-Indikatoren)		

Anmerkung. 1: Renner, 2010. 2: Schroth, 2015; Mickley, 2015. 3: Naescher, 2009. 4: Wolf, 1999. 5: Schroth, 2013. 6: Naescher, 2010. 7: Es ist möglich, dass die Bezeichnung „weitgehend“ eine bessere Bewertung meint als das „weitgehend“ der TBS-TK-Rezensionen, die eine Abwertung beinhalten. Ø = Durchschnitt. Ges. = Gesamt.

2.5.4 Bedeutungsvolle Aspekte bei der Testanwendung in der Sonderpädagogik

Überproportional häufig werden im sonderpädagogischen Kontext kognitiv schwache Kinder getestet. Über die durch die Testgütekriterien hinaus gehenden Qualitätsbelege werden weitere Aspekte bei der Anwendung von Intelli-

genztests vorgeschlagen, teilweise in Anlehnung an Nebengütekriterien, teilweise nicht erfasst mit Testgütekriterien. Die vorgeschlagenen Aspekte sollen die besonderen Bedürfnisse von Kindern mit sonderpädagogischem Unterstützungsbedarf berücksichtigen (Joél, 2017, S. 15 ff.).

- *Hohe Itemdichte*: Eine große Anzahl an Items verhindert Bodeneffekte, bei einer geringen Anzahl von Items sind differenzierte Aussagen erschwert.
- *Große Altersbandbreite*: Testverfahren sollten prinzipiell jüngere Kinder testen können, damit älteren, aber kognitiv schwächeren Kindern die einfacheren Aufgaben für die jüngeren Kinder zur Verfügung stehen (z. B. über die Umkehrregeln).
- *Große Normstichprobe*: Eine große Normstichprobe verhindert Verzerrungen, da in der Sonderpädagogik sinnbildlich oft am Rand der Gaußschen Kurve getestet wird und somit die hypothetische Vergleichsgruppe bei einer kleinen Normstichprobe zu klein sein könnte, um Artefakte auszuschließen.
- *Aktuelle Normstichprobe*: Praxis bei der Feststellung sonderpädagogischen Unterstützungsbedarfs ist der Einbezug eines Gesamt-IQ. Die Gefahr einer Stigmatisierung durch ein veraltetes und durch den Flynn-Effekt anfälliges Verfahren ist groß, da die Ergebnisse signifikant zu hoch sein könnten.
- *Praktikabilität*: Die Anwendung eines Intelligenztests ist oft ein lediglich gelegentlicher Part in der Arbeit von SonderpädagogInnen. Bestünde ein Test aus einer Regelflut, wäre der Test nicht nur mühsam zu erlernen, er müsste bei einer seltenen Anwendung dann erneut mühsam erlernt werden, auch die Durchführungsobjektivität könnte gefährdet sein.
- *Computerauswertung*: Die Auswertung über eine Software spart Zeit und erhöht die Auswertungsobjektivität (siehe Praktikabilität).
- *Freies und intuitives Erklären*: Günstig ist die Möglichkeit, kognitiv schwachen Kindern die Lernaufgaben frei vorstellen zu können. Ansonsten bestünde die Gefahr, dass die Anweisungen nicht kindgerecht angeboten werden und die Kinder nicht erfassen, worum es geht. Von Vorteil wären auch unbewertete Lernaufgaben, so dass bei einer sehr freien Instruktion die Durchführungsobjektivität nicht gefährdet ist (bei einer bewerteten Lernaufgabe müsste das standardisierte Vorgehen wieder in den Vordergrund rücken).
- *Motivierende Stimuli*: Die Testsituation ist eine Leistungssituation und Kinder mit sonderpädagogischem Unterstützungsbedarf hatten häufig unlustvolle Versagenerfahrungen in Leistungssituationen. Damit die Kinder in der Testsituation diese nicht mit vorher gemachten Misserfolgssituationen assoziieren, sollten gerade bei Kindern mit sonderpädagogischem Unterstützungsbedarf die Aufgaben motivierend, kindgerecht und spannend gestaltet sein, um die *Compliance* zu erhöhen.
- *Differentialdiagnostische Ableitungsmöglichkeiten*: Zur Ableitung von Fördermaßnahmen sind Stärken- und Schwächenanalysen mit eindimensiona-

len Verfahren kaum möglich. Pädagogische Ableitungen ermöglichen eher mehrdimensionale Tests, die neben dem Vergleich mit der Gesamtheit altersgleicher Kinder auch individuelle Stärken und Schwächen ermitteln.

Diese beschriebenen Aspekte berücksichtigen eingehender die Anforderungen an einen Intelligenztest, der mit Kindern mit sonderpädagogischem Unterstützungsbedarf durchgeführt wird. Vor allem Lernschwierigkeiten und körperliche Beeinträchtigungen der Kinder erschweren die Anwendung eines normierten standardisierten Tests. Ergänzend zu den oben beschriebenen Haupt- und Nebentestgütekriterien werden deshalb diese Aspekte erwähnt. Sicher wäre es lohnenswert, diese Aspekte bei der Konstruktion eines Intelligenztests für Kindergruppen im sonderpädagogischen Kontext zu berücksichtigen. In diese Untersuchung fließen ungünstige Rahmenbedingungen mit ein, denn aus diesen können Schwierigkeiten resultieren. Die Art und Weise, wie die Tests konstruiert worden sind, kann dessen Anwendung erschweren oder erleichtern und sich somit auf die Testatmosphäre auswirken.

Eine Einschätzung, ob die im Vordergrund stehenden Verfahren obige Aspekte günstig bzw. ungünstig berücksichtigen und somit Auswirkungen auf erlebte Schwierigkeiten haben können, fasst Tabelle 6⁴² zusammen.

Tabelle 6. Einschätzung ausgewählter Intelligenztests unter Berücksichtigung bedeutungsvoller Aspekte.

	K-ABC	KABC-II	CFT1-R	CFT20-R	WISC-IV	WPPSI-III	WNV ¹	SON-R 5½-17 ²	SON-R 6-40	IDS
Itemdichte	++	++	++	++	++	+	+	+	+	-
Altersbandbreite	+	++	o	+	+	o	++	+	+	o
Große Normstichprobe	+	-	++	+	-	-	- bis ++		o	o
Aktuelle Normstichprobe	--	++	++	o	o	+	++	--	++	+
Praktikabilität	o	--	+	+	--	o	+	+	+	o
Computerauswertung	o	++	+	+	+	+	n. v.	+	+	++
Freies/intuitives Erläutern	++	+	+	-	+	+	++	++	++	-
Motivierend	+	+	+	o	+	+	+	o	o	++
Stärken/Schwächen Analyse	++	++	-	-	++	++	o	o	-	+

Anmerkungen. Einschätzungen von „--“ (= negativ) bis „++“ (= positiv).

1 WNV: Normstichprobengrößen variieren stark. n. v. = nicht vorhanden.

2 SON-R 5½-17: keine deutsche Normierung.

42 Mangels genügend eigener Erfahrungen ohne SON-R 2½-7.

Als letzte Aspekte sollen Basisinformationen über die ausgewählten Intelligenztests zusammengefasst vorgestellt werden. Tabelle 7 gibt einen Überblick über die Testdauer, die Kosten, den Altersbereich und die Möglichkeit, per PC auszuwerten, der Tabelle 8 sind Einschätzungen zu entnehmen, für welche Fragestellungen bzw. Kindergruppen die Verfahren geeignet scheinen.

Mit diesen Übersichten sind somit die besonders interessierenden Tests sowohl in der praktischen Relevanz eingeschätzt als auch in der testtheoretischen Qualität.

Tabelle 7. Basisinformationen ausgewählter Intelligenztests (Stand: 4. 8. 2017).

	KABC-II	WISC-IV	CFT1-R	CFT20-R	IDS	SON-R 6-40	WNV	WPPSI-III
Testdauer Min.	ca. 90	ca. 70	ca. 20	ca. 30	ca. 90	ca. 60	30/50	
Kosten	1 565 €	1 562 €	118 €	232 €	1 164 €	2 076 €	1 123 €	1 110 €
PC-auswertung	inkl.	inkl.	298 €	289 €	inkl.	inkl.	n. v.	n. v.
Altersbereich in Jahren	3;0- 18;11	6;0- 16;11	5;3- 9;11	8;5- 60	5;0- 10;11	6;0- 40;11	4;0- 21;11	3;3- 7;2

Anmerkungen. Angaben entsprechend der Hinweise der Testverlage. CFT20-R Durchführung in Min. ohne kristalline Zusatztests. WNV: Testdauer unterschieden in Kurzversion/Langversion. WPPSI-III: Testdauer kann je nach Alter stark schwanken. n. v. = nicht vorhanden.

Tabelle 8. Einschätzungen nach Eignung ausgewählter Intelligenztests für häufige Fragestellungen.

Fragestellungen/ Kindergruppen	KABC-II	WISC-IV	CFT	IDS	SON-R 6-40	WNV
Sprache	ja	eher nein	ja	ja	ja	ja
Lernen	ja	ältere Kinder ja	ja	ab ca. 7 Jahren	ja	ja
Geistige Entwicklung	ja	evtl. ab 12-14 J.	eher nein	nein	ausprobieren	ausprobieren
geflüchtete Kinder	eher ja: SFI, evtl.: IVI	nein	eher ja	eher nein	ja	ja
(Hoch-)Begabung	ja	ja	ergänzend	ergänzend	ja	ja
Teilleistungs- störungen	ergänzend zu den TLS Tests ja					
Analyse Stärken/ Schwächen möglich?	ja	ja	nein	moderat ja	nein	eher nein

Anmerkungen. SFI = Sprachfrei-Index (nonverbale Durchführung); IVI = Individueller Verarbeitungsindex (Durchführung nach Lurija).

2.5.5 Rahmenbedingungen im Umgang mit Intelligenztests auf der Ebene der Bundesländer

Bildung ist Sache der Bundesländer und resultiert aus dem Art. 30 des Grundgesetzes (Kultusministerkonferenz, 2019). Aus diesem wird abgeleitet, dass Bildung zu den Befugnissen und Aufgaben der Länder gehöre. Dieses Privileg der Länder bedeutet, dass bildungspolitische Angelegenheiten unterschiedlich umgesetzt werden und kann im Zusammenhang mit dieser Arbeit bedeuten, dass die Anwendung von Intelligenztests im sonderpädagogischen Kontext vom jeweiligen Bundesland abhängen kann, in dem das Kind getestet wird. Es kann möglich sein, dass die bildungspolitischen Rahmenbedingungen der jeweiligen Bundesländer dazu führen, dass ein veralteter Test häufiger in dem einen Bundesland genutzt wird, während in dem anderen Bundesland aktuellere und aussagkräftigere Tests verwendet werden. Wäre dem so, hinge eine mehr oder weniger angemessene Testdiagnostik davon ab, wo das Kind wohnt. Hinzu könnten unterschiedliche Rahmenbedingungen dazu führen, dass TestanwenderInnen mehr oder weniger Zeit für die Vorbereitung der Tests erhalten, mehr oder weniger gut an die Tests kommen, mehr oder weniger auf eine Tradition in der Anwendung von Tests zurückgreifen können. Während in der Regel in Nordrhein-Westfalen (NRW) z. B. innerhalb weniger Wochen die sonderpädagogischen Gutachten geschrieben werden sollen, werden in Niedersachsen weitestgehend ganzjährig Gutachten geschrieben. Aus diesem Unterschied könnte eine geringere Verfügbarkeit der Testverfahren in NRW resultieren, da viele SonderpädagogInnen zur gleichen Zeit die Tests nutzen möchten.

Überregionale Fragen zur Bildung werden auf den Kultusministerkonferenzen (KMK) diskutiert. Im Zusammenhang mit der Intelligenzdiagnostik gibt es z. B. einen zurückliegenden Beschluss zur Eingliederung von Berechtigten nach dem Bundesvertriebenengesetz von 1971 in der Fassung von 1997 (Kultusministerkonferenz, 1997). Es wird hier für die Aufnahme in Sonderschulen die Anwendung von sprachfreien Intelligenztests empfohlen.

Es ist jedoch nicht bekannt, dass die aktuelle Anwendung von Intelligenztests bei einer Zusammenkunft der KultusministerInnen geregelt worden sei, dazu müsste eine Bedeutung über die Bundesländer hinweg vorliegen (Kultusministerkonferenz, 2019). Es gelten dementsprechend diesbezüglich die in den Ländern festgelegten Regularien, die auf den jeweiligen Bildungsservern eingesehen werden können. Um Unterschiede und Gemeinsamkeiten bezüglich der Anwendung von Intelligenztests zu skizzieren ist allerdings zu bedenken, dass auch bei Vorliegen entsprechender Richtlinien auf Landesebene nicht zwingend dessen kongruente Umsetzung auf regionaler Ebene vorliegen muss. So gibt es nach Sichtung der für die Darstellung von Unterschieden und Gemeinsamkeiten vorliegenden Unterlagen Hinweise, dass sich auch innerhalb eines Bundeslands die Anwendung von Intelligenztests regional unterscheiden kann. Eine

differenzierte Darstellung von Gemeinsamkeiten und Unterschieden getrennt nach Schulkreis bzw. Schulamt, im Grunde genommen getrennt nach Schule bzw. Förderzentrum, würde den Rahmen dieser Arbeit sprengen.

In dieser Arbeit sollen auf Bundeslandebene Hinweise und Richtlinien zum Thema vorgestellt werden, überwiegend entnommen den jeweiligen Bildungsservern, aber auch auf Grundlage von Rückmeldungen entsprechender Anfragen.

Bei einem Vergleich könnten aus dem unterschiedlichen Umgang mit Intelligenztests Rückschlüsse für Variablen gezogen werden, die zu vermehrten Schwierigkeiten bei der Anwendung der Tests führen könnten.

Es könnte z. B. interessant sein, ob die Freiheit, selbst über die Anwendung eines Intelligenztests bei der Gutachtenerstellung entscheiden zu können, zu weniger oder mehr Schwierigkeiten im Umgang mit den Tests führen. Ebenfalls interessant wäre das Ausmaß beschriebener Schwierigkeiten abhängig vom Umfang der zur Verfügung stehenden Tests.

Einerseits geben die ProbandInnen bereits beim Ausfüllen des Fragebogens einen Hinweis über das Bundesland, in dem sie arbeiten, so dass vergleichende Rückschlüsse über diesen Zugang bereits möglich sind. Dennoch stellen die Hinweise auf Länderebene zum Umgang mit Intelligenztests bei der sonderpädagogischen Gutachtenerstellung eine weitere Grundlage dar, die Ursachen von Schwierigkeiten zu erkennen. Auch wenn die pädagogische Arbeit innerhalb des Bildungssystems deutlichen Spielraum zulässt – resultierend aus unklaren oder wenigen Vorgaben – völlig losgelöst von bildungspolitischen Vorgaben auf Landesebene ist die Gutachtenerstellung kaum denkbar.

Im Folgenden wird für jedes Bundesland umrissen, welche Aussagen zur Anwendung von Intelligenztests getroffen werden. Die Grundlage für entsprechende Informationen war die Sichtung der jeweiligen Bildungsserver sowie Antworten aus formellen und informellen Anschreiben an persönlich bekannte SonderpädagogInnen, aber auch an offizielle Stellen wie Schulämter, Behörden usw. Die Anschreiben⁴³ waren bis auf Ausnahmen in der Anrede (abhängig vom persönlichen Bekanntheitsgrad) sinngemäß identisch:

Im Rahmen meiner Dissertation (Anwendung von Intelligenztests in der Sonderpädagogik) möchte ich den unterschiedlichen Umgang mit Intelligenztests in den Bundesländern skizzieren und möchte Sie um Hilfe bitten, mir kurz folgende Fragen zu beantworten:

- Gibt es Handreichungen oder ähnliche Schriften für SonderpädagogInnen, wie die Diagnostik zur Feststellung sonderpädagogischen Unterstützungsbedarfs gestaltet sein soll?

43 Gesendet Mitte Januar 2019 bis Mitte Februar 2019.

- Gibt es eine Liste von Intelligenztests, die im Rahmen des Feststellungsverfahrens durchgeführt werden sollen?
- Ist es den SonderpädagogInnen generell freigestellt, IQ-Tests durchzuführen oder nicht?
- Wer führt in XX⁴⁴ federführend die Gutachten durch?

Für jedes Bundesland wurden in der Regel vier bis zehn E-Mails gesendet, ca. jede zweite E-Mail wurde beantwortet, gelegentlich entwickelte sich ein Austausch zum Thema und es wurden weiterführende Informationen gesendet, z.B. regionale oder schulinterne Handreichungen, persönliche Erfahrungen mitgeteilt etc. Der folgende Abriss stellt eine Übersicht über den Umgang mit Intelligenztests dar, alphabetisch sortiert nach Bundesland. Auf die Darstellung der Angaben aus privaten E-Mails wird mangels Verifizierbarkeit nur selten zurückgegriffen, auch wenn sie Aufschluss geben über das regional unterschiedliche Vorgehen. Es sei jedoch darauf hingewiesen, dass die regionalen Bestimmungen der Schulämter nicht übereinstimmen müssen mit den entsprechenden Verordnungen der Länder und es gibt Hinweise für einen unterschiedlichen Umgang bezüglich der Anwendung von Intelligenztests, unabhängig von den Verordnungen der Länder. So „kann“ in Baden-Württemberg z.B. ein Intelligenztest durchgeführt werden, muss aber nicht (SBA-VO, 2016). Dem gegenüber stehen Angaben von SonderpädagogInnen, für die die Nutzung aus Intelligenztestergebnissen in ihrer Region zwingend sei. Widersprüche dieser Art scheinen nicht ungewöhnlich zu sein und werden auch kritisch als nicht vereinbar mit dem Schulgesetz beschrieben (LAG⁴⁵, 2016, S. 11).

In *Baden-Württemberg* gilt seit dem 1. August 2015 ein neues inklusives Schulgesetz (LAG, 2016, S. 3), in dem die Abschaffung der Sonderschulpflicht festgelegt ist. Das als „sehr allgemein“ kritisierte Gesetz (ebd., S. 3) wird konkretisiert durch die „Verordnung des Kultusministeriums über die Feststellung und Erfüllung des Anspruchs auf ein sonderpädagogisches Bildungsangebot“ (SBA-VO, 2016: Teil 2 Abschnitt 1 § 6 Abs. 2). Eingeleitet wird ein Feststellungsverfahren von der Schulaufsichtsbehörde, beauftragt für die sonderpädagogische Diagnostik wird eine Lehrkraft für Sonderpädagogik. Die beauftragte Lehrkraft ist inhaltlich „nicht an Weisungen“ gebunden (ebd.). Es wird erwähnt, dass die Diagnostik eine Schulleistungsprüfung und einen Intelligenztest beinhalten „kann“ (ebd.). Empfehlungen für die Anwendung konkreter Intelligenztests werden auf Landesebene nicht vorgenommen.

44 Für XX steht das jeweilige Bundesland; teilweise in den großen Flächenländern mit dem Hinweis XX, bzw. in Ihrem Schulkreis.

45 LAG: Landesarbeitsgemeinschaft Baden-Württemberg – Gemeinsam leben – gemeinsam lernen e. V.

Grundlage schulischer Angelegenheiten in *Bayern* ist das *Bayerische Gesetz über das Erziehungs- und Unterrichtswesen* (BayEUG) vom 31.5.2000, novelliert 2003. Mit einer Änderung des BayEUG zum 1.8.2011 (Bayerisches Staatsministerium für Unterricht und Kultus, 2014) „wird besonderer Wert auf die integrativen Bemühungen für Kinder mit Förderbedarf“ gelegt (ebd., S. 3). Lehrkräfte der Mobilen Sonderpädagogischen Dienste sind u.a. für die Diagnostik zuständig. In einer Broschüre über Kinder mit sonderpädagogischem Förderbedarf (Bayerisches Staatsministerium für Unterricht und Kultus, 2018) wird beschrieben, dass der Förderbedarf „mithilfe gezielter sonderpädagogischer Diagnostik festgestellt“ wird (ebd., S. 7). Aktuelle Richtlinien über die Gestaltung der Diagnostik mit Hilfe von Intelligenztests, gültig für das ganze Bundesland, konnten nicht erkannt werden. Es bleibt unklar, ob Tests angewendet werden müssen oder können und welche Tests bei Bedarf Anwendung finden.

Im Gegensatz dazu ist die Anwendung von Intelligenztests in *Berlin* eindeutig geregelt. In der aktuellen Auflage des *Leitfaden zur Feststellung sonderpädagogischen Förderbedarfs an Berliner Schulen* (Kern et al., 2017) werden konkrete Testverfahren benannt, abgestimmt auf die jeweiligen Unterstützungsbedarfe. Standardisierte Testverfahren werden u.a. als Grundlage der sonderpädagogischen Diagnostik genannt (ebd., S. 5), beispielsweise WISC-IV, der WNV, SON-R 6–40 oder KABC-II für den Unterstützungsbedarf *Lernen*, SON-R 6–40 oder die CFT-Tests für den Unterstützungsbedarf *Sprache*. Mit letzteren Tests sind auch alle SonderpädagogInnen ausgestattet, die in der Inklusion arbeiten. Mit den kürzeren Tests der CFT-Reihe kann eine Vorauswahl für eine ausführliche Überprüfung durch die Diagnostik- und Beratungslehrkräfte in den Beratungsstellen vorgenommen werden. Die in den Beratungsstellen⁴⁶ arbeitenden SonderpädagogInnen sind spezialisiert u. a. in der Anwendung von Intelligenztests. Dieses System hat mehrere Vorteile. Es gibt zwischen Testergebnissen eine bessere Vergleichbarkeit durch die beschriebene Auswahl der Tests, die spezialisierten SonderpädagogInnen in den SIBUZ sind routinierter in der Anwendung und die SonderpädagogInnen, die in der Inklusion arbeiten (z.B. Grund- oder Realschulen) sind bereits mit einem Test ausgestattet⁴⁷, welcher ohne Umwege über eine zentrale Testleihe unbürokratisch angewendet werden kann⁴⁸.

Berlin und das Berlin umgebende *Brandenburg* arbeiten bildungspolitisch eng zusammen und unterhalten z. B. gemeinsam das Landesinstitut für Schule und Medien Berlin-Brandenburg (LISUM) in Ludwigsfelde. Auch für Branden-

46 Über das Stadtgebiet verteilte SIBUZ: Schulpsychologische und Inklusionspädagogische Beratungs- und Unterstützungszentren.

47 Und in dessen Anwendung in Fortbildungen geschult.

48 Selbstredend unter Beachtung ethischer und gesetzlicher Regelungen, z.B. das Einverständnis der Sorgeberechtigten.

burger SonderpädagogInnen gibt es eine verbindliche Handreichung (MBSJ⁴⁹, 2018). Bereits in einer Unter-Überschrift wird von „verbindlich einzusetzenden diagnostischen Instrumenten (...)“ auf der ersten Seite gesprochen und somit festgelegt, dass die an späterer Stelle beschriebenen Testverfahren angewendet werden sollen (ebd., S. 1). Diese sind ebenfalls wie in Berlin detailliert aufgeschlüsselt je nach Unterstützungsbedarf und entsprechen bezüglich der Intelligenztests dem aktuellen Stand.

In *Bremen* ist es nicht verpflichtend, Intelligenztests durchzuführen. Auch wenn aktuelle Intelligenztests im Stadtgebiet zum Leihen vorhanden sind, gibt es keine offizielle Liste der anzuwendenden Verfahren wie in Berlin oder Brandenburg, abgestimmt auf die Unterstützungsbedarfe. Die Gutachten werden dezentral von SonderpädagogInnen geschrieben, welche organisiert sind in *Zentren für unterstützende Pädagogik* (ZuP), bei bestimmten Unterstützungsbedarfen von den *Regionalen Beratungs- und Unterstützungszentren* (ReBUZ). Orientierungsgebend bei der Verfassung eines Gutachtens ist die Erste Verordnung für unterstützende Pädagogik (EVuP, 2013⁵⁰). Hinweise auf eine Intelligenzdiagnostik, die Anwendung von Intelligenztests oder Hinweise auf die Anwendung von standardisierten Verfahren liegen nicht vor. Vielfach wird Bezug genommen auf die Förderdiagnostik zur Erkennung sonderpädagogischer Förderbedarfe (ebd., § 9f.).

Das Hamburgische Schulgesetz regelt in *Hamburg* den Ablauf einer Gutachtererstellung (§ 12 HmbSG Abs. 3) im Allgemeinen. Im Speziellen regeln Handreichungen die Anwendung von Intelligenztests. Ein vertieftes Diagnostikverfahren wird durch die Regionalen Bildungs- und Beratungszentren (ReBBZ) durchgeführt, unter anderem mit „geeigneten Testverfahren“ (Behörde für Schule und Berufsbildung, 2017, S. 2). Diese sind in weiteren Handreichungen spezifiziert, z.B. WNV, WISC-IV, SON-R 6–40, KABC-II oder IDS im *Diagnosebogen Lernen* (Behörde für Schule und Berufsbildung, 2016, S. 2), in dem ausdrücklich festgehalten ist, dass eines der Verfahren durchgeführt werden „muss“ (ebd., S. 2). In einer *Liste der Testverfahren zur Diagnostik bei Förderbedarf Lernen, Sprache sowie emotionale und soziale Entwicklung* (Behörde für Schule und Berufsbildung, 2014) sind für die in Hamburg arbeitenden SonderpädagogInnen neben Intelligenztests auch andere Tests aufgeführt, entspre-

49 MBSJ; Land Brandenburg: Ministerium für Bildung, Jugend und Sport.

50 In vielen privaten E-Mails wird sinngemäß für mehrere Bundesländer erwähnt, dass an neuen Bestimmungen gearbeitet wird. Einschränkend zu den Ausführungen kann angemerkt werden, dass nach Ende dieser Arbeit sich einige der Bestimmungen zur Feststellung sonderpädagogischen Unterstützungsbedarfs geändert haben können. Die EVuP für Bremen z.B. läuft Ende Juli 19 aus, eine Nachfolgeverordnung ist zum Zeitpunkt des Schreibens dieser Arbeit nicht erkennbar.

chend der Fragestellungen und inkl. einer Kurzbeschreibung der Tests und Angaben zu den Bezugsquellen.

Sonderpädagogische Gutachten werden in *Hessen* nicht mehr geschrieben, lediglich Stellungnahmen. Zwar gibt es Handreichungen mit regionaler Gültigkeit, die jedoch selbst erstellt sind und ohne Zitierfähigkeit (keine Hinweise auf die VerfasserInnen, keine Jahresangaben). In einer persönlichen E-Mail eines Mitarbeiters des Hessischen Kultusministeriums (D. Bognar, persönliche Kommunikation, 31.1.2019) wird mitgeteilt, dass es keine Listen mit Intelligenztest-Empfehlungen gibt und die Anwendung der Tests in der Entscheidung der Förderschullehrkraft liegt. Grundlage für die Erstellung der Stellungnahmen ist die *Verordnung über Unterricht, Erziehung und sonderpädagogische Förderung von Schülerinnen und Schülern mit Beeinträchtigungen oder Behinderungen* (VOSB) vom 15.6.2012. Da die Stellungnahmen nach Aktenlage erfolgen (§ 9 VOSB Abs. 1), sind dementsprechend keine Hinweise für selbst durchzuführende diagnostische Verfahren zu entnehmen.

In *Mecklenburg-Vorpommern* wiederum sind die Vorgaben zur Anwendung von Intelligenztests im Rahmen von Begutachtungen in einer Handreichung im Detail beschrieben. Die *Standards der Diagnostik* (Ministerium für Bildung, Wissenschaft und Kultur, 2015), empfehlen nicht nur konkrete Testverfahren, es wird in der Handreichung explizit zur Würdigung und Beachtung der zu den Testverfahren gehörenden Testgütekriterien aufgefordert (ebd., S. 12). So sollten die Verfahren Splitt-half-Retest- und Konsistenz-Reliabilitäten auf der Gesamtergebnisebene nicht unter $r = .91$ und externe Validitätskoeffizienten bei $r = .50$ liegen. Durchzuführende Testverfahren werden vorgeschlagen entsprechend der Förderbedarfe und entsprechend der Schulstufen. Für den Bereich *Lernen* wird z.B. für die Sekundarstufe 1 WISC-IV, KFT 4–12+R und AID 3 (ebd., S. 16) empfohlen. Darüber hinaus gibt es in einer Kriterienübersicht Hinweise, wie die Testergebnisse zu interpretieren und ob daraus Förderbedarfe abzuleiten sind, z.B. kein IQ von kleiner oder gleich 70 für den Förderbedarf *Lernen* (ebd., S. 17).

Die Verordnung zum Bedarf an sonderpädagogischer Unterstützung vom 22. Januar 2013⁵¹ regelt verbindlich den Rahmen für die Erstellung eines sonderpädagogischen Gutachtens in *Niedersachsen*. Dieser Verordnung sind keine Empfehlungen für die Anwendung von Intelligenztests zu entnehmen. In Stellungnahmen von MitarbeiterInnen der Landesschulbehörden wird darauf hingewiesen, dass es den SonderpädagogInnen freigestellt ist, Intelligenztests durchzuführen (J. Rath-Groneick, Regionalabteilung Osnabrück, private Kommunikation, 25.1.2019) und dessen Anwendung in den Hintergrund getreten sei

51 Nds.GVBl. Nr. 2/2013 S. 23; SVBl. 2/2013 S. 67 – VORIS 22410.

(D. Christmann, Regionalabteilung Lüneburg, private Kommunikation, 25.1.2019).

In *Nordrhein-Westfalen* ist die *Verordnung über die sonderpädagogische Förderung, den Hausunterricht und die Schule für Kranke* (AO-SF) Grundlage für die Erstellung eines Gutachtens⁵². Gutachten werden im Team (eine sonderpädagogische und eine weitere Lehrkraft der allgemeinbildenden Schulen) durchgeführt. Verbindliche Hinweise zur Anwendung von Intelligenztests sind für das Bundesland nicht zu entnehmen. Diese Verordnung als übergeordnetes Gerüst lässt Spielraum für die regionale Gestaltung im Umgang mit standardisierten Verfahren, sowohl auf der Ebene der fünf Regierungsbezirke als auch auf Ebene der Schulämter. In einer Handreichung der Bezirksregierung Münster für die Sekundarstufe I (2017) können z. B. Aussagen über „Testdurchführungen und Auswertungen“ vorgenommen werden (ebd., S. 20), verbunden mit dem Hinweis, dass „immer aktuelle Tests“ verwendet werden sollen (ebd., S. 21). In einem Leitfaden der Schulämter Bochum und Herne (Schulämter Bochum und Herne, 2015) wird bei Vorliegen einer intellektuellen Beeinträchtigung die Durchführung eines Intelligenztests empfohlen, es ist jedoch ausdrücklich erwähnt, dass ein „verbindliches Instrumentarium standardisierter Verfahren“ nicht vorgegeben wird (ebd., S. 39). In einem Leitfaden der Stadt Dortmund ist beschrieben, dass Intelligenztests eingesetzt werden können, es wird jedoch auch darauf hingewiesen, dass diese alleine nicht für eine differenzierte Wahrnehmung des Kinds ausreichen (Schulamt für die Stadt Dortmund, 2010, S. 13). Die AO-SF ermöglicht eine individuelle Gestaltung im Umgang mit Intelligenztests, allerdings scheint dies insbesondere in Nordrhein-Westfalen zu stark regionalen Unterschieden in der Anwendung zu führen.

In der *Schulordnung für die öffentlichen Sonderschulen* (SoSchulO RP, gültig ab 30.8.2006) wird der Ablauf zur Feststellung des sonderpädagogischen Förderbedarfs in § 11 in *Rheinland-Pfalz* erläutert. Bereits hier wird in Absatz 3 festgelegt, dass die Feststellung u. a. auch auf den Ergebnissen „anerkannter Testverfahren“ beruhen soll. Näheres zum Verfahrensablauf regelt eine *Handreichung zur Feststellung des sonderpädagogischen Förderbedarfs* (Bildungsministerium Rheinland-Pfalz, 2017), in dem auf die Anwendung von Intelligenztests nicht eingegangen wird. Auch für Rheinland-Pfalz sind landesweit einheitliche Regelungen zum Umgang mit Intelligenztests nicht erkennbar.

Im *Saarland* werden für „Menschen mit Behinderung“ (Ministerium für Bildung und Kultur, 2019) kurze Informationen ohne Bezug zur Diagnostik angeboten im Zusammenhang mit SchülerInnen mit sonderpädagogischem Förderbedarf. In einer Handreichung zur Feststellung sonderpädagogischen För-

52 Zuletzt geändert am 1. Juli 2016. AO-SF = Ausbildungsordnung sonderpädagogische Förderung.

derbedarfs (Bildungsserver Saarland, 2019) – orientiert am Schulpflichtgesetz (SchpflG § 6 Abs. 1 und 2) – wird dementsprechend die „sonderpädagogische Förderdiagnostik“ beschrieben als „Kind-Umfeld-Analyse“ (ebd., S. 1). Standardisierte Testverfahren finden keine Erwähnung.

In *Sachsen-Anhalt* ist die *Verordnung über die Förderung von Schülerinnen und Schülern mit sonderpädagogischem Bildungs-, Beratungs- und Unterstützungsbedarf* (SoPädFV ST 2013 Abschnitt 2 § 4, Fassung vom 8. August 2013) bindend bei der Feststellung sonderpädagogischen Förderbedarfs und bildet den übergeordneten Rahmen für die Gutachtenerstellung. Der Zeitraum von der Antragstellung bis zum 10. Januar eines Jahres bis zur Entscheidung der Landesschulbehörde bis zum 20. Mai eines Jahres (ebd.) verdeutlicht die Zeitspanne, in denen Tests ausgeliehen, geprobt bzw. gelernt und durchgeführt werden können. Das vom *Mobilien Sonderpädagogischen Diagnostischen Dienst* durchgeführte Gutachten orientiert sich an der Handreichung zur sonderpädagogischen Förderung in Sachsen-Anhalt (Kultusministerium Sachsen-Anhalt, o.J.), in denen der Einsatz eines Intelligenztests ausdrücklich bei entsprechenden Fragestellungen empfohlen wird (ebd., S. 24). Es wird jedoch auch darauf hingewiesen, den „Einfluss der Intelligenz auf den Lernerfolg nicht zu überschätzen“ (ebd., S. 24). Intelligenztests werden namentlich vorgeschlagen, z.B. der HAWIK-IV (ebd., S. 135), der SON-R 5½–17, der AID II (ebd., S. 146) oder der CFT 20 bzw. CFT 1 (ebd., S. 146). Für die meisten der vorgeschlagenen Intelligenztests liegen allerdings inzwischen neuere Fassungen vor.

Die Verordnung des Sächsischen Staatsministeriums für Kultus über Förderschulen im Freistaat *Sachsen* (SOFS⁵³) regelt in Abschnitt 2 von § 13 bis § 17 das Verfahren zur Feststellung sonderpädagogischen Förderbedarfs, durchgeführt in der Regel auch in Sachsen durch einen *Mobilien Sonderpädagogischen Dienst*. Dort ist beschrieben, dass die oberste Schulaufsichtsbehörde Vorgaben zu „einheitlichen landesweit einzusetzenden standardisierten Testverfahren veröffentlichen“ kann (SOFS Abs. 2 § 13). Als Download werden als Hilfestellung das *Handbuch zur Förderdiagnostik* bereitgestellt (Staatsministerium für Kultus Sachsen, 2005) sowie die *Material- und Methodensammlung zur Förderdiagnostik* (Sächsisches Staatsinstitut für Bildung und Schulentwicklung, 2005), dessen Anhang Listen von Testverfahren zu entnehmen sind, die zum Zeitpunkt der Veröffentlichung aktuell waren.

In *Schleswig-Holstein* ist die Landesverordnung über sonderpädagogische Förderung (SoFVO) vom 8. Juni 2018 Grundlage für die Gutachtenerstellung zur Feststellung sonderpädagogischen Förderbedarfs, näher geregelt in § 4 und durchgeführt von den Förderzentren des Landes. Es wird beschrieben, dass ein

53 SOFS = Schulordnung Förderschulen (SächsGVBI S. 317 vom 3.8.2004, letzte Fassung gültig ab 1.8.2019).

Gutachten alle Umstände berücksichtigt, die für eine sonderpädagogische Förderung von Bedeutung sind (SoFVO §4 Abs. 4), konkrete Hinweise über den Umgang mit standardisierten Verfahren sind der Verordnung nicht zu entnehmen. Offizielle Listen mit empfohlenen Intelligenztests gibt es nicht, jedoch besteht die Möglichkeit, auf Nachfrage Empfehlungen mit aktuellen Testverfahren beim Institut für Qualitätsentwicklung an Schulen (IQSH) zu erfragen (B. Ebert, Koordinator Diagnostik im Schulteam Sonderpädagogik des IQSH, 24.1. 2019). In der Broschüre Wissenswertes über Sonderpädagogik in Schleswig-Holstein (Institut für Qualitätsentwicklung an Schulen, 2016) werden normorientierte Verfahren wie Intelligenztests zur Feststellung des Förderbedarfs *Lernen* als eine Möglichkeit der „diagnostischen Herangehensweise“ genannt (ebd., S. 27).

Grundlage in *Thüringen* zur Feststellung des sonderpädagogischen Förderbedarfs ist der 3. Abschnitt der Thüringer Verordnung zur sonderpädagogischen Förderung (ThürSoFöV vom 6. April 2004, geändert am 26. Mai 2009). Hinweise zur Anwendung standardisierter Testverfahren sind in dieser Verordnung nicht enthalten. Konkretere Hinweise sind hier dem Thüringer *Diagnostikkonzept zur Qualitätssicherung* (Vernooij, 2013) zu entnehmen, in dem die Anwendung von Intelligenztests beschrieben wird (ebd., S. 11) und Hinweise zur Interpretation von Intelligenztestergebnissen vorgestellt werden (ebd., S. 15 f.). In der Durchführung möglicher Intelligenztests werden unter Angabe der jeweiligen Testdauer und kurzen Beschreibungen Tests tabellarisch aufgeführt (ebd., S. 20), jedoch gibt es für jeden der drei beschriebenen Intelligenztests inzwischen Nachfolgeversionen. In einer *Handreichung für den Gemeinsamen Unterricht* (Ministerium für Bildung, Wissenschaft und Kultur, 2013), wird unter Berufung auf das Thüringer Diagnostikkonzept nach Vernooij festgelegt, dass die Anwendung von standardisierten Testverfahren vereinheitlicht und unter Festlegung auf bestimmte Testverfahren vergleichbar wird (Ministerium für Bildung, Wissenschaft und Kultur, 2013, S. 12).

Zusammengefasst kann festgestellt werden, dass der offizielle Umgang mit Intelligenztests auf Landesebene sehr unterschiedlich geregelt ist. In einigen Bundesländern gibt es nur wenige bis gar keine Hinweise dazu, in anderen Bundesländern gibt es Listen mit Intelligenztests, die anzuwenden sind, in Mecklenburg-Vorpommern gar Vorgaben zu den Testgütekriterien der Tests, die zu beachten sind. In einigen der Bundesländer, die konkrete Tests empfehlen, fällt auf, dass die empfohlenen Tests veraltet sind.

Für einen Widerspruch von Gaus und Drieschners Feststellung, dass das Bildungssystem weniger einer internen Logik folgt, eher als labil chaotisches und lose gekoppeltes System bezeichnet werden kann (2014, S. 29), fehlt es bezüglich der Anwendung von Intelligenztests in der Tat an Befunden. Es kann also vom Bundesland abhängen, ob, in welchem Rahmen und womit ein Kind

getestet wird und es sind Unterschiede bezüglich empfundener Schwierigkeiten zwischen den Bundesländern möglich, dessen Erkennen für das Ableiten von Empfehlungen bezüglich des Umgangs mit Intelligenztests hilfreich sein kann.

2.5.6 Antwort- und Verzerrungstendenzen, Beobachtungsfehler und TestleiterInneneffekte

Unabhängig von der Leistung eines Kinds kann das Testergebnis von der Person beeinflusst werden, die testet, von der Situation, in der getestet wird und von Einstellungen des Kinds. Dies gilt sowohl bei der Beantwortung eines Fragebogens, aber auch bei der Anwendung eines Intelligenztests. Kubinger (2009b) konnte in einem Versuch nachweisen, dass TestleiterInnen Einfluss auf das Testergebnis haben können. Kinder wurden zweimal von unterschiedlichen Personen getestet, es kamen teils abweichende Ergebnisse heraus bis zu 6 IQ Punkte (die nicht aus dem Übungseffekt resultierten)⁵⁴.

Effekte, die ein Ergebnis beeinflussen können, müssen bei der Interpretation berücksichtigt werden. Sie stehen im engen Zusammenhang mit den Testergebnissen, da sie diese negativ beeinflussen. Diese Effekte gehören zweifelsohne zu den Schwierigkeiten bei der Anwendung standardisierter Tests.

Würde eine Klassenlehrerin ein Kind mit einem Persönlichkeitstest befragen und die Fragen vorlesen, würde ein Kind auf die Frage, *hast du schon einmal bei einer Klassenarbeit geschummelt*, evtl. sozial erwünscht antworten, um negative Konsequenzen zu vermeiden. Die *Soziale Erwünschtheit* ist eines der bedeutsamsten Effekte, die Testergebnisse verzerren können und könnte beschrieben werden mit einer Antworttendenz, die beeinflusst ist von dem Wunsch, den erwarteten Vorstellungen der TestleiterIn zu entsprechen. Besonders bei Persönlichkeitstests mit einer direkten Befragung der ProbandInnen ist die Gefahr dieser Verzerrung groß und entsprechend sollte den ProbandInnen versichert werden, dass jede Antwort legitim ist, dass es kein richtig oder falsch gibt, dass spontan geantwortet werden sollte usw. Doch auch bei vermeintlich standardisierten Intelligenztests kann *Soziale Erwünschtheit* auftreten. Auf die Frage, *was tust du, wenn du in einem Geschäft eine Geldbörse findest* (WISC-IV, Subtest *allgemeines Verständnis*), haben Kinder zuweilen sinngemäß in Testsituationen mit *soll ich dir sagen, was ich mache oder was du hören möchtest* geantwortet. *Selfenhancement* könnte als eine Form von *Sozialer Erwünschtheit* beschrieben werden, allerdings wäre die Verzerrung von dem bewussten oder

54 Es kann eingewendet werden, dass eine Differenz von 6 IQ-Punkten in der Regel keine wirklichen kritischen Differenzen darstellen müssen, sondern den grundsätzlich anzunehmenden Messungenauigkeiten geschuldet sein können.

unbewussten Wunsch geleitet, wie man sich selber gerne hätte oder wahrnimmt, auch wenn dies der Realität nicht entspricht.

In der folgenden Übersicht werden neben *Sozialer Erwünschtheit* und *Self-enhancement* Effekte beschrieben, die Einfluss nehmen können auf Testergebnisse. Bei der folgenden Liste Einfluss nehmender und in der Literatur meist als TestleiterInneneffekte oder Beobachtungsfehler beschriebenen Variablen handelt es sich um eine Auswahl der meist beschriebenen Effekte.

Doch zunächst soll genauer begründet werden, warum TestleiterInneneffekte bzw. Beobachtungsfehler aufgeführt werden und im Zusammenhang mit dieser Forschungsarbeit stehen. Es gibt zweifelsohne eine Vielzahl von Variablen, die Einfluss nehmen können auf das Testergebnis und nicht zuletzt diese Forschungsarbeit untersucht einige dieser Variablen. Im Sinne der Klassischen Testtheorie sind dies Messfehler, möglich sind auch bewusst herbeigeführte und den standardisierten Ablauf gefährdende Veränderungen durch TestleiterInnen usw.

TestleiterInneneffekte stellen eine weitere Gruppe von Gefährdungen dar, die ein Testergebnis weiter weg vom wahren Ergebnis führen könnten. Zwar ist es schwierig, Effekte wie die *Soziale Erwünschtheit* objektiv zu erfassen und für viele der weiter unten beschriebenen Effekte müsste man in die Köpfe der ProbandInnen schauen können, um bestimmen zu können, ob ein Effekt auftritt oder nicht. Dennoch bleibt unbestritten, dass TestleiterInneneffekte bzw. Beobachtungsfehler ein Ergebnis beeinflussen können. Zu einer umfassenden Würdigung von Testergebnissen sollten – so gut es geht – alle Variablen genannt werden, die in der Testsituation Einfluss nehmen können.

Ein weiterer Grund für die Beschäftigung mit den Effekten liegt darin begründet, dass diese Effekte sowohl in der Testsituation auftreten können als auch bei der Beantwortung des Fragebogens, welcher Grundlage dieser Untersuchung ist. Auch die SonderpädagogInnen, die sich an dieser Untersuchung beteiligen, sind evtl. durch die Effekte beeinflusst, z. B. bei einer *Tendenz zum Extremen*. Untersuchungsergebnisse dieser Forschungsarbeit können also beeinflusst sein durch die im Folgenden beschriebenen Effekte. Deshalb wäre es eine Vernachlässigung, würden die Beschreibungen von Untersuchungsergebnissen beeinflussenden Variablen ausgespart bleiben.

Zunächst sollen wichtige TestleiterInneneffekte bzw. Beobachtungsfehler beschrieben werden, im Anschluss soll eine Einordnung vorgenommen werden, aus der ersichtlich wird, ob diese Effekte in der Testsituation (SonderpädagogIn testet Kind) oder beim Ausfüllen des Fragebogens dieser Untersuchung auftreten können oder beides. Somit ist dieses letzte Kapitel des theoretischen Teils das Bindeglied zwischen dem Theorie- und Methodenteil.

- *Tendenz zur Mitte*: Bei mehrstufigen Skalen wie der Likert-Skala kennzeichnet dies ein Antwortverhalten, welches die mittlere Antwort bevorzugt. Dies

könnte aus einer Verunsicherung über die richtige Antwort resultieren und mit einer ungeraden Anzahl von Antwortmöglichkeiten verhindert werden. Resultiert die Tendenz zur Mitte aus der Einschätzung, keine passende Antwort unter den vorgegebenen Möglichkeiten gefunden zu haben, wäre eine ergänzende offene Antwortmöglichkeit (z. B. ein Feld mit der Beschriftung *Sonstiges* oder *Anmerkungen*) sinnvoll. Eine *Tendenz zur Mitte* lässt sich über eine statistisch festgestellte geringe Varianz nachweisen (Schmidt-Atzert & Amelang, 2012).

- *Tendenz zum Extremen*: Werden bei mehrstufigen Skalen die Antworten an den Rändern gewählt, würde dies die Varianz erhöhen (Häufigkeitsverteilungen hätten eine U-Form).
- *Skalenorientierung*: Es besteht eine geringe Tendenz bei Fragebögen, die Antwortkategorien links zu bevorzugen (Tourangeau, Rips & Rasinski, 2000), sozusagen der erste Eindruck⁵⁵. Deutlichere Effekte bestehen bei einer vertikalen Darstellung der Ratingskalen (Krosnick & Alwin, 1987).
- *Milde-Effekt*: Die Tendenz einer milden Beurteilung kann z. B. bei der Beantwortung eines Verhaltensfragebogens dazu führen, problembezogene Items milde zu bewerten mit der Folge eines negativ-falschen Ergebnisses (vorhandene Auffälligkeiten werden nicht erkannt). Bei einem Evaluationsbogen nach einem Seminar kann es bei einem sympathisch wirkenden Seminarleiter dazu führen, dass eine Veranstaltung zu positiv bewertet wird. Der *Milde-Effekt* wird mit einer Furcht vor einer zu negativen Beurteilung assoziiert (Schmidt-Atzert & Amelang, 2012).
- *Härte-Effekt (Strenge-Effekt)*: Bei einer strengen Bewertung ist es z. B. möglich, beim Ausfüllen eines Fragebogens diesen zu instrumentalisieren, um vorhandene Verhaltensauffälligkeiten eines Kinds unverhältnismäßig hervorzuheben. Bei einem Evaluationsbogen nach einem Seminar kann es bei einem unsympathisch wirkenden Seminarleiter dazu führen, dass eine Veranstaltung zu negativ bewertet wird. Würde die Bewertungstendenz (sowohl beim Milde- als auch beim Härte-Effekt) an den erwarteten Nutzen angepasst, kann dies mit strategischer Selbstdarstellung umschrieben werden (ebd., S. 249).
- *Rosenthal-Effekt (Versuchsleiter-Artefakt; Pygmalion-Effekt)*: Im Sinne einer selbsterfüllenden Prophezeiung (*selffulfilling prophecy*, Merton, 1948) entsprechen ProbandInnen den Erwartungen der TestleiterInnen. Rosenthal und Fode (1963) wiesen in einem Experiment nach, dass der Glaube von Studierenden an besondere Fähigkeiten von Ratten dazu führte, dass diese (willkürlich ausgewählten) Ratten durch die höhere Erwartungshaltung tatsächlich höhere Leistungen erzielten. Lernschwache und Intelligenzgem-

55 Wobei der *primacy-effekt* im Zusammenhang dieses Kapitels etwas anderes meint (s. u.).

derte Kinder werden in der Regel im Rahmen der sonderpädagogischen Unterstützung bedarfe *Lernen* und *Geistige Entwicklung* beschult, hier ist das Experiment von Rosenthal und Jacobson (1968) in einer Grundschule interessant: LehrerInnen wurde vorgetäuscht, nach einem Test besonders leistungsstarke und weniger leistungsstarke Kinder erkannt zu haben. Den Lehrkräften wurden jedoch imaginäre Ergebnisse mitgeteilt. Tatsächlich wurde ein Intelligenztest durchgeführt und acht Monate nach der Mitteilung an die Lehrkräfte über die erfundenen Leistungskompetenzen ein weiterer. Nur erklärbar durch die Erwartungshaltung – denn alle Rahmenbedingungen sind gleichgeblieben – hat sich der Gesamtwert der Intelligenztestung der vermeintlich starken SchülerInnen verbessert⁵⁶. Allerdings hat sich – wenn auch geringer – der Gesamtwert der vermeintlich schwachen SchülerInnen ebenfalls verbessert. Es ist möglich, dass sich der Skandal um Nenad Mihailovic (Süddeutsche Zeitung, 2017) mit dem *Rosenthal-Effekt* erklären lässt. Als Kind wurde Herr Mihailovic falsch mit einem Intelligenztest (K-ABC) getestet. Er erzielte weit unterdurchschnittliche Ergebnisse. Ihm wurde auf dem Testergebnis basierend der falsche Unterstützungsbedarf attestiert und infolgedessen jahrelang in einer Schule mit dem Schwerpunkt *Geistige Entwicklung* falsch beschult. Mit guter Aussicht auf Erfolg hat er das Land NRW verklagt⁵⁷. Erstaunlich ist, dass in den Jahren der Beschulung nicht erkannt wurde, dass Nenad Mihailovic weder eine Intelligenzminderung noch eine Lernbeeinträchtigung hatte, die Lehrkräfte aber auf Grund der Testergebnisse davon ausgehen mussten und möglicherweise eine entsprechende Erwartungshaltung hatten. Der *Rosenthal-Effekt* ist für die Testsituation im Rahmen der Intelligenzdiagnostik von besonderer Bedeutung, denn es könnte sein, dass die Kinder einer Erwartungshaltung oder einer Rollenzuweisung der TestleiterInnen entsprechen und agieren. Es könnte vorteilhaft sein, würden die TestleiterInnen weder das Kind gut kennen noch ausführliche Anamnese-Gespräche vorher geführt haben. Baudson (2011) misst dem *Rosenthal-* bzw. *Pygmalion-Effekt* weniger Bedeutung zu als er nach Veröffentlichung der entsprechenden Studien hatte und beschreibt ihn als im Durchschnitt gering und wenig stabil, wenn auch als nachgewiesen existent (ebd., S. 9). In der Medizin werden zur Vermeidung des *Rosenthal-Effekts* bevorzugt Doppel-Blindstudien durchgeführt.

- *Beobachterdrift*: Vor allem in Beobachtungssituationen kann dies mit einem zunehmend oder partiellen unaufmerksamen (in der Aufmerksamkeit abdriftenden) Zustand der TestleiterIn beschrieben werden. Bei der Anwendung eines Intelligenztests ist die Beobachtung jedoch auch wesentlicher

56 Am deutlichsten bei den gutaussehenden.

57 Anmerkung 28. 7. 19: Die Klage war erfolgreich.

Bestandteil und gibt für die Interpretation der Testergebnisse wichtige Hinweise, z. B., ob ein Kind aus gemachten Fehlern günstige Rückschlüsse zieht, ob es Problemlösungsstrategien entwickelt usw.

- *Reaktivität*: Beschreibt das Reagieren oder die Veränderung im Verhalten der ProbandInnen auf die TestleiterIn, z. B. auf die Kleidung oder das Geschlecht. Abhängig vom Geschlecht der TestleiterIn könnten ProbandInnen unterschiedlich agieren, z. B. eine geringere *Compliance* gegenüber Frauen zeigen, wenn das Kind aus einem sehr patriarchal geprägten Milieu kommt oder voreingenommen gegenüber Männern sein, wenn das Kind sexualisierte Gewalterfahrungen erlitten hatte und Männer mit dem Tätergeschlecht assoziiert.
- *Übungseffekt (Lerneffekt)*: Bei der wiederholten Anwendung von Intelligenztests kann es zu höheren Ergebnissen als Ergebnis von Lernprozessen und Übung kommen. Es ist möglich, dass sich Kinder an die Art und Weise bei der Bearbeitung von Items aus Subtests erinnern (weniger an die Items selbst) bei einer Retestung. Es ist aber auch möglich, dass Kinder eine gewisse Erfahrung im Umgang mit Testverfahren entwickeln und deshalb souveräner damit umgehen können. *Übungseffekte* bei Testwiederholungen sind nicht umstritten und vielfach belegt (u. a. Kubinger, 2009b; Bühner, Ziegler, Bohnes & Lauterbach, 2006; Hausknecht, Halpert, Di Paolo & Gerrard, 2007).
- *Primacy-/Regency-Effekt*: Dieser Effekt beschreibt den Einfluss auf eine BeurteilerIn, die am Anfang bzw. am Ende der Beobachtungssituation gemacht werden. Im Rahmen einer Intelligenzdiagnostik könnte dies nach dem einführenden Gespräch vor einer Testung, welches in der Regel zum Aufwärmen und Kennenlernen durchgeführt wird, die Erwartungshaltung der TestleiterIn beeinflussen, da ein erster Eindruck dazu führen könnte, erste Hypothesen bestätigt zu bekommen (Schmidt-Atzert & Amelang, 2012). Im Zusammenhang mit der Anwendung von Intelligenztests ist der *Primacy-Effekt* von größerer Bedeutung als der *Regency-Effekt*, da der letzte Eindruck weniger Einfluss nehmen kann bei der Anwendung eines standardisierten Tests.
- *Hawthorne-Effekt*: Dieser Effekt beschreibt Verhaltensänderungen bei dem Wissen, unter Beobachtung zu stehen, z. B. im Rahmen einer Intelligenztestung. Dieser Effekt geht zurück auf die berühmte Hawthorne Studie (Mayo, 1930, 1933; Roethlisberger & Dickson, 1939). In der Hawthorne Fabrik in den USA wurde in den 20er/30er Jahren festgestellt, dass Arbeiterinnen ihre Produktivität ohne Veränderung der Rahmenbedingungen steigerten als Folge des Glaubens daran, an einer wichtigen Studie teilzunehmen und unter Beobachtung zu stehen. Obwohl die Studie methodisch umstritten ist (Walter-Busch, 1989), bewirkte sie in der Arbeitspsychologie nicht nur ein Umdenken in der Betrachtung von Arbeit und Arbeitsbedingungen auf das

Wohlbefinden, sondern beschreibt in der psychologischen Diagnostik Verzerrungseffekte auf Grund des Gefühls, unter Beobachtung zu stehen.

- *Halo-Effekt (Hofeffekt)*: Dieser Effekt beschreibt das Überstrahlen (Schmidt-Atzert & Amelang, 2012, S. 320) eines Merkmals einer Person auf andere Merkmale, infolgedessen es zu vorschnellen oder falschen Urteilsbildungen kommen kann. Zeigt ein Kind z.B. sehr gute sprachliche Kompetenzen während einer Testung, kann es zu einer ungünstigen Verknüpfung über die Annahme intellektueller Kompetenzen führen, da vermeintlich angenommen wird, Kinder mit einem elaborierten Sprachcode sind intelligent.

Die beschriebenen Effekte können sowohl Einfluss bei der Anwendung eines Intelligenztests nehmen, aber auch bei der Beantwortung dieser Arbeit zu Grunde liegenden Fragebogens. Tabelle 9 schätzt mögliche Effekte auf diese beiden Situationen ein.

Tabelle 9. Mögliche Effekte, die bei der Anwendung von Intelligenztests bzw. bei der Beantwortung des Untersuchungsfragebogens auftreten können.

	Effekt kann bei Anwendung eines Intelligenztests auftreten	Effekt kann beim Ausfüllen des (Untersuchungs-)Fragebogens auftreten
Selbstenhancement	ja	ja
Soziale Erwünschtheit	ja	ja
Tendenz zur Mitte	bei Auswahlitems	ja
Tendenz zum Extremen	bei Auswahlitems	ja
Skalenorientierung	nein	gering
Milde-Effekt	nein	ja
Härte-Effekt (Strenge-Effekt)	nein	ja
Rosenthal-Effekt	ja	nein
Beobachterdrift	ja	nein
Reaktivität	ja	ja, wenn Untersucher persönlich bekannt
Übungseffekt	ja	ohne Bedeutung
Primacy-/Regency-Effekt	ja	nein
Hawthorne-Effekt	ja	ja
Halo-Effekt	ja (durch Änderung der Erwartungshaltung)	nein

3 Forschungsfragen

Die Anwendung von Intelligenztests in der Sonderpädagogik ist ein umfangreiches Thema und benötigt eine Reduzierung auf Aspekte, die im Rahmen einer wissenschaftlichen Studie zu bewältigen sind.

Intelligenz, Intelligenzdiagnostik, Arbeitsbedingungen in der Sonderpädagogik, die diagnostische universitäre Ausbildung von SonderpädagogInnen, die Möglichkeiten von Ableitungen pädagogischer (Förder-)Maßnahmen aus Testergebnissen, die generelle diagnostische Qualifikation von SonderpädagogInnen und die Legitimation von Intelligenztests im sonderpädagogischen Kontext sind nicht Gegenstand dieser Untersuchung, sondern jeweils Teilbereiche im weiteren Zusammenhang mit dem Forschungsgegenstand.

Die übergeordnete Frage lautet, ob in der Sonderpädagogik Intelligenztests unter Bedingungen durchgeführt werden, die die Durchführungs- und Auswertungsobjektivität gefährden und somit die Gefahr falsch ermittelter Testergebnisse mit der daraus resultierenden Gefahr der Stigmatisierung. In diesem Zusammenhang sind die eingangs erwähnten Bereiche von Bedeutung. Es wird z. B. nicht in Gänze untersucht werden können, wie genau die universitäre Ausbildung diagnostische Inhalte vermittelt, sondern ob und wie Basisbegriffe im Zusammenhang mit standardisierten Tests vermittelt werden und wie sich dies auf beschriebene Schwierigkeiten bei der Anwendung von Intelligenztests in der Praxis auswirkt.

Die von SonderpädagogInnen häufig eingesetzten Intelligenztests benötigen für die Anwendung nach den Regeln der Kunst (z. B. Zeit für die Durchführung, Zeit für die Vorbereitung, angemessene Testräume usw.) einen strukturellen Rahmen, aber auch eine angemessene universitäre Ausbildung.

Zudem sind für die Durchführung, Auswertung und Interpretation testtheoretische Kenntnisse sowie Kenntnisse über die den Tests zugrunde liegenden Intelligenzmodelle notwendig. Da die Anwendung von teils aufwändig konstruierten Intelligenztests nur ein kleiner Ausschnitt der vielfältigen Tätigkeitsfelder von SonderpädagogInnen darstellt, werden Problematiken bei der Anwendung von Intelligenztests im sonderpädagogischen Kontext angenommen und sind bereits beschrieben worden (Huber, 1999, 2000; Müller, 2009; Bundschuh, 2010; Staud & Staud, 2011).

Um evtl. notwendige Veränderungen in der universitären Ausbildung und eine evtl. notwendige Veränderung im strukturellen Rahmen zu diskutieren, in dem Intelligenztests angewendet werden, soll geklärt werden, ob angenommene Problematiken tatsächlich vorhanden sind. Sollte sich diese Forschungsfrage bestätigen, ist weiterhin zu analysieren, welche Problematiken im Detail vorliegen.

Über den deskriptiven Zugang werden Hinweise auf den Ist-Zustand im Umgang mit Intelligenztests vorgestellt. Es wird dargestellt, in welchem Ausmaß und unter welchen Bedingungen Intelligenztests Anwendung finden.

Über die Beschreibung eines Status hinaus sollen mit Hilfe der Inferenzstatistik verifizierbare Forschungsannahmen bzw. Hypothesen untersucht werden, die aus den Forschungsfragen resultieren:

Forschungsannahme 1

Es wird erwartet, dass die Aussagekraft eines Tests von der *Dimensionalität* des Verfahrens abhängt.

Mehrdimensionale Intelligenztests ermöglichen über die Generierung eines Gesamtwerts in Form des Generalfaktors der Intelligenz eine Interpretation entsprechend dem CHC-Modell auf *Stratum-II*-Ebene und somit neben dem normativen Vergleich eine Ermittlung individueller Stärken und Schwächen, aus denen sich Hinweise für pädagogische Maßnahmen bzw. Erklärungsmodelle für beobachtbares Verhalten ableiten lassen. Der unterstellten höheren Nützlichkeit mehrdimensionaler Tests für das pädagogische Handeln entsprechend wird erwartet, dass SonderpädagogInnen mehrdimensionale Intelligenztests als aussagekräftiger beurteilen. Obwohl das Ergebnis der Prüfung absehbar ist und trivial wirken könnte, ist die Prüfung dieser Annahme die Grundlage für spätere Diskussionen: sollten die an dieser Studie beteiligten SonderpädagogInnen mehrdimensionale Intelligenztests als signifikant aussagekräftiger einschätzen, dennoch aber überproportional häufig eindimensionale und somit weniger aussagekräftig eingeschätzte Verfahren einsetzen, wäre zu diskutieren, woran dies liegen könnte. Ohne Prüfung dieser Forschungsfrage wäre eine Ungenauigkeit in einer möglichen Argumentationskette vorprogrammiert. Würde unterstellt werden, dass eindimensionale Verfahren häufiger eingesetzt werden, obwohl mehrdimensionale augenscheinlich aussagekräftiger sind, könnte die unterstellte höhere, aber nicht geprüfte Aussagekraft mehrdimensionaler Tests hinterfragt werden.

Forschungsannahme 2

Es wird erwartet, dass komplexere Tests seltener angewendet werden.

Durch die vielfältigen Tätigkeitsfelder von SonderpädagogInnen stellt die Anwendung von Intelligenztests in der Regel eine Ausnahme dar. Komplexe Intelligenztests bestehen aus einer Vielzahl von Regeln, dessen Erlernen und bei einer seltenen Anwendung Wiedererlernen zeitliche Ressourcen bindet. Deshalb ist anzunehmen, dass einfacher zu lernende Intelligenztests präferiert werden.⁵⁸

58 Besonders interessant ist, ob dies auch für die KABC-II gilt. In allen von mir durchgeführten Seminaren 2016 und 2017 ist die KABC-II der einzige Intelligenztest, für den in-

Forschungsannahme 3

Es wird erwartet, dass SonderpädagogInnen eine Vorliebe für bestimmte Tests haben, auch wenn andere Tests vorhanden sind.

Bedingt durch die teils schwer zu erlernenden Verfahren wird angenommen, nicht die Auswahl eines Verfahrens zielgenau an das zu testende Kind anzupassen (z. B. ein nonverbaler Test für ein nicht deutschsprechendes Kind), sondern bestimmte Tests zu favorisieren, dessen Durchführungen bekannt sind und weniger Vorbereitung benötigen. Dies hätte allerdings zur Folge, dass das Kind sich an den Test anpassen müsste und nicht umgekehrt.

Forschungsannahme 4

Es wird erwartet, dass es Unterschiede in der Häufigkeit von Durchführungsfehlern und Beeinträchtigungen in der Testsituation zwischen den Bundesländern gibt.

Bedingt durch das föderale Bildungssystem ist es bundesweit nicht einheitlich geregelt, welche Tests unter welchen Rahmenbedingungen von wem durchgeführt werden. Während in Norddeutschland z. B. mehr Kinder inklusiv beschult werden, ist die Umsetzung der Inklusion in den südlicheren Bundesländern weniger weit vorangeschritten, was zur Folge hat, dass weniger SonderpädagogInnen im Gemeinsamen Unterricht, sondern häufiger in klassischen Kollegien innerhalb einer Förderschule arbeiten und deshalb einen besseren Zugang zu Intelligenztests und zum kollegialen Austausch haben. Die aus dem föderalen System resultierenden Unterschiede im Bildungsbereich können sich je nach politischer Lage zudem temporär innerhalb der Bundesländer verändern. So wird in Nordrhein-Westfalen nach einem Regierungswechsel 2017 erwogen, weniger Förderschulen aufzulösen bzw. geschlossene wiederzueröffnen. In Hamburg sollen im Gemeinsamen Unterricht tätige SonderpädagogInnen wieder komplexere Intelligenztests durchführen. Dies war bisher besonders qualifizierten SonderpädagogInnen vorbehalten.

Da es kaum möglich ist, die unterschiedlichen und sich zudem stetig ändernden (Arbeits-)Bedingungen in der Bildungspolitik zu erfassen, resultieren auch aus dieser Forschungsannahme ungerichtete Hypothesen.

Die Forschungsannahme 4 soll präzisiert und in folgende Teilannahmen überführt werden:

- 4.1: Abhängig vom Bundesland stehen unterschiedliche Tests zur Verfügung.

dividuelle Schulungen im sonderpädagogischen Kontext (Förderzentren, Förderschulen) nachgefragt worden sind. In den letzten 18 Monaten (Stand: 8/17) fanden 44 Tagesseminare ausschließlich zur KABC-II statt.

- 4.2: Abhängig vom Bundesland werden unterschiedlich die Durchführungsobjektivität gefährdende Veränderungen vorgenommen.
- 4.3: Abhängig vom Bundesland liegen unterschiedliche Beeinträchtigungen wie fehlende oder unvollständige Testmaterialien vor.
- 4.4: Abhängig vom Bundesland liegen unterschiedliche Freiheiten vor zu entscheiden, ob ein Intelligenztest durchgeführt werden soll.
- 4.5: Abhängig vom Bundesland liegen unterschiedliche Schwierigkeiten im Umgang mit Durchführungsregeln vor.
- 4.6: Abhängig vom Bundesland wird die Anwendung von Intelligenztests als schwierig bewertet.
- 4.7: Abhängig vom Bundesland wird die zur Verfügung stehende Zeit für die Anwendung als zu kurz bewertet.

Forschungsannahme 5

Es wird erwartet, dass sich das Alter der TestleiterInnen auf Schwierigkeiten bei der Anwendung von Intelligenztests und auf die Auswahl auswirken. Mit zunehmendem Alter werden weniger Schwierigkeiten angenommen.

Im Laufe des (Berufs-)Lebens entwickelt sich Erfahrungswissen. Dies betrifft sowohl die Anwendung von Intelligenztests, die Interpretation von Testergebnissen als auch das Erfahrungswissen über Kinder. Sollte sich diese Hypothese bestätigen, würde Erfahrungswissen für einen angemesseneren Umgang bei der Anwendung von Intelligenztests sprechen und es wäre zu überlegen, wie die Erfahrung forciert werden könnte. Allerdings liegen auch Hinweise vor, dass vor allem erfahrene TestanwenderInnen für Durchführungsfehler anfälliger sind (Lipsius et al., 2008).

Weiter kann angenommen werden, dass durch die Vertrautheit mit bekannten Verfahren eine Präferenz für ältere und somit vertrautere Verfahren vorliegt.

Die Forschungsannahme 5 soll präzisiert und in folgende Teilannahmen überführt werden:

- 5.1: Mit zunehmendem Alter der TesterInnen werden weniger Schwierigkeiten bei der Anwendung von Intelligenztests erwartet.
- 5.2: Mit zunehmendem Alter der TesterInnen werden seltener aktuelle Tests angewendet.

Forschungsannahme 6

Es wird erwartet, dass sich das Geschlecht nicht auf Schwierigkeiten bei der Anwendung von Intelligenztests auswirkt.

Es ist möglich, dass sich das Geschlecht des Testenden auf ein Kind auswirkt, z.B. könnte ein traumatisiertes geflüchtetes Kind gegenüber einem männlichen Tester befangen sein, da ein typischer Grund von Traumatisierung

gen geflüchteter Kinder die Konfrontation mit gewalttätigen Männermilieus war. Für mit dem Geschlecht zusammenhängende Schwierigkeiten auf Seiten des Testenden liegen allerdings keine Hinweise vor.

Forschungsannahme 7

Es wird erwartet, dass eine geringere universitär vermittelte Auseinandersetzung mit der Testdiagnostik mehr Schwierigkeiten bei der Anwendung von Intelligenztests in der Praxis nach sich zieht.

Das Curriculum der Fachbereiche für Sonderpädagogik ist nicht einheitlich geregelt und kann auch davon abhängen, ob eine eher ablehnende Haltung gegenüber der Statusdiagnostik bzw. eine zustimmende Haltung gegenüber der Förderdiagnostik vorliegt (siehe Eberwein, 1996; Kobi, 1977; Schlee, 2008). Obwohl ein Abgleich der Curricula der Fachbereiche mit den beschriebenen Schwierigkeiten bei der Anwendung von Intelligenztests mit dem Ziel einer Evaluation interessant wäre, würde dies den Rahmen der Studie überschreiten. Deshalb kann auch über die Qualität der universitären Ausbildung in den Bundesländern keine Rückschlüsse gezogen werden, da der Arbeitsplatz nicht übereinstimmen muss mit dem Bundesland, in dem studiert wurde und nach dem Studienort nicht gefragt werden wird. Generell sollen Rückschlüsse zwischen universitärer Ausbildung und beschriebenen Schwierigkeiten untersucht werden.

Die Forschungsannahme 7 soll präzisiert und in folgende Teilannahmen überführt werden:

- 7.1: Das Ausmaß an Schwierigkeiten bei der Anwendung von Intelligenztests hängt vom Ausmaß der in der universitären Ausbildung besuchten Seminare zur Testdiagnostik ab.
- 7.2: Das Ausmaß an Schwierigkeiten bei der Anwendung von Intelligenztests hängt vom Ausmaß der in der universitären Ausbildung referierten Inhalte zur Testdiagnostik ab.

Forschungsannahme 8

Es wird angenommen, dass TeilnehmerInnen an einer außeruniversitären Fortbildung zur Testdiagnostik weniger Schwierigkeiten bei der Anwendung von Testverfahren beschreiben.

Da angenommen werden kann, dass die Teilnahme an einer Fortbildung zur Testdiagnostik mit einer besonderen Affinität zum Thema verbunden ist, aus der bereits eine Auseinandersetzung mit Intelligenztests resultieren könnte, sind Unterschiede bei den beschriebenen Schwierigkeiten im Umgang mit Intelligenztests zu erwarten, abhängig von der Teilnahme an der Fortbildung. Es wäre jedoch auch möglich, dass besonders verunsicherte TeilnehmerInnen eine Fortbildung zum Thema wahrnehmen, während versierte TesterInnen sich sicher genug fühlen, auf Fortbildungen zu verzichten.

Forschungsannahme 9

Es wird angenommen, dass bei Nutzung eines PC-Auswertungsprogramms zur Auswertung der Testergebnisse die Auswertungsobjektivität erhöht wird.

Die in Studien beschriebenen Auswertungsfehler (Lipsius et al., 2008; Alfonso et al., 1998) dürften bei einer computergestützten Auswertung größtenteils entfallen, da sowohl Rechen- und Auswertungsfehler als auch Fehler beim Abgleich mit Normtabellen weitgehend entfallen. Kubinger empfiehlt grundsätzlich die Nutzung eines PC-Auswertungsprogramms (2009b, S. 46).

Forschungsannahme 10

Es wird angenommen, dass komplexere Tests fehleranfälliger sind.

Besteht ein Test aus einer Vielzahl von Durchführungs- und Auswertungsregeln, liegen mehr Möglichkeiten des Regelverstößes vor. Daraus resultiert die Annahme, dass Tests mit vielen Regeln mehr Anwendungsfehler bzw. einfach durchzuführende Tests weniger Fehler nach sich ziehen.

4 Methoden

Um die Schwierigkeiten bei der Anwendung von Intelligenztests durch SonderpädagogInnen zu erforschen, bietet sich eine direkte Befragung dieser mittels eines Fragebogens an. Schwerpunkt der Untersuchung wird also ein von SonderpädagogInnen zu beantwortender Fragebogen sein.

Nicht ausgeschlossen werden kann, dass eine Problematik bei der Beantwortung eines Fragebogens gar nicht als Problematik gewertet wird. Diese Gefahr soll durch folgendes Beispiel belegt werden:

Ein älteres Kind würde bei einigen Subtests mit altersentsprechenden Anfangsitems beginnen und die einfachen Aufgaben werden von vornherein übersprungen (z. B. beginnt ein elfjähriges Kind in dem Subtest *Bausteine zählen* der KABC-II mit Item 7, die Items 1–6 werden ohne Durchführung als richtig bewertet). Durch die Anwendung von Anfangsitems soll eine vermutete Unterforderung vermieden werden. Es besteht jedoch die Sonderregelung, dass vermeintlich ältere intelligenzgeminderte Kinder generell bei Item 1 beginnen können, auch wenn sie eigentlich mit einem späteren Anfangsitem auf Grund des biologischen Alters beginnen müssten. Hiermit soll eine Überforderung kognitiv schwacher Kinder vermieden werden. Häufig wird bei der Anwendung dieser Sonderregelung jedoch übersehen, dass alle Items vor dem eigentlichen Anfangsitem mit einem Punkt bewertet werden müssen, auch wenn das Kind bei obigem Beispiel eine oder mehrere der Aufgaben 1–6 nicht gelöst hat, sofern die eigentlichen Umkehraufgaben 7–9 richtig gelöst worden sind. Nur wenn dies nicht der Fall wäre, dürften die Aufgaben 1–6 nachträglich nicht mehr geändert werden. Die Anwendung dieser Sonderregelung (Beginn bei Item 1 bei vermeintlich kognitiv schwachen Kindern) ist eine verbreitete Praxis bei der Überprüfung auf den sonderpädagogischen Unterstützungsbedarf *Geistige Entwicklung*.

Es würde also die Anwendung der Umkehrregel als nicht problematisch beim Ausfüllen eines Fragebogens beschrieben werden, sofern einem eine mögliche falsche Anwendung gar nicht bewusst wäre.

Um der Gefahr entgegenzuwirken, dass ProbandInnen Schwierigkeiten aus Unkenntnis nicht angeben, wird der Fragebogen durch eine Überprüfung von ausgewerteten Testformularen ergänzt.

Bei der Überprüfung von Intelligenztestformularen werden ausgefüllte Formulare auf Korrektheit geprüft. Um die Relevanz zu erhöhen, werden ausschließlich Formulare überprüft, die im Verfahren zur Feststellung sonderpädagogischen Unterstützungsbedarfs angefertigt worden sind.

4.1 Fragebogenkonstruktion

Ausgehend von den Fragestellungen sind die Fragen zu Gruppen zusammengefasst:

- a) Allgemein: (Alter; Geschlecht; allgemeine Testerfahrungen etc.),
- b) Anwendung: (spezielle Testerfahrungen; Vorlieben für Tests; Verfügbarkeit der Tests etc.),
- c) Ausbildung: (belegte Seminare während des Studiums; universitäre Inhalte bzgl. Testdurchführung etc.),
- d) Beliefs: (Einstellungen zu Intelligenztests, z.B. *Mir fällt die Durchführung von Intelligenztests leicht*),
- e) Abschluss: (z.B. Arbeitsschwerpunkte; repräsentative Daten).

Einige Fragen aus dem Block *Beliefs* sind nicht Gegenstand dieser Forschungsarbeit, sondern münden in eine separate Untersuchung zusammen mit Prof. A. Castello (Universität Flensburg, Abteilung für Sonderpädagogik). Die Integration der Fragen dieser separaten Untersuchung hat praktische Gründe, da die Infrastruktur der Befragung genutzt werden konnte. Ergebnisse aus dieser Befragung werden in einer eigenen Veröffentlichung dargestellt.

Ebenfalls münden nicht Fragen zur Testung von geflüchteten Kindern in diese Forschungsarbeit, die für eine weitere Untersuchung in den Fragebogen aus oben genannten Gründen integriert worden sind (siehe Joél, 2018).

Der Fragebogen ist im Online-Material abgebildet⁵⁹. Folgende Herangehensweise wurde gewählt: Zunächst wurden in Form eines Brain-Storming lose Ideen für Fragen gesammelt, eher, um einen Anfang zu finden. Diese Fragen resultierten aus Beobachtungen, Vermutungen zu Schwierigkeiten und persönlichen Erfahrungen und basierten zunächst nicht auf einem theoretischen Fundament und den daraus abgeleiteten Fragestellungen. Die Idee, aus Interesse und Neugier interessante Fragen stellen zu wollen, kann somit wissenschaftstheoretisch als naiv bezeichnet werden (Pilshofer, 2001). Erst nach Erstellung des theoretischen Teils dieser Arbeit und den daraus abgeleiteten Fragestellungen fand eine Operationalisierung statt. Angelehnt an ein Mind-Map zur Thematik boten sich die vorliegenden Fragen an, die wie oben beschrieben gruppiert worden sind.

Bei einigen Fragen wurde ein dichotomes Antwortformat gewählt (*Entscheiden Sie nach eigenem Ermessen, Intelligenztests durchzuführen: Ja/Nein*), die meisten Fragen werden jedoch in Form von Ratingskalen mit fünf Antwortkategorien gestellt. Dies hat zwar den Nachteil, dass die Antwortkategorien zu

59 Ohne die Fragen, die in andere Studien münden.

weilen als nicht gleichabständig verstanden werden, somit also eher von Ordinal-, denn von Intervallskalierungen auszugehen ist (Bühner, 2011), sowie Verzerrungen auf Grund einer extremen Antworttendenz bzw. einer Tendenz zur Mitte auftreten könnten. Doch auch wenn Q(uestionnaire)-Daten verzerrungsanfällig sind (Beauducel & Leue, 2014), der entscheidende Vorteil ist neben der Praktikabilität der hohe Differenzierungsgrad.

Überwiegend werden Ratingskalen mit meist gebundenen Antwortformaten gewählt, deren Antwortkategorien durchweg verbalisiert sind zuungunsten von uni- bzw. bipolaren Antwortskalen. Die vorauszusetzende Fähigkeit für Rating-skalen, eigenes Verhalten angemessen reflektieren zu können, wird der Zielgruppe unterstellt.

Der Fragebogen ist online auszufüllen. Eine Befragung über das Internet spart nicht nur Zeit und Kosten, es reduziert für die ProbandInnen ebenfalls den Aufwand. Nach dem Öffnen einer E-Mail mit der Bitte um Teilnahme kann auf einen Link gedrückt und der Fragebogen bereits ausgefüllt werden. Zwischenschritte wie das Öffnen eines Briefs, dem Ausfüllen eines Bogens sowie der anschließenden Rücksendung mit dem Verpacken in ein Couvert, der Beschriftung des Couverts, der Suche nach einer Briefmarke und der Suche nach einem Briefkasten und der Weg dorthin entfallen⁶⁰. Weitere Vorteile sind die gute Erreichbarkeit der Zielgruppe, so könnten die angeschriebenen SonderpädagogInnen über eine E-Mail-Weiterleitung weitere potenzielle ProbandInnen informieren. Das unkomplizierte Bearbeiten der Fragebögen könnte zu einer hohen Rücklaufquote führen, da die investierten Mühen für die ProbandInnen gering sind. Gerade bei einer Befragung zu Schwierigkeiten könnte die Anonymität zu einer höheren Akzeptanz führen, als wenn eine Befragung in Gegenwart des Forschers stattfinden würde. Nachteile bei der Offenheit sind nicht zu befürchten. Tourangeau & Yan (2007) fanden bei Persönlichkeitsfragebögen diesbezüglich keinen Unterschied zwischen konventionellen Papier/Bleistift- und online Fragebögen. Weitere Vorteile sind die bereits digitalisierten Daten, so dass die Auswertung erleichtert ist und die Möglichkeit der adaptiven Fragenweiterleitung (wenn-dann Bedingungen, z. B. wenn Frage 5 mit *stimmt* beantwortet wurde, wird zu Frage 11 gesprungen). Bei einer online-Befragung stellt eine Befragung von ProbandInnen aus vielen Bundesländern keine Hürde dar. Weitere Vorteile sind die für eine objektive Beantwortung vorliegenden Rahmenbedingungen, denn die ProbandInnen füllen den Bogen bei freier Zeiteinteilung möglicherweise korrekter und auch emotionsloser aus, was im Sinne der Durchführungobjektivität ist. Allerdings kann dies auch zum Nachteil gezählt werden, denn die Rahmenbedingungen können in keiner Weise vom Un-

60 Auch bei einem Freiums Schlag müsste zumindest der Weg zum Briefkasten vorgenommen werden.

tersucher beeinflusst werden, so dass nicht bekannt ist, wie die ProbandInnen den Bogen ausgefüllt haben (am Handy; betrunken; mit einem Kind auf dem Arm; sich nebenbei streitend; bei laufendem Fernseher etc.).

Zu den weiteren Nachteilen von online-Befragungen zählt die geringere Einflussnahme auf den Pool der ProbandInnen. So könnten Personen teilnehmen, die gar nicht angeschrieben worden sind, z.B. KollegInnen der angeschriebenen SonderpädagogInnen. Es ist jedoch unerheblich, ob angeschriebene oder interessierte SonderpädagogInnen teilnehmen⁶¹. Nicht ausgeschlossen werden kann zudem, dass ProbandInnen mehrfach teilnehmen. Eine absichtlich falsche und somit die Untersuchung verfälschende Antwort kann nicht zu den Nachteilen aufgezählt werden, da dieses auch bei einer Face-to-face-Untersuchung möglich wäre.

Ein wichtiger Einwand gegen online Befragungen ist die Möglichkeit, dass nur bestimmte Personengruppen teilnehmen könnten, z.B. die dem Untersucher eher wohlgesonnenen Personen bzw. die über den Untersucher verärgerten. Diese Selektion könnte dazu führen, dass nur bestimmte ProbandInnen teilnehmen. Ein bekanntes Beispiel dafür ist die Prognose für die Wahl zum US-Präsidenten 1936 durch eine Zufallsstichprobe an 50 000 BürgerInnen, die genauer war als die Prognose von 2,3 Millionen BürgerInnen, die allerdings freiwillig teilnahmen aus einem Pool von 10 Millionen angeschriebenen BürgerInnen (Bandilla, 1999, S. 18).

Insgesamt stehen die vielen Vorteile einer online-Befragung in keinem Verhältnis zu den Nachteilen.

Der Fragebogen orientiert sich an folgenden Prämissen:

1. Der Datenschutz beinhaltet eine Zusicherung zu Beginn des Fragebogens über die Anonymität. Es wird versichert, dass keinerlei Rückschlüsse auf die ProbandInnen genommen werden. Der Server des Anbieters, über den der Fragebogen läuft (*surveymonkey*) befindet sich in den USA. Dies hat den Nachteil, dass dort weniger rigorose Datenschutzrichtlinien wie in Europa gelten. Deshalb wurde die Funktion deaktiviert, die die IP-Adressen der ProbandInnen aufzeichnet⁶². Ergänzend wurde versichert, dass nach Beendigung der Untersuchung lediglich Daten für die evtl. Prüfung der Dissertation archiviert werden, nach Beendigung der Prüfung diese mit einem Reißwolf vernichtet werden.
2. Selbst von den ProbandInnen vorzunehmende Verzweigungen müssen nicht vorgenommen werden. Hat z.B. eine ProbandIn auf die Frage *Ich habe noch*

61 Im Nachhinein war es sogar günstig, dass einige TeilnehmerInnen nicht an einer Fortbildung teilnahmen, da diese Personengruppe als Kontrollgruppe diente.

62 Zumindest der Untersucher hat keine Möglichkeit, die IP-Adresse zu erkennen.

niemals einen Intelligenztest durchgeführt mit *stimmt* geantwortet, sind die folgenden Fragen zur konkreten Anwendung überflüssig. Hinweise wie *bitte überspringen Sie die Fragen 10–15* gefährden das flüssige Antworten. Einen automatischen Sprung wurde bei der Programmierung des online-Fragebogens berücksichtigt. Bei der Antwort *stimmt* auf obige Frage springt der Fragebogen an eine passende Stelle, überflüssige Fragen zur konkreten Durchführung unterbleiben.

3. Zur Vermeidung von *Sozialer Erwünschtheit* bzw. *Akquieszenz*⁶³, aber auch von *Self-enhancement*⁶⁴ wurde auf eine Lügenskala verzichtet. Es wäre damit zu rechnen, dass entsprechende Fragen die *Compliance* verringern könnten, da SonderpädagogInnen dies rasch erkennen würden. Der angenommene negative Effekt würde den möglichen positiven nicht relativieren. Stattdessen wurde das Phänomen der *Sozialen Erwünschtheit* offen in der Einleitung angesprochen. Zur Vermeidung der sozialen Erwünschtheit wird auf wertende Fragestellungen verzichtet bzw. werden die Fragen neutral formuliert, um sozial erwünschte Antworten nicht herauszufordern (Pilshofer, 2001). Es ist anzunehmen, dass viele ProbandInnen den Autoren aus Fortbildungen kennen und deshalb evtl. wohlwollend gegenüberstehen⁶⁵ und deshalb entsprechend bewusst oder unbewusst sozial erwünscht agieren könnten. Es ist jedoch kaum möglich, aus den Fragestellungen die Einstellungen des Autors herauszulesen. Dementsprechend wüssten die ProbandInnen gar nicht, wie sozial erwünscht geantwortet werden könnte. Zudem wird erfragt, ob die ProbandInnen an einer Fortbildung des Autors teilgenommen haben. Die Auswertung wird im Gruppenvergleich (ehemalige TeilnehmerInnen vs. den Autoren nicht kennend) mögliche Unterschiede ermitteln.
4. Auf unterschiedlich gepolte Fragen wird ebenfalls verzichtet. Gelegentlich negativ gepolte Fragen könnten zwar der *Nein-sage-* bzw. *Ja-sage-Tendenz* entgegenwirken, würde aber die Beantwortung der Fragen weniger flüssig gestalten.
5. Ebenfalls unberücksichtigt bleibt zugunsten der Flüssigkeit des Fragebogens die Frage der Skalenorientierung (fängt man links mit der höchsten oder niedrigsten Ausprägung an; mit der negativsten oder positivsten Antwortkategorie etc.). Generell besteht eine leichte Tendenz, die Antwortkategorien links zu bevorzugen (Tourangeau, Rips & Rasinski, 2000), neuere Studien stellen diesen *general primacy effect* gar im Zusammenhang mit der Skalenorientierung, doch ist die Forschungslage nicht einheitlich (siehe Toepoel, 2008; Hofmans et al., 2007; Krebs & Hoffmeyer-Zlotnick, 2010).

63 Zustimmungstendenz unabhängig vom Inhalt.

64 Tendenz zur Selbsttäuschung bzw. zur selbsttäuschenden Überhöhung.

65 Bzw. ablehnend.

Berücksichtigt wurde jedoch die Studie von Krosnick & Alwin (1987), die stärkere Reihenfolge-Effekte bei einer vertikalen Darstellung der Ratingskalen feststellte. Aus diesem Grund sind alle Ratingskalen horizontal angeordnet.

6. Die Vermeidung einer Tendenz zur Mitte durch Rating-Skalen mit einer geraden Anzahl an Antwortmöglichkeiten wurde vernachlässigt, um eine Richtungsentscheidung nicht zu erzwingen, sofern ProbandInnen eine mittlere Antwort tatsächlich als gültig empfinden. Es wurden fünfstufige Rating-skalen verwendet. Menold und Bogner (2015) empfehlen generell fünf- bis siebenstufige Antwortkategorien, da mehr Kategorien die Eindeutigkeit verringern und weniger schlecht differenzieren. Krosnick und Fabrigar (1997; siehe auch Krosnick & Presser, 2010) postulieren eine verbesserte Reliabilität und Validität und einen guten Differenzierungsgrad bei fünf bis sieben Kategorien; O'Muircheartaigh, Krosnick & Helic (1999) stellten zudem eine Zunahme von Reliabilität und Validität bei der Verwendung einer Mittelkategorie fest. Es wird in Kauf genommen, dass eine ungerade Ratingskala die Tendenz zur Mitte verstärken könnte (Saris & Gallhofer, 2007).
7. In Anlehnung an Rohrmann (1978) werden für die Bezeichnung der fünf Stufen der Rating-Skalen Worte verwendet, die als gleichabständig empfunden werden (z.B. *völlig falsch* – *ziemlich falsch* – *unentschieden* – *ziemlich richtig* – *völlig richtig*). Dies hat auch methodische Gründe: würde bei der Wortwahl nicht sorgfältig vorgegangen werden, bestünde die Gefahr, dass kein Mittelwert wie bei einer Intervallskala ermittelt werden kann, denn bei einem überwiegend als ungleich empfundenen Abstand zwischen den fünf Antwortmöglichkeiten verbietet sich die Ermittlung eines Mittelwertes. Die Annahme, dass die Stufen der Skalen gleichabständig sind, ermöglicht zudem die Einordnung der Daten als metrisch, was Vorteile bei der inferenzstatistischen Auswertung nach sich zieht. Jedoch könnte kritisch hinterfragt werden, ob die 1978 von Rohrmann mit Hilfe einer nicht repräsentativen Stichprobe vorgeschlagenen Empfehlungen auf Grund eines möglichen Wandels im Sprach- und Wortgebrauch heutzutage noch Gültigkeit haben.
8. CFT1 und CFT1-R sind zusammengefasst, da nicht sicher davon ausgegangen werden kann, dass den ProbandInnen bewusst ist, ob sich Antworten auf den CFT1 oder den sehr ähnlichen CFT1-R beziehen. Aus diesem Grund und auf Grund der tatsächlich hohen Ähnlichkeit wurde zwischen CFT1 und CFT1-R nicht unterschieden, obwohl es sich strenggenommen um verschiedene Verfahren handelt.
9. Bei der Frage, ob jede der fünf Kategorien eine Beschriftung erhält, oder ob lediglich eine Skalenpolarität Verwendung findet, scheinen uni- (z.B. *nicht zufrieden* – *sehr zufrieden*) bzw. bipolare (z.B. *stimme zu* – *lehne ab*) Rating-skalen verlockend. So könnte eine Frage lauten: *die Testräume finde ich oft (...)* und am linken Rand der Ratingskala steht *ungenügend*, am rechten ge-

eignet, die mittleren drei Kategorien bleiben unbeschriftet (bipolare Rating-skala). Die Frage ist klar formuliert, der ProbandIn ist ebenso schnell klar, worum es geht. Menold und Bogner (2015) sprechen keine Empfehlungen bezüglich einer Skalenpolarität aus, da die Effekte wenig untersucht sind, schließen allerdings aus einer Metaanalyse zu Forschungsarbeiten zu Ratingskalen den Schluss, dass „die Ergebnisse der bisherigen Forschung (...) eher für die Verwendung vollverbalisierter Ratingskalen (...)“ spricht (ebd., S. 3). Diese Erkenntnisse nicht ignorierend wurde der Fragebogen, der zunächst verschiedene Formen von Ratingskalen vorsah, überarbeitet und die Ratingskalen so gestaltet, dass alle Antwortkategorien verbalisiert sind. Dieses Vorgehen vermeidet bei bipolaren Ratingskalen mit einer Mittelkategorie (die ja durchgehend vorhanden ist) Verwirrung bei einer möglichen Beantwortung der mittleren Kategorie. In diesem Zusammenhang merken Menold und Bogner an (2015), dass nach Kaplan (1972) sowie Dubois und Burns (1975) entweder Indifferenz (weder noch) oder Ambivalenz (teils-teils) zum Ausdruck kommen könnte, dies aber in der Auswertung nicht nachvollziehbar wäre, was genau eine mittlere Antwort aus Sicht der ProbandIn verdeutlichen soll.

10. Ebenfalls um die *Compliance* zu erhöhen, wurden die Fragen einfach gestaltet und der Umfang des Fragebogens so begrenzt, dass er in maximal 20 Minuten zu beantworten ist. Die Formulierung der Fragen orientierte sich daran, ob sie präzise, ausbalanciert und allgemein verständlich (Menold & Bogner, 2015) sind. Ein Fortschrittsbalken gibt auf jeder Seite des Onlineformulars in Prozent an, wie viele Fragen bereits beantwortet worden sind. Diese Hinweise sollen die Abbruchquote verringern. Dennoch soll der Umfang des Fragebogens zugunsten der erhöhten Mitwirkungsbereitschaft nicht zu kurz gestaltet sein. Dieses von Krosnick und Alwin (1987) als *Satisficing* bezeichnete Phänomen würde die Aussage- und Auswertungskraft reduzieren.
11. Auf eine Motivation in Form eines inzwischen üblichen Preisausschreibens bei einer Teilnahme wurde verzichtet. Zum einen, da genügend Kontakte zu Schulen und SonderpädagogInnen vorhanden sind und die Annahme begründet ist, aus diesem Pool von Kontakten ausreichend ProbandInnen zu erreichen; zum anderen, da ein Preisausschreiben mit dem Hinterlegen einer Adresse verbunden wäre. Dies spräche gegen die Wahrung der Anonymität.
12. Die in der Regel geschlossenen Fragen werden weitgehend durch eine offene Frage für Anmerkungen ergänzt. Damit soll vermieden werden, dass Antwortideen außerhalb des Korsetts der geschlossenen Fragen präsentiert werden können.⁶⁶

66 Es fließen allerdings keine offenen Antworten in die Auswertung mit ein.

13. Die Instruktion erklärt offen die Fragestellung (Anwendung von Intelligenztests in der Sonderpädagogik) und die Motivation (Dissertation, Verbesserung der Diagnostik); jedoch werden negativ konnotierte Worte wie Fehler oder Schwierigkeiten vermieden. Neben der kurzen Erläuterung der Fragestellung, Hinweisen zur Anleitung und zum Datenschutz wird um die Vermeidung der *Sozialen Erwünschtheit* gebeten und für die Teilnahme gedankt.
14. Kurze Fragen zum Geschlecht und Beruf werden am Anfang gestellt, soziodemografische Daten jedoch am Ende. Diese könnten am Anfang gestellt irritierend wirken, da sie als die Zusicherung der Anonymität unterlaufend empfunden werden könnten (Pilshofer, 2001).
15. Der Fragebogen ist auf der Plattform *surveymonkeys* abrufbar. Als *Tag* wurde lediglich das Wort „Intelligenztests“ gewählt. Damit kann ausgeschlossen werden, dass zufällig eine Suchmaschine einen Treffer anzeigt, z.B. wenn eine Suchmaschine eine Anfrage für „Intelligenztests“ erhält. Es gibt so viele Treffer auf diese Anfrage, dass ein Hinweis auf die Befragung ausgeschlossen werden kann. Selbst bei der Eingabe von „Surveymonkey“ „+“ „Intelligenztests“ gab es am 18. 1. 2017 keinen Treffer⁶⁷, obwohl der Fragebogen zur Probe bereits seit Monaten online war. Diese kostenpflichtige Plattform stellt ein professionelles Design zur Verfügung. Damit sollte verhindert werden, dass ein zusammengebastelt wirkender Fragebogen die *Compliance* verringert. Allerdings wurden auf grafische Elemente und technische Spielereien durchgehend verzichtet. So fanden z.B. Tourangeau, Couper & Conrad (2004), dass Stilmittel wie farbliche Übergänge innerhalb der Rating-skalen zur Vermeidung von Extremantworten führten.

In einem Pretest wurden 41 Personen gebeten, den Fragebogen auszufüllen. Ein Feedback sollte mögliche Mängel in der Lesbarkeit, Formulierung der Fragen, nicht erkannte (Rechtschreib-)Fehler, unlogische Zusammenhänge etc. erkennen. Um eine sonderpädagogische Betriebsblindheit zu verhindern, wurden auch Personen befragt, die nicht in (sonder-)pädagogischen Zusammenhängen arbeiten bzw. mit der Anwendung von Intelligenztests nichts zu tun haben.

Alle Sprungfunktionen des online-Fragebogens wurden deaktiviert, damit sichergestellt war, dass von den ProbandInnen alle Fragen bewertet werden konnten.

In Anlehnung an Pilshofer (2001) wurde im Pretest-Verfahren um die Beantwortung folgender Fragen gebeten:

1. Wie lange hat die Bearbeitung gedauert?
2. Wird das Layout als übersichtlich empfunden?

67 Geprüft wurden die ersten Ergebnisseiten.

3. Wirkt der Fragebogen insgesamt zu lang oder in bestimmten Bereichen ermüdend?
4. Ist bei den offenen Fragen genug Platz vorgesehen zum Beantworten?
5. Fühlt man sich bei einzelnen Fragen in eine bestimmte Richtung gedrängt?
6. Sind alle Fragen verständlich?
7. Sind alle Antworten in den vorgesehenen Antwortkategorien eindeutig unterzubringen oder fehlt z.B. eine Antwortmöglichkeit – oder kommen Zweifel auf bei der Zuordnung?
8. Sind Rechtschreib- oder Grammatikfehler vorhanden?
9. Sonstige Anmerkungen.

Das Pretest-Verfahren dauerte vom 19. 1. bis zum 7.2.2017. Es wurden 24 Fragebögen ausgefüllt. Insgesamt gab es kaum Anmerkungen zur Konstruktion des Fragebogens. Einige Verbesserungsvorschläge wurden gemacht, diese mündeten in folgende Modifikationen des Fragebogens:

- Auf die Frage, in welchen Räumen die Tests durchgeführt werden (z.B. *Klassenraum, spezieller Testraum*), wurde die Antwortmöglichkeit *sonstige Räume* eingefügt mit einem Textfeld zur Beschreibung sonstiger Testräume. Damit sollte eine Antwortmöglichkeit über die vorgegebenen Vorschläge hinaus ermöglicht werden.
- Auf die Frage, welche Tests bereits durchgeführt worden sind (z.B. KABC-II, IDS usw.) wurde die Rubrik *der Test ist mir unbekannt* hinzugefügt.
- Auf die Frage, welche Tests als aussagekräftig empfunden werden (z.B. KABC-II, IDS usw.) wurde die Rubrik *keine Angabe/Test unbekannt* hinzugefügt, um Verfälschungen durch Fehlantworten auf Grund mangelnder Antwortkategorien zu vermeiden.
- Auf die Frage, welche Tests zur Verfügung stehen (sich z.B. im Testschrank befinden), wurde die Rubrik *ich weiß nicht, welche Tests vorhanden sind* hinzugefügt.
- Auf die Frage nach Schwierigkeiten im Umgang mit Durchführungsregeln (z.B. Umkehrregeln, Abbruchregeln etc.) wurde die Rubrik *weiß nicht* hinzugefügt.
- Auf die Frage, ob die Erklärungen in den Handbüchern als verständlich empfunden werden (z.B. für die KABC-II usw.) wurde die Rubrik *weiß nicht* hinzugefügt.
- Die Frage *Testergebnisse aus Intelligenztests beeinflussen Eltern in ihren Planungen und Maßnahmen* wurde hinzugefügt.
- In einer Frage zur universitären Ausbildung wurde *Konstrukte referiert* der besseren Lesbarkeit halber in *Inhalte vorgestellt* geändert.

4.2 Vorannahmen für die Auswertung der Fragebögen

Die Auswertung der Fragebögen basiert bei einigen Hypothesenprüfungen auf der Annahme, dass die Verfahren in ein- bzw. mehrdimensionale Verfahren unterteilt werden können, sowie in der Annahme, dass die Verfahren unterschiedlich komplex sind. Diese Unterteilungen sollen begründet werden.

4.2.1 Ein- versus mehrdimensionale Intelligenztests

Die Annahme, mehrdimensionale Tests sind aussagekräftiger als eindimensionale, setzt eine Zuordnung in ein- bzw. mehrdimensionale Verfahren voraus. Alle hier primär untersuchten Intelligenztests basieren auf hierarchischen Intelligenzmodellen. Entsprechend dem Drei-Schichten-Modell der Intelligenz (*Three-Stratum-Theory*; Carroll, 1993), aus dem u. a. das derzeit bedeutsame CHC-Modell der Intelligenz resultiert (siehe Abbildung 3), können Fähigkeiten auf drei Ebenen dargestellt werden (siehe Kapitel 2.3.3.5: Das CHC-Modell als integrierendes Intelligenzmodell, Abbildung 1). Eindimensionale Tests prüfen auf der untersten Ebene (*Stratum I*) mit Hilfe von Subtests Fähigkeiten, die direkt auf der obersten Ebene in den Generalfaktor der Intelligenz münden (siehe Abbildung 4). Eigentlich wird bei eindimensionalen Tests auf der *Stratum-II*-Ebene einer von mehreren *broad abilities* überprüft (in der Regel die *fluide* Intelligenz) und diese mit dem Generalfaktor gleichgesetzt. Mehrdimensionale Verfahren hingegen bieten auf der mittleren Ebene (*Stratum II*) normierte Ergebnisse an, oft *Indices* genannt. Diese sind hilfreich zur Erkennung von individuellen Stärken und Schwächen der Kinder, scheinen also geeignet für die Ableitung (sonder-)pädagogischer Maßnahmen.

Abbildung 3. Drei-Schichten-Modell der Intelligenz (angelehnt an Carroll, 1993. *g* = Generalfaktor). Str. = Stratum.

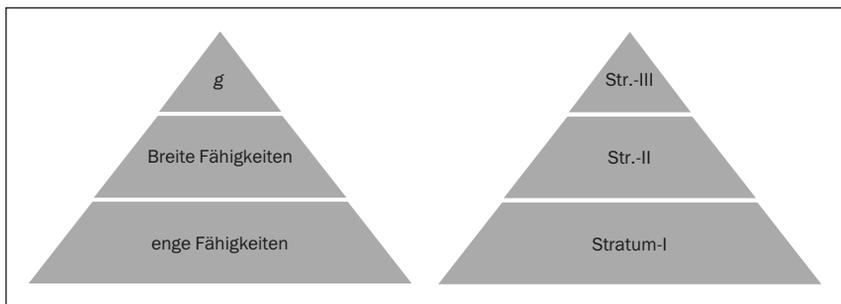
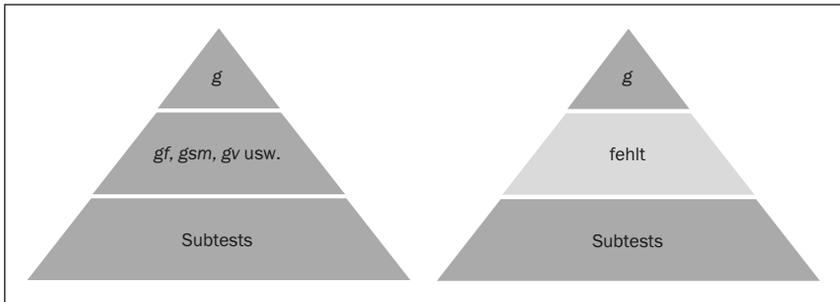


Abbildung 4. Übertragung des Drei-Schichten-Modells auf mehrdimensionale Intelligenztests (links) bzw. eindimensionale Intelligenztests (rechts), (angelehnt an Carroll, 1993). *gf* = fluide Intelligenz. *gsm* = Kurzzeitgedächtnis. *gv* = visuelle Intelligenz.



Zusammengefasst kann festgestellt werden, dass eindimensionale Tests vor allem einen Vergleich mit altersgleichen Personen ermöglichen, während darüber hinaus mehrdimensionale Tests auch intraindividuelle Analysen anbieten. Kriterium für die Einordnung zu einem mehr- bzw. eindimensionalen Test ist die Auswertungsmöglichkeit auf der *Stratum-II*-Ebene. Tabelle 10 verdeutlicht die sich daraus ergebende Zuordnung:

Tabelle 10. Zuordnung der Tests in eindimensionale- bzw. mehrdimensionale Intelligenztests.

Eindimensionale Verfahren	Mehrdimensionale Verfahren
CFT1/CFT1-R	K-ABC
CFT20-R	KABC-II
SON-R 6–40	WISC-IV
	WPPSI-III
	SON-R 2½–7

Anmerkung. CFT20-R: ohne Zahlenfolgentest und Wortschatztest, die in der Praxis häufig nicht durchgeführt werden.

Einige der primär untersuchten Testverfahren sind von dieser Zuordnung ausgenommen, was im Folgenden begründet werden soll:

Für die IDS werden laut Handbuch in dem die Intelligenz messenden Part des Tests in „aufsteigender Komplexität“ (Grob, Meyer & Hagmann-von Arx, 2009, S. 17) *Wahrnehmung, Aufmerksamkeit, Gedächtnis* und *Denken* erfasst. Obwohl diese als *Indices* im weiteren Sinne bezeichnet werden können, ist eine Auswertung dieser vier Bereiche mit Hilfe einer Normtabelle nicht möglich, so dass lediglich ein Gesamtwert ermittelt werden kann.

Gleiches gilt für den SON-R 5½–17 und die WNV. Weder für die vier übergeordneten Bereiche des SON-R 5½–17 (*Abstraktes Denken, Konkretes Denken,*

Räumliches Denken, Perzeption; Snijders, Tellegen & Laros, 1997) noch für die ausdrücklich Mehrdimensionalität postulierende WNV (z.B. *Visuell räumliche oder feinmotorische Fähigkeiten*; Petermann, 2014, S. 10) werden über die Auswertung eines Gesamtwerts und der Subtests Auswertungsmöglichkeiten auf *Stratum-II*-Ebene ermöglicht.

Wird entsprechend dem CHC-Modell ein *Indice* bzw. eine *breite Fähigkeit* postuliert, sollte diese nach Renner & Mickley (2015b) aus zumindest 2 sich unterscheidenden Subtests gebildet werden, um von einer gewissen Aussagekraft sprechen zu können. Auch dieses Kriterium ist bei keiner der drei zuletzt beschriebenen Tests erfüllt, da die Dimensionen teils nur auf einem Subtest basieren. Da die IDS, der WNV und der SON-R 5½–17 bezüglich der *Dimensionalität* weder eindeutig der einen noch der anderen Gruppe zuzuordnen sind, werden sie bei der Hypothesenprüfung nicht berücksichtigt, ob mehrdimensionale Tests als aussagekräftiger eingeschätzt werden.

4.2.2 Komplexe vs. weniger komplexe Intelligenztests

Um zu klären, ob komplexere Intelligenztests tatsächlich seltener angewendet werden, müssen die Verfahren nach ihrer *Komplexität* geordnet werden. Eine Vielzahl von Faktoren könnte die *Komplexität* definieren, z.B. die Interpretationsmöglichkeiten, die ein Test bietet, die Vielzahl von Signifikanzprüfungen auf der Ebene der *Indices* oder die Analysemöglichkeiten unter Einbezug optionaler Ergänzungstests.

Im Sinne dieser Arbeit wird die *Komplexität* eines Tests sehr reduziert definiert. Da es vor allem um Bedingungen geht, die die Durchführungsobjektivität gefährden könnten, wird die *Komplexität* an der Anzahl der Anwendungsregeln gemessen, denn es wird angenommen, dass auf Grund der Vielzahl an Regeln komplexere Tests gemieden werden, auch wenn sie zur Verfügung stünden.

Zur Bestimmung der *Komplexität* im Sinne dieser Arbeit wird die Anzahl aller Regeln und Hinweise überschlagen, die in der Eins-zu-Eins Testsituation zu beachten sind. Dies bezieht alle Subtests einer Testbatterie mit ein, unabhängig davon, ob es sich für bestimmte Altersgruppen um optionale oder Kerntests handelt. Da die Auswahl der Subtests für ein zu testendes Kind von einer Vielzahl von Faktoren abhängt (z.B. Alter, Fragestellung, Intelligenzmodell), wird eine Gewichtung nach Bedeutung der Subtests nicht vorgenommen.

Grundlage für die Zählung der Durchführungsregeln und Hinweise sind die Manuale⁶⁸ und basiert auf folgender Zählweise:

68 ausgenommen der SON-R 5 ½–17, da für diesen Test kein Manual zur Verfügung stand.

- Gezählt wurden Instruktions-, Durchführungsregeln und Hinweise. Bewertungsregeln wurden mitgezählt, wenn sie in der Testsituation beachtet werden müssen (und nicht z. B. in einer anschließenden Situation ohne Kind), da sie dann Bestandteil der Testsituation sind. Dies ist z. B. möglich, wenn eine Bewertung sofort vorgenommen werden muss, um zu ermitteln, ob eine Abbruch- oder Umkehrregel greift.
- Variieren allgemeine Regeln wie die Abbruch- oder die Umkehrregeln von Subtest zu Subtest, wurden sie jeweils addiert; sind allgemeine Regeln innerhalb eines Tests jeweils gleich für jeden Subtest anwendbar, wurden sie nur einmal gezählt.
- Spezifische Regeln, die sich von Subtest zu Subtest wiederholen, wurden nicht jedes Mal erneut gezählt, auch nicht, wenn der Wortlaut leicht variiert. So wurde z. B. davon ausgegangen, dass die in der WNV vielfach beschriebene Regel *lassen Sie dem Kind Zeit, die Bilder zu betrachten (bis zu einer Minute)* (Wechsler & Naglieri, 2006) von den TesterInnen nach spätestens zwei/drei Subtests internalisiert worden ist.
- Nicht mitgezählt wurden allgemeine Regeln zur Gestaltung der Testsituation (z. B. Testraum, Beleuchtung) oder Hinweise zum Beziehungsaufbau, der Bestimmung der Untertests etc.

Die Anzahl an Regeln und Hinweisen für die Testsituation wird nicht als exakt ermittelt postuliert. Dies wird auch damit begründet, dass die Regeln in den Manualen nicht immer kohärent und eindeutig erläutert werden und nicht immer eindeutig ist, ob ein Wortlaut ein Hinweis oder eine Erläuterung darstellt. Deshalb wird die Anzahl der Regeln als Überschlag bewertet und ein *ca.* hinzugefügt, dient also einer Orientierung zur Einordnung, siehe auch Abbildung 5 und Tabelle 11.

Gezählte Regeln:

K-ABC:	ca. 168 (16 Subtests)
KABC-II:	ca. 580 (18 Subtests)
CFT1/CFT1-R ⁶⁹ :	ca. 38 (6 Subtests)
CFT 20-R ⁷⁰ :	ca. 58 (10 Subtests)
CFT 20-R ⁷¹ :	ca. 47 (8 Subtests)
WISC-IV:	ca. 406 (15 Subtests)
WPPSI-III:	ca. 322 (14 Subtests)

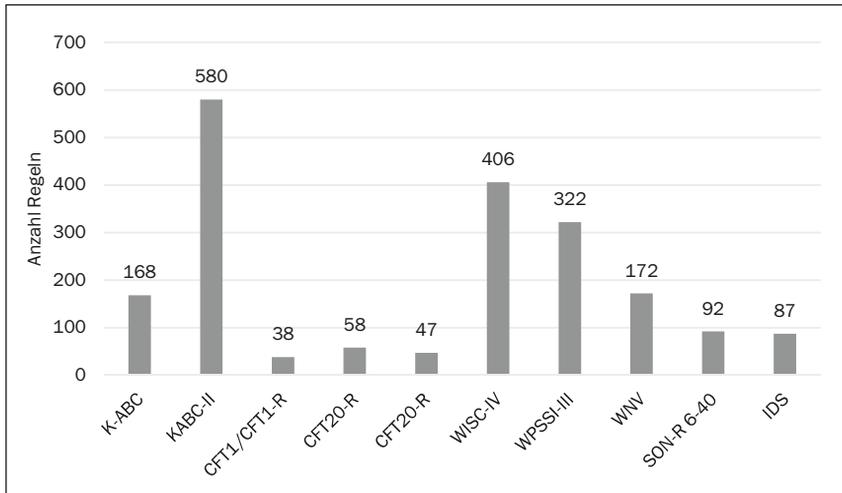
69 Grundlage: CFT1-R.

70 Inkl. *Zahlenfolgentest* und *Wortschatztest* (kristalline Intelligenz), es wird allerdings angenommen, dass deren Anwendung in der Praxis selten ist und eher der (fluide) Teil durchgeführt wird, für den die Tests der CFT-Reihe in der Regel genutzt werden.

71 Ohne *Zahlenfolgentest* und *Wortschatztest*.

WNV: ca. 172 (6 Subtests)
 SON-R 6–40: ca. 92 (4 Subtests)
 IDS⁷²: ca. 87 (7 Subtests)

Abbildung 5. Anzahl der Regeln und Hinweise der Intelligenztests.



Aus diesem Ergebnis resultiert folgende Kategorisierung:

Tabelle 11. Zuordnung der Tests nach Komplexität.

wenig komplex	leicht komplex	komplex	sehr komplex	außerordentlich komplex
CFT1/CFT1-R CFT20-R	SON-R 6–40 IDS	K-ABC WNV	WISC-IV WPPSI-III	KABC-II

4.2.3 Konstruktion eines Schwierigkeiten-Index

Für die Hypothesenprüfung ist die Ableitung eines *Schwierigkeiten-Index* sinnvoll. In diesem Index sollen Schwierigkeiten bei der Durchführung eines Intelligenztests zusammengefasst mit einem Kennwert erfasst werden, z. B. *Umkehr-*

72 Grundlage: Kognitive Entwicklung (Intelligenz), nicht weitere Entwicklungsbereiche wie Motorik, Sprache etc.

*regeln bereiten mir Schwierigkeiten*⁷³ (Q15/1⁷⁴) oder *ich wurde von einem Telefonanruf während der Testsituation gestört* (Q18/1). Mit Hilfe dieses Kennwerts könnte z. B. geprüft werden, ob ältere oder jüngere oder nordrhein-westfälische oder niedersächsische SonderpädagogInnen signifikant mehr oder weniger Schwierigkeiten benennen bei der Anwendung von Intelligenztests. Vor der Hypothesenprüfung soll an dieser Stelle die für die Prüfungen notwendige Skalenkonstruktion beschrieben werden. Inhaltliche Überlegungen führten zu einer Vorauswahl von 20 Items, die für eine erste Sichtung für die Bildung des Index geeignet schienen. Zur Prüfung wurde Cronbachs Alpha (Cronbach, 1951) gewählt. Alternativ böte sich auch eine Faktorenanalyse an, die auf Grund der verschiedenen Item-Merkmale (z. B. dichotome vs. nominal skalierte Merkmale) verworfen wurde. Die Vorauswahl beschränkt sich auf diejenigen Items mit einer fünfstufigen Likert-Skala. So wurden die Vergleichbarkeit und Konsistenz der Daten als geeigneter eingeschätzt. Grundlage der Indexbildung waren die vorliegenden Daten nach Abschluss der Erhebung, genauere Angaben zur Erhebung werden an späterer Stelle erläutert.

Das Verfahren Cronbachs Alpha (auch Alpha) ist vieldiskutiert, im Kern häufig bezüglich der empfohlenen Höhe der Werte, die als akzeptabel genannt werden, aber auch bezüglich der Manipulationsmöglichkeiten zur Erreichung (scheinbar) akzeptabler Werte. Bortz und Döring (2006) beschreiben z. B. die Möglichkeit über die Erhöhung der Anzahl von Items immer höhere Koeffizienten nach Cronbach zu erzielen, so dass ab einer gewissen Anzahl der Items zwangsläufig ein akzeptabler Wert erreicht wird, obwohl die Items wenig im Zusammenhang stehen.

Doch dies ist das Ziel: mit Hilfe von Cronbachs Alpha die interne Konsistenz mehrerer Variablen zu prüfen und in einem Index zu integrieren und somit den Zusammenhang zu belegen.

Die für das *Schwierigkeiten-Index* ausgewählten Items Q13, Q14, Q15, Q17, Q18 und teils Q28 werden aus Gründen der Nachvollziehbarkeit aufgeführt:

1. *Kommt es vor, dass einige Ihrer Intelligenztests nicht zur Verfügung stehen (z. B. ausgeliehen sind etc.)* (Q13/1)?
2. *Kommt es vor, dass die Testmaterialien unvollständig sind (z. B. fehlende Puzzleteile)* (Q13/2)?
3. *Kommt es vor, dass Formulare/Arbeitsbögen fehlen* (Q13/3)?
4. *Welche dieser Veränderungen haben Sie schon einmal vorgenommen: Durchführungszeiten geändert (z. B. nach Ablauf der regulären Durchführungszeit/Item einen Punkt gegeben bei richtiger Antwort)* (Q14/1)?

73 Sinngemäß umformuliert.

74 Angaben wie „Q15/1“ beziehen sich auf die Frage des Fragebogens. Q15/1 = Frage 15, erste Unterfrage.

5. *Welche dieser Veränderungen haben Sie schon einmal vorgenommen: Durchführungszeit ganz weggelassen (Q14/2)?*
6. *Welche dieser Veränderungen haben Sie schon einmal vorgenommen: Rückmeldungen gegeben, wenn diese nicht vorgesehen waren (z.B. richtig oder hast du richtig gelöst) (Q14/3)?*
7. *Folgende Durchführungsregeln bereiten mir Schwierigkeiten: Umkehrregeln (Q15/1).*
8. *Folgende Durchführungsregeln bereiten mir Schwierigkeiten: Abbruchregeln (Q15/2).*
9. *Folgende Durchführungsregeln bereiten mir Schwierigkeiten: Ausrechnen des Testalters (Q15/3).*
10. *Sind die Testräume geeignet? (Q17)?*
11. *Wurden Sie schon mal während einer Testung gestört: durch einen Telefonanruf (Q18/1)?*
12. *Wurden Sie schon mal während einer Testung gestört durch eine Person, die an die Tür geklopft hat (Q18/2)?*
13. *Wurden Sie schon mal während einer Testung gestört durch eine Person, die den Raum betreten hat (Q18/3)?*
14. *Wurden Sie schon mal während einer Testung gestört durch Geräusche (z.B. spielende Kinder, Verkehrslärm) (Q18/4)?*
15. *Wurden Sie schon mal während einer Testung gestört durch Lautsprecherdurchsagen (Q18/5)?*
16. *Bitte bewerten Sie folgende Aussagen: Im Rahmen meiner Arbeit steht mir zu wenig Vorbereitungszeit für das Lernen eines normierten Tests zur Verfügung (Q28/1).*
17. *Bitte bewerten Sie folgende Aussagen: Heutzutage habe ich weniger Zeit für die Anwendung eines Intelligenztests (Q28/2).*
18. *Bitte bewerten Sie folgende Aussagen: Heutzutage habe ich weniger Zeit für die Durchführung eines sonderpädagogischen Gutachtens (Q28/3).*
19. *Bitte bewerten Sie folgende Aussagen: Steht mir nicht genügend Vorbereitungszeit zur Verfügung für einen Test, bereite ich mich in der Freizeit vor (Q28/4).*
20. *Bitte bewerten Sie folgende Aussagen: Mir fällt die Durchführung von Intelligenztests leicht (Q28/5).*

Im ersten Schritt ist die Polung (negativ/positiv) angepasst (Q17, Q28/5), im zweiten Schritt sind die Fragengruppen (Q13, Q14, Q15: jeweils 3 Items; Q18, Q19: jeweils 5 Items) nach Cronbachs Alpha geprüft worden⁷⁵.

⁷⁵ Q17 besteht nur aus einem Item, deshalb wäre eine Gruppenprüfung sinnlos.

In einer ersten Zwischenbilanz wird die Fragengruppe Q14 als kritisch betrachtet (siehe Tabelle 12). Auch inhaltlich ist der Einbezug der drei Items zu Q14 nicht zwingend. Unter Entfernung weiterer kritischer Items (Q15/3, Q18/5, Q28/5) konnte der höchste Alpha Koeffizient von .773, basierend auf 14 Items ermittelt werden (siehe Tabelle 13).

Tabelle 12. Konstruktion Schwierigkeiten-Index: Prüfung nach Cronbachs Alpha für Fragengruppen.

Fragengruppe	Anzahl Items	Alpha	Cronbachs Alpha, wenn ein Item weggelassen
Q13	3	.732	.725
Q14	3	.573	.590
Q15	3	.669	.813
Q18	5	.687	.711
Q28	5	.710	.748

Anmerkungen. Q13: Schwierigkeiten durch fehlende Materialien. Q14: unerlaubte Veränderungen. Q15: Schwierigkeiten mit Regeln. Q18: Störungen in Testsituation. Q28: Einstellungen zur Anwendung von Tests.

Tabelle 13. Konstruktion Schwierigkeiten-Index: Prüfung nach Cronbachs Alpha nach einer ersten Bereinigung.

Item:	Skalenmittelwert, wenn Item weggelassen	Skalenvarianz, wenn Item weggelassen	Korrigierte Item-Skala-Korrelation	Cronbachs Alpha, wenn Item weggelassen
Q13/1	43.6856	37.387	.368	.762
Q13/2	42.8790	37.893	.400	.759
Q13/3	43.1207	37.287	.402	.758
Q15/1	43.3473	37.905	.312	.767
Q15/2	43.0045	38.232	.339	.764
Q17	43.4502	37.847	.414	.758
Q18/1	42.3803	40.653	.211	.772
Q18/2	43.3249	38.227	.407	.759
Q18/3	43.1956	37.922	.418	.758
Q18/4	44.0527	37.492	.408	.758
Q28/1	44.9579	35.588	.432	.755
Q28/2	44.4433	33.968	.514	.746
Q28/3	44.7424	34.471	.477	.751
Q28/4	45.7020	39.374	.293	.767

Anmerkungen. Q13: Schwierigkeiten durch fehlende Materialien. Q15: Schwierigkeiten mit Regeln. Q17: Eignung der Testräume. Q18: Störungen in Testsituation. Q28: Einstellungen zur Anwendung von Tests.

Da bei einer erhöhten Anzahl von Items der Koeffizient auch ohne Zusammenhang zwischen den Items steigen kann, würde ein höherer Koeffizient bei weniger Items für den tatsächlichen Zusammenhang und somit für die Qualität des ermittelten Index sprechen, dies war der Fall nach Wegnahme der Fragen Q15/3, Q18/5, Q28/5. Dennoch sollen die statistischen Kennzahlen nicht kritiklos übernommen werden, ansonsten bestünde auch hier die Gefahr, dass „der Koeffizient in gut und böse, ohne Verstand, aber mit dem Taschenrechner gehorsam abgehakt, befolgt und verfolgt werde“ (Kersting, 2006, S. 248). Zur endgültigen Bestimmung der Items, die zu dem *Schwierigkeiten-Index* gehören werden, soll eine abschließende inhaltliche Betrachtung vorgenommen werden. Folgende Fragen mit eigentlich akzeptablem Koeffizienten sind entfernt worden, da sie zu weit entfernt von dem Konstrukt *Schwierigkeiten* scheinen:

- *Steht mir nicht genügend Vorbereitungszeit zur Verfügung für einen Test, bereite ich mich in der Freizeit vor (Q28/4).*
- *Mir fällt die Durchführung von Intelligenztests leicht (Q28/5).*

Folgende Fragen scheinen jedoch gut Schwierigkeiten im Rahmen der Anwendung von Intelligenztests abzubilden, werden z. B. als Schwierigkeiten bei Schulungen zu Intelligenztests benannt und werden deshalb zunächst aufgenommen in den Index, auch wenn Alpha sich etwas verringert:

- *Folgende Durchführungsregeln bereiten mir Schwierigkeiten: Ausrechnen des Testalters (Q15/3).*
- *Wurden Sie schon mal während einer Testung gestört durch Lautsprecherdurchsagen (Q18/5)?*

Sowohl inhaltlich als auch unter Berücksichtigung akzeptabler Alpha-Werte ergäbe sich daraus eine Skala bestehend aus 14 Items. Leider sind die Hinweise zu der korrigierten Item-Skala-Korrelation (Trennschärfe) nicht für alle Items befriedigend, da unter .3.

Entsprechend der Empfehlung von Hemmerich (2015a, ohne Seitenangabe) sollten „Items mit einer Trennschärfe unter .3 verworfen (...) werden“. Zugunsten einer gültigen Skala und unter Ausschluss vermeintlich inhaltlich passender Items besteht der endgültige *Schwierigkeiten-Index* nun aus den aus Tabelle 14 ersichtlichen 11 Items.

Aus Gründen der Nachvollziehbarkeit werden in Tabelle 14 die Fragen wiederholt, der nun ermittelte Alpha nach Cronbach beträgt .743 ($N = 1074$).

Tabelle 14. Items des Schwierigkeiten-Index.

	Frage	Alpha, wenn Item weggelassen	Korr. Item-Skala-Korrelation
Q13/1	Wenn Sie testen möchten, kommt es vor, dass einige Ihrer Intelligenztests nicht zur Verfügung stehen (z. B. ausgeliehen sind etc.)?	.726	.377
Q13/2	Wenn Sie testen möchten, kommt es vor, dass die Testmaterialien unvollständig sind (z. B. fehlende Puzzleteile)?	.720	.429
Q13/3	Wenn Sie testen möchten, kommt es vor, dass Formulare/Arbeitsbögen fehlen?	.719	.429
Q15/1	Folgende Durchführungsregeln bereiten mir Schwierigkeiten: Umkehrregeln	.734	.318
Q15/2	Folgende Durchführungsregeln bereiten mir Schwierigkeiten: Abbruchregeln	.727	.353
Q17	Sind die Testräume geeignet?	.720	.428
Q18/2	Wurden Sie schon mal während einer Testung gestört ... durch eine Person, die an die Tür geklopft hat?	.720	.435
Q18/3	Wurden Sie schon mal während einer Testung gestört ... durch eine Person, die den Raum betreten hat?	.719	.440
Q18/4	Wurden Sie schon mal während einer Testung gestört ... durch Geräusche (z. B. spielende Kinder, Verkehrslärm)?	.722	.408
Q28/1	Bitte bewerten Sie folgende Aussagen: Im Rahmen meiner Arbeit steht mir zu wenig Vorbereitungszeit für das Lernen eines normierten Tests zur Verfügung.	.731	.354
Q28/2	Bitte bewerten Sie folgende Aussagen: Heutzutage habe ich weniger Zeit für die Anwendung eines Intelligenztests.	.727	.387

Anmerkung. Korr. = korrigierte.

4.3 Analyse von ausgewerteten Intelligenztestformularen

Neben dem Fragebogen werden Intelligenztestformulare analysiert, um insbesondere zu klären, ob tatsächlich Mängel in der Durchführungs- und Auswertungsobjektivität bei der Anwendung der Tests im sonderpädagogischen Alltag vorliegen. Die Testformulare sind zur Prüfung von mehreren Schulämtern zugesandt worden. Die Formulare sind während der Testanwendungen zum Zwecke der Dokumentation entstanden. Die Intelligenztests wurden ausschließlich von SonderpädagogInnen im Rahmen einer Prüfung durchgeführt, ob sonderpädagogischer Förderbedarf vorliegt oder nicht, somit war die Anwendung der Intelligenztests ein Hilfsmittel während der Gutachtenerstellung.

Im Zentrum des Interesses steht eine Auszählung entdeckter Fehler bei der Auswertung oder Durchführung von Intelligenztests. Es können selbstverständlich nur aus den Aufzeichnungen ableitbare Fehler dokumentiert werden. Nicht

nachvollziehbare Mängel während der Testdurchführung durch z.B. zu weit gehende Erläuterungen der Instruktionen oder durch das Geben von Rückmeldungen während der Testsituation, wenn dies nicht vorgesehen ist, entfallen ebenso wie Analysen darüber, welche genauen Veränderungen bei den Gesamtwerten entstehen würden ohne die jeweils entdeckten Fehler. Letzteres wäre auch nicht möglich, da die aus den Fehlern resultierenden Veränderungen nicht eindeutig zu bestimmen sind. Würde z.B. das Anfangsitem in einem Subtest falsch bestimmt werden und somit die Durchführung von Aufgaben durch das Kind entfallen, könnte im Nachhinein keine Bewertung für irrtümlich ausgelassene Items vermutet werden. Ohne diese genaue Bewertung kann aber kein Rohwert und somit kein exakter Abgleich mit einer Normstichprobe stattfinden. Dennoch ergeben sich zuweilen aus den Analysen der Testformulare Hinweise über veränderte Testergebnisse bei einer korrekten Durchführung.

Für die Analyse der Formulare wurden die Testprotokolle zunächst sortiert und mit einem Deckblatt versehen, auf dem übersichtlich alle nach der Durchsicht entstandenen Daten eingetragen worden sind. Bei Bedarf wurden Namen mit einem Stift geschwärzt. Die Formulare wurden mit Hilfe der Testmanuale gesichtet. Alle zehn Tests, die für die Auszählung in Frage kommen, sind dem Autoren aus eigener Praxis bekannt und für alle Tests lagen die vollständigen Testmaterialien vor. Geprüft sind die Formulare auf folgende Kriterien:

- *Alter*: prüft, ob das Alter am Testtag richtig ausgerechnet worden ist. Wäre dies nicht der Fall, bestünde die Gefahr, die Rohwerte mit der falschen Altersnormstichprobe zu vergleichen. Da beim Berechnen des Testalters rigide Vorgaben eingehalten werden müssen und weder auf- noch abgerundet werden darf, ist das Berechnen des Testalters fehleranfällig.
- *Test*: welcher Test wurde verwendet.
- *Computerauswertung*: wurden die Ergebnisse manuell oder mit Hilfe eines PC-Programms berechnet.
- *Punkte/Auswertung*: prüft, ob die Punkte richtig gezählt worden sind. Da eine falsche Bewertung zu falsch gezählten Punkten führen kann, diese beiden Aspekte sich also bedingen, sind beide Aspekte zu einem Kriterium zusammengefasst worden.
- *Umkehrregel*: liegt ein Anfangsitem nicht bei Aufgabe 1 und hat das Kind bei einem der altersentsprechenden Anfangsitems einen Fehler gemacht, muss entsprechend der je nach Test unterschiedlichen Umkehrregeln zu den einfacheren Aufgaben umgekehrt werden.
- *Abbruchregel*: Jeder Test, gelegentlich sogar Subtests innerhalb einer Testbatterie haben unterschiedliche Abbruchregeln. Diese bestimmen, nach wie vielen Fehlern in Folge der Subtest abgebrochen werden muss.
- *Anfangsitem*: prüft, ob das altersentsprechende Anfangsitem gewählt wurde. Bei adaptiven Testverfahren (z.B. SON-R 6–40) werden Anfangsitems auch

abhängig von den Leistungen vorheriger Testaufgaben bestimmt. Die Leistung des Kinds in einem Durchgang wird somit adaptiert für den nächsten Durchgang, um eine evtl. Unter- bzw. Überforderung zu vermeiden.

- *Gesamtfehler*: Gesamtheit aller oben beschriebenen fünf Fehlerkategorien.
- *Auswirkungen*: es ist in einigen Fällen möglich, Aussagen darüber zu machen, ob gefundene Fehler zu einem veränderten Ergebnis geführt hätten. Ab einem gefundenen Fehler/ProbandIn werden Aussagen gemacht über entweder mögliche Auswirkungen auf das Testergebnis oder keine Auswirkungen. In einigen Fällen ist es möglich, dass sich die Testergebnisse bei korrekter Auswertung teils über eine Standardabweichung hinaus verändern würden. Unabhängig von den statistischen Berechnungen werden Beispiele an späterer Stelle qualitativ vorgestellt.

Die Formulare sollen auf die Anzahl gemachter Fehler und Fehlerarten untersucht werden. Geprüft werden ausschließlich Formulare, die folgenden Kriterien entsprechen:

- es handelt sich um einen dieser Intelligenztests: CFT-Reihe, WNV, KABC-II, K-ABC, SON-R 5½-17, SON-R 6-40, WPPSI-III, WISC-IV, IDS,
- die Kopien sind anonymisiert (z.B. Abkleben der Namen; Überschreiben mit einem schwarzen Stift etc.),
- der Intelligenztest wurde von einer Sonderpädagogin bzw. einem Sonderpädagogen durchgeführt,
- der Intelligenztest wurde im Rahmen zur Feststellung sonderpädagogischen Unterstützungsbedarfs durchgeführt,
- das Formular ist vollständig kopiert (wenn möglich inkl. evtl. Zusatzbögen).

Angeschrieben worden sind aus dem beruflichen Kontext bekannte SchulrätInnen bzw. Einzelpersonen. Die beteiligten Einrichtungen erhalten als Gegenleistung für eine zukünftige Fortbildung vergünstigte Konditionen und nach Abschluss dieser Untersuchung eine schriftliche Rückmeldung über Auffälligkeiten und entdeckte Fehler bzw. Fehlerarten für die Umsetzung gezielter Verbesserungen. So kann beispielsweise in einem Schulkreis darauf hingewiesen werden, dass beim SON-R 6-40 überproportional häufig das Anfangsitem falsch bestimmt worden ist bzw. beim Subtest *Wahrnehmung Visuell* der IDS häufig die Auswertungsregeln falsch angewendet worden sind.

Allen Einrichtungen wurde ein anonymes Vorgehen zugesichert, dies beinhaltet keine Nennung des jeweiligen Schulamts.

4.4 Beschreibung der Stichprobe: Fragebogen

Insgesamt 6 339 E-Mails mit der Bitte um Teilnahme an der Online-Befragung wurden versendet an ehemalige TeilnehmerInnen von Fortbildungen zu Intelligenztests. Überwiegend nahmen SonderpädagogInnen an den Fortbildungen teil, seltener PsychologInnen und andere Berufsgruppen. Hinzu kommen ca. 85⁷⁶ versendete E-Mails an Schulleitungen von Schulen, in denen Inhouse-Seminare stattfanden mit der Bitte um Weiterleitung an das Kollegium. Da nicht bekannt ist, wie viele Weiterleitungen vorgenommen worden sind, kann eine genaue Rücklaufquote nicht ermittelt werden.

Es lagen bis zum Februar 2018 1 323 ausgefüllte Fragebögen vor. Danach wurde das Portal für den Fragebogen geschlossen. Ca. zwei Drittel der Fragebögen wurden im April und Mai 2017 ausgefüllt. Aus strategischen Gründen wurden einige E-Mails mit der Bitte um Teilnahme erst nach einer Sichtung der ersten Daten versendet. So sollte die Möglichkeit erhalten bleiben, auch nach einer möglichen Entdeckung methodischer Mängel für einen modifizierten Fragebogen weitere potentielle TeilnehmerInnen anschreiben zu können. Da weder Mängel noch andere Probleme nach Sichtung der ersten Daten festgestellt werden konnten, wurden weitere E-Mails mit der Bitte um Teilnahme versendet, so dass ca. ein letztes Drittel der Fragebögen im Oktober und November 2017 ausgefüllt worden sind.

Von den 1 323 ausgefüllten Fragebögen sind einige von der Auswertung ausgeschlossen worden:

- Alle Fragebögen, die auf die Frage nach der Profession nicht mit Sonderpädagoge/Sonderpädagogin geantwortet haben.
- Alle Fragebögen, die nicht komplett beantwortet worden sind. Da die Anzahl der ProbandInnen als ausreichend groß eingeschätzt wird, konnte auf die unvollständig ausgefüllten Fragebögen verzichtet werden. Interessante Rückschlüsse, die gelegentlich über die Abbruchquote bei Studien gezogen werden könnten (z. B. Studien über Impulsivität), sind im Rahmen dieser Studie nicht zu erwarten und also ohne Relevanz.
- Fragebögen, die bei der Beantwortung der Frage nach dem Geschlecht nicht eindeutig mit Mann oder Frau geantwortet haben. Die in den ersten Tagen nach der Veröffentlichung des Fragebogens bestehende Antwortmöglichkeit *andere Angaben zum Geschlecht* führte bei einigen Teilnehmenden offensichtlich zur Verwirrung und wurde mit Antworten wie *Sonderschullehrer, sehr engagiert, Lehrerin und Psychologin, kinderlieb und ehrgeizig* und weiteren beantwortet. Sehr zügig wurde das konservative Antwortformat bei der

76 „Ca.“ deshalb, da eine doppelte Anschrift nicht ausgeschlossen werden kann.

Frage nach dem Geschlecht *Mann* bzw. *Frau* gewählt und weitere Antwortmöglichkeiten herausgenommen. Eine Verzerrung durch diese Veränderung im Fragebogen nach wenigen Tagen wird ausgeschlossen, zumal die nicht eindeutigen Antworten (s.o.) aus der Auswertung herausgenommen worden sind.

- Fragebögen ohne Altersangaben, da für die entsprechenden ProbandInnen keine Gewichtung möglich ist,
- Alle Fragebögen, die nach dem Februar 2018 beantwortet worden sind.

In die Auswertung sind 1077 vollständig ausgefüllte und ausschließlich von SonderpädagogInnen ausgefüllte Fragebögen übernommen worden, davon 943 Sonderpädagoginnen und 134 Sonderpädagogen. Das Durchschnittsalter der Gesamtstichprobe beträgt gerundet 45 Jahre, die Standardabweichung 9.27 Jahre. Die jüngste Teilnehmerin ist 26 Jahre, die älteste Teilnehmerin 66 Jahre alt. 108 der 1077 ProbandInnen nahmen niemals an einer Fortbildung zu standardisierten Verfahren teil, diese Personengruppe diente als Kontrollgruppe. Durch eine Anpassung der Gewichtungsfaktoren (siehe Kapitel 5.1) für die ersten sieben Fragestellungen verringert sich die Gesamtfallzahl auf 1037. Unberücksichtigt blieben die ProbandInnen, die weder der Versuchs- noch der Kontrollgruppe eindeutig zuzuordnen waren (siehe Tabelle 16).

4.5 Beschreibung der Stichprobe: Formularprüfung

Aus sechs Schulämtern bzw. Schulberatungszentren sind 271 Intelligenztest-Formulare zur Verfügung gestellt worden. Hinzu kommen Formulare von fünf Einzelpersonen. Die Formulare sollten untersucht werden auf Anzahl gemachter Fehler, Fehlerarten und Hinweise auf die verwendeten Testverfahren geben. Geprüft worden sind ausschließlich Formulare, die den in Kapitel 4.3 beschriebenen Kriterien entsprachen.

28 Formulare fließen nicht in die Auswertung mit ein, vor allem auf Grund unvollständiger Angaben. Insgesamt wurden 248 Formulare überprüft.

Allen Einrichtungen wurde ein anonymes Vorgehen zugesichert (siehe Anschreiben/Datenschutzerklärung im Online-Material), dies beinhaltet keine Nennung des jeweiligen Schulamts. Vier beteiligte Schulämter befinden sich in Nordrhein-Westfalen, eines in Brandenburg und eine Einrichtung ist ein regionales Bildungs- und Beratungszentrum (ReBBZ) in Hamburg, vergleichbar mit einem Schulamt.

5 Ergebnisse

Die Darstellung der Ergebnisse unterteilt sich in die Auswertung der Fragebögen, sowie in der Analyse der Testformulare, die auf Fehler überprüft worden sind.

Für beide Forschungszweige werden die Ergebnisse zunächst deskriptivstatistisch und zusammenfassend vorgestellt. Schwerpunkt dieses Kapitels wird jedoch die anschließende Vorstellung der Ergebnisse aus den inferenzstatistischen Berechnungen bzw. Hypothesenprüfungen sein.

Für die Auswertung der Fragebögen wurden die Daten von dem Server *surveymonkey* zur Datenerfassung heruntergeladen und über das Programm Excel in SPSS konvertiert. Das Erstellen einer Datenmaske als Grundlage für die Berechnungen wurde als fehleranfällig angenommen. Deshalb wurde besondere Sorgfalt verwendet und es wurden mehrere Kontrollen vorgenommen, damit z. B. ein Verrutschen von Zeilen nicht zu invaliden Ergebnissen führt. Zunächst sind alle Rohdaten in numerische Daten umgewandelt worden. Mehrere Ergebniszeilen sind danach mehrfach abgeglichen worden mit den heruntergeladenen Rohdaten, um ein Verrutschen der Zeilen zu verhindern. Nachdem die fertige Datenmaske vorlag, wurden im Rahmen einer letzten Kontrolle erneut die Rohdaten vom Server heruntergeladen und mit der Datenmaske in Stichproben abgeglichen. Dabei wurden besonders die letzten ProbandInnen geprüft (ab ProbandIn 1050), da bei fehlerfreien Daten davon ausgegangen werden konnte, dass davor keine Zeile verrutscht war, die Werteangaben in der Variablenansicht von SPSS nicht fehlerhaft und auch die Zuordnungen zu den Fragen fehlerfrei sind.

Lediglich für eine Plausibilitätsprüfung der Daten wurde eine Interkorrelationsmatrix erstellt, auch um explorativ Zusammenhänge verschiedener Variablen zu prüfen.

Für die Interkorrelationsmatrix sind alle Variablen nach den Skalenniveaus sortiert und gruppiert worden, metrische und ordinale Daten sind zusammengefügt, nominale Skalen wurden dummykodiert. Bei metrischen und ordinalen Skalen wurde mit der Produktmomentkorrelation nach Pearson geprüft, ob Zusammenhänge bestehen. Bei der Kombination binär/binär wurden Phi-Koeffizienten sowie χ^2 - bzw. Fisher-Tests berechnet. Für die Kombination binär/metrisch wurden punktbiseriale Korrelationskoeffizienten berechnet. Nicht nachvollziehbare Ergebnisse konnten nicht identifiziert werden, die evtl. eher für eine schlechte Datenqualität als für eine plausible Ableitung von Hypothesen sprechen könnten.

Der Mittelwert der Gewichtungen ist nicht gleich 1, deshalb konstruiert SPSS 24 etwas abweichende Fallzahlen für eine gültige Auswertung. Wenn nicht anders erwähnt, basieren die folgenden Ergebnisse aus gewichteten Daten.

5.1 Gewichtungen

Bei der Fragebogenauswertung sind die Daten nach Geschlecht und Alter entsprechend den Angaben des statistischen Bundesamtes zur Schulstatistik des Schuljahres 2016/2017 (Destatis, 2017) gewichtet worden.

Für die Gewichtungen sind die Altersgruppen des statistischen Bundesamts übernommen worden. Tabelle 15 stellt dar, wie die 2017 erfassten berufstätigen SonderpädagogInnen alters- und geschlechtsverteilt waren.

Tabelle 15. Teil- und vollzeitbeschäftigte SonderpädagogInnen im Schuljahr 2016/2017 (Destatis, 2017).

Alter	gesamt	< 30 Jahre	30–34 Jahre	35–39 Jahre	40–44 Jahre	45–49 Jahre	50–54 Jahre	55–59 Jahre	60–64 Jahre	>64 Jahre
gesamt	68 134	4 166	8 442	7 962	9 121	8 833	9 282	11 006	8 956	357
männl.	15 485	546	1 741	1 864	2 101	2 151	1 895	2 592	2 442	153
weibl.	52 649	3 620	6 701	6 098	7 020	6 682	7 387	8 414	6 514	204

Für die deskriptivstatistischen Auswertungen sind die Gewichtungen nach Alter und Geschlecht verwendet worden ($N = 1077$).

Für die inferenzstatistischen Auswertungen sind die Gewichtungen angepasst worden. Da vielfach zwischen den Ergebnissen der Kontroll- und der Versuchsgruppe unterschieden wird, bleiben alle ProbandInnen unberücksichtigt, die nicht eindeutig zuzuordnen sind.

Die Kontrollgruppe besteht aus den ProbandInnen, die noch nicht an einer Fortbildung zu Intelligenztests teilgenommen haben (Q25), also weder als besonders testaffin angenommen werden noch mit dem Untersucher in Kontakt gekommen sind. Deshalb wird die Kontrollgruppe als repräsentativer für die Gesamtheit angenommen, die Gewichtungen erhöhen somit zusätzlich diese Annahme. Die Gesamtstichprobe verringert sich um 40 ProbandInnen, da diese auf die Frage Q25 (an Fortbildungen teilgenommen) nicht antworteten, die Antwort für die Bestimmung der Versuchs- bzw. Kontrollgruppe allerdings notwendig war. Nach dieser Bereinigung betrug $N = 1037$.

Tabelle 16 stellt die aus diesen Daten resultierenden Gewichtungsfaktoren bei den Berechnungen dar.

Tabelle 16. Anzahl ProbandInnen (N) und daraus resultierende Gewichtungsfaktoren (GF).

Alter		< 30 Jahre	30–34 Jahre	35–39 Jahre	40–44 Jahre	45–49 Jahre	50–54 Jahre	55–59 Jahre	60–64 Jahre	> 64 Jahre
Gewichtung	männl.	N: 3 GF: 2.88	18 1.53	20 1.47	24 1.38	24 1.42	16 1.87	18 2.28	11 3.51	n. v.
	weibl.	N: 38 GF: 1.501	117 0.91	143 0.67	201 0.55	156 0.68	120 0.97	111 1.20	56 1.84	1 3.22
Versuchs- gruppe	männl.	N: 1 GF: 7.44	14 1.70	19 1.34	18 1.59	24 1.22	15 1.72	15 2.36	10 3.33	n. v.
	weibl.	N: 21 GF: 2.35	99 0.92	120 0.69	175 0.55	143 0.64	103 0.98	99 1.16	52 1.71	1 2.78
Kontroll- gruppe	männl.	N: 2 GF: 0.43	4 0.69	1 2.95	4 0.83	n. v.	n. v.	3 1.34	1 3.87	n. v.
	weibl.	N: 17 GF: 0.34	14 0.76	19 0.51	17 0.65	10 1.06	7 1.67	6 2.22	3 3.44	n. v.

Anmerkung. Obere zwei Zeilen: Gewichtung nach Alter und Geschlecht. Untere vier Zeilen: Gewichtung nach Alter und Geschlecht, angepasst an die Möglichkeit der Unterscheidung von Versuchs- und Kontrollgruppe. n. v. = nicht vorhanden.

5.2 Gesamt-, Versuchs- und Kontrollgruppe

Die Auswahl der ProbandInnen für die Beantwortung des Fragebogens resultiert überwiegend aus Anfragen an ehemalige TeilnehmerInnen von Diagnostikseminaren. Verzerrungen aus einer selektiven Stichprobe sollen durch den Vergleich mit einer Kontrollgruppe umgangen werden. Diese besteht aus ProbandInnen, die niemals an einer außeruniversitären Fortbildung zur Testdiagnostik teilgenommen haben und entsprechend weder von der Person beeinflusst sind, die die Fortbildung durchgeführt hat (in der Regel der Autor dieser Arbeit) noch von den Inhalten der Fortbildung oder den Motiven für die Teilnahme an der Fortbildung (z. B. Vertiefung in die KABC-II).

Somit kann angenommen werden, dass die ProbandInnen der Kontrollgruppe repräsentativer für die Gesamtheit der SonderpädagogInnen stehen. Die Unterscheidung wird relevant ab der inferenzstatistischen Auswertung.

Es gilt folgende Begriffsbestimmung für die inferenzstatistischen Auswertungen:

Gesamtgruppe: Alle ProbandInnen ($N = 1037$)

Versuchsgruppe: alle ProbandInnen, die an einer außeruniversitären Fortbildung teilgenommen haben ($N = 929$)

Kontrollgruppe: alle ProbandInnen, die nicht an einer außeruniversitären Fortbildung teilgenommen haben ($N = 108$).

5.3 Deskriptivstatistische Auswertung

5.3.1 Auswertung Fragebögen

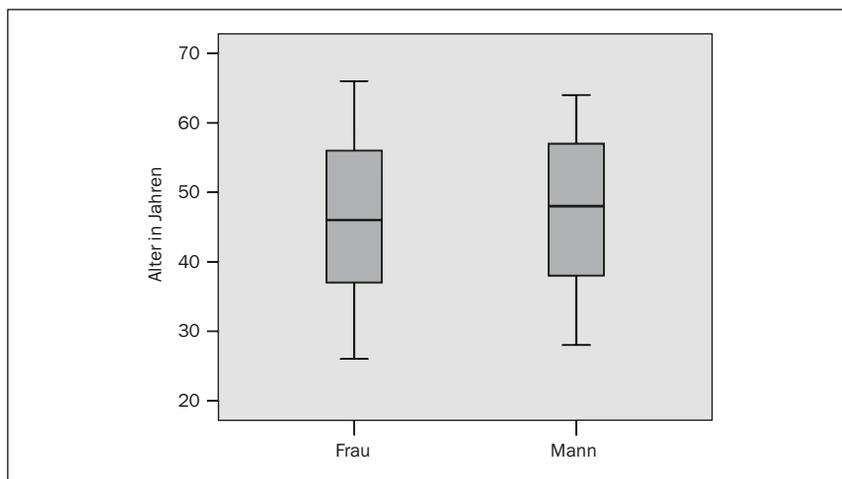
Zu allen Fragen des Fragebogens werden Ergebnisse auf den folgenden Seiten in einer ausgewählten Übersicht dargestellt.

Fragen zur Profession (Q1), Geschlecht (Q2), Alter (Q3), Förderschwerpunkt (Q43) und Schulort (Q44, Q45):

Ausgeschlossen von den Auswertungen sind alle ProbandInnen, die nicht Sonderpädagoge/Sonderpädagogin als Profession angaben, demnach blieben 943 (88 %) Sonderpädagoginnen und 134 (12 %) Sonderpädagogen. Dies entspricht nicht der Verteilung des statistischen Bundesamtes (Destatis, 2017) (weiblich 77,27 %, männlich 22,73 %). Im Folgenden werden alle statistisch auswertbaren Daten analysiert. Ausgenommen sind deshalb qualitative Anmerkungen zu den Fragen, dessen durchaus interessante Auswertung unabhängig von dieser Studie in Fachartikel münden könnten.

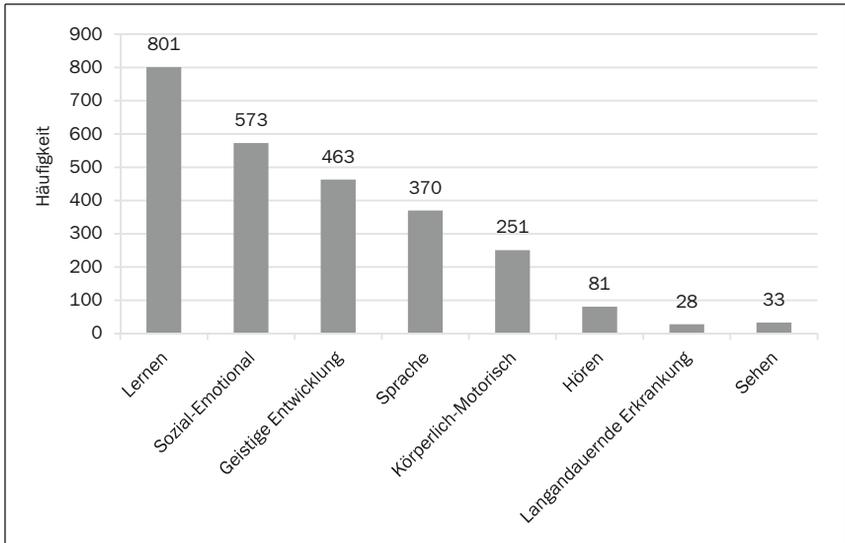
Das Alter der jüngsten Teilnehmerin ist 26 Jahre, das der ältesten Teilnehmerin 66 Jahre (siehe Abbildung 6). Das Durchschnittsalter aller ProbandInnen beträgt 44,55 Jahre ($SD = 9,27$ Jahre).

Abbildung 6. Darstellung Alter und Geschlecht.



Bei der Angabe zum Förderschwerpunkt⁷⁷ (Abbildung 7) waren Mehrfachnennungen möglich (Abbildung 7), da in vielen Regionen Schwerpunkte zusammengelegt worden sind (z.B. arbeiten in NRW viele SonderpädagogInnen im kombinierten Förderschwerpunkt *Lernen/Sozial-Emotional/Sprache*). Dies sind auch die Förderschwerpunkte, die neben dem Schwerpunkt *Geistige Entwicklung* am häufigsten angegeben worden sind.

Abbildung 7. Förderschwerpunkte, in denen gearbeitet wird, in Anzahl der ProbandInnen.

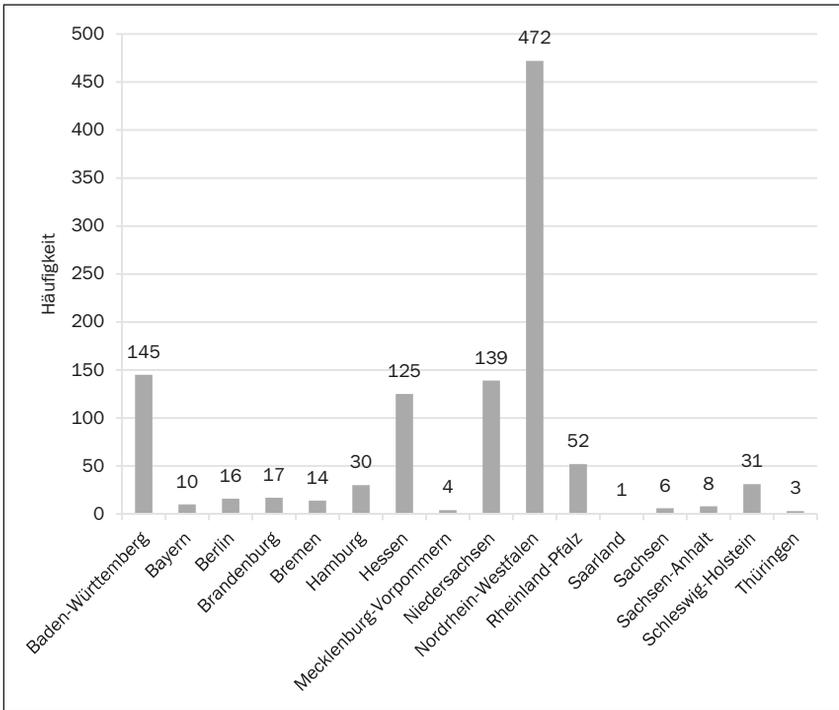


353 ProbandInnen arbeiten in einer Großstadt ab 100 000 EinwohnerInnen, 74 in einer Millionstadt, 351 in einer mittelgroßen Stadt (unter 100 000) und 270 in Kleinstädten und kleineren Gemeinden.

Die mit Abstand größte Gruppe der Teilnehmenden kommt aus Nordrhein-Westfalen (N = 472). Gültige Bundesländer-Vergleiche sind neben Nordrhein-Westfalen ebenfalls anzunehmen für Hessen (N = 125), Niedersachsen (N = 139) und Baden-Württemberg (N = 145). Kleinere Stichproben liegen für Hamburg (N = 30), Rheinland-Pfalz (N = 52) und Schleswig-Holstein (N = 31) vor (siehe Abbildung 8).

77 Angaben zum Förderschwerpunkt, Schulort, Alter und Geschlecht sind nicht gewichtet ausgewertet worden.

Abbildung 8. Die Arbeitsorte der Teilnehmenden in Anzahl der ProbandInnen.



Zusammenfassung aller Ja/Nein, Ja/Nein/Weiß Nicht bzw.

Ja/Nein/Anmerkungen Fragen (Q4, Q6, Q9, Q10, Q23, Q25)⁷⁸:

83.7 Prozent (N = 900) aller ProbandInnen werden in Zukunft Intelligenztests durchführen, 97.2 Prozent (N = 1045) haben bereits einen Intelligenztest durchgeführt, von 72 Prozent (N = 774) wird erwartet, Intelligenztests durchzuführen, jedoch 60.2 Prozent (N = 647) gaben an, nach eigenem Ermessen Intelligenztests durchzuführen. An der Universität probierten 60.6 Prozent (N = 651) Intelligenztests aus (siehe Tabelle 17).

Besondere Bedeutung erhält die Frage nach der Teilnahme an einer außer-universitären Fortbildung (Q25). Ungewichtet und also real verneinten dies 108 Personen. Dieser Personenkreis wird für eine Kontrollgruppe genutzt werden. Gruppenvergleiche könnten Auskunft geben über bessere Rückschlüsse der Ergebnisse auf die Grundgesamtheit der SonderpädagogInnen.

78 Hinweis: da der Mittelwert der Gewichtungen nicht gleich 1 ist, sondern .9976, konstruiert SPSS etwas kleinere Fallzahlen für eine gültige Auswertung. Daraus resultieren von N 1 077 abweichende Angaben.

Tabelle 17. Übersicht überwiegend dichotomer Fragen.

Nr.	Frage	Ja	Nein	weiß nicht	Anmer- kungen	fehlend
Q4	Ich werde in Zukunft Intelligenztests durchführen.	900 83.7 %	33 3.0 %	132 12.3 %	–	10 9.0 %
Q6	Haben Sie bereits mit einem Kind einen Intelligenztest durchgeführt?	1045 97.2 %	25 2.3 %	–	–	5 4.0 %
Q9	Wird von Ihnen erwartet, einen Intelligenztest durchzuführen?	774 72.0 %	155 14.4 %	–	111 10.3 %	35 3.2 %
Q10	Entscheiden Sie nach eigenem Ermessen, Intelligenztests durchzuführen?	647 60.2 %	232 21.6 %	–	–	196 18.2 %
Q23	Haben Sie an der Uni Intelligenztests durchgeführt?	651 60.6 %	291 27.0 %	49 4.6 %	–	84 7.8 %
Q25	Haben Sie an einer außeruniversitären Fortbildung zu Intelligenztests teilgenommen?	922 85.8 %	116 10.8 %	–	–	37 3.4 %

Testergebnisse aus folgenden Tests sind aussagekräftig (Q5)⁷⁹:

Die eingeschätzte Aussagekraft der Intelligenztests (siehe Abbildung 9, nächste Seite) ist wichtig für weitere Fragestellungen, z.B. ob die aussagekräftigsten oder eher die einfach durchzuführenden Tests angewendet werden. Die zwei mehrdimensionalen Tests KABC-II ($MW = 1.77$) und WISC-IV ($MW = 2.15$) sowie die IDS ($MW = 2.18$) und der eindimensionale SON-R 6–40 ($MW = 2.26$) werden als am aussagekräftigsten eingeschätzt; die Tests der CFT-Reihe als am wenigsten aussagekräftig (CFT1-R: $MW = 2.97$; CFT20-R: $MW = 2.84$), obwohl die Tests der CFT-Reihe und der SON-R 6–40 objektiv betrachtet jeweils lediglich einen interpretierbaren Hinweis auf einen Gesamtwert ermitteln, über dieses Ergebnis hinaus jedoch keine weiteren Interpretationen möglich sind.

Fragen zur Anwendung der Testverfahren (Q7, Q8, Q12):

Abbildung 10 (nächste Seite) zeigt, welche Tests bereits durchgeführt worden sind (Q7). Bei mehreren Tests (WPPSI-III, WNV, SON-R 6–40, IDS) ist die Antwort *nie* (blauer Balken) häufiger angegeben als die anderen Antwortmöglichkeiten zusammen (*1–3-mal*, *4–7-mal*, *mehr als 7-mal*). Allein beim CFT1/CFT1-R ist die Antwort *nie* nicht die am häufigsten genannte Antwort. Abbildung 10 verdeutlicht, dass der WPPSI-III (*nie*: 85,61 %) und der WNV (*nie*: 92,5 %) selten angewendet worden sind. Neben dem CFT1/CFT1-R (*mehr als 7-mal*: 34,40 %) und dem CFT20-R (*mehr als 7-mal*: 27,57 %) wurde die KABC-II ebenfalls *häufiger als 7-mal* durchgeführt (20,33 %). Dies sind auch die drei am

79 Likert-Skala: *außerordentlich* (1) – *ziemlich* (2) – *mittelmäßig* (3) – *kaum* (4) – *gar nicht* (5).

Abbildung 9. Eingeschätzte Aussagekraft der Intelligenztests (von *außerordentlich* (1) bis *gar nicht* (5)).

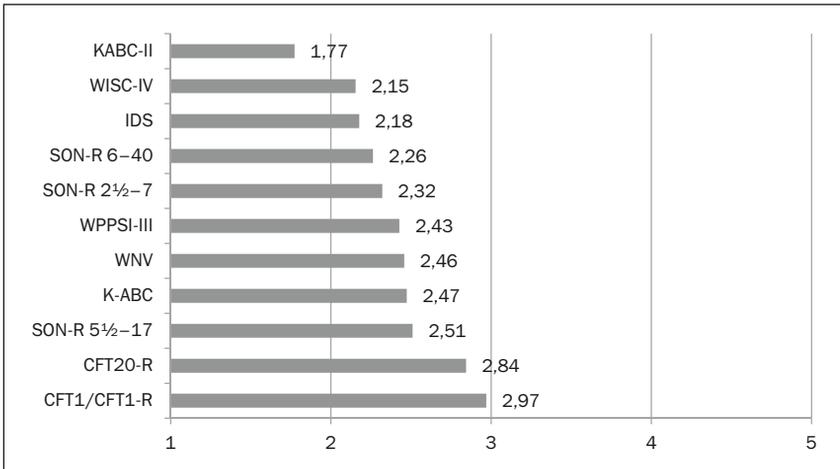
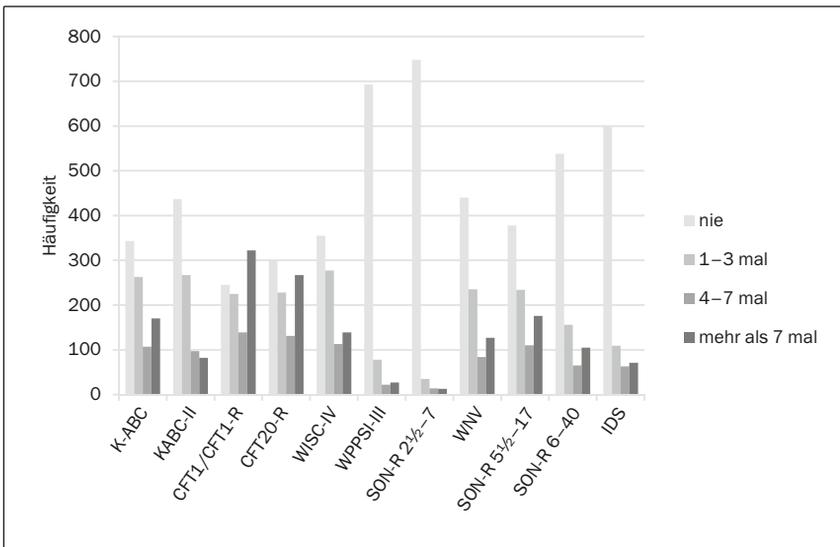


Abbildung 10. Häufigkeit der bisherigen Anwendung von Intelligenztests.

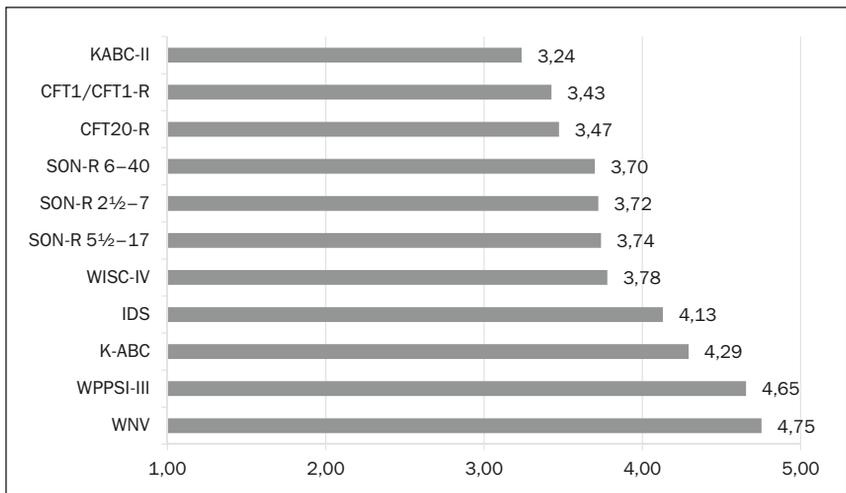


häufigsten angewendeten Tests. Somit wurden die als am wenigsten aussagekräftigen eingeschätzten Tests (Q5; CFT1/CFT1-R; CFT20-R) und der als am aussagekräftigsten eingeschätzte Test (KABC-II) am häufigsten durchgeführt.

Während die vorherige Frage (Q7) auch nach früher angewendeten Verfahren fragt, die aktuell nicht mehr Anwendung finden dürften (z. B. K-ABC), wird mit Q8 (*wenn ich teste, nehme ich folgende Tests (...)*⁸⁰) nach aktuell verwendeten Intelligenztests gefragt. Besonders im Zusammenhang mit der eingeschätzten Aussagekraft und den zur Verfügung stehenden Tests sind hier Rückschlüsse über die Anwendung von Intelligenztests in der Sonderpädagogik möglich.

Aus den Antwortmöglichkeiten von *immer* bis *nie* und den daraus resultierenden Mittelwertvergleichen (siehe Abbildung 11) ist ein Vergleich zwischen den aktuell verwendeten Tests möglich. Am häufigsten werden aktuell die KABC-II ($MW = 3.24$) und die beiden Tests der CFT-Reihe angewendet (CFT1/CFT1-R: $MW = 3.42$; CFT20-R: $MW = 3.47$), am seltensten der WPPSI-III ($MW = 4.65$) und der WNV ($MW = 4.75$). Es fällt auf, dass der veraltete SON-R 5½-17 aktuell häufiger angewendet wird als der WISC-IV ($MW = 3.78$) oder die IDS ($MW = 4.13$).

Abbildung 11. Mittelwertvergleich für die Häufigkeit in der Anwendung aktuell genutzter Intelligenztests von *immer* (1) bis *nie* (5).



Q12 fragte nach den zur Verfügung stehenden Tests. Diese Frage ist u. a. interessant für die Prüfung des Zusammenhangs zwischen zur Verfügung stehen-

80 Likert-Skala: *immer* (1) – *oft* (2) – *gelegentlich* (3) – *selten* (4) – *nie* (5)

der Tests und tatsächlich angewendeter Tests. Am häufigsten stehen der CFT1/CFT1-R (N = 694), der CFT20-R (N = 660) und die KABC-II (N = 625) zur Verfügung, am seltensten der WPPSI-III (N = 153) und der WNV (N = 75) bei N = 1077 (siehe Abbildung 12).

Abbildung 12. Intelligenztests, die zur Verfügung stehen (Mehrfachnennungen möglich).

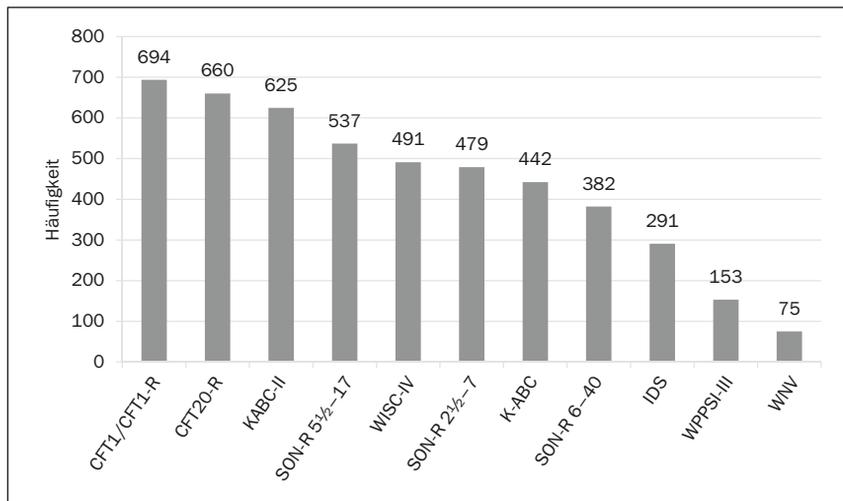


Tabelle 18. Anzahl der zur Verfügung stehenden Tests.

Anzahl Tests	Häufigkeit	Prozent	Kumulierte Prozente
0	45	4.4	4.4
1	48	4.7	9.0
2	85	8.2	17.3
3	168	16.4	33.7
4	201	19.6	53.3
5	157	15.3	68.5
6	118	11.5	80.1
7	102	9.9	90.0
8	55	5.4	95.4
9	34	3.3	98.7
10	11	1.1	99.8
11	2	0.2	100.0
Gesamt	1 028	100.0	

Interessant ist auch, wie viele Intelligenztests insgesamt zur Verfügung stehen. Tabelle 18 zeigt, dass in der Regel mehrere Tests zur Auswahl stehen, ca. die Hälfte der SonderpädagogInnen verfügen über vier oder mehr Intelligenztests.

Fragen zu Schwierigkeiten bei der Anwendung der Tests (Q13–Q18):

Q13 ermittelte äußere Umstände, die die Anwendungen von Intelligenztests erschweren. Bei den erfragten Möglichkeiten liegt am häufigsten das Fehlen eines Tests vor, wenn getestet werden soll ($MW = 3,45$; $SD = 0,99$), seltener das Fehlen von Formularen in Testsituationen ($MW = 4,00$; $SD = 0,93$) und am seltensten werden unvollständige Materialien beschrieben ($MW = 4,23$; $SD = 0,84$).

Im Gegensatz zu der vorherigen Frage (Q13), die nach von außen bedingten Schwierigkeiten bei der Anwendung von Intelligenztests fragt, werden mit Q14 bewusst herbeigeführte Regelverletzungen erfragt, die die Durchführungsobjektivität gefährden und bereits von Huber (2000) festgestellt worden sind. Die Mittelwertvergleiche bezeugen moderat bis selten vorgenommene Abweichungen bei unerlaubt gegebenen Feedbacks ($MW = 3,94$; $SD = 0,87$), dem Verändern von Durchführungszeiten ($MW = 4,42$; $SD = 0,78$) und beim gänzlichen Weglassen der Durchführungszeiten ($MW = 4,71$; $SD = 0,66$). Abbildung 13 veranschaulicht die von außen bedingten Schwierigkeiten, Abbildung 14 die bewusst herbeigeführten Veränderungen.

Abbildung 13. Von außen bedingte Schwierigkeiten während der Testsituationen (Wenn sie testen möchten, kommt es vor (...)).

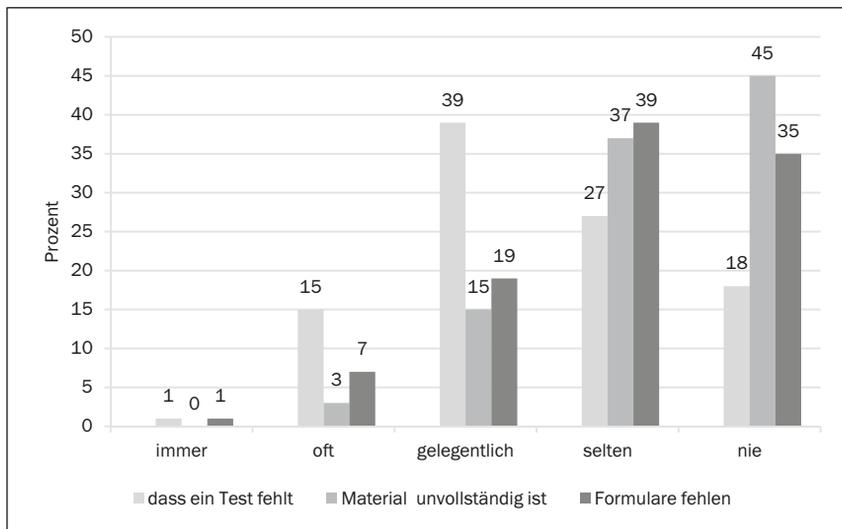
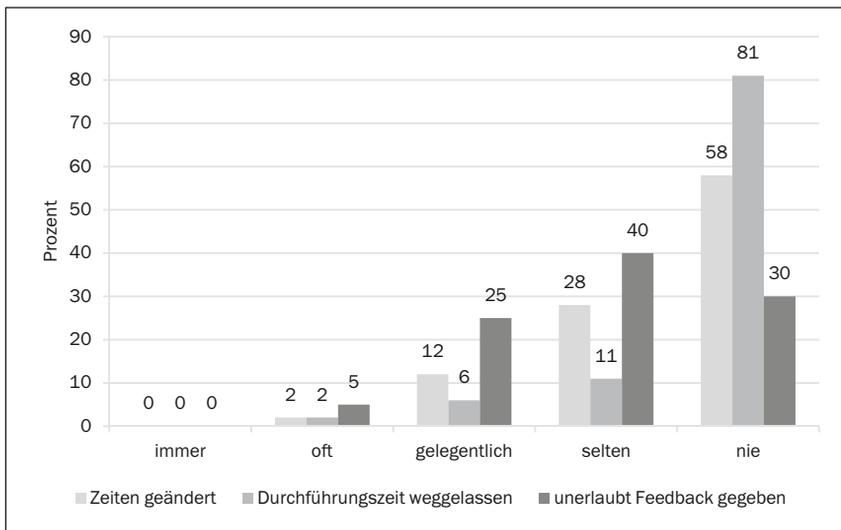


Abbildung 14. Veränderungen, die bei Testanwendungen vorgenommen worden sind (Welche dieser Veränderungen haben Sie schon einmal vorgenommen?).



Q15 fragte, welche Durchführungsregeln Schwierigkeiten bereiten (siehe Tabelle 19). Es ist interessant, ob gelegentlich geäußerte Schwierigkeiten bei der Durchführung auch tatsächlich bestätigt werden, andererseits interessiert beim Vergleich mit gefundenen Fehlern nach der Prüfung von Formularen (angefertigt bei der Anwendung von Intelligenztests im Rahmen der Ermittlung sonderpädagogischen Unterstützungsbedarfs), ob die Häufigkeit von gemachten Fehlern mit den subjektiv empfundenen Schwierigkeiten bei der Anwendung korrelieren. Wäre dies so, müssten bei der Umkehrregel ($MW = 3.75$; $SD = 0.98$) häufiger Fehler vorliegen als bei der Anwendung von Abbruchregeln ($MW = 4.09$; $SD = 0.89$) oder beim Berechnen des Testalters ($MW = 4.47$; $SD = 0.83$)⁸¹. Lediglich 26 Prozent der Befragten gab an, bei den Umkehrregeln *gar nicht* Schwierigkeiten zu empfinden.

81 Es sei erneut darauf hingewiesen, dass subjektiv wenig empfundene Schwierigkeiten gegebenenfalls nicht positiv korrelieren müssen mit der Anzahl von gemachten Fehlern; immer dann, wenn ProbandInnen aus Mangel an genauen Kenntnissen über die entsprechenden Regeln irrtümlich zu der Annahme verleitet werden könnten, die Regeln korrekt anzuwenden.

Tabelle 19. Durchführungsregeln, die als schwierig empfunden werden.

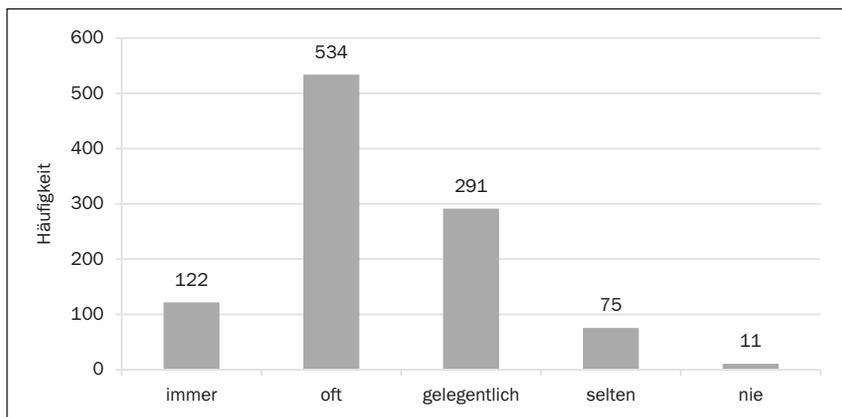
	außerordentlich	ziemlich	mittelmäßig	kaum	gar nicht
Umkehrregeln	$n = 7$ 1%	$n = 86$ 10%	$n = 235$ 28%	$n = 291$ 35%	$n = 221$ 26%
Abbruchregeln	$n = 6$ 1%	$n = 48$ 5%	$n = 169$ 17%	$n = 401$ 40%	$n = 375$ 38%
Testalter berechnen	$n = 6$ 1%	$n = 28$ 3%	$n = 96$ 10%	$n = 231$ 23%	$n = 645$ 64%

Überwiegend finden die Testanwendungen in wechselnden Räumen statt (Q16; $n = 758$), seltener in einem speziellen Testraum ($N = 323$) oder im Klassenzimmer ($N = 151$). 31 SonderpädagogInnen gaben an, im LehrerInnenzimmer zu testen.

Obwohl es schwer vorstellbar ist, dass eine Testung im LehrerInnenzimmer optimale Testbedingungen bietet, ist dies bei entsprechender Vorbereitung nicht ausgeschlossen. Deshalb ist eher die Frage nach der empfundenen Eignung der Testräume interessant (Q17). Immerhin 8,3 Prozent der Befragten fanden die Räume *selten* oder *nie* geeignet ($MW = 2.34$; $SD = 0.82$).

Von allen Befragten ($N = 1033$) gaben 122 (11.81%) *immer*, 534 (51.69%) *oft* an, in geeigneten Testräumen zu testen, siehe Abbildung 15.

Abbildung 15. Eignung der Testräume (Q17: Sind die Testräume geeignet?).



Q18 fragte nach Störungen während der Testsituationen, welche vielfältig bestätigt worden sind (siehe Tabelle 20). Am häufigsten kamen Störungen durch Geräusche vor ($MW = 3.10$; $SD = 0.88$), durch Personen, die geklopft haben ($MW = 3.80$; $SD = 0.77$) und durch Personen, die den Raum betraten ($MW =$

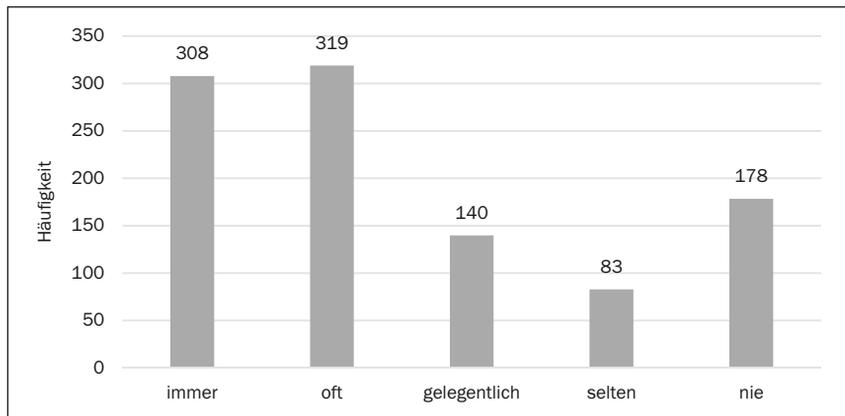
3.92; $SD = 0.79$). Seltener wurden Störungen durch Lautsprecherdurchsagen ($MW = 4.32$; $SD 0.85$) und durch Telefonanrufe ($MW = 4.73$; $SD = 0.60$) angegeben.

Tabelle 20. Störungen während der Testdurchführung (Q18).

	N	immer	oft	gelegentlich	selten	nie
Telefon	1004	$n = 0$ 0%	$n = 10$ 1%	$n = 47$ 4.68%	$n = 149$ 14.84%	$n = 798$ 79.48%
an Tür geklopft	1027	$n = 1$ 0.09%	$n = 41$ 3.99%	$n = 304$ 29.60%	$n = 502$ 48.88%	$n = 180$ 17.53%
Person betritt Raum	1027	$n = 4$ 0.39%	$n = 31$ 3.02%	$n = 243$ 23.66%	$n = 514$ 50.05%	$n = 235$ 22.88%
Geräusche	1037	$n = 16$ 1.54%	$n = 241$ 23.24%	$n = 472$ 45.52%	$n = 242$ 23.34%	$n = 66$ 6.36%
Lautsprecher	1025	$n = 5$ 0.49%	$n = 25$ 2.44%	$n = 150$ 14.63%	$n = 296$ 28.88%	$n = 549$ 53.56%

30 Prozent der ProbandInnen gaben an *immer* mit Computerauswertungen auszuwerten (Q19: *Werten Sie die Intelligenztests mit Computerauswertungen aus?*⁸², siehe Abbildung 16), 31 Prozent *oft*, der Mittelwert liegt mit 2.52 ($SD = 1.43$) im Sinne der Durchführungsobjektivität hoch (*gelegentlich*: 14%; *selten*: 8%; *nie*: 17%, $N = 1024$).

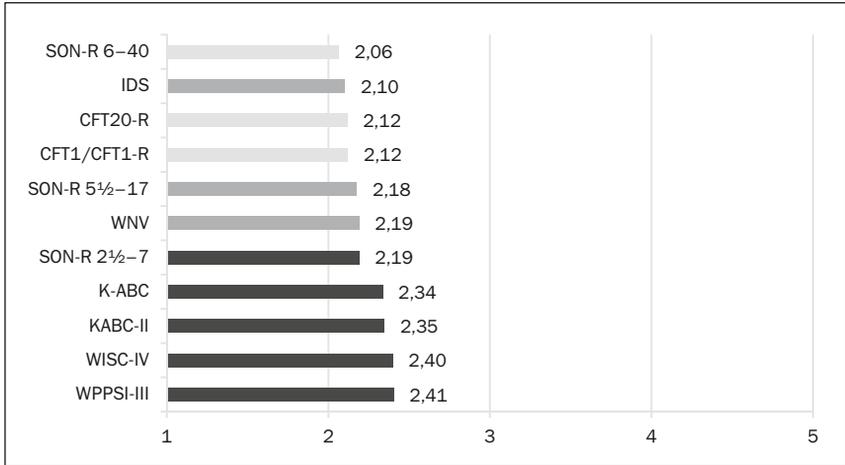
Abbildung 16. Auswertungshäufigkeit mit Computerprogrammen.



82 Likert-Skala: *immer* (1) – *oft* (2) – *gelegentlich* (3) – *selten* (4) – *nie* (5).

Bei der Frage, ob die Erklärungen in den Handbüchern der Tests verständlich sind (Q20)⁸³, zeigte sich beim Mittelwertvergleich, dass die Handbücher der eindimensionalen Tests als verständlicher, die der mehrdimensionalen Tests als weniger verständlich empfunden werden. Es liegen insgesamt jedoch nur geringe Mittelwertunterschiede vor (siehe Abbildung 17).

Abbildung 17. Mittelwerte Verständlichkeit der Handbücher (dunkelgrau: mehrdimensional; hellgrau: eindimensional; grau: ohne Zuordnung).



Fragen zur universitären Ausbildung (Q21, Q22):

Tabelle 21 und Tabelle 22 veranschaulichen die Fragen, welche bedeutsam sind zur Prüfung des Zusammenhangs von Schwierigkeiten bei der Anwendung der Tests und der vorherigen universitären Ausbildung. Auf die Frage, wie viele universitäre Seminare bzw. Vorlesungsreihen zur Testdiagnostik besucht worden sind (Q21; $N = 1074$), antworteten 36 ProbandInnen, dass sie *kein Seminar* dazu belegten (3.35%). Die meisten ProbandInnen belegten *zwei Seminare* (24.40%) mit abnehmender Tendenz (*drei Seminare*: 15.74%; *vier Seminare*: 8.19%). Bei der Antwortmöglichkeit *mehr als vier Seminare* hingegen gibt es wieder eine Zunahme (20.39%).

83 Likert-Skala: *außerordentlich* (1) – *ziemlich* (2) – *mittelmäßig* (3) – *kaum* (4) – *gar nicht* (5).

Tabelle 21. Anzahl belegter Seminare bzw. Vorlesungsreihen zur Testdiagnostik.

keines	1	2	3	4	mehr als 4	fehlend
<i>n</i> = 36	<i>n</i> = 148	<i>n</i> = 262	<i>n</i> = 169	<i>n</i> = 88	<i>n</i> = 219	<i>n</i> = 152
3.35 %	13.78 %	24.40 %	15.74 %	8.19 %	20.39 %	14.15 %

Bei der Frage zur universitären Auseinandersetzung mit zentralen Begriffen der Testdiagnostik (Q22) wurde angegeben, sich damit im Rahmen des Studiums beschäftigt zu haben, die Zustimmungen befinden sich im oberen Quartil bis auf das Konstrukt Vertrauens-/Konfidenzintervall (Zustimmung: 70.08 %). Dieses Konstrukt hatte bei den Verneinungen auch den einzigen zweistelligen Prozentwert (11.48 %).

Tabelle 22. Angaben zur universitären Auseinandersetzung mit Basisbegriffen der Testdiagnostik.

Basisbegriff	N	Ja	Nein	weiß nicht
Standardabweichung	1063	<i>n</i> = 910 85.61 %	<i>n</i> = 65 6.11 %	<i>n</i> = 88 8.28 %
Durchführungsobjektivität	1063	<i>n</i> = 876 82.41 %	<i>n</i> = 67 6.31 %	<i>n</i> = 120 11.29 %
Vertrauens-/Konfidenzintervall	1063	<i>n</i> = 745 70.08 %	<i>n</i> = 122 11.48 %	196 18.44 %
Messungengenauigkeit/-fehler	1062	<i>n</i> = 825 77.68 %	<i>n</i> = 92 8.66 %	<i>n</i> = 145 13.65 %
Gaußsche Kurve	1061	<i>n</i> = 872 82.19 %	<i>n</i> = 67 6.31 %	<i>n</i> = 122 11.50 %

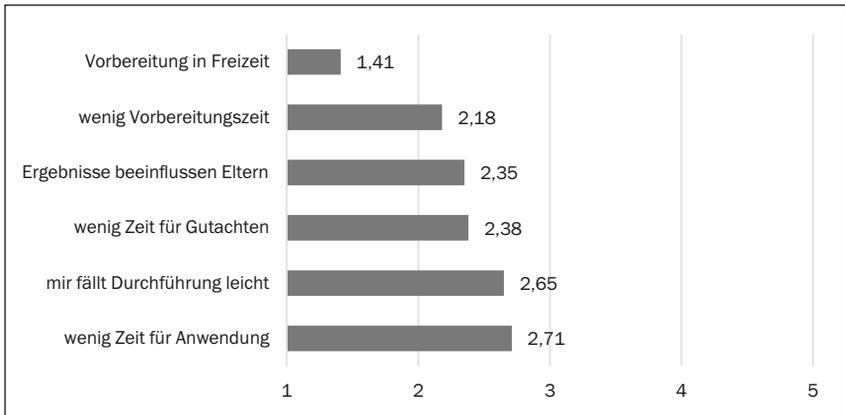
Fragen zu Einschätzungen im Rahmen der Anwendung von Intelligenztests (Q28⁸⁴)

Die höchste Zustimmung (siehe Abbildung 18) erhielt die Frage, ob Tests in der Freizeit vorbereitet werden, wenn nicht genügend Zeit zur Verfügung steht (*MW* = 1.41; *SD* = 0.73). Für die Frage nach zu wenig zur Verfügung stehender Zeit während der Arbeitszeit gab es die zweithöchste Zustimmung (*MW* = 2.18; *SD* = 1.12). Der Mittelwert auf die Frage, ob Ergebnisse Eltern in ihren Planungen und Maßnahmen beeinflussen, beträgt 2.35 (*SD* = 0.74), auf die Frage nach zu wenig Zeit für die Gutachtererstellung wurde ein Mittelwert von 2.38 (*SD* = 1.21) berechnet. Geringere Zustimmung erhielt die Frage, ob die Durchführung

84 Likert-Skala: *völlig richtig* (1) – *ziemlich richtig* (2) – *unentschieden* (3) – *ziemlich falsch* (4) – *völlig falsch* (5).

von Intelligenztests leicht falle ($MW = 2.65$; $SD = 0.94$) und die Frage, ob zu wenig Zeit für die Anwendung der Tests zur Verfügung stehe ($MW = 2.71$; $SD = 1.21$)

Abbildung 18. Grad der Zustimmung zu den Aussagen von Q28.



5.3.2 Auswertung Testformulare

Aus sechs Schulämtern bzw. Schulberatungszentren sind Intelligenztest-Formulare zur Verfügung gestellt worden. Hinzu kommen Formulare von fünf Einzelpersonen. Die Formulare sollten untersucht werden auf Anzahl gemachter Fehler, Fehlerarten und Hinweise auf die verwendeten Testverfahren geben. Die Auszählung wurde konservativ vorgenommen. Die Aufzeichnungen in den Testformularen wurden von den SonderpädagogInnen in der ursprünglichen Testsituation nicht in dem Bewusstsein angefertigt, dass diese später ausgewertet werden würden und nachvollziehbar sein sollten. Es ist möglich, dass Notizen oder Ergänzungen im Formular irrtümlich als Fehler fehlinterpretiert werden könnten, obwohl sie im Rahmen eines persönlichen Protokollierungsstils legitim wären. Dieses Phänomen ist dem Autor aus vielen Testungen bekannt. Ein Beispiel soll dies belegen. Bei der Durchführung des WISC-IV beginnen ältere Kinder oft nicht mit Item 1, sondern mit einem altersgerechten Anfangsitem. Die vorher liegenden Aufgaben werden in der Regel als richtig gelöst bewertet, auch wenn sie nicht durchgeführt worden sind. Wären in den Formularaufzeichnungen Notizen über richtig gelöste Aufgaben vor dem altersentsprechenden Anfangsitem, könnte vermutet werden, dass die Aufgaben vor dem Anfangsitem zu Unrecht durchgeführt worden sind. Es wäre allerdings auch möglich, dass nach der Testung die eigentlich nicht durchzuführenden Items

als korrekt markiert worden sind um die Auszählung zu erleichtern und die Aufzeichnungen somit den Anschein erwecken, Items wären durchgeführt worden, obwohl sie nicht durchgeführt worden sind.

In die Auszählung der Fehler fließen also nur die Fehler ein, die zweifelsohne welche sind. Es ist möglich, dass Fehler nicht erkannt werden. So wäre es bei obigem Beispiel möglich, dass der TesterIn nicht bekannt ist, dass ältere Kinder in vielen Subtests nicht mit Item 1 beginnen⁸⁵.

Wie im vorherigen Kapitel soll die deskriptivstatistische Auswertung eine Übersicht der Daten vorstellen, während die inferenzstatistische Auswertung genauere Analysen durchführt. Die sechs Bezirke werden nicht benannt in NRW 1, Brandenburg usw. Damit soll verhindert werden, voreilige Rückschlüsse zu ziehen, denn die sechs teilnehmenden Schulämter sind zufällig gewählt und repräsentieren nicht die zu den Schulämtern gehörenden Bundesländer. Die sechs Schulämter sind durchnummeriert, die fünf Einzelpersonen werden als solche benannt. In die folgenden Darstellungen fließen ausschließlich ausgewertete Fragebögen mit ein ($N = 248$). Das ReBBZ in Hamburg ist vergleichbar einem Schulamt und wird als Schulamt 6 geführt.

Übersicht der Daten:

In 39,1 Prozent (97 von 248) der überprüften Formulare konnten keine Fehler entdeckt werden, 60,9 Prozent der anderen Formulare waren dementsprechend fehlerhaft. In den 151 fehlerhaften Formularen konnten insgesamt 367 Fehler erkannt werden, im Durchschnitt also 2,43 Fehler/fehlerhaftem Formular. Mit Einbezug der Formulare, für die keine Fehler entdeckt worden sind, wäre der Durchschnitt 1,48 Fehler/Formular. In einem Fall konnten elf, in einem anderen Fall zehn Fehler pro Formular bemerkt werden (siehe Abbildung 19).

Am häufigsten sind die Rohwerte falsch bestimmt worden bzw. auf Grund von Auswertungsfehlern kam es zu falsch ermittelten Rohwerten/Punkten (42,51 % der Gesamtfehler; siehe Abbildung 20), 21,25 Prozent wendeten die Abbruch-, 12,53 Prozent die Umkehrregel falsch an. Das Anfangsitem wurde in 20,44 Prozent der Fälle falsch bestimmt und eher selten das Alter falsch berechnet (3,27 %).

85 Es wäre übrigens auch möglich, dass eine TesterIn bei einem vermeintlich intelligenzgeminderten Kind grundsätzlich bei Item 1 die Testung begonnen hat, was als Durchführungsabweichung nicht verboten wäre. Diese Begründung für einen grundsätzlichen Start bei Item 1 wäre nachträglich nicht nachvollziehbar. Auch in diesem Fall würde die konservative Auszählung falsch entdeckte Fehler verhindern.

Abbildung 19. Häufigkeit gemachter Fehler (N = 248).

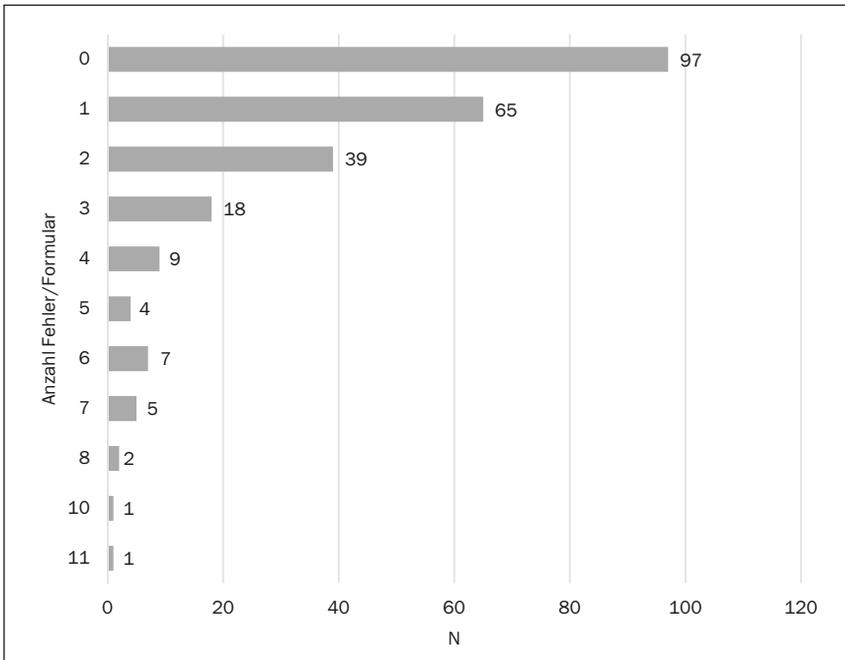
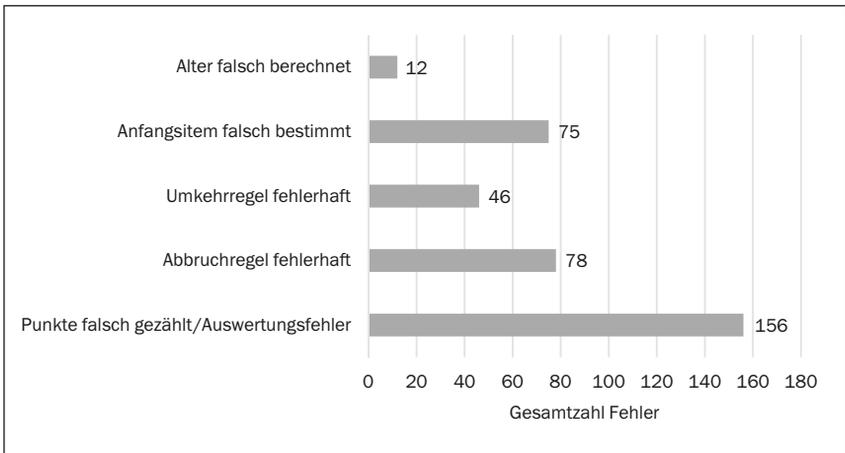


Abbildung 20. Fehlerarten (N = 151; Mehrfachfehler möglich).

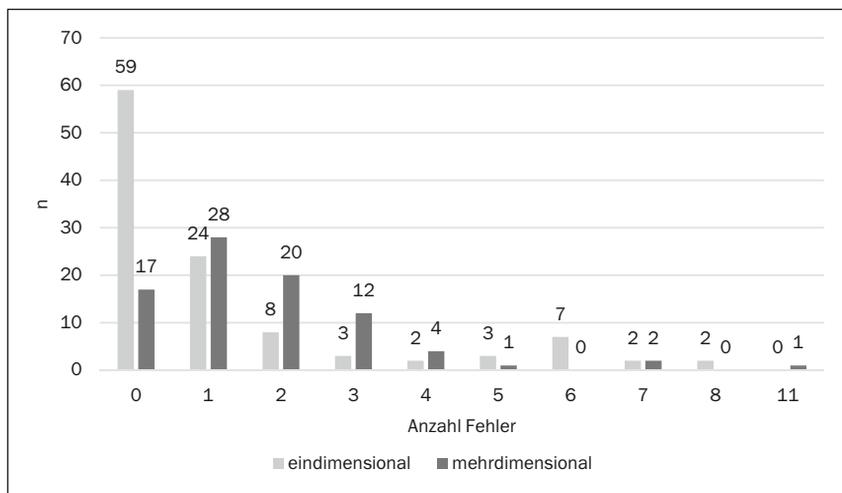


Fehlerhäufigkeiten abhängig von der Dimensionalität der Tests:

110 überprüfte Formulare sind den eindimensionalen Tests (CFT1/CFT1-R, CFT20-R, SON-R 6–40), 85 den mehrdimensionalen Tests zuzuordnen (K-ABC, KABC-II, WISC-IV, WPPSI-III⁸⁶; siehe Abbildung 21). Im Durchschnitt wurden bei den eindimensionalen Tests 1.31 Fehler gemacht, bei den mehrdimensionalen Tests im Durchschnitt 1.76 Fehler. Von den 110 eindimensionalen Tests waren 59 fehlerfrei (53,64%), von den 85 mehrdimensionalen 17 fehlerfrei (20%).

Bei ausschließlicher Betrachtung der fehlerhaften Formulare lagen bei den eindimensionalen geprüften Formularen 2.82 Fehler/Test, bei den mehrdimensionalen 2.21 Fehler/Test vor. Dieses Ergebnis wird an späterer Stelle zu diskutieren sein, da es interessante Rückschlüsse auf die Anwendungspraxis zulässt.

Abbildung 21. Häufigkeit fehlerhafter bzw. fehlerfreier Formulare im Vergleich nach Dimensionalität.

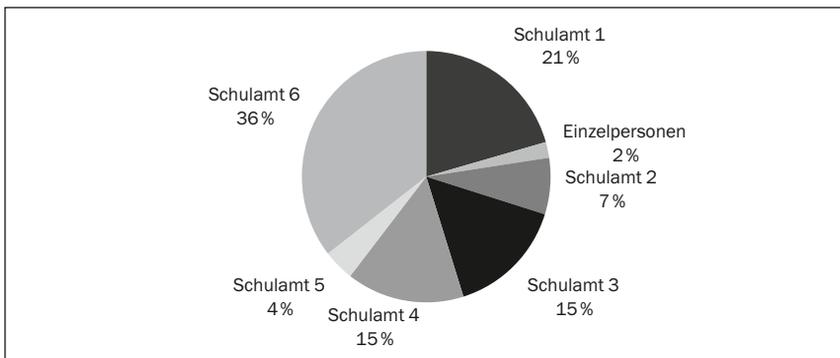


Unterschiede zwischen den beteiligten Schulämtern:

Von den sechs Schulämtern (siehe Abbildung 22) und fünf Einzelpersonen sind vor allem die Schulämter 1, 3, 4 und 6 interessant, da diese über 85 Prozent der Formulare stellten (Schulamt 1: $n = 51$; Schulamt 2: $n = 18$; Schulamt 3: $n = 38$; Schulamt 4: $n = 38$; Schulamt 5: $n = 10$; Schulamt 6: $n = 88$; Einzelpersonen: $n = 5$).

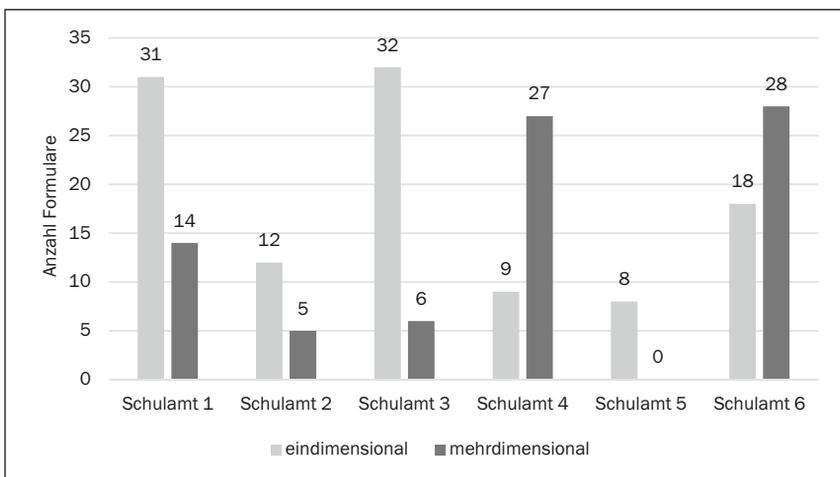
86 SON-R 2½–7 als ebenfalls mehrdimensionaler Test entfällt, da keine Formulare ausgewertet worden sind.

Abbildung 22. Verteilung der ausgewerteten Formulare.



Aussagekräftigere mehrdimensionale und weniger aussagekräftige eindimensionale Intelligenztests werden im Rahmen der Erstellung eines sonderpädagogischen Gutachtens unterschiedlich häufig angewendet. Ohne Einbezug der fünf Tests, die von Einzelpersonen durchgeführt worden sind, gibt Abbildung 23 Hinweise darauf, ob ein Kind mit mehr oder weniger aussagekräftigen Tests begutachtet wurde.

Abbildung 23. Vergleich der Anwendungen von mehr- bzw. eindimensionalen Tests in unterschiedlichen Schulämtern.



Bei den Schulämtern wird deutlich, dass häufiger entweder die eindimensionalen Tests präferiert werden (Schulämter 1, 2, 3 und 5) oder die mehrdimensionalen (Schulämter 4 und 6). Ein Vergleich zwischen den Schulämtern und der

Gesamtzahl gemachter Fehler wäre für sich genommen irreführend. Da die eindimensionalen Tests auch einfacher durchzuführen sind und weniger Regeln beinhalten, die Schulämter aber unterschiedlich verteilt ein- bzw. mehrdimensionale Tests durchführen, ist ein dezidierter Vergleich unterschieden nach *Dimensionalität* aufschlussreicher (Tabelle 23). In diesem Zusammenhang ist besonders interessant, wie oft Tests fehlerfrei durchgeführt und wie viele Fehler pro fehlerhaftem Formular im Durchschnitt gemacht worden sind.

Bei den eindimensionalen Tests reicht die Spannweite von 43.33 bis 75.00 Prozent richtig durchgeführter Tests, bei den mehrdimensionalen Tests von 0 bis 50.00 Prozent richtig durchgeführter Tests (Schulamnt 1: (N = 45), durchschnittliche Fehler eindimensionale Tests (N = 31): 2.28, durchschnittliche Fehler mehrdimensionale Tests (N = 14): 2.09; Schulamt 2: (N = 17), durchschnittliche Fehler eindimensionale Tests (N = 12): 4, durchschnittliche Fehler mehrdimensionale Tests (N = 5): 2.8; Schulamt 3: n = 38, durchschnittliche Fehler eindimensionale Tests (N = 32): 4, durchschnittliche Fehler mehrdimensionale Tests (N = 6): 5.34; Schulamt 4: (N = 36), durchschnittliche Fehler eindimensionale Tests (N = 9): 1, durchschnittliche Fehler mehrdimensionale Tests (N = 27): 2.34; Schulamt 5: (N = 8), durchschnittliche Fehler eindimensionale Tests (N = 8): 2.28; Schulamt 6: (N = 46), durchschnittliche Fehler eindimensionale Tests (N = 18): 1.2, durchschnittliche Fehler mehrdimensionale Tests (N = 28) = 1.34).

Bis auf eine Ausnahme sind mehrdimensionale Tests fehleranfälliger gewesen. Das hier abweichende Ergebnis im Schulamt 3 wird als Zufallsergebnis angenommen, da lediglich 6 mehrdimensionale Tests durchgeführt worden sind.

In allen Schulämtern mit Ausnahme des Schulamts 1, die zur Auswertung Formulare aus sowohl ein- als auch mehrdimensionalen Tests zur Verfügung gestellt haben, wurden bezogen auf die fehlerhaften Formulare durchschnittlich mehr Fehler bei den mehrdimensionalen Tests gemacht. Der größte Unterschied ist bei Schulamt 4 (Differenz = 1.34 Fehler), der geringste beim Schulamt 1 festzustellen (Differenz = 0.03 Fehler). Tabelle 23 zeigt in der Übersicht die Fehlerhäufigkeiten aller Formulare und unterschieden nach Dimensionalität und Schulamt.

Tabelle 24 zeigt den Vergleich der durchschnittlichen Gesamtfehler unter Einbezug aller Formulare (also auch der fehlerfreien) für die Schulämter mit einer höheren Fallzahl. Bezogen auf alle überprüften Formulare kann z. B. für das Schulamt 6 festgestellt werden, dass weniger als ein Fehler sowohl bei den ein- als auch bei den mehrdimensionalen Tests gefunden worden sind⁸⁷.

87 Interessant könnte im abschließenden Kapitel ein Vergleich zwischen dem Konzept des Schulamts 6 (ein ReBBZ aus Hamburg) und den Konzepten der anderen Schulämter mit deutlich mehr festgestellten Fehlern sein.

Tabelle 23. Darstellung der Fehlerhäufigkeiten aller Formulare, unterschieden nach Dimensionalität und Schulamt.

		<i>n</i>	fehlerfreie Formulare	Anzahl Fehler	Gesamtdurchschnitt	Ø fehlerhafte Formulare
Schulamnt 1	Gesamt	45	15 (34.10 %)	66	1.47	2.20
	eindimensional	31	13 (43.33 %)	41	1.32	2.28
	mehrdimensional	14	2 (14.29 %)	25	1.79	2.09
Schulamnt 2	Gesamt	17	7 (41.18 %)	34	2.00	3.40
	eindimensional	12	7 (58.34 %)	20	1.67	4.00
	mehrdimensional	5	0	14	2.80	2.80
Schulamnt 3	Gesamt	38	17 (44.74 %)	88	2.32	4.20
	eindimensional	32	14 (43.75 %)	72	2.25	4.00
	mehrdimensional	6	3 (50.00 %)	16	2.67	5.34
Schulamnt 4	Gesamt	36	7 (19.44 %)	67	1.86	2.31
	eindimensional	9	6 (66.67 %)	3	0.34	1.00
	mehrdimensional	27	1 (3.70 %)	64	2.37	2.34
Schulamnt 5	Gesamt	8	6 (75.00 %)	2	0.25	1.00
	eindimensional	8	6 (75.00 %)	2	0.25	1.00
	mehrdimensional	0				
Schulamnt 6	Gesamt	46	23 (50.00 %)	30	0.65	1.30
	eindimensional	18	13 (72.22 %)	6	0.34	1.20
	mehrdimensional	28	10 (35.71 %)	24	0.86	1.34

Anmerkungen. Fehlerfreie Formulare gibt die Anzahl von Formularen ohne gefundene Fehler an. Gesamtdurchschnitt gibt den Durchschnitt aller Fehler im Verhältnis zu allen Formularen an. Ø fehlerhafte Formulare = Durchschnitt der Fehler im Verhältnis zu allen fehlerhaften Formularen.

Tabelle 24. Vergleich fehlerhafter Formulare, unterschieden nach Dimensionalität für Schulämter mit einer höheren Fallzahl.

	Gesamt auswertbare Formulare			eindimensional			mehrdimensional		
	<i>n</i>	Ø Fehler	SD	<i>n</i>	Ø Fehler	SD	<i>n</i>	Ø Fehler	SD
Schulamnt 1	51	1.43	1.65	31	1.32	1.66	14	1.79	1.72
Schulamnt 3	38	2.32	3.02	32	2.25	2.81	6	2.67	4.27
Schulamnt 4	38	1.76	1.62	9	0.33	0.50	27	2.37	1.52
Schulamnt 6	88	0.97	1.43	18	0.33	0.59	28	0.86	0.80

Erläuterungen. Gesamt auswertbare Formulare bezieht auch Formulare mit ein, die nicht eindeutig mehr- bzw. eindimensionalen Tests zuzuordnen sind. SD = Standardabweichung. Ø = Durchschnitt.

Der größte Unterschied zwischen der durchschnittlichen Fehlerhäufigkeit zwischen ein- und mehrdimensionalen Tests liegt beim Schulamt 4 vor (Durchschnitt Fehler eindimensionale Tests = 0.33, $SD = 0.50$; bei mehrdimensionalen Tests = 2.37, $SD = 1.52$)

Es bietet sich an dieser Stelle ein kurzer Blick auf signifikante Unterschiede bezüglich der Fehlerhäufigkeit zwischen den Schulämtern an, auch wenn keine Hypothesen zu dieser Frage im nächsten Kapitel geprüft werden sollen. Es fällt jedoch auf, dass z. B. Schulamt 6 positiv hervorsteht.

Bei den eindimensionalen Tests gibt es signifikante Unterschiede zwischen den Schulämtern sowohl bei den eindimensionalen ($N = 90$), ($H(3) = 9.36$, $p = .025$)⁸⁸ als auch bei den mehrdimensionalen Tests ($N = 75$), ($H(3) = 17.05$, $p = .001$).

Bereits in der deskriptiven Auswertung sticht Schulamt 6 sowohl bei den ein- als auch bei den mehrdimensionalen Tests positiv hervor, Schulamt 4 zumindest bei den eindimensionalen Tests.

Verglichen mit den anderen Schulämtern macht Schulamt 6 signifikant weniger Fehler sowohl bei den eindimensionalen ($U(18, 72) = 431$, $z = -2.37$, $p = .018$)⁸⁹ als auch bei den mehrdimensionalen Tests ($U(28, 47) = 330.5$, $z = -3.71$, $p < .001$).

Bei Schulamt 4 konnte hingegen bei den eindimensionalen Tests im Vergleich mit den anderen Schulämtern bezüglich der Anzahl der Gesamtfehler keine Signifikanz festgestellt werden ($U(9, 80) = 265.5$, $z = -1.40$, $p = .163$).

Fehleranfälligkeiten der Tests:

Tabelle 25 zeigt die Häufigkeit und die Fehlerarten, unterschieden nach Testverfahren. Bei den Gesamtfehlern beträgt die durchschnittliche Fehlerzahl 1.48 Fehler. Bei den Tests mit einer mindestens zweistelligen Fallzahl liegen über dem Durchschnitt die IDS ($MW = 1.79$), der WISC-IV ($MW = 1.75$) und der SON-R 6–40 ($MW = 1.55$). Unter dem Durchschnitt liegen die KABC-II ($MW = 1.21$) und der WNV ($MW = 0.9$).

Unterschieden nach Fehlerart fällt bei den falsch gezählten Punkten bzw. der falschen Auswertung negativ die IDS ($MW = 1.73$ Fehler/Test; Gesamtdurchschnitt = 0.63 Fehler/Test), positiv die KABC-II ($MW = 0.53$) auf. Bei den Abbruchregeln sticht negativ der WISC-IV ($MW = 0.6$, Gesamtdurchschnitt = 0.31), bei den Umkehrregeln ebenfalls negativ der WISC-IV hervor ($MW = 0.57$; Gesamtdurchschnitt = 0.19).

88 Gruppenvergleich der vier Schulämter mit höheren Fallzahlen mit dem Kruskal-Wallis-Test, asymptotische Signifikanz, zweiseitig.

89 Mann-Whitney-U-Test, asymptotische Signifikanz, zweiseitig, Vergleich der vier Schulämter mit einer höheren Fallzahl.

Bei dem falsch bestimmten Anfangsitem sind bei dem SON-R 6–40 annähernd viermal mehr Fehler ($MW = 1.03$) als im Gesamtdurchschnitt ($MW = 0.28$) vorhanden.

Tabelle 25. Anzahl Fehler und durchschnittliche Fehlerzahl, unterschieden nach Fehlerart.

	N	Gesamtfehler	Durchschnitt	Add./Bew.	Durchschnitt	Abbruchregel	Durchschnitt	Umkehrregel	Durchschnitt	Anfangsitem	Durchschnitt
CFT20	9	4	0.44	4	0.44						
WNV	31	28	0.9	9	0.29	12	0.39	3	0.1		
CFT1	36	39	1.08	37	1.03						
KABC-II	19	23	1.21	10	0.53	7	0.37	5	0.26	1	0.05
SON-R 6–40	65	101	1.55	30	0.46	4	0.06			67	1.03
WISC-IV	60	105	1.75	29	0.48	36	0.6	34	0.57	1	0.02
IDS	19	34	1.79	33	1.73						
WPPSI	4	11	2.75	3	0.75	6	1.5	1	0.25		
SON-R 5½–17	3	11	3.67	1	0.33	5	1.67				
K-ABC	2	11	5.5			8	4	3	1.5		
Gesamt	248	367	1.48	156	0.63	78	0.31	46	0.19	69	0.28

Anmerkung. Add./Bew. = Addition/Bewertung.

Mögliche Auswirkungen fehlerhafter Auswertungen auf die Testergebnisse:

Ein Fehler muss nicht zwangsläufig zu veränderten Ergebnissen führen. Einige Beispiele sollen dies belegen:

- Ein falsch berechnetes Testalter ist ohne Auswirkung, wenn die Normtabelle sich nicht ändert. Ist ein Kind am Testtag 8;3 Jahre alt und umfasst eine Normtabelle den Altersbereich 8;3–8;5 Jahre, würde die richtige Normtabelle bei einem falsch berechneten Alter von 8;5 Jahren dennoch genutzt werden.
- Würden Items unter Missachtung der Abbruchregel irrtümlich durchgeführt werden, das Kind erzielte aber keine Punkte bei den zu viel durchgeführten Aufgaben, wäre dies ohne Auswirkungen.
- Würden mehrere Rohwerte zu dem Bereich eines standardisierten Werts gehören (z. B. die Rohwerte 44–48 gehörten zu dem standardisierten Wert *Wertpunkt* 9), würde dies nicht zu einer Verfälschung führen, würde der

falsch addierte Rohwert zu dem Rohwert-Bereich gehören (z. B. statt korrekt Rohwert 45 den Rohwert 48 falsch berechnet).

Es wird unterschieden zwischen möglichen Auswirkungen und keinen Auswirkungen auf die Testergebnisse. In einigen Fällen kann sicher ausgeschlossen werden, dass die gefundenen Fehler zu einer Veränderung der Testergebnisse führen, in allen anderen Fällen wäre dies möglich. Die Formulierung *mögliche Auswirkung* ist bewusst vorsichtig gewählt und auch hier wird konservativ vorgegangen, um SonderpädagogInnen nicht zu Unrecht Testergebnisse verfälschende Mängel bei der Auswertung zu attestieren.

Für 151 fehlerhafte Fragebögen konnten mögliche Auswirkungen bestimmt werden ($N = 248$). Für 32 (21.2%) konnten Auswirkungen ausgeschlossen werden, bei 119 (78.8%) sind Auswirkungen möglich. Bezogen auf die Gesamtzahl würde dies bedeuten, dass bei 248 geprüften Testformularen 48 Prozent Auswirkungen auf die Ergebnisse durch eine fehlerhafte Anwendung möglich sind.

Da ausschließlich Testformulare aus Testdurchführungen in die Auswertung dieser Arbeit einfließen, die im Rahmen eines Gutachtens zur Feststellung sonderpädagogischen Förderbedarfs angefertigt worden sind, können durch Auswertungs- bzw. Durchführungsfehler resultierende Auswirkungen auf die Testergebnisse auch Auswirkungen auf die aus den Testergebnissen abgeleiteten Schlussfolgerungen für die Fragestellungen der Begutachtung resultieren.

Einige Beispiele von nicht möglichen, sondern tatsächlichen Auswirkungen der Fehler auf Testergebnisse bzw. besonders markante Fehler sollen qualitativ skizziert werden:

- Bezirk 1, Fall 23, WISC-IV, *Matrizen Test*: Wertpunkt = 2 falsch, Wertpunkt = 6 richtig⁹⁰. Der *Matrizen Test* misst vor allem die *fluide* Intelligenz (analytisch, abstrakt logisches Denken), einem Kernbereich des Generalfaktors. Ein Wertpunkt von 2 würde umgerechnet IQ 60 bedeuten, der richtige Wertpunkt 6 umgerechnet IQ 80⁹¹. Der falsche Wert könnte als ein Hinweis auf den Unterstützungsbedarf *geistige Entwicklung* interpretiert werden, der korrekte Wert ein Hinweis auf den Unterstützungsbedarf *Lernen*. Die *fluide*

90 Der standardisierte Wert bzw. die Skalierung Wertpunkt hat eine Mitte = 10 und eine Standardabweichung = 3, somit einen Normbereich von Wertpunkt (WP) 7–13.

91 Mit dem standardisierten Wert Intelligenz-Quotient (IQ) werden in der Regel Gesamtergebnisse angezeigt. Die Umrechnung eines Subtestergebnisses in IQ ist problematisch, da mit dem IQ die allgemeine Intelligenz assoziiert wird, ihm also eine hohe Bedeutung beigemessen wird. Diese Umrechnung ist zusätzlich problematisch, da ein Gesamt-IQ selten der Durchschnitt der Subtestergebnisse darstellt und niedrige Subtestergebnisse meist zu Gesamtergebnissen führen, die noch niedriger sind als der Durchschnitt der Teilergebnisse. Die Umrechnung des standardisierten Werts Wertpunkt in den standardisierten Wert IQ dient in diesem Abschnitt der Veranschaulichung.

Intelligenz ist einer der wichtigsten und bestuntersuchtsten Bereiche in der Intelligenzforschung und wird z.B. in den Tests der CFT-Reihe nachvollziehbar Grundintelligenz genannt,

- Bezirk 1, Fall 36, CFT1-R Kurzform, Gesamtwert: T-Wert⁹² 32 statt T-Wert 29,
- Bezirk 3, Fall 18, SON-R 6–40, Gesamtwert: mindestens IQ = 99 statt IQ = 90⁹³,
- Bezirk 3, Fall 21, SON-R 6–40, *Zeichenmuster*: Standardwert⁹⁴ = 3 statt Standardwert = 6,
- Bezirk 3, Fall 27, CFT1-R: fünf der sechs Subtests sind falsch durchgeführt bzw. ausgewertet worden; das Testalter ist um zwei Monate falsch berechnet worden,
- Bezirk 3, Fall 39, K-ABC: von 11 Subtests sind alle 11 falsch durchgeführt worden,
- Bezirk 4, Fall 1, WISC-IV, *Wortschatz*: Wertpunkt = 6 statt Wertpunkt = 8
- Bezirk 4, Fall 11, WISC-IV: Symbolsuche mit Aufgaben für sechs- bis siebenjährige Kinder durchgeführt, obwohl das Kind bereits 9 Jahre alt war. Der falsch ermittelte Wert von Wertpunkt 15 (Durchschnitt andere Tests: WP 7,9) beeinflusst Gesamt-IQ (falscher Wert: IQ 87) maßgeblich,
- Bezirk 4, Fall 19, WISC-IV, *Zahlensymboltest*: Wertpunkt = 5 statt Wertpunkt = 9 (entspricht IQ = 75 statt IQ = 95),
- Bezirk 4, Fall 22, WISC-IV, *Symbolsuche*: Wertpunkt = 11 statt Wertpunkt = 7 (entspricht umgerechnet in IQ = 105 statt IQ = 85),
- Bezirk 6, Fall 7, WISC-IV, *Wortschatz*: Wertpunkt = 5 statt Wertpunkt = 2,
- Bezirk 6, Fall 18, WNV, *Matrizen-Test*: T-Wert 43 statt T-Wert = 28 (entspricht IQ = 90 statt IQ = 67),
- Bezirk 6, Fall 20, WNV, *Zahlen-Symbol-Test*: T-Wert = 49 statt T-Wert = 61,
- Bezirk 6, Fall 21, WNV, *Bilder-Ordnen*: T-Wert = 49 statt T-Wert = 40,
- Bezirk 6, Fall 25, WNV: Testalter um ein Jahr verrechnet (10;10 Jahre statt 9;10 Jahre), dadurch Abgleich der Rohwerte mit falscher Normtabelle,
- Bezirk 6, Fall 47, KABC-II, *Dreiecke*: Skalenwert⁹⁵ = 6 statt Skalenwert = 1, dies entspricht einem umgerechneten IQ von 80 statt IQ 55.

92 Der standardisierte Wert bzw. die Skalierung T-Wert hat eine Mitte = 50 und eine Standardabweichung = 10, somit einen Normbereich von T-Wert 40–60.

93 Tatsächlich ist eine noch höhere Abweichung möglich.

94 Der standardisierte Wert bzw. die Skalierung Standardwert hat für diesen Test eine Mitte = 10 und eine Standardabweichung = 3, somit einen Normbereich von Standardwert 7–13.

5.4 Inferenzstatistische Auswertungen

In diesem Kapitel werden die Ergebnisse aus den Prüfungen der Forschungsfragen vorgestellt.

Für die Hypothesenprüfungen sind die weiter oben beschriebenen Forschungsfragen in falsifizierbare Hypothesen überführt worden. Das Signifikanzniveau wurde auf .05 festgelegt, wenn nicht anders angegeben sind die Hypothesen zweiseitig getestet worden. Die Gewichtungen sind angepasst, um eine Kontrollgruppe einbeziehen zu können (siehe Kapitel 5.1), $N = 1037$ für die ersten acht Forschungsfragen.

Zur Prüfung von Unterschieden, Zusammenhängen, auf Mitte und auf Streuung sind mit Hilfe der Statistik-Software SPSS 24 verschiedene statistische Verfahren genutzt worden, die in einer kurzen Übersicht dargestellt werden.

Die Qualität der Daten bestimmte die Auswahl der Verfahren. Konnten bei den Unterschiedshypothesen Annahmen über die Verteilung der Daten getroffen werden, sind parametrische, sonst nichtparametrische Verfahren gewählt worden. Bei den Zusammenhangsanalysen wurde bei intervallskalierten Merkmalen die Pearson-Korrelation und bei ordinalskalierten Merkmalen die Spearman-Korrelation gewählt (Riepl, 2013, o.S.). Eine ausführlichere Auseinandersetzung mit Cronbachs Alpha wurde in Kapitel 4.2.3 (Konstruktion eines *Schwierigkeiten-Index*) vorgenommen. Wenn nicht anders angegeben, wurde zweiseitig geprüft.

Die für die inferenzstatistischen Berechnungen genutzten Verfahren sind (alphabetisch geordnet):

- *Bonferroni-Korrektur*: Zur Vermeidung von Fehlern 1. Art wird diese Korrektur vielfach angewendet. Ansonsten wäre anzunehmen, dass bei dem grundsätzlich gewähltem Signifikanzniveau von 5 Prozent bei 100 Tests immerhin in fünf Fällen die Nullhypothese mit einem signifikanten Ergebnis zu Unrecht abgelehnt werden würde (Hemmerich, 2015b, o.S.). Der Nachteil dieses sehr konservativen Verfahrens ist die Erhöhung der Wahrscheinlichkeit falsch negativer Ergebnisse (ebd., o.S.). Um sicherzustellen, dass aus den Ergebnissen abgeleitete Empfehlungen gegenüber Institutionen auf einer soliden Datenbasis stehen, wird dieser Nachteil bewusst in Kauf genommen. Somit wird also auch in Kauf genommen, dass weniger der erstellten Alternativhypothesen zu signifikanten Ergebnissen führen.
- *Chi-Quadrat-Test*: Die mit Hilfe von Kreuztabellen ermittelten Zusammenhänge zwischen kategorialen Variablen können mit dem Chi-Quadrat-Test

95 Der standardisierte Wert bzw. die Skalierung Skalenwert hat eine Mitte = 10 und eine Standardabweichung = 3, somit einen Normbereich von Skalenwert 7–13.

darauf überprüft werden, ob die Zusammenhänge auch in der Grundgesamtheit bestehen (Brosius, 2017, S. 219). Angewendet wird dieser Test ab einer erwarteten Häufigkeit von größer 5, was bei den hohen Fallzahlen der Studie kein Problem darstellte. Es ist allerdings auch möglich, dass eine sehr große Fallzahl zu signifikanten Ergebnissen führen kann, obwohl die Unterschiede eher klein sind. In dieser Arbeit wird der Chi-Quadrat-Test nach Pearson verwendet.

- *Einfaktorielle Varianzanalyse (ANOVA⁹⁶)*: Zur Vermeidung von Alphafehler-Kumulierungen werden beim Vergleich mehrerer Gruppen (z.B. Bundesländern) nicht mit Hilfe von t-Tests die Mittelwerte von jeweils zwei Stichproben aus dem Pool der Gruppen berechnet, sondern beim Vergleich mehrerer Gruppen Varianzanalysen durchgeführt (Raab-Steiner & Benesch, 2015, S. 158). Dies vermeidet zudem die Gefahr einer verminderten Power (ebd., S. 158). Die einfaktorielle Varianzanalyse wurde genutzt beim Vorliegen von stetigen Merkmalen und kam lediglich zur Prüfung der Hypothese 4.6 zum Einsatz.
- *Friedman-Test*: Diese einfaktorielle Varianzanalyse mit Meßwiederholung zur Prüfung von Unterschieden erwartet Ordinaldaten bei abhängigen Stichproben (Pospeschill, 1996, S. 240) und ermittelt Ränge bzw. Rangreihen. Der Friedman-Test wird verwendet, wenn mehr als zwei Gruppen vorhanden sind. Verkürzt prüft der Friedman-Test, ob sich die zentralen Tendenzen einer Variable zwischen mehreren abhängigen Gruppen bzw. Messzeitpunkten unterscheiden (Universität Zürich, 2018, o.S.).
- *Kruskal-Wallis-Test*: Bei diesem Test handelt es sich wie beim Friedman-Test ebenfalls um eine einfaktorielle Varianzanalyse für mehr als zwei Gruppen und im Gegensatz zum Friedman-Test ohne Meßwiederholung (Pospeschill, 1996, S. 237). Ermittelt werden Unterschiede von zentralen Tendenzen. Das Pendant für Untersuchungen von zwei Gruppen ist der Mann-Whitney-U-Test.
- *Levene-Test*: Dieser Signifikanztest misst im Rahmen dieser Arbeit bei der Anwendung anderer Testverfahren (z.B. t-Test für unabhängige Stichproben), ob eine Varianzhomogenität/-gleichheit vorliegt. Davon ist abhängig, welcher Test in Folge für die Signifikanzprüfungen genutzt werden muss. Bei einer einfaktoriellen Varianzanalyse (ANOVA) wird z.B. bei vorliegender Varianzhomogenität die berechneten Signifikanzangaben des Tukey-Tests (post-hoc Verfahren, ähnelt dem t-Test), beim t-Test für unabhängige Stichproben würden bei einer Varianzungleichheit die Angaben des Welch-Tests genutzt werden.

96 ANOVA = Analysis of Variances.

- *Mann-Whitney-U-Test*: Dieser Test wird genutzt, wenn die Voraussetzungen wie bei dem Kruskal-Wallis-Test vorliegen (mindestens Ordinaldaten, zwei unabhängige Stichproben), jedoch für die Unterschiedsprüfung bei zwei Gruppen. Eine Normalverteilung ist nicht notwendig (Raab-Steiner & Benesch, 2015, S. 130). Im Rahmen dieser Arbeit wurde der Mann-Whitney-U-Test häufig genutzt für Unterschiedsprüfungen zwischen der Versuchs- und der Kontrollgruppe.
- *Korrelationstests*: Im Zusammenhang dieser Studie wurden Korrelationskoeffizienten nach Pearson (bei intervallskalierten Merkmalen) bzw. nach Spearman (bei ordinalskalierten Merkmalen; Variablen müssen nicht normalverteilt sein) berechnet. Ziel ist die Beschreibung von Zusammenhängen zwischen zwei Variablen.
- *t-Test*: Bei der Annahme einer Normalverteilung und dem Vorliegen von metrischen Daten kommt der t-Test für unabhängige Stichproben in Frage, um Mittelwertvergleiche vornehmen zu können. Bei der Anwendung des t-Tests ist die Verwendung der Ergebnisse abhängig von der Prüfung der Varianzhomogenität (siehe Levene-Test).
- *Wilcoxon-Test*: Dieser Test vergleicht zwei abhängige Stichproben auf ihre zentrale Tendenz (Bortz, 1995, S. 144). Im Gegensatz zum t-Test für verbundene Stichproben benötigt der Wilcoxon-Test keine Normalverteilung und Ordinaldaten genügen als Voraussetzung.

5.4.1 Empfundene Aussagekraft der Tests

Hypothese 1

H0: Die empfundene Aussagekraft eindimensionaler Tests unterscheidet sich nicht von der empfundenen Aussagekraft mehrdimensionaler Tests.

H1: Die empfundene Aussagekraft eindimensionaler Tests unterscheidet sich von der empfundenen Aussagekraft mehrdimensionaler Tests.

Zur Prüfung der Hypothese wurde der Wilcoxon-Test gewählt. Der t-Test für verbundene Stichproben konnte ausgeschlossen werden, da dieser eine Normalverteilung für die Kategorie *Dimensionalität* voraussetzen würde, was nicht der Fall ist.

Nach der Anwendung des Wilcoxon Tests ist das Ergebnis eindeutig mit $T(977) = 30398$, $z = -19.989$, $p < .001$. Grundlage war die Einordnung der Tests in *Dimensionalität* (eindimensional und mehrdimensional) und die Frage Q5 (*Testergebnisse aus folgenden Tests sind aussagekräftig (...)* (fünfstufige Rating-skala: *außerordentlich* (1), *ziemlich* (2), *mittelmäßig* (3), *kaum* (4), *gar nicht* (5)).

Tabelle 26. Vergleich Aussagekraft (Q5) und Dimensionalität.

	N	Mittlerer Rang
Negative Ränge	680 ^a	431.19
Positive Ränge	124 ^b	245.15
Bindungen	173 ^c	
Gesamt	977	

Anmerkungen. a.: mehrdimensional < eindimensional. b.: mehrdimensional > eindimensional. c.: mehrdimensional = eindimensional. Ein höherer mittlerer Rang bedeutet eine geringer eingeschätzte Aussagekraft.

Es liegt ein signifikanter Unterschied vor. Die Nullhypothese wird verworfen und die Alternativhypothese angenommen. Die Aussagekraft eines Tests hängt von der *Dimensionalität* ab, denn der ermittelte signifikant höhere mittlere Rang bei den negativen Rängen bedeutet, dass eindimensionalen Tests eine geringere Aussagekraft zugeschrieben wird als mehrdimensionalen Tests (siehe Tabelle 26).

Die Auswahl der ProbandInnen resultiert überwiegend aus Anfragen an ehemalige TeilnehmerInnen von Diagnostikseminaren. Verzerrungen aus einer selektiven Stichprobe sollen durch den Vergleich mit einer Kontrollgruppe umgangen werden. Diese besteht aus ProbandInnen, die niemals an einer außer-universitären Fortbildung zur Testdiagnostik teilgenommen haben und entsprechend weder beeinflusst sind von der Person, die die Fortbildung durchgeführt hat (in der Regel der Autor dieser Arbeit) noch von den Inhalten der Fortbildung oder den Motiven für die Teilnahme an der Fortbildung (z. B. Vertiefung in die KABC-II).

Tabelle 27. Prüfung Aussagekraft (Q5) und Dimensionalität, getrennt nach Kontroll- und Versuchsgruppe.

	Versuchsgruppe		Kontrollgruppe	
	N	Rang	N	Rang
Negative Ränge	629 ^a	397.87	51 ^a	33.85
Positive Ränge	112 ^b	220.09	12 ^b	24.13
Bindungen	155 ^c		18 ^c	
Gesamt	896		81	

Anmerkungen. a.: mehrdimensional < eindimensional. b.: mehrdimensional > eindimensional. c.: mehrdimensional = eindimensional. Ein höherer mittlerer Rang bedeutet eine geringer eingeschätzte Aussagekraft.

Getrennt nach Kontroll- und Versuchsgruppe (Vergleich mittlere Ränge siehe Tabelle 27) und geprüft mit dem Wilcoxon-Test bleiben die Ergebnisse signifi-

kant mit jeweils $p < .001$ (Versuchsgruppe $T(896) = 2465.5$, $z = -19.394$, $p < .001$; Kontrollgruppe $T(81) = 289.5$, $z = -4.925$, $p < .001$).

5.4.2 Unterschiede zwischen *Komplexität* und Anwendungshäufigkeit

Hypothese 2

H0: Es besteht kein Unterschied in der Anwendungshäufigkeit zwischen verschiedenen komplexen Intelligenztests.

H1: Es besteht ein Unterschied in der Anwendungshäufigkeit zwischen verschiedenen komplexen Intelligenztests.

Entsprechend der Anzahl der Regeln wurde die *Komplexität* der Tests bestimmt: je mehr Regeln, desto komplexer in der Anwendung. Daraus resultieren fünf Gruppen:

<i>wenig komplex:</i>	CFT1/CFT1-R, CFT20-R
<i>leicht komplex:</i>	SON-R 6–40, IDS ⁹⁷
<i>komplex:</i>	K-ABC, WNV
<i>sehr komplex:</i>	WISC-IV, WPPSI-III
<i>außerordentlich komplex:</i>	KABC-II

Verglichen wurden diese 5 Kategorien mit Q8 (*Wenn ich teste, nehme ich folgende Tests (...);* fünfstufige Ratingskala: *immer* (1), *oft* (2), *gelegentlich* (3), *selten* (4), *nie* (5)). Zur Prüfung der Hypothese wurde der Friedman-Test (nicht-parametrische ANOVA) nach Rang genutzt. Da die Kategorie *außerordentlich komplex* aus nur einem Item besteht, kann die parametrische ANOVA für Messwiederholungen nicht verwendet werden, da stetige Zielgrößen verwendet werden müssten. Das Ergebnis ist eindeutig mit $F(4) = 293.98$, $p < .001$.

Es gibt einen signifikanten Zusammenhang zwischen der *Komplexität* und der Anwendungshäufigkeit. Zur inhaltlichen Beurteilung sollen die genaueren Zusammenhänge zwischen den fünf Kategorien mit post-hoc Verfahren geprüft werden.

Am häufigsten insgesamt werden Tests aus den Gruppen *wenig komplex* und *außerordentlich komplex* angewendet. Doch im Gegensatz zu den anderen paarweisen Vergleichen (siehe Tabelle 28) gibt es hier keinen signifikanten Unterschied im Vergleich bezüglich der Anwendungshäufigkeit. Dies bedeutet, dass weder CFT1-R bzw. CFT20-R (*wenig komplex*) seltener oder häufiger angewendet werden wie KABC-II (*außerordentlich komplex*).

97 Nur Intelligenzteil.

Tabelle 28. Unterschiede zwischen Komplexität und Anwendungshäufigkeit der Tests (Q8/Komplexität).

	Teststatistik	Sig.	korr. Sig.
wenig komplex – außerordentlich komplex	-0.062	.510	1.000
wenig komplex – leicht komplex	-0.452	.000	.000
wenig komplex – sehr komplex	-0.780	.000	.000
wenig komplex – komplex	-1.206	.000	.000
außerordentlich komplex – leicht komplex	0.390	.000	.000
außerordentlich komplex – sehr komplex	0.718	.000	.000
außerordentlich komplex – komplex	1.144	.000	.000
leicht komplex – sehr komplex	-0.328	.000	.005
leicht komplex – komplex	-0.754	.000	.000
sehr komplex – komplex	0.426	.000	.000

Anmerkungen. Signifikanzwerte wurden von der Bonferroni-Korrektur für mehrere Tests angepasst. Sig. = Signifikanz, korr. = korrigierte.

Leicht komplexe Tests (mittlerer Rang = 2,95) werden signifikant ($p = .005$) häufiger durchgeführt als *sehr komplexe* Tests (mittlerer Rang = 3,28). Bei allen anderen paarweisen Vergleichen liegt ein signifikanter Unterschied von $p < .001$ vor (mittlere Ränge siehe Tabelle 29): *wenig komplexe* Tests werden häufiger als *leicht* und *sehr komplexe* und *komplexe* Tests angewendet. Bis auf die KABC-II werden die Tests mit den wenigsten Regeln (CFT-Reihe) gegenüber den anderen Tests bevorzugt. Die KABC-II, die der einzige Test des Labels *außerordentlich komplex* ist, wird häufiger als *leicht* und *sehr komplexe* und *komplexe* Tests durchgeführt, *leicht komplexe* Tests häufiger als *komplexe* und *sehr komplexe* häufiger als *komplexe* Tests. Letzteres Ergebnis und die Präferenz für die KABC-II geben einen Hinweis, dass nicht grundsätzlich die vermeintlich *leicht* durchzuführenden Tests bevorzugt werden.

Tabelle 29. Mittlere Ränge Vergleich Q8 und Komplexität für Gesamt-, Kontroll- und Versuchsgruppe.

	mittlerer Rang Gesamt	mittlerer Rang Versuchsgruppe	mittlerer Rang Kontrollgruppe
wenig komplex	2.50 (1)	2.51 (1)	2.35 (1)
leicht komplex	2.95 (3)	2.95 (3)	2.96 (3)
komplex	3.71 (5)	3.72 (5)	3.52 (5)
sehr komplex	3.28 (4)	3.29 (4)	3.21 (4)
außerordentlich komplex	2.56 (2)	2.53 (2)	2.95 (2)

Anmerkung. In Klammern die Reihenfolge.

Für die Frage, ob ein Unterschied in der Anwendungshäufigkeit bei verschiedenen komplexen Intelligenztests vorliegt, scheint eine Prüfung getrennt nach Kontroll- und Versuchsgruppe sinnvoll.

Nach Anwendung der nichtparametrischen Varianzanalyse nach Friedman bei verbundenen Stichproben konnten signifikante Unterschiede sowohl für die Kontrollgruppe ($F(4) = 15.81, p = .003$), wie für die Versuchsgruppe ($F(4) = 282.81, p < .001$) festgestellt werden. Für beide Teilgruppen besteht ein Zusammenhang zwischen *Komplexität* und Anwendung der Tests. Die Darstellung der mittleren Ränge verdeutlicht, dass die Reihenfolge in allen Gruppen gleich ist: am häufigsten werden die *wenig komplexen* (CFT1/CFT1-R/CFT20-R) und der *außerordentlich komplexe* (= KABC-II) Test durchgeführt. Ob diese Ergebnisse signifikant sind, verdeutlicht die vergleichende Tabelle 30.

Obwohl die mittleren Ränge in der Reihenfolge gleich sind und nicht von *wenig komplex* zu *außerordentlich komplex* verlaufen, praktisch von leicht zu schwer in der Anwendung, sind die Unterschiede zwischen den fünf Kategorien beim Gegenüberstellen zwischen der Kontroll- und Versuchsgruppe different. Es gibt marginale Unterschiede. In der Gesamtgruppe werden *leicht komplexe* gegenüber den *sehr komplexen* Tests bevorzugt ($p = .005$), bei der Versuchsgruppe beträgt $p = .006$. Alle anderen Ergebnisse ähneln⁹⁸ den oben beschriebenen Ergebnissen der Gesamtgruppe.

In der Gegenüberstellung der drei Gruppen (Gesamtstichprobe, Kontroll- und Versuchsgruppe, siehe Tabelle 30) wird ersichtlich, dass es deutliche Unterschiede zwischen der Versuchs-⁹⁹ und Kontrollgruppe gibt.

Unter Berücksichtigung der Bonferroni-Korrektur kann für die Kontrollgruppe festgestellt werden, dass lediglich die *wenig komplexen* (mittlerer Rang = 2.35) gegenüber den *komplexen* (mittlerer Rang = 3.52) Tests bevorzugt werden ($p = .008$). Unter Vernachlässigung der konservativen (Hemmerich, 2015b) Bonferroni-Korrektur, die zwar falsch positive Ergebnisse verhindern hilft (Fehler 1. Art), aber auch die Gefahr falsch negativer Ergebnisse erhöht (Fehler 2. Art), ergeben sich folgende Ergebnisse: *wenig komplexe* Tests werden signifikant häufiger durchgeführt als *sehr komplexe* ($p = .015$) und *komplexe* ($p = .001$) Tests und tendenziell häufiger als die *außerordentlich komplexe* KABC-II ($p = .087$) und *leicht komplexe* ($p = .081$) Tests. Dies bedeutet, dass unter Vernachlässigung der Bonferroni-Korrektur in der Kontrollgruppe die Verfahren mit den wenigsten Anwendungsregeln und somit leichtesten Tests signifikant bzw. tendenziell signifikant häufiger angewendet werden als die Tests mit mehr Anwendungsregeln und somit vermeintlich schwereren Tests.

98 Die auf den ersten Blick identischen Ergebnisse können sich in Nachkommastellen unterscheiden, so dass scheinbar gleiche Ergebnisse nicht angenommen werden.

99 Und somit praktisch mit der Gesamtgruppe.

Tabelle 30. Unterschiede zwischen Anwendungshäufigkeiten und verschiedenen komplexen Tests (Q8/Komplexität), unterschieden nach Gesamt-, Vergleichs- und Kontrollgruppe.

	Gesamtgruppe		Versuchsgruppe		Kontrollgruppe	
	Sig.	Korr. Sig.	Sig.	Korr. Sig.	Sig.	Korr. Sig.
wenig komplex – außerordentlich komplex	.510	1.000	.837	1.000	.087	.871
wenig komplex – leicht komplex	.000	.000	.000	.000	.081	.808
wenig komplex – sehr komplex	.000	.000	.000	.000	.015	.145
wenig komplex – komplex	.000	.000	.000	.000	.001	.008
außerordentlich komplex – leicht komplex	.000	.000	.000	.000	.972	1.000
außerordentlich komplex – sehr komplex	.000	.000	.000	.000	.463	1.000
außerordentlich komplex – komplex	.000	.000	.000	.000	.101	1.000
leicht komplex – sehr komplex	.000	.005	.001	.006	.485	1.000
leicht komplex – komplex	.000	.000	.000	.000	.108	1.000
sehr komplex – komplex	.000	.000	.000	.000	.364	1.000

Anmerkungen. Korr. Sig. = korrigierte Signifikanz (angepasst mit Bonferroni-Korrektur).

Ergänzend und abschließend wurde mit dem Mann-Whitney-U-Test geprüft, ob Unterschiede in der Anwendung zwischen den Kategorien der *Komplexität* zwischen der Kontrollgruppe und der Versuchsgruppe vorliegen (siehe Tabellen B1 und B2). Als Gruppierungsvariable diente Q25 (*Haben Sie an einer außeruniversitären Fortbildung zu Intelligenztests teilgenommen? Ja/Nein*).

Tendenziell wird die KABC-II (= *außerordentlich komplex*) von der Kontrollgruppe (mittlerer Rang: 435,12) weniger genutzt als von der Versuchsgruppe (mittlerer Rang: 384,26, $U(718, 57) = 17777$, $p = .088$). Für die vier anderen Variablen konnten keine Unterschiede festgestellt werden.

5.4.3 Unterschiede zwischen Verfügbarkeit und Vorlieben für Tests

Hypothese 3

H0: Es besteht kein Unterschied zwischen der Verfügbarkeit von Intelligenztests und der Vorliebe für einen bestimmten Intelligenztest.

H1: Es besteht ein Unterschied zwischen der Verfügbarkeit von Intelligenztests und der Vorliebe für einen bestimmten Intelligenztest.

Die Verwendung der Tests wurde unter der Bedingung geprüft, dass die Tests auch tatsächlich zur Verfügung stehen.

Tabelle 31. Unterschiede zwischen Verfügbarkeit und Vorlieben für Tests.

	K-ABC KABC-II	CFT1/ CFT1-R	CFT20-R	WISC-IV	WPPSI-III	WNV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40	IDS
K-ABC:										
neg. Rang <i>n</i>	99	103	87	77	18	17	94	87	63	54
pos. Rang <i>n</i>	22	64	45	42	11	2	25	32	21	18
Bindungen <i>n</i>	67	74	56	56	25	5	51	77	31	29
Gesamt <i>N</i>	188	241	188	175	54	24	170	196	115	101
<i>z</i>	-7.661	-3.228	-4.113	-4.005	-1.362	-3.186	-6.047	-5.177	-4.477	-4.941
asympt. Sig.	<.001	.001	<.001	<.001	.173	.001	<.001	<.001	<.001	<.001
KABC-II:										
neg. Rang <i>n</i>		108	102	65	9	4	57	58	71	43
pos. Rang <i>n</i>		176	161	125	40	31	100	136	70	66
Bindungen <i>n</i>		76	66	75	14	12	65	52	64	53
Gesamt <i>N</i>		360	329	265	63	47	222	246	205	162
<i>z</i>		-5.110	-5.005	-4.498	-4.516	-4.298	-3.580	-5.824	-8.388	-2.059
asympt. Sig.		<.001	<.001	.001	<.001	<.001	<.001	<.001	.402	.040
CFT1/CFT1-R:										
neg. Rang <i>n</i>			61	95	20	12	113	110	101	77
pos. Rang <i>n</i>			64	128	40	22	76	101	62	54
Bindungen <i>n</i>			367	84	23	10	79	93	75	52
Gesamt <i>N</i>			492	307	83	44	268	304	238	183
<i>z</i>			-.334	-1.234	-2.124	-1.975	-3.207	-.532	-2.948	-2.262
asympt. Sig.			.738	.217	.034	.048	.001	.594	.003	.024
CFT20-R										
neg. Rang <i>n</i>				90	13	19	82	72	106	65
pos. Rang <i>n</i>				119	34	14	68	95	62	54
Bindungen <i>n</i>				93	28	10	66	90	63	45
Gesamt <i>N</i>				302	75	43	216	257	231	164
<i>z</i>				-1.099	-2.476	-.579	-1.543	-1.644	-3.443	-1.976
asympt. Sig.				.272	.006	.563	.123	.100	.001	.048
WISC-IV										
neg. Rang <i>n</i>					3	9	67	69	83	48
pos. Rang <i>n</i>					30	16	49	62	59	43
Bindungen <i>n</i>					46	13	56	70	37	32
Gesamt <i>N</i>					79	38	172	201	179	123
<i>z</i>					-3.767	-1.076	-2.053	-.749	-2.399	-.062
asympt. Sig.					<.001	.282	.040	.454	.016	.950

	K-ABC KABC-II	CFT1/ CFT1-R	CFT20-R	WISC-IV	WPPSI-III	WNV	SON-R 2 ½-7	SON-R 5 ½-17	SON-R 6-40	IDS
WPPSI-III										
neg. Rang <i>n</i>						6	34	30	34	26
pos. Rang <i>n</i>						2	10	12	4	4
Bindungen <i>n</i>						9	19	20	10	12
Gesamt <i>N</i>						17	63	62	48	42
<i>z</i>						-1.137	-3.627	-2.806	-4.589	-3.237
asympt. Sig.						.256	<.001	.005	<.001	.001
WNV										
neg. Rang <i>n</i>							6	6	11	12
pos. Rang <i>n</i>							6	13	5	9
Bindungen <i>n</i>							17	14	8	14
Gesamt <i>N</i>							29	33	24	35
<i>z</i>							-.954	-1.043	-1.704	-.704
asympt. Sig.							.340	.297	.088	.482
SON 2										
neg. Rang <i>n</i>								44	51	45
pos. Rang <i>n</i>								53	39	42
Bindungen <i>n</i>								189	86	24
Gesamt <i>N</i>								286	176	111
<i>z</i>								-1.706	-1.737	-.480
asympt. Sig.								.088	.082	.631
SON 5										
neg. Rang <i>n</i>									57	49
pos. Rang <i>n</i>									29	30
Bindungen <i>n</i>									91	32
Gesamt <i>N</i>									177	111
<i>z</i>									-4.020	-1.732
asympt. Sig.									<.001	.083
SON-R 6-40										
neg. Rang <i>n</i>										24
pos. Rang <i>n</i>										45
Bindungen <i>n</i>										46
Gesamt <i>N</i>										115
<i>z</i>										-2.548
asympt. Sig.										.011

Anmerkungen Tabelle 31. Ist der negative Rang höher, wird bei vorliegender Signifikanz der Test signifikant häufiger angewendet, der in der oberen Zeile steht; ist der positive Rang höher als der negative, wird bei vorliegender Signifikanz der in der linken Spalte untersuchte Test bevorzugt.

Paarweise Vergleiche (Q8: *Wenn ich teste, nehme ich folgende Tests (...) vs. Q12: Folgende Tests stehen mir zur Verfügung (...)*)¹⁰⁰ mit dem Wilcoxon-Test zwischen den Intelligenztests ergaben folgende Ergebnisse, unterteilt in Testverfahren, siehe Tabelle 31.

- *K-ABC*: Bis auf den Vergleich mit *WPPSI-III* ($T = 156.5$, $z = -1.362$, $p = .173$) werden die übrigen neun Tests signifikant häufiger eingesetzt ($p < .001$ bis $p = .001$) bei gleichzeitiger Verfügbarkeit.
- *KABC-II*: Bis auf den Vergleich mit *SON-R 6–40* ($T = 5408$, $z = 0.838$, $p = .402$) sind die Ergebnisse ähnlich eindeutig. *KABC-II* wird signifikant häufiger als *IDS* angewendet ($T = 1618.5$, $z = 2.548$, $p = .040$) und noch deutlich häufiger als die anderen acht Tests ($p < .001$ bis $p = .001$).
- *CFT1/CFT1-R*: Dieser Test wird signifikant häufiger verwendet als *WPPSI-III* ($T = 1195$, $z = 2.124$, $p = .034$), *WNV* ($T = 410.5$, $z = 1.975$, $p = .048$) und *K-ABC* ($p = .001$) ($T = 8992.5$, $z = 3.228$, $p = .001$), und seltener verwendet als *KABC-II* ($T = 13250$, $z = -5.110$, $p < .001$), *SON-R 2½–7* ($T = 6628.5$, $z = -3.207$, $p = .001$), *SON-R 6–40* ($T = 4934$, $z = -2.948$, $p = .003$) und *IDS* ($T = 3354.5$, $z = -2.262$, $p = .024$).
- *CFT20-R*: Dieser Test wird signifikant häufiger verwendet als *K-ABC* ($T = 6163.5$, $z = 4.113$, $p < .001$) und seltener als *KABC-II* ($p = .001$) ($T = 11262$, $z = -5.005$, $p < .001$), *SON-R 6–40* ($T = 4964$, $z = -3.443$, $p = .001$) und *IDS* ($T = 2837$, $z = -1.976$, $p = .048$).
- *WISC-IV*: Dieser Test wird signifikant häufiger verwendet als *K-ABC* ($T = 5067$, $z = 4.055$, $p < .001$) und *WPPSI-III* ($T = 486.5$, $z = 3.767$, $p < .001$) und seltener als *KABC-II* ($T = 5696$, $z = -4.498$, $p < .001$), *SON-R 2½–7* ($T = 2660.5$, $z = -2.053$, $p = .040$) und *SON-R 6–40* ($T = 3915$, $z = -2.399$, $p = .016$).
- *WPPSI-III*: Entweder gab es keine signifikanten Unterschiede oder der Test wird signifikant seltener als andere eingesetzt. Dies gilt für *KABC-II* ($T = 164.5$, $z = -4.516$, $p < .001$), für *CFT1/CFT1-R* ($T = 635$, $z = -2.124$, $p = .034$), *CFT20-R* ($T = 309$, $z = -2.746$, $p = .006$), *SON-R 2½–7* ($T = 190$, $z = -3.627$, $p < .001$), *SON-R 6–40* ($T = 58$, $z = -4.589$, $p < .001$) und *IDS* ($T = 77.5$, $z = -3.237$, $p = .001$).
- *WNV*: Da dieser Test nur selten vorhanden ist, sind die Fallzahlen geringer, so dass vorsichtiger interpretiert werden sollte. *WNV* wird signifikant häufiger durchgeführt als *K-ABC* ($T = 172.5$, $z = 3.186$, $p = .001$) und seltener als *KABC-II* ($T = 56$, $z = -4.298$, $p < .001$) und *CFT1/CFT1-R* ($T = 184.5$, $z = -1.975$, $p = .048$).

100 Likert-Skala: *immer* (1) – *oft* (2) – *gelegentlich* (3) – *seltener* (4) – *nie* (5).

- SON-R 2½-7: Dieser Test wird signifikant häufiger durchgeführt als K-ABC ($T = 5814$, $z = 6.047$, $p < .001$), CFT1/CFT1-R ($T = 11326$, $z = 3.207$, $p = .001$), WISC-IV ($T = 4125.5$, $z = 2.053$, $p = .040$) und WPPSI-III ($T = 800$, $z = 3.627$, $p < .001$) und seltener als KABC-II ($T = 4191$, $z = -3.580$, $p < .001$).
- SON-R 5½-17: Dieser Test wird signifikant häufiger durchgeführt als K-ABC ($T = 5481.5$, $z = 5.177$, $p < .001$) und WPPSI-III ($T = 672$, $z = 2.806$, $p = .005$) und seltener als KABC-II ($T = 4956$, $z = -5.824$, $p > .001$) und SON-R 6-40 ($T = 952.5$, $z = -4.020$, $p < .001$).
- SON-R 6-40: Dieser Test wird signifikant häufiger durchgeführt als IDS ($T = 1618.5$, $z = 2.548$, $p = .011$), K-ABC ($T = 2777$, $z = 4.477$, $p < .001$), CFT1/CFT1-R ($T = 8432$, $z = 2.948$, $p = .003$), CFT20-R ($T = 9232$, $z = 3.443$, $p = .001$), WISC-IV ($T = 6238$, $z = 2.399$, $p = .016$), WPPSI-III ($T = 683$, $z = 4.589$, $p < .001$) und SON-R 5½-17 ($T = 2788$, $z = 4.020$, $p < .001$).
- IDS: Dieser Test wird signifikant häufiger als K-ABC ($T = 2183$, $z = 4.941$, $p < .001$), CFT1/CFT1-R ($T = 5291.5$, $z = 2.262$, $p = .024$), CFT20-R ($T = 4303$, $z = 1.976$, $p = .048$) ($T = 156.5$, $z = -1.362$, $p = .173$) und WPPSI-III ($T = 387.5$, $z = 3.237$, $p = .001$) und seltener als KABC-II ($T = 2325$, $z = -2.059$, $p = .040$) durchgeführt.

Bis hierhin wurde geprüft, ob bei zwei vorhandenen Tests einer präferiert wird. Es ist jedoch auch möglich, dass gleichzeitig drei oder mehr Intelligenztests zur Verfügung stehen. Im Folgenden wurde geprüft, ob Tests bevorzugt angewendet werden, wenn mehr als ein weiterer zur Verfügung steht. Unter der Bedingung des Vorhandenseins mehrerer Tests würden entsprechend eingesetzte Filter dazu führen, dass die Fallzahl sich evtl. erheblich reduziert ab der Prüfung von drei oder mehr gleichzeitig verfügbarer Tests. Um aussagekräftige Ergebnisse zu erzielen, wird auf die Prüfung bei zu geringen Fallzahlen (z.B. einstellige) verzichtet. Die geringste verwendete Fallzahl beträgt 25. Die Auswahl der Kombinationen basierten auf inhaltlichen Überlegungen und beschränken sich auf Grund der Vielzahl von Möglichkeiten auf ausgewählte Kombinationen¹⁰¹. Bei der Annahme, dass die *schnellen* Tests präferiert werden gegenüber den aufwändigen Tests, sind z.B. Vergleiche zwischen ein- bzw. mehrdimensionalen Tests sinnvoll.

Der paarweise Vergleich mit dem Friedman-Test nach Rang ermittelte für unterschiedliche Kombinationen folgende Ergebnisse:

101 Durch die große Anzahl an Kombinationsmöglichkeiten wurde zudem darauf verzichtet, zu prüfen, ob bestimmte Tests präferiert werden, wenn andere Tests vorhanden sind *und* andere Tests *nicht* zur Verfügung stehen. Es ist also durchaus möglich, dass bei einer Prüfung zwischen vier zur Verfügung stehenden Tests auch noch weitere Tests zur Verfügung stehen.

Prüfung unter der Bedingung, dass eine Vielzahl von Tests zur Verfügung stehen:

In dieser ersten Bedingung ist von Interesse, ob bei einer größeren Auswahl an Tests bestimmte bevorzugt angewendet werden.

Lediglich zwei ProbandInnen gaben an, über alle elf der für diese Arbeit besonders interessierenden Tests zu verfügen. Deshalb sind in der ersten Berechnung K-ABC, WPPSI-III, IDS und WNV ausgelassen worden, so dass immerhin 39 SonderpädagogInnen¹⁰²angaben, über die übrigen acht Tests zu verfügen. Dies ist die maximal mögliche Anzahl von Tests, für die eine Auswertung sinnvoll ist unter der Bedingung, dass möglichst viele Tests zur Verfügung stehen. Beim Vergleich der mittleren Ränge konnte ein signifikanter Unterschied bezüglich der Anwendungshäufigkeit ermittelt werden ($F(6) = 12.82, p = .046$), nach der Bonferroni-Korrektur allerdings kein signifikanter Unterschied zwischen zwei Tests.

Eine weitere Kombination von vielen Tests entstände bei der Auslassung von KABC-II, K-ABC, IDS, WNV und WPPSI-III ($N = 54$). Auch hier konnte für die verbliebenen sechs Tests keine Signifikanz festgestellt werden ($F(5) = 1.51, p = .912$). Dies ist auch der Fall, wenn WNV, IDS, SON-R 6–40 und WPPSI-III ausgelassen werden ($F(6) = 7.42, p = .283$)

Liegt also eine Auswahl mehrerer Tests vor, konnte nicht festgestellt werden, dass einer der Tests bevorzugt angewendet wird.

Die Ergebnisse dieser Bedingungen sind unter Berücksichtigung kleinerer Fallzahlen zu interpretieren.

Prüfung unter der Bedingung, dass die ein- bzw. mehrdimensionalen Tests zur Verfügung stehen:

Die für diese Arbeit vorgenommene Einteilung in ein- (CFT1/CFT1-R, CFT20-R, SON-R 6–40) bzw. mehrdimensionale Tests (K-ABC, KABC-II, WISC-IV, WPPSI-III, SON-R 2½–7) ermittelt eine zu geringe Fallzahl, so dass bei Auslassung der K-ABC 25 Personen gleichzeitig über die anderen Verfahren verfügen. Es gibt für diesen Fall signifikante Unterschiede ($F(6) = 25.68, p < .001$), die auch nach der strengen Bonferroni-Korrektur für den Vergleich zwischen den Tests vorliegen. Die KABC-II (mittlerer Rang = 3.16; $p = .002$), der SON-R 6–40 (mittlerer Rang = 3.32; $p = .005$) und der SON-R 2½–7 (mittlerer Rang = 3.54, $p = .018$) werden signifikant häufiger als der WPPSI-III (mittlerer Rang = 5.58) angewendet.

102 Unter Berücksichtigung der Gewichtungen.

Prüfung unter der Bedingung, dass die KABC-II und eindimensionale Tests zur Verfügung stehen:

Stehen neben der KABC-II alle eindimensionalen Tests zur Verfügung, gibt es keinen signifikanten Unterschied in der Anwendung ($F(3) = 1.58, p = .663$)

Prüfung unter der Bedingung, dass der WISC-IV und eindimensionale Tests zur Verfügung stehen:

Auch für diese Bedingung konnte kein signifikanter Unterschied festgestellt werden ($F(3) = 2.70, p = .439$).

Prüfung unter der Bedingung, dass die mehrdimensionalen Tests zur Verfügung stehen:

Bei der Prüfung zwischen KABC-II, WISC-IV und WPPSI-III ergaben sich signifikante Unterschiede ($F(2) = 16.25, p < .001$). Die KABC-II (mittlerer Rang = 1.72) wird signifikant häufiger angewendet ($p = .006$) als der WPPSI-III (mittlerer Rang = 2.35).

Bei der Prüfung zwischen KABC-II, WISC-IV und SON-R 2½-7 ergaben sich ebenfalls signifikante Unterschiede ($N = 116; F(2) = 18.12, p < .001$). Die KABC-II (mittlerer Rang = 1.73) wird signifikant häufiger als der SON-R 2½-7 (mittlerer Rang = 2,07; $p = .026$) und der WISC-IV (mittlerer Rang = 2.20; $p = .001$) eingesetzt.

Bei der Prüfung zwischen WISC-IV, SON-R 2½-7 und WPPSI-III liegen ebenfalls Unterschiede vor ($N = 47; F(2) = 12.39, p = .002$). Der SON-R 2½-7 (mittlerer Rang: 1.74) wird signifikant häufiger als der WPPSI-III (mittlerer Rang = 2.33; $p = .014$) eingesetzt.

Prüfung unter der Bedingung, dass die eindimensionalen Tests zur Verfügung stehen:

Bei der Prüfung von CFT1/CFT1-R, CFT20-R und SON-R 6-40 gibt es zwar insgesamt eine Signifikanz ($N = 190; F(2) = 6.98, p = .030$), allerdings nach der strengen Bonferroni-Korrektur nicht zwischen den Tests.

Prüfung unter der Bedingung, dass die neuesten und die ältesten Tests zur Verfügung stehen:

Im Vergleich zwischen den aktuellsten Tests KABC-II und SON-R 6-40 mit den ältesten Tests K-ABC und SON-R 5½-17 gibt es insgesamt ($N = 59$) einen signifikanten Unterschied ($F(3) = 22.76, p < .001$). Die KABC-II (mittlerer Rang = 2.19; $p = .002$) und der SON-R 6-40 (mittlerer Rang = 2.21; $p = .003$) werden signifikant häufiger als die K-ABC eingesetzt (mittlerer Rang = 3.04). Allerdings sei erwähnt, dass kein signifikanter Unterschied in der Anwendung zwischen dem SON-R 5½-17 (mittlerer Rang = 2.55) und dem SON-R 6-40 ($p = .923$) und zwischen der KABC-II und dem SON-R 5½-17 ($p = .806$) attestiert werden

kann. Dies wäre noch nicht einmal ohne Bonferroni-Korrektur der Fall (KABC-II vs. SON-R 5½-17: $p = .134$; SON-R 6-40 vs. SON-R 5½-17: $p = .154$).

Unter der Bedingung, dass die nonverbalen bzw. sprachfairen Tests zur Verfügung stehen:

Die sechs Verfahren, die auch die Möglichkeit der sprachfreien (SON-R 2½-7, SON-R 6-40, WNV) bzw. sprachfairen Testungen ermöglichen (CFT1/CFT1-R, CFT20-R) sind zu selten gemeinsam vorhanden ($N = 6$), so dass der WNV als wenig verbreiteter Test ausgelassen worden ist. Unter dieser Bedingung gibt es keine signifikanten Unterschiede ($N = 80$, $p = .670$). Stehen mehrere sprachfreie bzw. -faire Tests zur Verfügung, konnte kein Hinweis ermittelt werden, dass ein Test der SON- bzw. CFT-Reihe bevorzugt angewendet werden würde.

Hypothesenprüfung mit der Kontrollgruppe:

Die besondere Stellung der KABC-II resultiert möglicherweise aus einer selektiven Auswahl der Stichprobe, die überwiegend aus ehemaligen TeilnehmerInnen von Diagnostik-Fortbildungen besteht. An den Fortbildungen nahm die Vorstellung der KABC-II eine prominente Stellung ein. Es könnte also eine besondere Affinität zu diesem Test angenommen werden. Dies ist ein Beispiel dafür, dass entsprechende Verzerrungen bei der Interpretation der Ergebnisse berücksichtigt werden müssen.

In der Gegenüberstellung zwischen der Kontrollgruppe – ProbandInnen, die noch nie an einer außeruniversitären Fortbildung zum Thema teilnahmen – und den ProbandInnen, die an einer Fortbildung teilnahmen, sind durch die geringere Fallzahl der Kontrollgruppe lediglich Prüfungen für die Annahme sinnvoll, dass jeweils zwei Tests gleichzeitig zur Verfügung stehen. Tabelle 32 zeigt die Ergebnisse der Prüfung an, Tabelle 33, wie häufig die Tests vorhanden sind. Geprüft worden sind alle Kombinationen für die elf Intelligenztests, beschrieben werden jedoch lediglich die Kombinationen mit einer Fallzahl im zweistelligen Bereich zur Erhöhung der statistischen Power, auch wenn der verwendete Wilcoxon-Test für kleine Fallzahlen geeignet ist. Da die Gesamtfallzahl sehr groß ist, besteht allerdings keine Notwendigkeit, auf Prüfungen mit sehr kleinen Fallzahlen zurückzugreifen. Zur Vermeidung von Artefakten wird deshalb diese Grenze festgelegt.

Die Unterschiede zwischen der Versuchs- und Kontrollgruppe sind beachtlich. Für einige Tests liegen zu wenige Übereinstimmungen vor, so dass die Ergebnisse der Prüfungen für WNV, WPPSI-III und IDS entfallen. Die Signifikanzprüfungen werden zusammengefasst mit den p-Werten dargestellt.

Tabelle 32. Unterschiede zwischen Verfügbarkeit und Vorlieben für Tests, unterschieden nach Versuchs- und Kontrollgruppe – Fortsetzung nächste Seite.

	Versuchsgruppe						Kontrollgruppe							
	KABC-II	CFT1/CFT1-R	CFT20-R	WISC-IV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40	KABC-II	CFT1/CFT1-R	CFT20-R	WISC-IV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40
K-ABC														
neg. Rang <i>n</i>	94	94	80	73	88	83	5	9	7	4	6	4		
pos. Rang <i>n</i>	18	57	41	37	22	30	4	7	4	5	3	2		
Bindungen <i>n</i>	64	72	51	53	43	66	3	2	5	3	8	11		
Gesamt <i>N</i>	176	223	172	163	153	179	12	18	16	12	17	17		
<i>z</i>	-7.6	-3.2	-4.1	-4.4	-6.0	-5.0	-1.1	-0.7	-0.7	-0.2	-0.8	-1.4		
asympt. Sig.	<.001	.001	<.001	<.001	<.001	<.001	.280	.511	.500	.810	.399	.163		
KABC-II														
neg. Rang <i>n</i>		104	96	60	48	51	67		4	6	5	9	7	3
pos. Rang <i>n</i>		167	150	118	92	128	67		9	11	7	8	8	4
Bindungen <i>n</i>		70	62	74	64	49	54		6	4	1	1	3	10
Gesamt <i>N</i>		341	308	252	204	228	188		19	21	13	18	18	17
<i>z</i>		-5.2	-4.9	-4.2	-3.3	-5.6	-0.9		-0.2	-1.1	-1.6	-1.1	-1.5	-0.9
asympt. Sig.		<.001	<.001	<.001	.001	<.001	.360		.804	.279	.102	.271	.133	.660
CFT1-R														
neg. Rang <i>n</i>			58	91	106	101	95			3	4	7	9	6
pos. Rang <i>n</i>			55	120	66	95	59			9	8	10	6	3
Bindungen <i>n</i>			345	79	73	81	66			22	5	6	12	9
Gesamt <i>N</i>			458	290	245	277	220			34	17	23	27	18
<i>z</i>			-1.0	-0.9	-3.5	-0.7	-2.9			-1.8	-1.4	-0.7	-0.4	-0.5
asympt. Sig.			.920	.323	.001	.497	.004			.071	.169	.537	.665	.590
CFT20-R														
neg. Rang <i>n</i>				82	74	63	95				8	8	9	11
pos. Rang <i>n</i>				113	62	88	59				6	6	7	3
Bindungen <i>n</i>				87	59	78	54				6	7	12	9
Gesamt <i>N</i>				282	195	229	208				20	21	28	23
<i>z</i>				-1.6	-1.4	-1.8	-3.1				-1.6	-0.7	-0.1	-1.6
asympt. Sig.				.118	.159	.078	.002				.101	.497	.958	.111

	Versuchsgruppe				Kontrollgruppe									
	KABC-II	CFT1/CFT1-R	CFT20-R	WISC-IV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40	KABC-II	CFT1/CFT1-R	CFT20-R	WISC-IV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40
WISC-IV														
neg. Rang <i>n</i>					59	60	75					8	9	8
pos. Rang <i>n</i>					47	59	56					2	3	3
Bindungen <i>n</i>					50	67	35					6	3	2
Gesamt <i>N</i>					156	186	166					16	15	13
<i>z</i>					-1.7	-.2	-1.9					-1.7	-1.7	-2.0
asymp. Sig.					.092	.855	.052					.084	.087	.041
SON-R 2½-7														
neg. Rang <i>n</i>						38							6	
pos. Rang <i>n</i>						47							6	
Bindungen <i>n</i>						167							22	
Gesamt <i>N</i>						252							34	
<i>z</i>						-1.6							-0.5	
asymp. Sig.						.108							.597	
SON-R 5½-17														
neg. Rang <i>n</i>							52							5
pos. Rang <i>n</i>							26							3
Bindungen <i>n</i>							78							13
Gesamt <i>N</i>							156							21
<i>z</i>							-3.8							-1.3
asymp. Sig.							<.001							.196

Anmerkungen. Ist der negative Rang höher, wird bei vorliegender Signifikanz der Test signifikant häufiger angewendet, der in der oberen Zeile steht; ist der positive Rang höher als der negative, wird bei vorliegender Signifikanz der in der linken Spalte untersuchte Test bevorzugt. Auslassungen: zu geringe Fallzahl. *z*: *z*-Wert aus Gründen der Übersichtlichkeit auf eine Dezimalstelle reduziert.

Tabelle 33. Übersicht vorhandene Intelligenztests in der Versuchs- und Kontrollgruppe.

	K-ABC	KABC-II	CFT1/CF T1-R	CFT20-R	WISC-IV	WPPSI-III	WNV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40	IDS
Kontrollgruppe <i>n</i> = 101	44	45	54	57	46	23	4	58	62	36	22
Versuchsgruppe <i>n</i> = 927	369	562	617	580	419	117	68	396	450	338	259

- *K-ABC*¹⁰³: Während in der Versuchsgruppe für jede geprüfte Kombination die anderen Tests präferiert werden ($p < .001$ bis $p = .001$), konnten für die Kontrollgruppe keine Signifikanzen festgestellt werden. Unabhängig von der Qualität, des Alters, der *Komplexität* oder der *Dimensionalität* der anderen Tests wird die K-ABC nicht seltener oder häufiger von der Kontrollgruppe genutzt, auch wenn bessere und aussagekräftigere Tests zur Verfügung stehen.
- *KABC-II*: Ähnliches gilt für die KABC-II, sie wird in der Kontrollgruppe nicht signifikant häufiger oder seltener angewendet bei gleichzeitigem Vorhandensein anderer Tests. In der Versuchsgruppe wird die KABC-II gegenüber den anderen Tests bevorzugt ($p < .001$ bis $p = .001$). Eine Ausnahme bildet lediglich der SON-R 6–40 ($p = .360$).
- *CFT1/CFT1-R*: In der Kontrollgruppe liegen keine Signifikanzen vor. In der Versuchsgruppe wird der CFT1/CFT1-R signifikant häufiger als die K-ABC ($p = .001$) und seltener als die KABC-II ($p < .001$), SON-R 2½–7 ($p = .001$) und SON-R 6–40 ($p = .004$) eingesetzt.
- *CFT20-R*: Es liegen keine signifikanten Ergebnisse für die Kontrollgruppe vor. Bei der Versuchsgruppe werden die K-ABC ($p < .001$) seltener und die KABC-II ($p < .001$) und der SON-R 6–40 ($p = .002$) häufiger eingesetzt.
- *SON-R 2½–7*: Es liegen keine signifikanten Ergebnisse für die Kontrollgruppe vor, in der Versuchsgruppe wird die K-ABC ($p < .001$) seltener und die KABC-II ($p = .001$) häufiger eingesetzt.
- *SON-R 5½–17*: Während hier ebenfalls keine Unterschiede in der Anwendung in der Kontrollgruppe festgestellt werden konnten, wird in der Versuchsgruppe die K-ABC ($p < .001$) seltener und die KABC-II ($p = .001$) häufiger eingesetzt.
- *SON-R 6–40*: Der einzig ermittelte signifikante Unterschied für die Kontrollgruppe konnte beim Vergleich mit dem WISC-IV ($p = .041$) festgestellt werden, welcher seltener angewendet wird. In der Versuchsgruppe werden der SON-R 5½–17 ($p < .001$), der CFT1/CFT1-R ($p = .004$) und der CFT20-R ($p = .002$) seltener angewendet.

5.4.4 Unterschiede in der Anwendung der Tests abhängig vom Bundesland

Im Vergleich zwischen den Bundesländern werden Unterschiede bezüglich der Anwendung und der mit der Anwendung verbundenen Problematiken angenommen. Der Vergleich zwischen den Bundesländern wird mit Hilfe von sie-

103 Kein Vergleich mit dem SON-R 6–40, da die Fallzahl zu niedrig ist.

ben Hypothesen geprüft. Obwohl ProbandInnen aus allen Bundesländern geantwortet haben, werden lediglich die Bundesländer verglichen, deren Fallzahl akzeptable Auswertungen erwarten lassen. Bei der letztmalig erhobenen Statistik für die deutsche Bevölkerung am 31. 12. 2015 (Destatis, 2015) sind 82.18 Millionen in Deutschland lebende Personen angegeben. Die für die Hypothesenprüfungen in Frage kommenden Länder repräsentieren zusammen ca. 51.48 Millionen Menschen (ca. 62.64% der Gesamtbevölkerung). Es sind Baden-Württemberg (N = 130), Hamburg (N = 29), Hessen (N = 109), Niedersachsen (N = 143), Nordrhein-Westfalen (N = 465), Rheinland-Pfalz (N = 51) und Schleswig-Holstein (N = 31)¹⁰⁴.

Hypothese 4.1

H0: Es besteht kein Zusammenhang zwischen dem Bundesland und der Verfügbarkeit der Intelligenztests.

H1: Es besteht ein Zusammenhang zwischen dem Bundesland und der Verfügbarkeit der Intelligenztests.

Für die Prüfung wurden der Chi-Quadrat-Unabhängigkeitstest und Kreuztabellen genutzt. Der Chi-Quadrat-Test akzeptiert auch kleinere Fallzahlen bei mindestens fünf Beobachtungen je Zelle, so dass die Ergebnisse aus Hamburg und Schleswig-Holstein problemlos ausgewertet werden konnten.

Die Prüfung ergab für jeden Intelligenztest signifikante Unterschiede zwischen dem jeweiligen Bundesland im Vergleich mit den anderen Bundesländern mit ausreichend hohen Fallzahlen, siehe zusammengefasst in Tabelle 34. Tatsächlich stehen abhängig vom Bundesland unterschiedliche Tests zur Verfügung.

Tabelle 34. Signifikanzprüfung mit dem Chi-Quadrat-Test nach Pearson, ob unterschiedliche Tests abhängig vom Bundesland zur Verfügung stehen.

K-ABC	KABC-II	CFT1/ CFT1-R	CFT20-R	WISC-IV	WPPSI- III	WNV	SON-R 2½-7	SON-R 5½-17	SON-R 6-40	IDS
$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p = .001$	$p = .028$	$p < .001$	$p = .002$	$p < .001$	$p < .001$

Anmerkung. Vergleich für ausgewählte Bundesländer mit ausreichend hohen Fallzahlen.

Ergänzend wurden die Berechnungen für alle 16 Bundesländer ausgeführt, die Unterschiede sind marginal. Veränderungen liegen vor für WPPSI-III ($\chi^2(15, N = 1025) = 44.52, p < .001$) und WNV ($\chi^2(15, N = 1024) = 112.91, p = .001$).

104 Abweichungen von den im deskriptiven Kapitel angegebenen Stichproben resultieren aus den weiter oben beschriebenen Gewichtungen.

Dass in den Bundesländern unterschiedliche Tests zur Verfügung stehen ist interessant, gibt aber noch keine Hinweise darauf, welche Unterschiede im Einzelnen vorliegen und welche Schlussfolgerungen aus diesen Unterschieden gezogen werden könnten. Die häufige Anwendung veralteter Tests in einem Bundesland bei gleichzeitigem Mangel an aussagekräftigen Tests gäbe z.B. einen Hinweis auf notwendige Veränderungen in der Anschaffungspraxis.

Exemplarisch soll die Verfügbarkeit der KABC-II und des SON-R 5½–17 detailliert vorgestellt werden. Dies ist interessant, da die KABC-II derzeit einer der aktuellsten und aussagekräftigsten Tests, der SON-R 5½–17 hingegen einer der ältesten und deutlich weniger aussagekräftigen Tests ist.

So zeigen sich bei der Verfügbarkeit der KABC-II deutliche Unterschiede zwischen den Bundesländern, siehe Tabelle 35.

Tabelle 35. Verfügbarkeit der KABC-II und des SON-R 5½–17 in ausgewählten Bundesländern.

	KABC-II			SON-R 5½–17		
	nicht vorhanden	vorhanden	Gesamt	nicht vorhanden	vorhanden	Gesamt
BW	32	98	130	70	61	131
	25 %	75 %	100 %	53 %	47 %	100 %
HH	18	11	29	23	6	29
	63 %	38 %	100 %	79 %	21 %	100 %
HE	21	89	110	50	60	110
	19 %	81 %	100 %	46 %	54 %	100 %
NI	73	70	143	77	66	143
	51 %	49 %	100 %	54 %	46 %	100 %
NRW	215	250	465	206	259	465
	46 %	53,8 %	100 %	44 %	56 %	100 %
RP	17	34	51	24	27	51
	33 %	66,7 %	100 %	47 %	53 %	100 %
SH	2	29	31	20	11	31
	7 %	93 %	100 %	64 %	36 %	100 %
Ges.	378	581	959	470	490	960
	39 %	61 %	100 %	49 %	51 %	100 %

Anmerkung. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein. Ges. = Gesamt.

Während die KABC-II im Durchschnitt bei 61 Prozent Verfügbarkeit liegt, sind dies deutlich mehr in Baden-Württemberg (75%) und weniger in Niedersachsen

(49%). Während also in Baden-Württemberg drei von vier SonderpädagogInnen über die KABC-II verfügen, sind dies in Niedersachsen zwei von vier. Auch ein kurzer Blick auf die Verfügbarkeit des SON-R 5½–17 zeigt Unterschiede:

Beim SON-R 5½–17 ist vor allem interessant, dass dieser veraltete Test insgesamt noch zur Hälfte vorhanden ist (51%), in Hamburg (20%) und Schleswig-Holstein (35%) jedoch unterdurchschnittlich selten¹⁰⁵.

Ob die Durchschnittsunterschiede signifikant sind, wurde mit dem Chi-Quadrat-Test nach Pearson geprüft, sowohl für die KABC-II und den SON-R 5½–17 als auch für die anderen Testverfahren. Tabelle 36 gibt einen Überblick über signifikante Unterschiede.

Tabelle 36. Signifikanzprüfung mit dem Chi-Quadrat-Test nach Pearson bezüglich der Verfügbarkeit der Tests in ausgewählten Bundesländern, verglichen mit der Gesamtheit der übrigen Bundesländer.

	K-ABC	KABC-II	CFT1/ CFT1-R	CFT20-R	WISC-IV	WPPSI-III	WNV	SON-R 2½–7	SON-R 5½–17	SON-R 6–40	IDS
BW	.000	.000	.008	.000	.424	.002	.138	.000	.427	.028	.004
HH	.003	.019	.713	.051	.065	.107	.138	.003	.001	.318	.000
HE	.006	.000	.000	.000	.000	.824	.831	.295	.293	.000	.000
NI	.142	.008	.147	.001	.617	.361	.034	.042	.340	.000	.015
NRW	.120	.002	.000	.000	.000	.005	.003	.169	.001	.000	.029
RP	.299	.256	.044	.013	.022	.011	.403	.032	.651	.306	.097
SH	.098	.000	.289	.001	.472	.679	.412	.321	.105	.045	.311

Anmerkungen. Fettdruck = Test ist signifikant häufiger vorhanden. Kursivdruck = Test ist signifikant seltener vorhanden. Aus Gründen der Übersichtlichkeit wird ausschließlich *p* angegeben. BW = Baden-Württemberg, HH = Hamburg, HE = Hessen, NI = Niedersachsen, NRW = Nordrhein-Westfalen, RP = Rheinland-Pfalz, SH = Schleswig-Holstein.

Tabelle 36 verdeutlicht, dass die unterschiedliche Verfügbarkeit (seltener oder häufiger vorhanden als in anderen Bundesländern) häufig vorkommt. Lediglich in Schleswig-Holstein ist dies nicht der Fall: sieben Tests sind weder häufiger noch seltener vorhanden, zwei Tests häufiger und ein Test seltener, ein weiterer Test tendenziell seltener.

105 Wäre dieser Unterschied signifikant, wäre dies lediglich ein Hinweis, dass dieser veraltete Test seltener vorhanden ist und darf nicht gleichgesetzt werden mit der Annahme, dass dann eher aktuellere Tests angewendet werden anstatt veralteter Tests. Es wäre auch möglich, dass weder alte noch neue Tests vorhanden sind.

Die Ergebnisse im Detail für die ausgewählten Bundesländer:

- In Baden-Württemberg verfügen die SonderpädagogInnen signifikant häufiger über die Tests K-ABC ($\chi^2(1, N = 1027) = 28.43, p < .001$), KABC-II ($\chi^2(1, N = 1027) = 16.32, p < .001$), WPPSI-III ($\chi^2(1, N = 1028) = 9.26, p = .002$), SON-R 2½-7 ($\chi^2(1, N = 1027) = 21.49, p < .001$) und IDS ($\chi^2(1, N = 1024) = 8.16, p = .004$) und seltener über die Tests CFT1/CFT1-R ($\chi^2(1, N = 1028) = 7.04, p = .008$), CFT20-R ($\chi^2(1, N = 1028) = 31.59, p < .001$) und SON-R 6-40 ($\chi^2(1, N = 1028) = 4.85, p = .028$).
- In Hamburg verfügen die SonderpädagogInnen signifikant häufiger über die IDS ($\chi^2(1, N = 1028) = 14.71, p < .001$) und signifikant seltener über die Tests K-ABC ($\chi^2(1, N = 1028) = 8.59, p = .003$), KABC-II ($\chi^2(1, N = 1028) = 5.50, p = .019$), SON-R 2½-7 ($\chi^2(1, N = 1028) = 8.83, p = .003$) und SON-R 5½-17 ($\chi^2(1, N = 1028) = 10.12, p = .001$).
- In Hessen verfügen die SonderpädagogInnen signifikant häufiger über die K-ABC ($\chi^2(11, N = 1027) = 7.53, p = .006$), KABC-II ($\chi^2(11, N = 1028) = 24.35, p < .001$), CFT1/CFT1-R ($\chi^2(1, N = 1028) = 27.97, p < .001$), CFT20-R ($\chi^2(1, N = 1028) = 35.92, p < .001$), WISC-IV ($\chi^2(1, N = 1029) = 22.08, p < .001$), SON-R 6-40 ($\chi^2(15, N = 1024) = 112.91, p = .001$) und IDS ($\chi^2(1, N = 1029) = 22.53, p < .001$). Es fällt auf, dass es keinen Test gibt, über den hessische SonderpädagogInnen signifikant seltener verfügen.
- In Niedersachsen verfügen die SonderpädagogInnen signifikant häufiger über den CFT20-R ($\chi^2(1, N = 1028) = 11.66, p = .001$) und signifikant seltener über die KABC-II ($\chi^2(1, N = 1028) = 7.00, p = .008$), den WNV ($\chi^2(1, N = 1028) = 4.51, p = .034$), den SON-R 2½-7 ($\chi^2(1, N = 1029) = 4.15, p = .042$), den SON-R 6-40 ($p < .001$) ($\chi^2(1, N = 1027) = 33.29, p < .001$) und die IDS ($\chi^2(1, N = 1029) = 5.94, p = .015$).
- In Nordrhein-Westfalen verfügen die SonderpädagogInnen signifikant häufiger über den SON-R 5½-17 ($\chi^2(1, N = 1029) = 11.60, p = .001$) und den SON-R 6-40 ($\chi^2(1, N = 1027) = 14.31, p < .001$), allerdings seltener über sieben der elf untersuchten Tests: KABC-II ($\chi^2(1, N = 1028) = 9.80, p = .002$), CFT1/CFT1-R ($\chi^2(1, N = 1028) = 28.85, p < .001$), CFT20-R ($\chi^2(1, N = 1028) = 28.85, p < .001$), WISC-IV ($\chi^2(1, N = 1027) = 21.83, p < .001$), WPPSI-III ($\chi^2(1, N = 1028) = 7.84, p = .005$), WNV ($\chi^2(1, N = 1028) = 8.97, p = .003$) und IDS ($\chi^2(1, N = 1027) = 4.77, p = .029$).
- In Rheinland-Pfalz verfügen die SonderpädagogInnen signifikant häufiger über die Tests CFT1/CFT1-R ($\chi^2(1, N = 1027) = 4.06, p = .044$), CFT20-R ($\chi^2(1, N = 1027) = 6.13, p = .013$), WISC-IV ($\chi^2(1, N = 1027) = 5.21, p = .022$) und WPPSI-III ($\chi^2(1, N = 1028) = 6.43, p = .011$) und seltener über den SON-R 2½-7 ($\chi^2(1, N = 1028) = 4.61, p = .032$).
- In Schleswig-Holstein verfügen die SonderpädagogInnen signifikant häufiger über die KABC-II ($\chi^2(1, N = 1028) = 15.74, p < .001$) und den CFT20-R

$(\chi^2(1, N = 1028) = 10.84, p = .001)$. Seltener vorhanden ist der SON-R 6–40
 $(\chi^2(1, N = 1028) = 4.00, p = .045)$.

Hypothese 4.2

H0: Es besteht kein Unterschied zwischen den Bundesländern und vorgenommenen Veränderungen bei der Anwendung von Intelligenztests, die die Durchführungsobjektivität gefährden.

H1: Es besteht ein Unterschied zwischen den Bundesländern und vorgenommenen Veränderungen bei der Anwendung von Intelligenztests, die die Durchführungsobjektivität gefährden.

Für diese Hypothesenprüfung werden die drei Items aus Q14 mit den Bundesländern verglichen:

- Q14: Welche dieser Veränderungen haben Sie schon einmal vorgenommen?¹⁰⁶
 - Q14/1: Durchführungszeiten geändert (z.B. nach Ablauf der regulären Durchführungszeit/Item einen Punkt gegeben bei richtiger Antwort)
 - Q14/2: Durchführungszeit ganz weggelassen
 - Q14/3: Rückmeldungen gegeben, wenn diese nicht vorgesehen waren (z. B. *richtig* oder *hast du richtig gelöst*)

Mittel der Wahl war aufgrund des ordinalen Skalenniveaus der Daten die ein-faktorielle ANOVA nach Kruskal-Wallis (Kruskal & Wallis, 1952). Dieser Test wird auch für weitere Hypothesenprüfungen genutzt werden. Der für die Mes-sung zentraler Tendenzen bei mehreren unabhängigen Stichproben verwendete Kruskal-Wallis-Test (Universität Zürich, 2016) ermittelt signifikante Unter-schiede zwischen unabhängigen Stichproben bei nicht vorgeschriebenen Min-dest-Fallzahlen. Für Q14/1 kann die Alternativhypothese bestätigt werden ($H(15) = 38.42, p = .001$), ebenso für Q14/2 ($H(15) = 34.82, p = .003$), nicht je-doch für Q14/3 ($H(15) = 20.91, p = .140$).

Bei einer Reduzierung auf die sieben Bundesländer mit einer ausreichend hohen Fallzahl¹⁰⁷ wären die Ergebnisse ähnlich (Q14/1: $H(6) = 22.52, p = .001$, Q14/2: $H(6) = 23.65, p = .001$, Q14/3: $H(6) = 11.93, p = .063$). Lediglich bei Q14/3 liegt nun eine Tendenz vor.

Über die Hypothesenprüfung hinaus sollen die post-hoc Verfahren für Q14/1 und Q14/2 die Grundlage für spätere Diskussionen bieten. So wäre es in-teressant, in welchen Bundesländern welche die Durchführungsobjektivität ge-

106 Likert-Skala: *immer* (1) – *oft* (2) – *gelegentlich* (3) – *seltener* (4) – *nie* (5).

107 Geringste Fallzahl: Hamburg (N = 29).

fährdenden Veränderungen signifikant häufiger vorgenommen werden, um so eine argumentative Grundlage für modifizierte Richtlinien bei der Anwendung von Intelligenztests anbieten zu können.

Für die Berechnungen werden die ermittelten Signifikanzen mit der Bonferroni-Korrektur überprüft. Diese Methode zur Vermeidung von Typ-I Fehlern (Nullhypothese ist wahr, wird aber zurückgewiesen) erhöht die Möglichkeit, einen Typ-II Fehler zu begehen (Nullhypothese wird angenommen, obwohl sie falsch ist). Teilweise werden bei der Ergebnisdarstellung sowohl die mit der Bonferroni-Korrektur korrigierten und nicht korrigierten Signifikanzen dargestellt, für die Hypothesenprüfungen werden jedoch in der Regel die mit der Korrektur ermittelten Signifikanzen genutzt. Aus den Hypothesenprüfungen sollen in der anschließenden Interpretation und Diskussion Empfehlungen für den Umgang mit Intelligenztests abgeleitet werden. Falsch abgeleitete Empfehlungen (Typ-I Fehler) werden für diese Forschungsarbeit als gravierender angenommen als nicht abgeleitete Empfehlungen (Typ-II Fehler). Es wäre z. B. anmaßend, einem Bundesland bei der Richtliniengestaltung im Umgang mit Intelligenztests zu empfehlen, sorgfältiger die Durchführungsobjektivität zu wahren, obwohl sie evtl. gar nicht verletzt ist (Typ-I Fehler).

Für das Item Q14/1 (Zeiten geändert) werden ebenfalls mit dem Kruskal-Wallis-Test Signifikanzen zwischen den sieben Bundesländern mit einer akzeptablen Fallzahl ermittelt (siehe Tabelle 37 und Tabelle 38).

Tabelle 37. Mittlere Ränge Bundesländer für Q14/1 (Durchführungszeiten weggelassen).

BW	HH	HE	NI	NRW	RP	SH
481.86	496.87	585.75	521.45	560.63	441.91	575.50

Anmerkung. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein.

Entsprechend der Antwortmöglichkeiten (*immer* (1), *oft* (2), *gelegentlich* (3), *selten* (4), *nie* (5)) bedeutet ein niedrigerer mittlerer Rang eine häufigere Verletzung der Durchführungsobjektivität. SonderpädagogInnen aus Hessen (mittlerer Rang = 585,75) verändern am wenigsten, SonderpädagogInnen aus Rheinland-Pfalz (mittlerer Rang = 441,91) am häufigsten die Durchführungszeiten, obwohl dies nicht erlaubt ist (Q14/1).

Signifikant häufiger werden die Regeln zu den Durchführungszeiten verletzt von SonderpädagogInnen aus Rheinland-Pfalz gegenüber denen aus Hessen ($p = .019$) und NRW ($p = .040$) und von SonderpädagogInnen aus Baden-Württemberg gegenüber denen aus NRW ($p = .047$) und Hessen ($p = .034$).

Tabelle 38. Signifikanzprüfung mit dem Kruskal-Wallis-Test für Q14/1 (Durchführungszeiten geändert) im Vergleich der Bundesländer.

Vergleich Länder	Teststatistik	Standardfehler	Sig.	Korr. Sig.
RP – BW	39.949	42.878	.351	1.000
RP – HH	54.954	61.804	.374	1.000
RP – NI	79.536	42.408	.061	1.000
RP – NRW	118.715	38.219	.002	.040
RP – SH	-133.588	58.332	.022	.462
RP – HE	143.840	43.378	.001	.019
BW – HH	-15.006	54.990	.785	1.000
BW – NI	-39.588	31.664	.211	1.000
BW – NRW	-78.766	25.786	.002	.047
BW – SH	-93.639	51.057	.067	1.000
BW – HE	-103.891	32.952	.002	.034
HH – NI	-24.582	54.625	.653	1.000
HH – NRW	-63.760	51.441	.215	1.000
HH – SH	-78.633	67.735	.246	1.000
HH – HE	-88.885	55.381	.108	1.000
NI – NRW	-39.178	24.996	.117	1.000
NI – SH	-54.051	50.663	.286	1.000
NI – HE	64.303	32.338	.047	.982
NRW – SH	-14.873	47.213	.753	1.000
NRW – HE	25.125	26.609	.345	1.000
SH – HE	10.252	51.478	.842	1.000

Anmerkungen. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein. Korr. Sig.: korrigierte Signifikanz.

Für das Item Q14/2 (Zeiten weggelassen) werden ebenfalls mit dem Kruskal-Wallis-Test Signifikanzen zwischen den sieben Bundesländern mit einer akzeptablen Fallzahl ermittelt (siehe Tabelle 39 und Tabelle 40).

SonderpädagogInnen aus Hessen (mittlerer Rang = 557,60) lassen am wenigsten, SonderpädagogInnen aus Rheinland-Pfalz (mittlerer Rang = 446,36) am häufigsten die Durchführungszeiten ganz weg, obwohl dies nicht erlaubt ist (Q14/2).

Tabelle 39. Mittlere Ränge Bundesländer für Q14/2 (Durchführungszeiten geändert).

BW (1)	HH (6)	HE (7)	NI (9)	NRW (10)	RP (11)	SH (15)
492.64	506.75	557.60	538.30	554.90	446.36	497.17

Anmerkungen. BW = Baden-Württemberg, HH = Hamburg, HE = Hessen, NI = Niedersachsen, NRW = Nordrhein-Westfalen, RP = Rheinland-Pfalz, SH = Schleswig-Holstein.

Tabelle 40. Signifikanzprüfung mit dem Kruskal-Wallis-Test für Q14/2 (Durchführungszeiten weggelassen) im Vergleich der Bundesländer.

Vergleich Länder	Teststatistik	Standardfehler	Sig.	Korr. Sig.
RP – BW	46.279	33.258	.164	1.000
RP – SH	-50.803	44.980	.259	1.000
RP – HH	60.386	46.648	.195	1.000
RP – NI	91.932	32.958	.005	.111
RP – NRW	108.539	29.753	.000	.006
RP – HE	111.240	33.749	.001	.021
BW – SH	-4.524	39.096	.908	1.000
BW – HH	-14.108	41.005	.731	1.000
BW – NI	-45.653	24.322	.061	1.000
BW – NRW	-62.260	19.762	.002	.034
BW – HE	-64.961	25.384	.010	.220
SH – HH	9.583	50.976	.851	1.000
SH – NI	41.129	38.842	.290	1.000
SH – NRW	57.736	36.161	.110	1.000
SH – HE	60.437	39.515	.126	1.000
HH – NI	-31.545	40.762	.439	1.000
HH – NRW	-48.153	38.217	.208	1.000
HH – HE	-50.854	41.404	.219	1.000
NI – NRW	-16.607	19.253	.388	1.000
NI – HE	19.308	24.990	.440	1.000
NRW – HE	2.701	20.578	.896	1.000

Anmerkungen. BW = Baden-Württemberg, HH = Hamburg, HE = Hessen, NI = Niedersachsen, NRW = Nordrhein-Westfalen, RP = Rheinland-Pfalz, SH = Schleswig-Holstein, Korr. Sig.: korrigierte Signifikanz.

Signifikant häufiger werden die Durchführungszeiten ganz weggelassen von SonderpädagogInnen aus Rheinland-Pfalz gegenüber denen aus Hessen ($p = .021$) und NRW ($p = .006$) und von SonderpädagogInnen aus Baden-Württemberg gegenüber denen aus NRW ($p = .034$).

Für eine Gegenüberstellung zwischen der Versuchs- und Kontrollgruppe wurde zunächst geprüft, ob für ein Bundesland ProbandInnen im zweistelligen Bereich vorhanden sind. Die Fallzahlen der für eine Prüfung in Frage kommenden Bundesländer sind der Tabelle 41 zu entnehmen.

Tabelle 41. Bundesländer mit ausreichend hohen Fallzahlen für eine Gegenüberstellung Kontroll- und Versuchsgruppe (Hypothese 4.2).

		Ja n	Nein n	Gesamt n
Baden-Württemberg	Kontrollgruppe	11	90	101
	Versuchsgruppe	119	807	927
Niedersachsen	Kontrollgruppe	22	79	101
	Versuchsgruppe	120	806	927
Nordrhein-Westfalen	Kontrollgruppe	39	62	101
	Versuchsgruppe	426	501	927
Rheinland-Pfalz	Kontrollgruppe	18	83	101
	Versuchsgruppe	33	894	927

Anmerkung. Ja = Schule befindet sich in dem Bundesland.

Der Kruskal-Wallis-Test ermittelte für die Versuchsgruppe ähnlich den Ergebnissen für die Gesamtgruppe signifikante Unterschiede für Q14/1 ($H(3) = 9.16$, $p = .027$) und für Q14/2 ($H(3) = 18.09$, $p < .001$), nicht aber für Q14/3 ($H(3) = 5.30$, $p = .151$). Dies bedeutet, dass abhängig vom Bundesland die Durchführungszeiten zur Bearbeitung von Testaufgaben abweichend den Vorgaben der Manuale verändert worden sind (Q14/1) und abweichend von den Vorgaben der Manuale abhängig vom Bundesland signifikant unterschiedlich häufig die Durchführungszeiten weggelassen worden sind. Für das unerlaubte Geben von Feedbacks während der Testsituation konnte in der Häufigkeit kein signifikanter Unterschied zwischen den Bundesländern festgestellt werden.

Die genauere Analyse ermittelte folgende Unterschiede zwischen den Bundesländern für Q14/1 und Q14/2 (da für Q14/3 keine signifikanten Unterschiede festgestellt worden sind, entfällt die Darstellung der mittleren Ränge für dieses Item), veranschaulicht in Tabelle 42 und Tabelle 43.

Grundlage soll die nach Bonferroni korrigierte Signifikanz sein. Demnach verändern SonderpädagogInnen aus Baden-Württemberg signifikant häufiger die Durchführungszeiten ($p = .034$) gegenüber denen aus Nordrhein-Westfalen

und lassen auch häufiger ganz die Durchführungszeiten weg gegenüber den SonderpädagogInnen aus Niedersachsen ($p = .037$) und Nordrhein-Westfalen ($p = .007$). Auch SonderpädagogInnen aus Rheinland-Pfalz lassen häufiger die Durchführungszeiten ganz weg gegenüber denen aus Niedersachsen ($p = .039$) und aus Nordrhein-Westfalen ($p = .025$).

Tabelle 42. Versuchsgruppe: mittlere Ränge für Q14/1 und Q14/2 (Durchführungszeiten geändert bzw. weggelassen).

	Baden-Würtemb.	Niedersachsen	Nordrhein-Westf.	Rheinland-Pfalz
mittlerer Rang Q14/1	355.00	388.71	409.10	358.46
mittlerer Rang Q14/2	353.53	403.79	401.14	328.34

Tabelle 43. Versuchsgruppe: Signifikanzprüfung mit dem Kruskal-Wallis-Test für Q14/1 und Q14/2 (Durchführungszeiten geändert bzw. weggelassen) im Vergleich der Bundesländer.

	Q14/1		Q14/2	
	Sig.	Korr. Sig.	Sig.	Korr. Sig.
BW – RP	.924	1.000	.363	1.000
BW – NI	.169	1.000	.006	.037
BW – NRW	.006	.034	.001	.007
RP – NI	.406	1.000	.007	.039
RP – NRW	.128	.767	.004	.025
NI – NRW	.297	1.000	.857	1.000

Anmerkungen. BW = Baden-Württemberg. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. Korr. Sig.: korrigierte Signifikanz.

In der Kontrollgruppe konnten keine signifikanten Unterschiede zwischen den Bundesländern mit ausreichend hohen Fallzahlen festgestellt werden (Q14/1: $H(3) = 6.91$, $p = .075$; Q14/2: $H(3) = 4.20$, $p = .240$; Q14/3: $H(3) = 2.63$, $p = .452$). Dies sagt jedoch nichts darüber aus, ob in der Kontrollgruppe weniger oder häufiger gegenüber den SonderpädagogInnen aus der Versuchsgruppe Regeln verletzt werden.

Abschließend ermittelte der Mann-Whitney-U-Test (siehe Tabelle 44 und Tabelle 45) unter Einbezug aller Bundesländer signifikante Unterschiede zwischen Kontroll- und Versuchsgruppe für Q14/1 ($U(1051,105) = 44247.00$, $z = -3.82$, $p < .001$) und Q14/2 ($U(1042,105) = 47973.50.00$, $z = -3.04$, $p = .002$), aber keinen signifikanten Unterschied für Q14/3 ($U(1058,104) = 50209.00$, $z = -1.56$, $p = .119$).

Tabelle 44. Mittlere Ränge für Q14/1, Q14/2, Q14/3 für die Kontroll- und Versuchsgruppe*.

		<i>n</i>	mittlerer Rang
Q14/1 (Zeiten verändert)	Versuchsgruppe	1051	588.90
	Kontrollgruppe	105	474.40
	Gesamt	1156	
Q14/2 (Zeiten weggelassen)	Versuchsgruppe	1042	580.46
	Kontrollgruppe	105	509.89
	Gesamt	1147	
Q14/3 (unerlaubtes Feedback gegeben)	Versuchsgruppe	1058	586.04
	Kontrollgruppe	104	535.28
	Gesamt	1162	

Anmerkung. * Größere Fallzahlen durch Gewichtungen möglich.

Tabelle 45. Gegenüberstellung signifikante Unterschiede zwischen Kontroll- und Versuchsgruppe für Q14/1, Q14/2 und Q14/3.

	Q14/1	Q14/2	Q14/3
Mann-Whitney-U	44247.000	47973.500	50209.000
<i>z</i>	-3.824	-3.044	-1.558
asymptotische Signifikanz (2-seitig)	.000	.002	.119

Anmerkung. *z* = *z*-Wert.

Ehemalige TeilnehmerInnen an Seminaren zur Intelligenzdiagnostik verändern signifikant seltener die Durchführungszeiten als SonderpädagogInnen, die an keiner entsprechenden Fortbildung teilnahmen (Q14/1: $p < .001$) und lassen signifikant seltener die Durchführungszeiten ganz weg (Q14/2: $p = .002$). Bezüglich dem nicht vorgesehenen Geben von Feedbacks (Q14/3) gibt es keinen Unterschied zwischen Kontroll- und Versuchsgruppe ($p = .119$).

Hypothese 4.3

H0: Bei den Bundesländern bestehen keine Unterschiede bei den Beeinträchtigungen während der Anwendung von Intelligenztests durch fehlende oder unvollständige Materialien.

H1: Bei den Bundesländern bestehen Unterschiede bei den Beeinträchtigungen während der Anwendung von Intelligenztests durch fehlende oder unvollständige Materialien.

Q13 besteht aus drei Items, die für die Hypothesenprüfung vorgesehen sind: *Wenn Sie testen möchten kommt es vor (...)*¹⁰⁸

- Q13/1: (...) *dass einige Ihrer Intelligenztests nicht zur Verfügung stehen (z. B. ausgeliehen sind etc.)?*
- Q13/2: (...) *dass die Testmaterialien unvollständig sind (z. B. fehlende Puzzle-teile)?*
- Q13/3: (...) *dass Formulare/Arbeitsbögen fehlen?*

Bei der Berechnung für alle 16 Bundesländer ergeben sich nach der Anwendung des Kruskal-Wallis-Test hohe Signifikanzen für Q13/1 ($H(15) = 74.57, p < .001$), Q13/2 ($H(15) = 74.19, p < .001$) und für Q13/3 ($H(15) = 51.26, p < .001$). Bei der Berechnung für die sieben Bundesländer mit einer akzeptablen Fallzahl sind die Ergebnisse für p ebenfalls signifikant.

Die Ergebnisse im Detail werden der Übersicht halber tabellarisch dargestellt¹⁰⁹.

Tabelle 46. Mittlere Ränge Bundesländervergleich für Q13/1, Q13/2, Q13/3.

	BW	HH	HE	NI	NRW	RP	SH
Q13/1 (Test weg)	635.34	789.38	534.78	561.99	500.27	528.46	476.93
Q13/2 (Material unvollständig)	648.78	670.84	591.60	444.10	534.09	442.97	434.96
Q13/3 (Formulare fehlen)	642.21	714.59	490.46	509.44	523.03	598.14	486.72

Anmerkungen. Farblich markiert sind die beiden höchsten (hellgrau) und niedrigsten (grau) Ergebnisse. Niedriger mittlerer Rang = höhere Beeinträchtigung. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein.

Die mittleren Ränge (siehe Tabelle 46) zeigen, dass die erfragten Beeinträchtigungen bei jedem der drei Items in Schleswig-Holstein (Q13/1: mittlerer Rang = 476.93; Q13/2: mittlerer Rang = 434.96; Q13/3: mittlerer Rang = 486.72) am höchsten und ebenso bei jedem der drei Items in Hamburg am niedrigsten sind (Q13/1: mittlerer Rang = 789.38; Q13/2: mittlerer Rang = 670.84; Q13/3: mittlerer Rang = 714.59). Erwähnt werden sollte das Ergebnis für Baden-Württemberg, welches für alle drei Items ebenfalls wenige Beeinträchtigungen ermittelte

108 Likert-Skala: immer (1) – oft (2) – gelegentlich (3) – selten (4) – nie (5).

109 Die Begründung für über die Hypothesenprüfung hinausgehende Berechnungen werden weiter oben dargestellt und werden im Folgenden nicht erneut vorgenommen.

(Q13/1: mittlerer Rang = 635.34; Q13/2: mittlerer Rang = 648.78; Q13/3: mittlerer Rang = 642.21).

Ebenfalls aus Gründen der Übersicht werden für die drei Items jeweils nur die Signifikanzen und die korrigierten Signifikanzen (angepasst mit der Bonferroni-Korrektur) in Tabelle 47 vorgestellt. Grundlage für die Prüfungen bleiben die korrigierten Signifikanzangaben.

Tabelle 47. Signifikanzprüfung mit dem Kruskal-Wallis-Test für Q13/1 (Tests fehlen), Q13/2 (Tests unvollständig), Q13/3 (Formulare fehlen) mit Angaben zu den Bundesländern, für die signifikant höhere Beeinträchtigungen nachgewiesen worden sind.

	Q13/1			Q13/2			Q13/3		
	Sig.	korr. Sig.	gr. Beein.	Sig.	korr. Sig.	gr. Beein.	Sig.	korr. Sig.	gr. Beein.
SH – NRW	.649	1.000		.046	.973		.475	1.000	
SH – RP	.425	1.000		.896	1.000		.081	1.000	
SH – HE	.301	1.000		.004	.083		.946	1.000	
SH – NI	.123	1.000		.864	1.000		.802	1.000	
SH – BW	.004	.090		.000	.001	SH	.005	.098	
SH – HH	.000	.000	SH	.001	.016	SH	.002	.036	SH
NRW – RP	.513	1.000		.024	.497		.076	1.000	
NRW – HE	.231	1.000		.042	.872		.254	1.000	
NRW – NI	.024	.505		.001	.014	NI	.403	1.000	
NRW – BW	.000	.000	NRW	.000	.000	NRW	.000	.000	NRW
NRW – HH	.000	.000	NRW	.009	.195		.000	.009	NRW
RP – HE	.877	1.000		.001	.025	RP	.024	.494	
RP – NI	.455	1.000		.980	1.000		.036	.761	
RP – BW	.021	.432		.000	.000	RP	.318	1.000	
RP – HH	.000	.002	RP	.000	.007	RP	.076	1.000	
HE – NI	.440	1.000		.000	.000	NI	.774	1.000	
HE – BW	.005	.103		.100	1.000		.000	.000	HE
HE – HH	.000	.000	HE	.164	1.000		.000	.003	HE
NI – BW	.034	.711		.000	.000	NI	.000	.001	NI
NI – HH	.000	.002	NI	.000	.001	NI	.000	.005	NI
BW – HH	.009	.179		.696	1.000		.229	1.000	

Anmerkungen. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein. Korr. Sig.: korrigierte Signifikanz. Gr. Beein.: größere Beeinträchtigung in dem jeweiligen Bundesland.

Entsprechend der mittleren Ränge für die Items Q13/1, Q13/2 und Q13/3 konnten keine signifikanten Beeinträchtigungen für Hamburg und Baden-Württemberg ermittelt werden, so dass für diese Bundesländer weniger Beeinträchtigungen angenommen werden können. Im Ländervergleich konnten bei gemeinsamer Betrachtung der drei Items sieben signifikante Ergebnisse bezüglich beschriebener höherer Beeinträchtigungen für Niedersachsen, fünf für Nordrhein-Westfalen, je vier für Rheinland-Pfalz und Schleswig-Holstein und drei für Hessen berechnet werden. Für diese Länder sind Beeinträchtigungen anzunehmen, für Niedersachsen bei jeder der drei Items.

Die Auswertung, getrennt nach Kontroll- und Versuchsgruppe, wurde für vier Bundesländer mit einer akzeptablen Fallzahl vorgenommen und ebenfalls mit dem Kruskal-Wallis-Test geprüft. Für die Versuchsgruppe liegen Signifikanzen für alle drei Items vor (Q13/1: $H(3) = 21.61, p < .001$; Q13/2: $H(3) = 33.59, p < .001$; Q13/3: $H(3) = 20.65, p < .001$).

Für die Kontrollgruppe liegt eine Signifikanz von $H(3) = 11.06, p = .011$ für Q13/2 vor, jedoch nicht für Q13/1 ($H(3) = 4.43, p = .219$) und Q13/3 ($H(3) = 4.30, p = .231$). Unterschiedlich häufig im Vergleich zwischen den Bundesländern stehen in der Versuchsgruppe Tests nicht zur Verfügung (Q13/1), sind unvollständig (Q13/2) oder es fehlen Formulare (Q13/3). In der Kontrollgruppe gibt es Unterschiede zwischen den Bundesländern bei der Frage nach fehlenden Formularen (Q13/2). In der Tabelle 48 werden die Unterschiede mit Hilfe der mittleren Ränge ersichtlich.

Tabelle 48. Mittlere Ränge der Versuchsgruppe für Q13/1 (Tests fehlen), Q13/2 (Tests unvollständig), Q13/3 (Formulare fehlen) und der Kontrollgruppe für Q13/2.

		Baden- Württemberg	Niedersachsen	Nordrhein- Westfalen	Rheinland-Pfalz
Versuchsgruppe	m. Rang Q13/1	462.77	415.78	368.32	384.85
	m. Rang Q13/2	473.42	325.43	392.03	367.38
	m. Rang Q13/3	461.00	357.65	379.33	436.82
Kontrollgruppe*	m. Rang Q13/2	62.27	44.52	50.71	32.39

Anmerkung. m. Rang = mittlerer Rang; * Angaben über mittlere Ränge für Q13/1 und Q13/3 entfallen, da keine Signifikanzen vorhanden sind.

In der Versuchsgruppe stehen nordrhein-westfälischen SonderpädagogInnen häufiger Tests nicht zur Verfügung (Q13/1: $p < .001$), sind die Testmaterialien häufiger unvollständig (Q13/2: $p < .001$) und es fehlen häufiger Formulare (Q13/3: $p = .001$) im Vergleich mit baden-württembergischen SonderpädagogInnen. Diese sind auch seltener von unvollständigen Testmaterialien betroffen (Niedersachsen: Q13/2: $p < .001$; Rheinland-Pfalz: $p = .034$) und seltener im

Vergleich mit niedersächsischen SonderpädagogInnen von fehlenden Formularen (Q13/3: $p = .001$).

In der Kontrollgruppe konnte lediglich eine Signifikanz ermittelt werden: baden-württembergische SonderpädagogInnen bemängeln seltener unvollständige Testunterlagen (Q13/2: $p = .012$) im Vergleich zu denen aus Rheinland-Pfalz (siehe Tabelle 49).

Tabelle 49. Signifikanzprüfung für Q13/1(Tests fehlen), Q13/2 (Tests unvollständig), Q13/3 (Formulare fehlen) der Vergleichsgruppe und Q13/2 der Kontrollgruppe mit Angaben zu den Bundesländern, für die signifikant höhere Beeinträchtigungen nachgewiesen worden sind.

			BW – RP	BW – NI	BW – NRW	RP – NI	RP – NRW	NI – NRW
Versuchs- gruppe	Q13/1	Sig.	.078	.048	.000	.434	.646	.026
		korr. Sig.	.469	.288	.000	1.00	1.00	.158
		Beeintr.			NRW			
	Q13/2	Sig.	.006	.000	.000.	.275	.483	.001
		korr. Sig.	.034	.000	.000	1.00	1.00	.008
		Beeintr.	RP	NI	NRW			NRW
	Q13/3	Sig.	.535	.000	.000	.043	.107	.303
		korr. Sig.	1.00	.001	.001	.256	.643	1.00
		Beeintr.		NI	NRW			
Kontroll- gruppe	Q13/2	Sig.	.002	.056	.179	.129	.011	.349
		korr. Sig.	.012	.338	1.00	.771	.064	1.00
		Beeintr.	RP					

Anmerkungen. BW = Baden-Württemberg. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. Korr. Sig.: korrigierte Signifikanz. Beeintr.: größere Beeinträchtigung in dem jeweiligen Bundesland.

Hypothese 4.4

H0: Es besteht kein Zusammenhang zwischen dem Bundesland und der Freiheit der SonderpädagogInnen, über die Anwendung eines Intelligenztests zu entscheiden.

H1: Es besteht ein Zusammenhang zwischen dem Bundesland und der Freiheit der SonderpädagogInnen, über die Anwendung eines Intelligenztests zu entscheiden.

Der für die Prüfung zwischen Bundesland und Q10 (*Entscheiden Sie nach eigenem Ermessen, Intelligenztests durchzuführen: Ja/Nein*) verwendete Chi-Quadrat-Test ergab eine hohe Signifikanz von $\chi^2(15, N = 847) = 88.96, p < .001$.

Die Reduzierung auf die sieben Bundesländer mit einer akzeptablen Fallzahl ergibt nach der Anwendung des Chi-Quadrat-Tests ebenfalls hohe Signifikanzen ($\chi^2(6, N = 800) = 88.96, p < .001$; Häufigkeiten siehe Tabelle 50).

Tabelle 50. Bundesländer-Vergleich Q10 (Entscheiden Sie nach eigenem Ermessen, einen Intelligenztest durchzuführen).

		Ja	Nein	Gesamt
BW	Anzahl	76	31	107
	% innerhalb von Bundesland	71 %	29 %	100 %
HH	Anzahl	17	8	25
	% innerhalb von Bundesland	68 %	32 %	100 %
HE	Anzahl	85	9	94
	% innerhalb von Bundesland	90 %	10 %	100 %
NI	Anzahl	118	11	129
	% innerhalb von Bundesland	92 %	8 %	100 %
NRW	Anzahl	270	105	375
	% innerhalb von Bundesland	72 %	28 %	100 %
RP	Anzahl	21	22	43
	% innerhalb von Bundesland	49 %	51 %	100 %
SH	Anzahl	15	12	27
	% innerhalb von Bundesland	56 %	44 %	100 %
Gesamt	Anzahl	602	198	800
	% innerhalb von Bundesland	75 %	25 %	100 %

Anmerkungen. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein.

Für die Analyse der signifikanten Unterschiede wurden die sieben Bundesländer mit einer akzeptablen Fallzahl jeweils in Beziehung gesetzt zu den übrigen Bundesländern.

Danach haben SonderpädagogInnen aus Hessen ($p < .001 \chi^2(1, N = 847) = 15.47, p < .001$) und Niedersachsen ($p < .001 \chi^2(1, N = 848) = 25.05, p < .001$) signifikant häufiger die Möglichkeit, selbst über die Anwendung eines Intelligenztests zu entscheiden.

Dies trifft nicht zu für rheinland-pfälzische ($p < .001 \chi^2(1, N = 848) = 14.27, p < .001$) und für schleswig-holsteinische SonderpädagogInnen ($p = .032 \chi^2(1, N = 849) = 4.61, p = .032$).

Bei der Gegenüberstellung von Kontroll- und Versuchsgruppe für vier Bundesländer mit einer zweistelligen Fallzahl (Niedersachsen, Rheinland-Pfalz, Ba-

den-Württemberg, Nordrhein-Westfalen) liegen jeweils signifikante Unterschiede vor. Verglichen wurde erneut mit dem Chi-Quadrat-Test nach Pearson. Für die Versuchsgruppe konnte ein signifikanter Unterschied von $\chi^2(3, N = 581) = 31.63, p < .001$, für die Kontrollgruppe ein signifikanter Unterschied von $\chi^2(3, N = 72) = 8.14, p = .043$ festgestellt werden.

Der Vergleich der vier Bundesländer mit einer zweistelligen Fallzahl ermittelt ein signifikantes Ergebnis für Niedersachsen: niedersächsische SonderpädagogInnen haben sowohl in der Kontrollgruppe ($\chi^2(1, N = 72) = 5.86, p = .016$) als auch in der Versuchsgruppe ($\chi^2(1, N = 580) = 19.49, p < .001$) eine größere Freiheit, darüber selbst zu entscheiden, ob sie Intelligenztests durchführen.

Hypothese 4.5

H0: Zwischen den Bundesländern gibt es keine Unterschiede im Umgang mit Durchführungsregeln bei der Anwendung von Intelligenztests.

H1: Zwischen den Bundesländern gibt es Unterschiede im Umgang mit Durchführungsregeln bei der Anwendung von Intelligenztests.

Zur Prüfung dieser Hypothese wurde der Kruskal-Wallis-Test für folgende Items verwendet: *Folgende Durchführungsregeln bereiten mir Schwierigkeiten:*¹¹⁰

- Q15/1: *Umkehrregeln*
- Q15/2: *Abbruchregeln*
- Q15/3: *Ausrechnen des Testalters*

Für kein Item kann die Nullhypothese verworfen werden nach Auswertung aller Bundesländer (Q15/1: $H(15) = 21.48, p = .122$; Q15/2: $H(15) = 19.67, p = .185$; Q15/3: $H(15) = 21.58, p = .119$). Nach der Auswertung für die sieben Bundesländer mit höheren Fallzahlen ergibt sich eine Signifikanz für Q15/3 ($H(6) = 16.60, p = .010$; Q15/1: $H(6) = 11.26, p = .081$; Q15/2: $H(6) = 9.49, p = .148$).

Hessische SonderpädagogInnen gaben signifikant mehr Probleme beim Ausrechnen des Alters am Testtag ($p = .012$) im Vergleich zu denen aus Rheinland-Pfalz an.

Bei der Gegenüberstellung von Kontroll- und Versuchsgruppe für die Bundesländer mit einer ausreichend hohen Fallzahl (Niedersachsen, Rheinland-Pfalz, Baden-Württemberg, Nordrhein-Westfalen) konnte für keines der Items eine Signifikanz in der Versuchsgruppe festgestellt werden (Q15/1: $H(3) = 4.50, p = .212$; Q15/2: $H(3) = 2.12, p = .548$; Q15/3: $H(3) = 2.85, p = .415$).

110 Likert-Skala: *außerordentlich (1), ziemlich (2), mittelmäßig (3), kaum (4), gar nicht (5)*.

In der Kontrollgruppe konnten Signifikanzen bei Q15/1 ($H(3) = 8.87, p = .031$) und Q15/2 ($H(3) = 11.29, p = .010$), bei Q15/3 lediglich eine Tendenz ($H(3) = 6.30, p = .098$) berechnet werden.

Rheinland-pfälzische SonderpädagogInnen haben signifikant weniger Schwierigkeiten mit der Umkehrregel (Q15/1; mittlerer Rang: 44.96; $p = .037$) als nordrhein-westfälische (mittlerer Rang = 28.83) und weniger Schwierigkeiten mit der Abbruchregel (Q15/2; mittlerer Rang: 59.28; $p = .005$) als niedersächsische (mittlerer Rang: 32.38).

Bei der Frage, ob es generell einen Unterschied zwischen Kontroll- und Versuchsgruppe gibt, ermittelte der Mann-Whitney-U-Test keinen Unterschied zwischen beiden Gruppen (Q15/1: $U(883, 73) = 29934.00, z = -1.06, p = .291$; Q15/2: $U(1025, 101) = 49125.50, z = -0.901, p = .367$; Q15/3: $U(1031, 99) = 48702.00, z = -0.875, p = .381$).

Hypothese 4.6

H0: Zwischen den Bundesländern gibt es keine Unterschiede bei den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

H1: Zwischen den Bundesländern gibt es Unterschiede bei den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

Geprüft wurde mit einer einfaktoriellen Varianzanalyse (ANOVA), da stetige Variablen vorliegen. Im Gegensatz zum Kruskal-Wallis-Test ist eine ANOVA weniger konservativ und besser geeignet, tatsächliche Unterschiede zu entdecken.

In Tabelle 51 kann festgestellt werden, dass die Mittelwerte für den Schwierigkeiten-Index¹¹¹ nicht sehr weit auseinanderliegen von $MW = 3.38$ (Niedersachsen, $SD = 0.60$) bis $MW = 3.60$ (Baden-Württemberg, $SD = 0.52$).

Der Levene-Test prüft, ob die Varianzen homogen sind, was die Voraussetzung für die Anwendung der ANOVA wäre. Die Varianzhomogenität ist gegeben, da keine Signifikanz vorliegt ($p = .596$). Obwohl die Mittelwerte dicht zusammen liegen, ist dennoch eine Signifikanz festzustellen ($F(6,950) = 2.729, p = .012$), so dass der Grad der empfundenen Schwierigkeit bei der Anwendung von Intelligenztests tatsächlich vom Bundesland abhängt (siehe Tabelle 52).

Da eine Varianzhomogenität vorliegt, konnte für die paarweisen Vergleiche der Tukey-Test verwendet werden. Dieser Test ist ein Posthoc-Verfahren und ähnelt dem t-Test, hält aber das Fehlerniveau bei ca. 5% (Hain, 2018). Zur Vermeidung von Typ-I Fehlern sind die Signifikanzen korrigiert mit der Bonferoni-Korrektur.

111 Hinweis: je geringer der Mittelwert, desto mehr Schwierigkeiten werden beschrieben.

Tabelle 51. Vergleich von sieben Bundesländern mit akzeptablen Fallzahlen für empfundene Schwierigkeiten bei der Anwendung von Intelligenztests.

	N	MW	SD	Standardfehler	95 % Konfidenzintervall für MW		Min	Max
					Untergrenze	Obergrenze		
BW	130	3.6028	0.520	0.046	3.5127	3.6928	2.00	4.90
HH	29	3.4713	0.705	0.131	3.2027	3.7398	1.00	4.45
HE	109	3.5859	0.496	0.047	3.4918	3.6799	2.36	4.64
NI	143	3.3821	0.598	0.050	3.2831	3.4810	1.00	4.56
NRW	465	3.4698	0.543	0.025	3.4203	3.5193	1.00	4.73
RP	51	3.5340	0.553	0.077	3.3786	3.6894	2.18	4.73
SH	31	3.3972	0.440	0.079	3.2360	3.5584	2.64	4.45
Gesamt	958	3.4892	0.550	0.018	3.4544	3.5240	1.00	4.90

Anmerkungen. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein. MW = Mittelwert. SD = Standardabweichung. Min = Minimum. Max = Maximum.

Tabelle 52. Übersicht einfaktorielle Varianzanalyse für den Vergleich von sieben Bundesländern mit akzeptablen Fallzahlen und empfundenen Schwierigkeiten.

	Quadratsumme	df	Mittel der Quadrate	F	Signifikanz
Zwischen den Gruppen	4.890	6	.815	2.729	.012
Innerhalb der Gruppen	283.678	950	.299		
Gesamt	288.568	956			

Anmerkungen. df = Freiheitsgrad. F = Teststatistik.

Tabelle B3 verdeutlicht die Ergebnisse der Signifikanzprüfung. Baden-Württemberg beschreibt weniger Schwierigkeiten im Vergleich mit Niedersachsen ($p = .016$). Im Vergleich zwischen Niedersachsen und Hessen beschreiben hessische SonderpädagogInnen tendenziell weniger Schwierigkeiten ($p = .053$).

Beim Gegenüberstellen zwischen Kontroll- und Versuchsgruppe mit der einfaktoriellen Varianzanalyse ANOVA konnte unter Einbezug der vier Bundesländer mit mindestens zweistelliger Fallzahl für die Versuchsgruppe eine Signifikanz von $F(3, 694) = 3,761, p = .011$, für die Kontrollgruppe keine Signifikanz ($F(3, 86) = 1.221, p = .307$) ermittelt werden. Für beide Gruppen ermittelte der Levene-Test eine Varianzhomogenität (Kontrollgruppe: $p = .534$; Versuchsgruppe: $p = .971$), so dass als Post-Hoc Verfahren der Tukey-Test genauere Angaben über Unterschiede zwischen den vier Bundesländern berechnet.

In der Versuchsgruppe gaben niedersächsische SonderpädagogInnen signifikant mehr Schwierigkeiten im Vergleich mit denen aus Baden-Württemberg ($p = .007$) und SonderpädagogInnen aus Nordrhein-Westfalen gegenüber denen aus Baden-Württemberg ($p = .048$; siehe Tabelle B4) an.

Der t-Test für unabhängige Stichproben ermittelte abschließend, ob die Kontroll- oder Versuchsgruppe mehr Schwierigkeiten empfindet bei der Anwendung von Intelligenztests unter Berücksichtigung der Gesamtfallzahl. Da der Levene-Test keine Varianzgleichheit festgestellt hat, wurde das Ergebnis des Welch-Tests ermittelt. ProbandInnen aus der Kontrollgruppe ($MW = 3,3701$; $SD = .754$) empfinden tendenziell mehr Schwierigkeiten bei der Anwendung von Intelligenztests gegenüber denen aus der Versuchsgruppe ($MW = 3.5027$; $SD = .540$; $t(111.472) = 1.719$, $p = .088$).

Hypothese 4.7

H0: Es gibt keinen Unterschied zwischen den Bundesländern bei der zur Verfügung stehenden Zeit für die Anwendung eines Intelligenztests.

H1: Es gibt Unterschiede zwischen den Bundesländern bei der zur Verfügung stehenden Zeit für die Anwendung eines Intelligenztests.

Folgende Items waren für die Hypothesenprüfung vorgesehen (Q28): *Bitte bewerten Sie folgende Aussagen:*

- Q28/1: *Im Rahmen meiner Arbeit steht mir zu wenig Vorbereitungszeit für das Lernen eines normierten Tests zur Verfügung.*
- Q28/2: *Heutzutage habe ich weniger Zeit für die Anwendung eines Intelligenztests.*
- Q28/3: *Heutzutage habe ich weniger Zeit für die Durchführung eines sonderpädagogischen Gutachtens.*
- Q28/4: *Steht mir nicht genügend Vorbereitungszeit zur Verfügung für einen Test, bereite ich mich in der Freizeit vor.*

Die vier Items wurden geprüft mit dem Kruskal-Wallis-Test. Für jedes Item sind signifikante Unterschiede ermittelt worden bei Einbezug der Gesamtstichprobe (Q28/1: $H(15) = 27.41$, $p = .026$, Q28/2: $H(15) = 30.08$, $p = .012$, Q28/3: $H(15) = 53.76$, $p < .001$, Q28/4: $H(15) = 35.02$, $p < .001$).

Unter Einbezug der sieben Bundesländer mit einer akzeptablen Fallzahl sind die Signifikanzen noch eindeutiger (Q28/1: $H(6) = 22.84$, $p = .001$; Q28/2: $H(6) = 18.03$, $p = .006$; Q28/3: $H(6) = 40.09$, $p < .001$; Q28/4: $H(6) = 22.08$, $p = .001$).

Grundlage für die folgende Analyse sind die nach der Bonferroni-Korrektur angepassten Signifikanzangaben, dennoch werden die nicht korrigierten Signifikanzen ebenfalls dargestellt, um Hinweise auf Unterschiede anzudeuten. Der

Vergleich der mittleren Ränge ist in Tabelle 53 dargestellt, die Ergebnisse der Signifikanzprüfungen in Tabelle 54 und Tabelle 55.

Die mittleren Ränge zeigen höhere, bzw. niedrigere Problematiken an. Die fünfstufige Ratingskala lautete *völlig richtig* (1), *ziemlich richtig* (2), *unentschieden* (3), *ziemlich falsch* (4) und *falsch* (5).

Tabelle 53. Mittlere Ränge für Q28/1, Q28/2, Q28/3, Q28/4 im Bundesländervergleich.

	BW	HH	HE	NI	NRW	RP	SH
Q28/1 (wenig Vorbereitungszeit)	598.94	549.76	609.41	608.95	515.76	605.98	584.94
Q28/2 (wenig Anwendungszeit)	595.99	491.36	617.99	486.41	554.15	591.34	552.44
Q28/3 (wenig Zeit für Gutachten)	552.82	398.53	652.26	469.71	559.67	654.30	566.49
Q28/4 (Vorbereitung in Freizeit)	584.89	610.53	633.97	566.37	531.02	537.47	513.39

Anmerkungen. Farblich markiert sind die beiden höchsten (hellgrau) und niedrigsten (grau) Ergebnisse. Niedriger mittlerer Rang = höhere Beeinträchtigung. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein.

SonderpädagogInnen aus Nordrhein-Westfalen (mittlerer Rang = 515,76) und Hamburg (mittlerer Rang = 549,76) hatten den höchsten Zustimmungsgrad auf die Frage, ob zu wenig Vorbereitungszeit für das Lernen eines Intelligenztests zur Verfügung steht (Q28/1); hessische (mittlerer Rang = 609,41) und niedersächsische (mittlerer Rang = 608,95) beantworteten diese Frage am wenigsten zustimmend.

Interessant könnte eine genauere Analyse in der anschließenden Diskussion aus dem Ergebnis für hamburgische SonderpädagogInnen resultieren. Denn neben der ebenfalls hohen Zustimmung auf die Frage, ob zu wenig Zeit für die Anwendung eines Tests zur Verfügung steht (Q28/2: mittlerer Rang = 491,36), und auf die Frage, ob es generell zu wenig Zeit für die Erstellung von Gutachten gibt (Q28/3: mittlerer Rang = 398,53), stimmt diese Personengruppe auf die Frage, ob in der Freizeit die Tests vorbereitet werden würden, neben hessischen SonderpädagogInnen (mittlerer Rang = 633,97) am wenigsten zu.

Unterschiedlich antworten auch die SonderpädagogInnen aus Niedersachsen. Neben der hohen Zustimmung über zu wenig Zeit für die Anwendung eines Tests (mittlerer Rang = 486,41) und zu wenig Zeit für die Gutachtenerstellung (mittlerer Rang = 469,41) gab es wenig Zustimmung auf die Frage nach zu wenig Vorbereitungszeit zum Lernen eines Tests.

Hessische Lehrkräfte stimmten bei allen vier Fragen am wenigsten (Q28/1: mittlerer Rang = 609,41, Q28/2: mittlerer Rang = 617,99; Q28/4: mittlerer Rang = 633,97) bzw. am zweitwenigsten zu (Q28/3: mittlerer Rang = 652,26).

Tabelle 54. Signifikanzprüfung mit dem Kruskal-Wallis-Test für Q28/1 (weniger Vorbereitungszeit) und Q28/2 (wenig Zeit für Anwendung).

	Q28/1		Q28/2			
	Sig.	Korr. Sig.	gr. Beeintr.	Sig.	Korr. Sig.	gr. Beeintr.
NRW – HH	.538	1.000		.263	1.000	
NRW – SH	.191	1.000		.975	1.000	
NRW – BW	.003	.068		.147	1.000	
NRW – RP	.031	.654		.383	1.000	
NRW – NI	.001	.014	NRW	.015	.315	
NRW – HE	.002	.035	NRW	.035	.738	
HH – SH	.635	1.000		.418	1.000	
HH – BW	.405	1.000		.082	1.000	
HH – RP	.399	1.000		.140	1.000	
HH – NI	.313	1.000		.934	1.000	
HH – HE	.319	1.000		.037	.785	
SH – BW	.806	1.000		.453	1.000	
SH – RP	.746	1.000		.555	1.000	
SH – NI	.671	1.000		.251	1.000	
SH – HE	.672	1.000		.265	1.000	
BW – RP	.881	1.000		.922	1.000	
BW – NI	.772	1.000		.002	.039	NI
BW – HE	.774	1.000		.554	1.000	
RP – NI	.949	1.000		.026	.547	
RP – HE	.943	1.000		.584	1.000	
NI – HE	.990	1.000		.000	.006	NI

Anmerkungen. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein. Korr. Sig.: korrigierte Signifikanz. Gr. Beeintr.: größere Beeinträchtigung in dem jeweiligen Bundesland.

Auf die Frage, ob im Rahmen der Arbeit zu wenig Vorbereitungszeit für das Lernen eines normierten Tests zur Verfügung steht (Q28/1), bejahen dies signifikant mehr SonderpädagogInnen aus Nordrhein-Westfalen gegenüber denen aus Niedersachsen ($p = .014$) und aus Hessen ($p = .035$). Eine Tendenz gibt es gegenüber denen aus Baden-Württemberg ($p = .068$).

Auf die Frage, ob heutzutage weniger Zeit für die Anwendung eines Tests zur Verfügung stehe, bejahten dies signifikant häufiger niedersächsische SonderpädagogInnen gegenüber denen aus Baden-Württemberg ($p = .039$) und Hessen ($p = .006$).

Tabelle 55. Signifikanzprüfung mit dem Kruskal-Wallis-Test für Q28/3 (wenig Zeit für Gutachten) und Q28/4 (Vorbereitung in Freizeit).

	Q28/3		Q28/4			
	Sig.	Korr. Sig.	gr. Beeintr.	Sig.	Korr. Sig.	gr. Beeintr.
HH – NI	.229	1.000		.369	1.000	
HH – BW	.010	.201		.605	1.000	
HH – NRW	.004	.079		.085	1.000	
HH – SH	.025	.516		.118	1.000	
HH – HE	.000	.001	HH	.640	1.000	
HH – RP	.000	.003	HH	.192	1.000	
NI – BW	.017	.363		.523	1.000	
NI – NRW	.001	.024	NI	.123	1.000	
NI – SH	.090	1.000		.264	1.000	
NI – HE	.000	.000	NI	.024	.513	
NI – RP	.000	.002	NI	.460	1.000	
BW – NRW	.810	1.000		.023	.481	
BW – SH	.812	1.000		.135	1.000	
BW – HE	.007	.145		.109	1.000	
BW – RP	.033	.690		.230	1.000	
NRW – SH	.898	1.000		.691	1.000	
NRW – HE	.002	.043	NRW	.000	.001	NRW
NRW – RP	.026	.548		.855	1.000	
SH – HE	.141	1.000		.013	.269	
SH – RP	.181	1.000		.659	1.000	
HE – RP	.966	1.000		.017	.351	

Anmerkungen. BW = Baden-Württemberg. HH = Hamburg. HE = Hessen. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. SH = Schleswig-Holstein. Korr. Sig.: korrigierte Signifikanz. Gr. Beeintr.: größere Beeinträchtigung in dem jeweiligen Bundesland.

Auf die Frage, ob heutzutage weniger Zeit für die Durchführung eines sonderpädagogischen Gutachtens zur Verfügung steht (Q28/3), bejahten dies signifikant häufiger SonderpädagogInnen aus Hamburg gegenüber denen aus Hessen ($p = .001$) und aus Rheinland-Pfalz, SonderpädagogInnen aus Niedersachsen

gegenüber denen aus Nordrhein-Westfalen ($p = .024$), aus Hessen ($p < .001$) und aus Baden-Württemberg ($p = .002$) sowie SonderpädagogInnen aus Nordrhein-Westfalen gegenüber denen aus Hessen ($p = .017$).

Auf die Frage, ob die Vorbereitungszeit für die Anwendung eines Tests in der Freizeit vorgenommen wird, wenn ansonsten nicht genügend Zeit zur Verfügung steht (Q28/4), bejahten diese Frage signifikant häufiger SonderpädagogInnen aus Nordrhein-Westfalen gegenüber denen aus Hessen ($p = .001$).

Auch für diese Hypothesenprüfung ist eine Gegenüberstellung zwischen Kontroll- und Versuchsgruppe sinnvoll, da die Kontrollgruppe als repräsentativer für die Gesamtheit der SonderpädagogInnen angenommen wird. Die Prüfung mit dem Kruskal-Wallis-Test für die vier in Frage kommenden Bundesländer mit einer zweistelligen Fallzahl (Niedersachsen, Rheinland-Pfalz, Baden-Württemberg, Nordrhein-Westfalen) berechnete folgende Ergebnisse:

In der Versuchsgruppe gab es bei Q28/1 ($H(3) = 13,37, p = .004$), Q28/2 ($H(3) = 12,34, p = .006$) und Q28/3 ($H(3) = 12,14, p = .007$) signifikante Ergebnisse, bei Q28/4 jedoch nur eine Tendenz ($H(3) = 7,15, p = .067$). In der Kontrollgruppe gab es eine Signifikanz bei Q28/3 ($H(3) = 10,74, p = .013$), nicht jedoch bei Q28/1 ($H(3) = 6,14, p = .105$), Q28/2 ($H(3) = 1,34, p = .720$) und Q28/4 ($H(3) = 1,531, p = .675$).

Dies bedeutet, dass es in der Kontroll- und Versuchsgruppe signifikante Unterschiede zwischen den vier Bundesländern bei der Frage nach der zur Verfügung stehenden Zeit für die Durchführung eines Gutachtens gibt (Q28/3: *Heutzutage habe ich weniger Zeit für die Durchführung eines sonderpädagogischen Gutachtens*) und ausschließlich in der Versuchsgruppe signifikante Unterschiede bei der Frage nach zu wenig Vorbereitungszeit beim Lernen eines Tests (Q28/1) und zu wenig Zeit bei der Anwendung eines Tests (Q28/2). Tabelle 56 zeigt den Vergleich der mittleren Ränge, Tabelle 57 die Ergebnisse der Signifikanzprüfung. Keine signifikanten Unterschiede gibt es in beiden Gruppen bei der Frage nach der Vorbereitung für das Lernen der Tests in der Freizeit (Q28/4: *Steht mir nicht genügend Vorbereitungszeit zur Verfügung für einen Test, bereite ich mich in der Freizeit vor.*).

Tabelle 56. Q28/1, Q28/2, Q28/3, Q28/4: mittlere Ränge Bundesländervergleich.

		BW	NI	NRW	RP
Versuchsgruppe	Q28/1 (wenig Vorbereitungszeit)	446.72	444.68	386.08	437.33
	Q28/2 (wenig Anwendungszeit)	444.71	351.56	410.71	422.11
	Q28/3 (wenig Zeit für Gutachten)	417.67	346.65	416.09	444.83
Kontrollgruppe	Q28/3 (Vorbereitung in Freizeit)	41.35	43.29	50.06	68.58

Anmerkungen. Niedriger mittlerer Rang = höhere Beeinträchtigung. BW = Baden-Württemberg. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz.

Tabelle 57. Signifikanzprüfung mit dem Kruskal-Wallis-Test für die Versuchsgruppe (Q28/1, Q28/2, Q28/3) und für die Kontrollgruppe (Q28/3).

			BW-RP	BW-NI	BW-NRW	RP-NI	RP-NRW	NI-NRW
Versuchsgruppe	Q28/1	Sig.	.812	.940	.005	.852	.154	.006
		korr. Sig.	1.00	1.00	.028	1.00	.925	.038
		Beeintr.			NRW			NRW
	Q28/2	Sig.	.573	.001	.120	.079	.756	.007
		korr. Sig.	1.00	.004	.722	.473	.100	.041
		Beeintr.		NI				NI
	Q28/3	Sig.	.498	.009	.942	.014	.433	.001
		korr. Sig.	1.00	.053	1.00	.086	1.00	.008
		Beeintr.		(NI)		(NI)		NI
Kontrollgruppe	Q28/3	Sig.	.007	.837	.323	.003	.018	.328
		korr. Sig.	.043	1.00	1.00	.018	.107	1.00
		Beeintr.	BW			NI		

Anmerkungen. BW = Baden-Württemberg. NI = Niedersachsen. NRW = Nordrhein-Westfalen. RP = Rheinland-Pfalz. Korr. Sig.: korrigierte Signifikanz. Gr. Beeintr.: größere Beeinträchtigung in dem jeweiligen Bundesland. Q28/1 = wenig Vorbereitungszeit. Q28/2 = wenig Anwendungszeit. Q28/3 = wenig Zeit für Gutachten = Q28/3 = Vorbereitung in Freizeit.

Nordrhein-westfälische SonderpädagogInnen beschrieben in der Versuchsgruppe signifikant weniger zur Verfügung stehende Zeit zum Lernen eines Tests (Q28/1) gegenüber denen aus Baden-Württemberg ($p = .028$) und Niedersachsen ($p = .038$). Niedersächsische SonderpädagogInnen beschrieben weniger zur Verfügung stehende Zeit für die Anwendung eines Tests (Q28/2) gegenüber denen aus Baden-Württemberg ($p = .004$) und Nordrhein-Westfalen ($p = .041$) und ebenfalls gegenüber denen aus Nordrhein-Westfalen weniger zur Verfügung stehende Zeit für das Schreiben eines Gutachtens (Q28/3; $p = .008$).

In der Kontrollgruppe beschrieben baden-württembergische SonderpädagogInnen weniger zur Verfügung stehende Zeit für das Schreiben eines Gutachtens gegenüber denen aus Rheinland-Pfalz ($p = .043$) und für die gleiche Frage SonderpädagogInnen aus Niedersachsen ebenfalls gegenüber denen aus Rheinland-Pfalz ($p = .018$).

Generell und unter Einbezug der Gesamtstichprobe kann kein Unterschied zwischen der Kontroll- und Versuchsgruppe ermittelt werden. Der Mann-Whitney-U-Test berechnete nicht signifikante Ergebnisse für Q28/1 ($U(1087, 113) = 60778.50$, $z = -0.191$, $p = .849$), Q28/2 ($U(1083, 112) = 59455.50$, $z = -0.353$, $p = .724$), Q28/3 ($U(1082, 113) = 60632.00$, $z = -0.149$, $p = .882$) und Q28/4 ($U(1086, 112) = 59638.00$ und $z = -0.421$, $p = .674$).

5.4.5 Zusammenhänge zwischen Alter, empfundenen Schwierigkeiten und Anwendung der Tests

Hypothese 5.1

H0: Es gibt keinen Zusammenhang zwischen dem Alter der TesterInnen und den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

H1: Je älter die TesterInnen sind, desto weniger Schwierigkeiten werden bei der Anwendung von Intelligenztests empfunden.

Der Zusammenhang zwischen zunehmendem Alter und dem Schwierigkeiten-Index wurde mit der Pearson-Korrelation für stetige Variablen geprüft. Der Korrelationskoeffizient nach Pearson lag bei $r(1026) = .058$, $p = .065$. Das Vorzeichen des Korrelationskoeffizienten ist positiv, daraus resultiert aus dem gleichgerichteten Ergebnis ein höherer Wert im Schwierigkeiten-Index bei zunehmendem Alter, da im Fall gerichteter Hypothesen der zweiseitige p-Wert halbiert werden kann (Gehring & Weins, 2009, S. 285 f.) und somit $p = .033$ ergibt.

Da ein höherer Wert im Schwierigkeiten-Index gleichbedeutend mit weniger Schwierigkeiten einhergeht, bedeutet das Ergebnis, dass tatsächlich mit zunehmendem Alter weniger Schwierigkeiten bei der Anwendung von Intelligenztests auftreten und es von Vorteil ist, älter zu sein.

Beim Vergleich zwischen Kontroll- und Versuchsgruppe sind die Ergebnisse für die Versuchsgruppe ähnlich bei der ungerichteten Prüfung ($r(925) = .057$, $p = .083$); in der Kontrollgruppe ist hingegen weder eine Tendenz noch eine Signifikanz festzustellen ($r(99) = .062$, $p = .537$). In der Kontrollgruppe kann kein Zusammenhang zwischen erlebten Schwierigkeiten und dem Alter attestiert werden.

Hypothese 5.2

H0: Es gibt keinen Zusammenhang zwischen dem Alter der SonderpädagogInnen und der Anwendung der Testverfahren.

H1: Es gibt einen Zusammenhang zwischen dem Alter der SonderpädagogInnen und der Anwendung der Testverfahren.

Der Zusammenhang zwischen dem Alter und der Anwendung der Tests (Q8: *Wenn ich teste, nehme ich folgende Tests (...): immer (1), oft (2), gelegentlich (3), selten (4), nie (5)*) wurde mit der Spearman-Korrelation für ordinale Daten geprüft.

Bei den signifikanten Ergebnissen sind alle Vorzeichen negativ. Daraus resultiert, dass bei den vorliegenden Signifikanzen mit zunehmendem Alter die Tests häufiger durchgeführt werden.

Tatsächlich stehen die beiden ältesten Verfahren K-ABC ($rSp(696) = -.196$, $p < .001$) und SON-R 5½-17 ($rSp(763) = -.138$, $p < .001$) in einem Zusammenhang mit dem Alter; je älter, desto häufiger werden diese Tests angewendet. Um ein Verfahren als alt oder veraltet zu bezeichnen, sind die Gestaltung der Stimuli und vor allem die letztmalig erstellte Normstichprobe maßgeblich. Die K-ABC verfügt über veraltete Stimuli¹¹² und wurde 1986–1989 geeicht (Rollett & Preckel, 2011). Ebenfalls trifft dies für den SON-R 5½-17 zu, deren ausschließlich in den Niederlanden vorgenommene Normierung von 1984/1985 Grundlage der Auswertung ist.

Doch die Daten sollten dennoch nicht zu einer voreiligen Annahme der Alternativhypothese verleiten, denn entweder gab es keinen Unterschied zwischen Anwendung und Alter, oder aber ältere SonderpädagogInnen führen generell häufiger die Verfahren durch, sowohl veraltete als auch aktuellere Verfahren. Für den aktuellsten Test, die KABC-II (Erscheinung: 2015, Normierung: 2013/2014) gibt es keine Signifikanz ($rSp(773) = .009$, $p = .795$), ebenso wenig für die IDS ($rSp(678) = .051$, $p = .187$) und für den WNV ($rSp(608) = .036$, $p = .372$). Für alle anderen Tests ist festzustellen, dass ältere SonderpädagogInnen generell häufiger die Verfahren anwenden (CFT1/CFT1-R: $rSp(674) = -.202$; $p < .001$, CFT20-R: $rSp(638) = -.296$; $p < .001$, WISC-IV: $rSp(766) = -.180$; $p < .001$, WPPSI-III: $rSp(638) = -.129$; $p = .001$, SON-R 2½-7: $rSp(755) = -.174$; $p < .001$, SON-R 6-40: $rSp(728) = -.128$; $p = .001$).

Beim Vergleich zwischen Kontroll-, Versuchs- und Gesamtgruppe (siehe Tabelle 58) sind die Unterschiede zwischen Versuchs- und Gesamtgruppe marginal. Leichte Veränderungen in den Signifikanzen liegen bei der KABC-II (Gesamtgruppe: $rSp(773) = .009$, $p = .795$; Versuchsgruppe: $rSp(716) = .023$, $p = 533$.), dem SON-R 6-40 (Gesamtgruppe: $rSp(728) = -.128$; $p = .001$; Versuchsgruppe: $rSp(668) = -.109$; $p = .005$) und dem SON-R 5½-17 (Gesamtgruppe: $rSp(763) = -.138$, $p < .001$; Versuchsgruppe: $rSp(690) = -.124$, $p = .001$) vor, doch die aus den Ergebnissen resultierenden Aussagen bleiben gleich.

In der Kontrollgruppe werden signifikant häufiger der CFT1/CFT1-R ($rSp(64) = -.542$; $p < .001$), der CFT20-R ($rSp(64) = -.481$; $p < .001$), der SON-R 2½-7 ($rSp(64) = -.322$; $p = .008$), der SON-R 5½-17 ($rSp(71) = -.277$; $p = .018$) und der SON-R 6-40 ($rSp(58) = -.342$; $p = .007$) durchgeführt, je älter die SonderpädagogInnen sind.

112 Z. B. ein Bild von Charlie Chaplin, der Anfang der 80er sicherlich bekannter war bei den Kindern und Jugendlichen der Normstichprobe damals.

Tabelle 58. Spearman-Korrelation zwischen der Anwendung der Testverfahren (Q8) und dem Alter.

		Gesamtgruppe	Versuchsgruppe	Kontrollgruppe
K-ABC	Korrelationskoeffizient	-.196**	-.198**	-.183
	Sig.	.000	.000	.188
	N	698	645	53
KABC-II	Korrelationskoeffizient	.009	.023	-.146
	Sig.	.795	.533	.279
	N	775	718	57
CFT1/1-R	Korrelationskoeffizient	-.202**	-.176**	-.542**
	Sig.	.000	.000	.000
	N	876	810	66
CFT20-R	Korrelationskoeffizient	-.296**	-.277**	-.481**
	Sig.	.000	.000	.000
	N	840	774	66
WISC-IV	Korrelationskoeffizient	-.180**	-.174**	-.215
	Sig.	.000	.000	.115
	N	768	713	55
WPPSI-III	Korrelationskoeffizient	-.129**	-.138**	.002
	Sig.	.001	.001	.987
	N	640	592	48
WNV	Korrelationskoeffizient	.036	.044	-.019
	Sig.	.372	.298	.901
	N	610	566	44
SON-R 2½-7	Korrelationskoeffizient	-.174**	-.161**	-.322**
	Sig.	.000	.000	.008
	N	757	691	66
SON-R 5½-17	Korrelationskoeffizient	-.138**	-.124**	-.277
	Sig.	.000	.001	.018
	N	765	692	73
SON-R 6-40	Korrelationskoeffizient	-.128**	-.109**	-.342**
	Sig.	.001	.005	.007
	N	730	670	60
IDS	Korrelationskoeffizient	.051	.048	.184
	Sig.	.187	.230	.216
	N	680	633	47

Anmerkung. ** Korrelation ist auf 0,01 Niveau signifikant (zweiseitig). Sig. = signifikant.

5.4.6 Unterschiede zwischen Geschlecht und empfundenen Schwierigkeiten bei der Anwendung der Tests

Hypothese 6

H0: Es gibt einen Unterschied zwischen den Geschlechtern und den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

H1: Es gibt keinen Unterschied zwischen den Geschlechtern und den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

Die Prüfung wurde mit dem t-Test durchgeführt, da die Werte für den Schwierigkeiten-Index stetig und für das Geschlecht dichotom sind.

Der Unterschied der Mittelwerte im Schwierigkeiten-Index von den Frauen ($MW = 3.46$) und den Männern ($MW = 3.61$) ist signifikant. Der Levene-Test ermittelt eine Verletzung der Varianzhomogenität ($p = .021$), so dass die Prüfung unter der Bedingung nicht gleicher Varianzen mit dem Welch-Test vorgenommen wurde. Es liegt eine hohe Signifikanz von $t(433.744) = -4.005$, $p < .001$ vor. Auch wenn der tatsächliche Unterschied in den Mittelwerten nicht sehr hoch ist und auch wenn angenommen werden kann, dass die Signifikanz trotz der geringen Unterschiede in den Mittelwerten auch auf Grund der sehr großen Stichprobe vorliegt, ist das Ergebnis eindeutig: Männer beschreiben signifikant weniger Schwierigkeiten bei der Anwendung von Intelligenztests als Frauen.

Bei der Gegenüberstellung zwischen Kontroll- und Versuchsgruppe liegt in der Versuchsgruppe bei vorliegender Varianzgleichheit eine Signifikanz von $t(925) = -3.487$, $p = .001$ vor: Männer ($MW = 3.617$) beschreiben weniger Schwierigkeiten als Frauen ($MW = 3.469$).

In der Kontrollgruppe liegt bei ungleicher Varianz und dem daraus resultierenden Welch-Test eine Tendenz vor von $t(98.994) = 1.684$, $p = .095$: Männer ($MW = 3,508$) beschreiben tendenziell weniger Schwierigkeiten als Frauen ($MW = 3.341$).

5.4.7 Zusammenhänge zwischen Schwierigkeiten bei der Anwendung der Tests und der universitären Ausbildung

Hypothese 7.1

H0: Es gibt keinen Zusammenhang zwischen der Anzahl an besuchten universitären Seminaren zur Testdiagnostik und dem Ausmaß an erlebten Schwierigkeiten bei der Anwendung von Intelligenztests.

H1: Es gibt einen Zusammenhang zwischen der Anzahl an besuchten universitären Seminaren zur Testdiagnostik und dem Ausmaß an erlebten Schwierigkeiten bei der Anwendung von Intelligenztests.

Konkrete Problematiken erfragen Q14/1, Q14/2, Q14/3, Q15/1, Q15/2, Q15/3, die aus Gründen der Übersichtlichkeit vorgestellt werden:

Welche dieser Veränderungen haben Sie schon einmal vorgenommen?

- Q14/1: *Durchführungszeiten geändert (z.B. nach Ablauf der regulären Durchführungszeit/Item einen Punkt gegeben bei richtiger Antwort)?*
- Q14/2: *Durchführungszeit ganz weggelassen?*
- Q14/3: *Rückmeldungen gegeben, wenn diese nicht vorgesehen waren (z.B. richtig oder hast du richtig gelöst)?*

Folgende Durchführungsregeln bereiten mir Schwierigkeiten:

- Q15/1: *Umkehrregeln*
- Q15/2: *Abbruchregeln*
- Q15/3: *Ausrechnen des Testalters*

Gepprüft wird ebenfalls Q28/5 (*Mir fällt die Durchführung von Intelligenztests leicht. (völlig richtig (1), ziemlich richtig (2), unentschieden (3), ziemlich falsch (4) und falsch (5)¹¹³*). Die Items aus Q14 konnten fünfstufig mit *immer (1), oft (2), gelegentlich (3), selten (4), nie (5)*; die Items von Q15 fünfstufig mit *außerordentlich (1), ziemlich (2), mittelmäßig (3), kaum (4), gar nicht (5)* beantwortet werden.

Gepprüft wurde mit der Spearman-Korrelation zwischen diesen sieben Items und den Angaben zu den besuchten universitären Seminaren bzw. Vorlesungsreihen (Q21; *keine, 1, 2, 3, 4, mehr als 4*).

Q14 fragt nach bewusst vorgenommenen und die Durchführungsobjektivität gefährdenden Abweichungen in der Durchführung von Intelligenztests. Diese stehen nicht im direkten Zusammenhang mit Inhalten der universitären Ausbildung. Da es sich aber um Problematiken bei der Anwendung der Tests handelt, soll ein Zusammenhang ergebnisoffen nicht ausgeschlossen werden.

Es sind keine signifikanten Unterschiede zwischen der Anzahl belegter Seminare und der Änderung von Durchführungsregeln (Q14/1: $r_{Sp}(991) = .010$; $p = .756$), dem bewussten Weglassen von Durchführungszeiten (Q14/2: $r_{Sp}(986) = -.016$; $p = .616$) und dem Geben unerlaubter Feedbacks (Q14/3: $r_{Sp}(996) = .007$; $p = .817$) festgestellt worden.

Q15 fragt nach Schwierigkeiten im Umgang mit grundlegenden Anwendungsregeln. Diese stehen in einem Zusammenhang zwischen der Anzahl belegter Seminare und dem Grad der angegebenen Schwierigkeit, sowohl für Q15/1 (Umkehrregeln: $r_{Sp}(820) = .074$; $p = .035$), Q15/2 (Abbruchregeln: $r_{Sp}(962) = .093$; $p = .004$) als auch für Q15/3 (Testalter berechnen: $r_{Sp}(970) =$

113 Anmerkung: im Gegensatz zu den anderen Items liegt hier eine andere Polung vor.

.126; $p < .001$). Je mehr Seminare besucht worden sind, desto weniger Schwierigkeiten mit der Umkehr- und Abbruchregel und mit dem Ausrechnen des Testalters am Testtag werden erlebt.

Einen ebenfalls signifikanten Zusammenhang gibt es zwischen Q28/5 (*Mir fällt die Durchführung von Intelligenztests leicht*) und der Anzahl belegter Uniseminare $rSp(1027) = -.201$; $p < .001$). Je mehr Uniseminare bzw. Vorlesungsreihen zur Testdiagnostik belegt worden sind, desto leichter fällt SonderpädagogInnen die Anwendung von Intelligenztests.¹¹⁴

Vergleich Kontroll-, Gesamt- und Versuchsgruppe:

Bei der Gegenüberstellung von der Gesamt- und Versuchsgruppe gibt es bei der Umkehrregel (Q15/1) Unterschiede. Hier hat die Gesamtgruppe mehr Schwierigkeiten, je weniger Seminare belegt worden sind ($rSp(820) = .074$; $p = .035$), bei der Versuchsgruppe gibt es keinen signifikanten Unterschied ($rSp(761) = .055$; $p = .128$). Geringe Unterschiede gibt es zudem bei den Abbruchregeln (Q15/2; Gesamtgruppe: $rSp(962) = .093$; $p = .004$; Versuchsgruppe: $rSp(886) = .080$; $p = .018$). Die ermittelten Signifikanzen sind ansonsten zwischen Versuchs- und Gesamtgruppe in der Bedeutung ähnlich.

In der Kontrollgruppe verändern SonderpädagogInnen weniger (sic) unerlaubt die Durchführungszeiten während einer Testung, je weniger Seminare belegt worden sind (Q14/1: $rSp(85) = -.223$; $p = .038$) und haben tendenziell Schwierigkeiten mit den Umkehrregeln, je weniger Seminare belegt worden sind (Q15/1: $rSp(57) = .247$; $p = .059$).

In allen drei Gruppen (siehe Tabelle 59) gibt es einen signifikanten Zusammenhang zwischen Q28/5 und der Anzahl belegter Seminare (Gesamtgruppe: $rSp(1027) = -.204$; $p < .001$; Versuchsgruppe: $rSp(934) = -.182$; $p < .001$; Kontrollgruppe: $rSp(91) = -.343$; $p < .001$). Allen SonderpädagogInnen fällt die Anwendung von Intelligenztests leichter, je mehr Seminare sie belegt haben.

Hypothese 7.2

H0: Es gibt keinen Unterschied zwischen dem Ausmaß der in der universitären Ausbildung referierten Inhalte zur Testdiagnostik und den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

H1: Es gibt einen Unterschied zwischen dem Ausmaß der in der universitären Ausbildung referierten Inhalte zur Testdiagnostik und den empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests.

114 Hinweis: da diese Frage anders gepolt ist, bedeutet eine negative Korrelation, dass mit zunehmender Anzahl belegter Seminare die Anwendung von Intelligenztests als leichter empfunden wird.

Tabelle 59. Spearman-Korrelation Q14/1, Q14/2, Q14/3, Q15/1, Q15/2, Q15/3 mit Q21 für die Gesamt-, Versuchs- und Kontrollgruppe.

		Gesamtgruppe	Versuchsgruppe	Kontrollgruppe
Q14/1 (Zeiten anders)	Korrelationskoeffizient	.010	.029	-.223
	Sig.	.756	.387	.038
	<i>n</i>	993	906	87
Q14/2 (Zeit weggelassen)	Korrelationskoeffizient	-.016	-.028	.054
	Sig.	.616	.398	.619
	<i>n</i>	988	901	87
Q14/3 (unerlaubtes Feedback)	Korrelationskoeffizient	.007	.012	-.046
	Sig.	.817	.724	.677
	<i>n</i>	998	912	86
Q15/1 (Umkehrregeln)	Korrelationskoeffizient	.074	.055	.247
	Sig.	.035	.128	.059
	<i>n</i>	822	763	59
Q15/2 (Abbruchregeln)	Korrelationskoeffizient	.093**	.080	.178
	Sig.	.004	.018	.107
	<i>n</i>	964	881	83
Q15/3 (Testalter berechnen)	Korrelationskoeffizient	.126**	.123**	.115
	Sig.	.000	.000	.298
	<i>n</i>	972	888	84
Q28/5 (Anwendung fällt leicht)	Korrelationskoeffizient	-.204**	-.182**	-.343**
	Sig.	.000	.000	.001
	<i>n</i>	1029	936	93

Anmerkung. ** Korrelation ist auf 0,01 Niveau signifikant (zweiseitig). Sig. = Signifikanz.

Der bedingt durch die aus der großen Fallzahl resultierenden Annahme einer Normalverteilung der Stichprobenverteilung resultierende t-Test für zwei unabhängige Stichproben verglich die Angaben zu den universitären Inhalten (Q22: *Wurden im Rahmen der universitären Ausbildung folgende Inhalte vorgestellt (Ja/Nein): Standardabweichung (Q22/1), Durchführungsobjektivität (Q22/2), Vertrauens-/Konfidenzintervall (Q22/3), Messungenauigkeit/Messfehler (Q22/4) und Gaußsche Kurve der Normalverteilung (Q22/5)*) mit dem Schwierigkeiten-Index.

Tabelle 60 fasst die Mittelwertangaben, Tabelle B5 im Online-Material die Ergebnisse zusammen.

Tabelle 60. Mittelwertvergleiche Schwierigkeiten-Index für Q22/1–Q22/5: Wurden im Rahmen der universitären Ausbildung folgende Inhalte vorgestellt.

		N	Mittelwert Schwierigkeiten-Index	SD	Standardfehler Mittelwert
Standardabweichung	Ja	874	3.5086	0.55070	.01863
	Nein	64	3.3304	0.67571	.08473
Durchführungsobjektivität	Ja	845	3.5272	0.54820	.01886
	Nein	64	3.2623	0.64402	.08030
Vertrauens-/Konfidenzintervall	Ja	718	3.5275	0.53215	.01987
	Nein	118	3.3757	0.58265	.05365
Messungengenauigkeit	Ja	790	3.5023	0.55316	.01968
	Nein	91	3.3805	0.63703	.06692
Gaußsche Kurve der Normalverteilung	Ja	835	3.5065	0.55374	.01917
	Nein	64	3.3366	0.66472	.08332

Anmerkung. Je niedriger der Mittelwert, desto mehr beschriebene Schwierigkeiten. SD = Standardabweichung.

Der Mittelwertvergleich ermittelt für jedes Item einen höheren Mittelwert für den Schwierigkeiten-Index und somit weniger beschriebene Schwierigkeiten bei der Anwendung von Intelligenztests, wenn die jeweiligen Inhalte im Rahmen der universitären Ausbildung referiert worden sind:¹¹⁵ Dies trifft zu für die Inhalte *Standardabweichung* (Ja: MW = 3.51/SD = 0.551, Nein: MW = 3.330/SD = 0.676), *Durchführungsobjektivität* (Ja: MW = 3.53/SD = 0.548, Nein: MW = 3.26/SD = 0.644), *Vertrauens-/Konfidenzintervall* (Ja: MW = 3.53/SD = 0.532, Nein: MW = 3.38/SD = 0.583) *Messgenauigkeit/-fehler* (Ja: MW = 3.50/SD = 0.554, Nein: MW = 3.38/SD = 0.637) und für die *Gaußsche Kurve der Normalverteilung* (Ja: MW = 3.51/SD = 0.554, Nein: MW = 3.37/SD = 0.665).

Zur Prüfung vorhandener Signifikanzen prüfte der Levene-Test vorab für jedes Item, ob eine Varianzgleichheit vorliegt, was lediglich bei Q22/3 annähernd der Fall war (Q22/1: $p = .003$; Q22/2: $p = .009$; Q22/3: $p = .060$; Q22/4: $p = .009$; Q22/5: $p = .017$). Die bei Q23/3 vorliegende Tendenz zur Signifikanz ($p = .060$) ist unerheblich, da sowohl bei Varianzgleichheit ($t(834) = 2.83$, $p = .005$) als auch bei Ungleichheit ($t(150.74) = 2.65$, $p = .009$) ein signifikanter Mittelwertunterschied vorliegt (siehe Tabelle B5).

Die Prüfungen nach dem Welch-Test bei vorliegender Ungleichheit der Varianzen ermittelte signifikante Unterschiede in den Mittelwerten bei Q22/1

115 Es kann jedoch nicht ausgeschlossen werden, dass die jeweiligen Inhalte referiert worden sind, aber nicht erinnert werden können.

(Standardabweichung): $t(68.79) = 2.05$, $p = .044$ und bei Q22/2 (Durchführungsobjektivität): $t(70.48) = 3.21$, $p = .002$.

Tendenzen liegen vor bei Q22/4 (Messungenaugigkeit/Messfehler): $t(105.70) = 1.75$, $p = .084$ und bei Q22/5 (Gaußsche Kurve): $t(69.43) = 1.99$, $p = .051$.

Nach der Zusammenfassung der fünf Items zu einer gemeinsamen Variablen ergab die Prüfung mit der Spearman-Korrelation (siehe Tabelle B6) eine Signifikanz zwischen der gemeinsamen Variablen (alle Items der Fragengruppe Q22) mit dem Schwierigkeiten-Index ($r_{Sp}(1126) = .115$; $p < .001$).

Q23 (*Haben Sie an der Uni Intelligenztests ausprobiert?*) wurde mit Hilfe des t-Tests mit dem Schwierigkeiten-Index verglichen. SonderpädagogInnen, die diese Frage bejahten, erzielten im Mittelwertvergleich einen höheren Wert (gleichbedeutend mit weniger Schwierigkeiten) von $MW = 3,53$ ($SD = 0.542$) im Vergleich zu denen, die dies verneinten ($MW = 3,43$; $SD = 0.621$). Obwohl die Unterschiede gering sind, liegt auch hier eine Signifikanz vor nach der Prüfung mit dem Welch-Test bei nicht vorhandener Varianzgleichheit ($t(451.93) = 2.30$, $p = .022$).

Zur Verwendung des Schwierigkeiten-Index muss bedacht werden, dass einzelne Fragen dieses Index keinen Bezug zur universitären Ausbildung, sondern zu schulischen Rahmenbedingungen (z.B. *Wurden Sie schon einmal gestört durch Geräusche (...)*) oder zur Eignung der Testräume etc. haben. Die trotz dieser Einschränkungen signifikanten Ergebnisse deuten auf einen Zusammenhang zwischen der universitären Ausbildung und erlebten Schwierigkeiten bei der Anwendung von Intelligenztests hin, obwohl einige dieser Schwierigkeiten nicht in direktem Zusammenhang mit der Ausbildung stehen. Deshalb soll eine detailliertere Auswertung Zusammenhänge prüfen zwischen den Items, die Problematiken erfragen und den Items, die nach der universitären Ausbildung fragen.

Der im Anschluss für die post-hoc Prüfung durchgeführte Mann-Whitney-U-Test prüfte Zusammenhänge zwischen den drei Items aus Q14, den drei Items aus Q15 und Q28/5 mit den Items, die nach der universitären Ausbildung fragten (fünf Items aus Q22 und Q23). Es sei darauf hingewiesen, dass ein niedriger Wert für Q28/5 (*Mir fällt die Anwendung von Intelligenztests leicht*) im Gegensatz zu den anderen Testvariablen positiv bewertet wird, da aus einer hohen Zustimmung zu den anderen Testvariablen höhere Problematiken resultieren.

Zur besseren Übersicht werden die Items kurz dargestellt:

Welche dieser Veränderung haben Sie schon einmal vorgenommen:

- *Durchführungszeiten geändert (Q14/1)*
- *Durchführungszeit ganz weggelassen (Q14/2)*
- *Rückmeldungen gegeben, wenn diese nicht vorgesehen waren (Q14/3)*

Folgende Durchführungsgesetze bereiten mir Schwierigkeiten:

Umkehrgesetze (Q15/1) – Abbruchgesetze (Q15/2) – Ausrechnen des Testalters (Q15/3)

Mir fällt die Durchführung von Intelligenztests leicht (Q28/5).

Wurden im Rahmen der universitären Ausbildung folgende Inhalte vorgestellt:

Standardabweichung (Q22/1) – Durchführungsobjektivität (Q22/2) – Vertrauens-/Konfidenzintervall (Q22/3) – Messungenauigkeit/Messfehler (Q22/4) – Gaußsche Kurve der Normalverteilung (Q22/5)

Haben Sie an der Uni Intelligenztests ausprobiert (Q23)?

Tabelle 61. Mann-Whitney-U-Test zur Prüfung von Zusammenhängen zwischen Problematiken während der Testanwendung und universitären Inhalten.

		Q14/1 (Zeiten geändert)	Q14/2 (Zeiten weggelassen)	Q14/3 (unerlaubte Feedbacks)	Q15/1 (Umkehrregel schwierig)	Q15/2 (Abbruchregel schwierig)	Q15/3 (Testalter ausrechnen schwierig)	Q28/5 (Tests fallen mir leicht)
Q22/1 (Standardabweichung)	U	35411	33893	31399	21076	30448	33360	34469
	z	-.164	-.966	-1.888	-2.276	-.959	-.282	-1.584
	a. Sig.	.869	.334	.059	.023	.337	.778	.113
Q22/2 (Durchführungsobjektivität)	U	33092	33249	30407	19107	26247	25576	32687
	z	-.337	-.546	-1.537	-2.831	-2.139	-3.137	-1.663
	a. Sig.	.736	.585	.124	.005	.032	.002	.096
Q22/3 (Konfidenzintervall)	U	53850	51932	45277	32208	45620	45487	48373
	z	-.165	-1.059	-3.321	-2.935	-2.213	-2.828	-3.320
	a. Sig.	.869	.290	.001	.003	.027	.005	.001
Q22/4 (Messfehler)	U	45074	44434	38999	29507	38602	39772	48048
	z	-.309	-.814	-2.679	-1.892	-1.510	-1.572	-.254
	a. Sig.	.757	.416	.007	.058	.131	.116	.800
Q22/5 (Gaußsche Kurve)	U	30809	31546	28533	19444	28228	29180	34708
	z	-.713	-.651	-2.145	-1.447	-1.153	-1.189	-.393
	a. Sig.	.476	.515	.032	.148	.249	.234	.694
Q23 (Tests ausprobiert)	U	109298	108116	110050	72030	101560	94551	106834
	z	-.295	-.090	-.472	-1.399	-.714	-3.290	-2.957
	a. Sig.	.768	.929	.637	.162	.475	.001	.003

Anmerkungen. U = Mann-Whitney-U-Test. z = z-Wert. a. Sig. = asymptotische Signifikanz.

Es sind 13 signifikante Unterschiede und drei Tendenzen berechnet worden (siehe Tabelle 61). Jede dieser Signifikanzen gibt einen Hinweis darauf, dass aus nicht referierten (oder erinnerten) Inhalten während der universitären Ausbildung Problematiken resultieren bzw. die Anwendung von Intelligenztests leichter fällt, wenn Inhalte im Zusammenhang mit der Anwendung von Intelligenztests an der Universität referiert worden sind.

SonderpädagogInnen tendieren dazu, ein unerlaubtes Feedback in der Testsituation zu geben (Q14/3: $U(983, 73) = 31399.50, z = -1.888, p = .059$), wenn das Konstrukt der Standardabweichung nicht referiert wurde und haben außerdem dann signifikant mehr Schwierigkeiten mit der Umkehrregel (Q15/1: $U(815, 62) = 21075.50, z = -2.276, p = .023$).

Wenn an der Universität die Durchführungsobjektivität nicht referiert wurde, resultierten daraus signifikant mehr Schwierigkeiten bei der Anwendung der Umkehrregel (Q15/1: $U(791, 61) = 19107.00, z = -2.831, p = .005$), der Abbruchregel (Q15/2: $U(933, 66) = 26246.50, z = -2.139, p = .032$) und dem Ausrechnen des Testalters (Q15/3: $U(932, 68) = 25576.00, z = -3.137, p = .002$). Ebenfalls liegt dann eine Tendenz vor, die Anwendung von Intelligenztests als *fällt nicht leicht* einzuschätzen (Q28/5: $U(978, 75) = 32687, z = -1.663, p = .096$).

Wurde das Vertrauens- bzw. Konfidenzintervall nicht an der Universität referiert (oder erinnert), werden signifikant häufiger unerlaubte Feedbacks in der Testsituation gegeben (Q14/3: $U(813, 134) = 45277.00, z = -3.321, p = .001$), die Umkehr- (Q15/1: $U(696, 111) = 32208.00, z = -2.935, p = .003$) und Abbruchregeln (Q15/2: $U(798, 129) = 45619.50, z = -2.213, p = .027$) sowie das Ausrechnen des Testalters (Q15/3: $U(798, 131) = 45487.00, z = -2.828, p = .005$) als schwierig empfunden und generell die Anwendung von Intelligenztests als weniger leicht (Q28/5: $U(828, 140) = 48373, z = -3.320, p = .001$).

Wurden Inhalte zur Messgenauigkeit bzw. zum Messfehler¹¹⁶ nicht referiert, werden signifikant häufiger Feedbacks gegeben (Q14/3: $U(884, 104) = 38999.00, z = -2.679, p = .007$) und es liegt eine Tendenz zu Schwierigkeiten beim Umgang mit der Umkehrregel vor (Q15/1: $U(734, 91) = 29506.50, z = -1.892, p = .058$).

Signifikant häufiger werden Feedbacks (Q14/3: $U(939, 71) = 28532.50, z = -2.145, p = .032$) gegeben, wenn die Gaußsche Kurve der Normalverteilung nicht referiert worden ist.

Signifikant mehr Schwierigkeiten liegen beim Ausrechnen des Testalters vor, wenn Tests an der Universität nicht ausprobiert worden sind (Q15/3: $U(700, 304) = 94551.00, z = -3.290, p = .001$). SonderpädagogInnen fällt die Anwendung von Tests signifikant leichter (Q28/5: $U(732, 327) = 106833.50, z = -2.957, p = .003$), wenn Intelligenztests an der Universität erprobt werden konnten.

116 Als Kernkonstrukt der Klassischen Testtheorie.

Die drei Items aus Q14 (Schwierigkeiten bei den Umkehr- und Abbruchregeln und dem Ausrechnen des Testalters) beziehen sich auf Inhalte, die in Fortbildungen zur Anwendung von Intelligenztests referiert werden. Es ist möglich, dass TeilnehmerInnen aus diesen Seminaren weniger Schwierigkeiten bezüglich dieser Konstrukte empfinden (obwohl vielfältige Signifikanzen festgestellt worden sind, s. o.), unabhängig davon, ob es an der Universität gelehrt wurde oder nicht, da sie es ja vorgestellt bekamen im Rahmen der Seminare. Es wäre interessant, ob die Ergebnisse ähnlich sind für die ProbandInnengruppe, die nicht an Seminaren zu standardisierten Testverfahren teilnahmen. Deshalb werden Zusammenhänge im Rahmen dieser Hypothesenprüfung erneut unter der Bedingung geprüft, dass die ProbandInnen nie an einer Fortbildung zu standardisierten Testverfahren teilnahmen.

Beim Vergleich zwischen Gesamt-, Kontroll- und Versuchsgruppe wurde zunächst der t-Test für zwei unabhängige Stichproben getrennt für die Kontroll- und Versuchsgruppe durchgeführt (Q22: *Wurden im Rahmen der universitären Ausbildung folgende Inhalte vorgestellt: Standardabweichung (Q22/1), Durchführungsobjektivität (Q22/2), Vertrauens-/Konfidenzintervall (Q22/3), Messungenauigkeit/Messfehler (Q22/4) und Gaußsche Kurve der Normalverteilung (Q22/5)*) und verglichen mit dem Schwierigkeiten-Index. Aus Gründen der Übersichtlichkeit werden die Mittelwerte, die Standardabweichungen und die Ergebnisse der Signifikanzprüfungen für die drei Gruppen zusammengefasst dargestellt in Tabelle 62. Die Wahl des post-hoc Tests ist abhängig von der Prüfung der Varianzgleichheit mit dem Levene-Test.

Die zu einer Variablen zusammengeführten fünf Items der Fragengruppe Q22 sind anschließend getrennt nach Versuchs- und Kontrollgruppe mit der Spearman-Korrelation mit dem Schwierigkeiten-Index in Verbindung gebracht worden. Während für die Gesamtgruppe eine Korrelation von $r_{Sp}(1126) = .115$; $p < .001$ festgestellt wurde, hat sich dies in der Versuchsgruppe bestätigt ($r_{Sp}(1021) = .107$; $p = .001$), in der Kontrollgruppe wurde eine Tendenz mit $r_{Sp}(102) = .173$; $p = .077$ ermittelt. Die jeweils gleichgerichteten Ergebnisse bedeuten, dass je häufiger mit *Ja* geantwortet wurde (*Ja* wurde mit 1, *Nein* mit 0 umgepolt), der Schwierigkeiten-Index höher ist (was weniger Schwierigkeiten bedeutet). Je mehr der beschriebenen Inhalte also referiert worden sind, desto weniger Schwierigkeiten werden erlebt (Versuchs- und Gesamtgruppe) bzw. tendenziell erlebt (Kontrollgruppe).

Beim Vergleich zwischen Q23 (*Haben Sie an der Uni Intelligenztests ausprobiert*) und dem Schwierigkeiten-Index gibt es ein signifikantes Ergebnis nach der Prüfung mit dem t-Test für die Gesamtgruppe ($t(451.93) = 2.30$, $p = .022$ bei ungleicher Varianz), für die Versuchsgruppe ($t(813) = 2.238$, $p = .026$ bei gleicher Varianz), aber keinen Unterschied bei der Kontrollgruppe ($t(31.80) = 0.927$, $p = .361$ bei ungleicher Varianz).

Tabelle 62. Vergleich Gesamt-, Versuchs- und Kontrollgruppe: Uni-Inhalte vs. Schwierigkeiten-Index.

	Uni	Gesamtgruppe			Versuchsgruppe			Kontrollgruppe					
		n	MW	SD	Sig.	n	MW	SD	Sig.	n	MW	SD	Sig.
Q22/1	Ja	874	3.51	0.551	.044	787	3.52	0.529	.047	87	3.41	0.713	.673
	Nein	64	3.33	0.676		58	3.35	0.674		6	3.29	0.752	
Q22/2	Ja	845	3.53	0.548	.002	765	3.53	0.019	.002	80	3.46	0.708	.521
	Nein	64	3.26	0.644		59	3.26	0.638		6	3.26	0.776	
Q22/3	Ja	718	3.53	0.532	.005	650	3.53	0.529	.009	68	3.52	0.569	.279
	Nein	118	3.38	0.583		109	3.39	0.574		9	3.29	0.720	
Q22/4	Ja	790	3.50	0.553	.084	716	3.51	0.535	.075	74	3.40	0.702	.853
	Nein	91	3.38	0.637		84	3.38	0.634		7	3.35	0.735	
Q22/5	Ja	835	3.51	0.554	.051	760	3.52	0.538	.048	75	3.42	0.695	.740
	Nein	64	3.34	0.665		57	3.34	0.647		7	3.33	0.863	

Anmerkungen. Q22/1 = Standardabweichung. Q22/2 = Durchführungsobjektivität. Q22/3 = Vertrauens-/Konfidenzintervall. Q22/4 = Messgenauigkeit. Q22/5 = Gaußsche Kurve. Uni = wurde das Konstrukt referiert. MW = Mittelwert. SD = Standardabweichung. Sig. = Signifikanz.

Es sei erneut daran erinnert, dass einige Items, die zum Schwierigkeiten-Index gehören, keinen Bezug zur universitären Ausbildung haben. Deshalb soll eine detailliertere Analyse mögliche Unterschiede zwischen Kontroll- und Versuchsgruppe ermitteln. Aus Gründen der Übersichtlichkeit werden die Ergebnisse der Signifikanzprüfung in den Tabellen 63 und 64, sowie B7, B8 und B9 dargestellt.

Wurde im Rahmen der universitären Ausbildung das Konstrukt Standardabweichung (Q22/1) referiert, wurden in der Versuchsgruppe weniger Rückmeldungen (Q14/3) während der Testsituation gegeben ($U(893, 66) = 25444.50$, $z = -1.965$, $p = .049$; keine Signifikanzen in der Kontrollgruppe). Wurde die Durchführungsobjektivität (Q22/2) an der Universität thematisiert, hatten die SonderpädagogInnen aus der Versuchsgruppe weniger Schwierigkeiten mit der Umkehrregel (Q15/1: $U(731, 57) = 16537.50$, $z = -2.711$, $p = .007$), mit der Abbruchregel (Q15/2: $U(851, 60) = 21546.00$, $z = -2.160$, $p = .031$) und mit dem Ausrechnen des Testalters (Q15/3: $U(853, 61) = 19652$, $z = -3.776$, $p < .001$).

Die meisten Unterschiede konnten erkannt werden, wenn die zum Vertrauens- bzw. Konfidenzintervall (Q22/3) zugehörigen Inhalte nicht referiert worden sind. Für diesen Fall hatten die SonderpädagogInnen der Versuchsgruppe signifikant mehr Schwierigkeiten bei der Anwendung der Umkehrregel (Q15/1: $U(643, 106) = 28291.50$, $z = -2.925$, $p = .003$), der Abbruchregel (Q15/2: $U(725, 121) = 38812.00$, $z = -2.169$, $p = .030$) und dem Ausrechnen des Testalters (Q15/3: $U(725, 122) = 37964.50$, $z = -2.972$, $p = .003$) und haben dann auch

häufiger unerlaubte Rückmeldungen in der Testsituation gegeben (Q14/3: $U(738, 126) = 38302.50, z = -3.357, p = .001$). Die Anwendung von Intelligenztests wird für diesen Fall auch signifikant leichter empfunden (Q28/5: $U(752, 130) = 40650.00, z = -3.268, p = .001$). In der Kontrollgruppe konnten keine signifikanten Unterschiede festgestellt werden.

Tabelle 63. Vergleich Universitäts-Inhalte (Q2, Q23) mit Q14/1, Q14/2, Q14/3 und Q15/1, getrennt für die Kontroll- und Versuchsgruppe (Mann-Whitney-U-Test).

	Uni	Q14/1		Q14/2		Q14/3		Q15/1	
		Rang	Sig.	Rang	Sig.	Rang	Sig.	Rang	Sig.
Q22/1 Vers.	Ja	479.2	.553	473.5	.521	485.4	.049	409.2	.052
	Nein	461.1		488.3		419.0		349.4	
Q22/1 Kontr.	Ja	48.5	.203	48.8	.296	49.0	.965	35.4	.161
	Nein	61.1		57.7		48.6		23.1	
Q22/2 Vers.	Ja	468.3	.389	464.4	.723	472.5	.079	400.4	.007
	Nein	442.4		472.5		414.9		319.1	
Q22/2 Kontr.	Ja	44.9	.158	45.6	.523	45.1	.607	32.9	.405
	Nein	58.7		50.6		50.5		25.4	
Q22/3 Vers.	Ja	431.5	.746	426.3	.332	443.6	.001	384.0	.003
	Nein	424.8		441.6		367.5		320.4	
Q22/3 Kontr.	Ja	41.6	.290	42.4	.832	42.4	.624	29.9	.594
	Nein	50.2		43.8		38.3		25.8	
Q22/4 Vers.	Ja	451.4	.727	447.1	.655	460.9	.003	388.7	.083
	Nein	443.0		455.5		382.2		346.4	
Q22/4 Kontr.	Ja	41.3	.051	42.2	.271	42.1	.617	30.1	.416
	Nein	57.4		49.8		46.4		24.4	
Q22/5 Vers.	Ja	460.9	.499	458.6	.727	467.9	.032	393.0	.222
	Nein	440.3		450.4		397.7		355.2	
Q22/5 Kontr.	Ja	43.5	.987	44.0	.514	43.3	.741	29.6	.330
	Nein	43.4		39.1		402		21.5	
Q23 Vers.	Ja	467.2	.946	467.7	.833	473.6	.581	404.1	.140
	Nein	468.3		465.4		463.6		379.2	
Q23 Kontr.	Ja	45.9	.930	46.8	.552	45.5	.980	32.9	.756
	Nein	46.4		43.8		45.4		31.1	

Anmerkungen. Q22/1 = Standardabweichung. Q22/2 = Durchführungsobjektivität. Q22/3 = Vertrauens-/Konfidenzintervall. Q22/4 = Messgenauigkeit. Q22/5 = Gaußsche Kurve. Uni = wurde das Konstrukt referiert. Q14/1 = Durchführungszeiten geändert. Q14/2 = Durchführungszeiten weggelassen. Q14/3 = unerlaubtes Feedback gegeben. Q15/1 = Umkehrregeln. Sig. = Signifikanz. Vers. = Versuchsgruppe. Kontr. = Kontrollgruppe.

Tabelle 64. Vergleich Universitäts-Inhalte (Q22, Q23) mit Q15/2, Q15/3 und Q28/5 getrennt für die Kontroll- und Versuchsgruppe (Mann-Whitney-U-Test).

	Uni	Q15/2		Q15/3		Q28/5	
		Rang	Sig.	Rang	Sig.	Rang	Sig.
Q22/1 Vers.	Ja	467.7	.425	469.3	.370	488.3	.113
	Nein	441.1		442.5		541.5	
Q22/1 Kontr.	Ja	48.4	.594	45.7	.089	52.1	.640
	Nein	42.9		60.6		57.1	
Q22/2 Vers.	Ja	460.7	.031	465.0	.000	474.7	.066
	Nein	389.6		353.1		534.8	
Q22/2 Kontr.	Ja	44.7	.779	42.6	.168	48.5	.948
	Nein	41.8		54.2		49.1	
Q22/3 Vers.	Ja	430.5	.030	432.6	.003	430.6	.001
	Nein	381.8		372.7		504.8	
Q22/3 Kontr.	Ja	41.4	.627	41.4	.902	42.7	.377
	Nein	37.4		42.3		49.9	
Q22/4 Vers.	Ja	444.0	.144	447.0	.041	463.3	.794
	Nein	405.3		398.2		470.3	
Q22/4 Kontr.	Ja	41.8	.734	41.0	.218	45.8	.810
	Nein	38.9		50.1		47.9	
Q22/5 Vers.	Ja	452.5	.249	453.7	.103	474.9	.680
	Nein	415.1		406.3		461.8	
Q22/5 Kontr.	Ja	42.1	.938	40.2	.277	45.6	.884
	Nein	41.4		48.5		44.3	
Q23 Vers.	Ja	459.6	.443	477.3	.000	464.9	.006
	Nein	446.0		415.9		515.2	
Q23 Kontr.	Ja	44.1	.926	42.8	.244	48.0	.268
	Nein	43.6		49.2		54.8	

Anmerkungen. Q22/1 = Standardabweichung. Q22/2 = Durchführungsobjektivität. Q22/3 = Vertrauens-/Konfidenzintervall. Q22/4 = Messgenauigkeit. Q22/5 = Gaußsche Kurve. Uni = wurde das Konstrukt referiert. Q15/2 = Abbruchregeln. Q15/3 = Testalter ausrechnen. Q28/5 = Testanwendung fällt leicht. Sig. = Signifikanz. Vers. = Versuchsgruppe. Kontr. = Kontrollgruppe. Q28/5: andere Polung (niedriger Wert = Anwendung von Tests fällt leichter).

Wurde das Konstrukt Messgenauigkeit/-fehler (Q22/4) nicht an der Universität referiert, wurden häufiger unerlaubte Rückmeldungen gegeben in der Versuchsgruppe (Q14/3: $U(808, 96) = 32038.50$, $z = -2.952$, $p = .003$) und das Ausrechnen des Testalters bereitete mehr Schwierigkeiten (Q15/3: $U(792, 91) =$

32052.50, $z = -2.039$, $p = .041$), während in der Kontrollgruppe keine Signifikanzen vorliegen.

Ebenfalls keine signifikanten Unterschiede sind für die Kontrollgruppe berechnet worden, wenn an der Universität die Gaußsche Kurve der Normalverteilung (Q22/5) nicht referiert worden ist, während in der Versuchsgruppe häufiger Rückmeldungen in Testsituationen gegeben worden sind (Q14/3: $U(861, 41) = 23373.00$, $z = -2.146$, $p = .032$).

Sind an der Universität Intelligenztests ausprobiert worden (Q23), fällt die Anwendung von Tests (Q28/5: $U(662, 298) = 88286.50$, $z = -2.771$, $p = .006$) und das Ausrechnen des Testalters (Q15/3: $U(635, 281) = 77251.50$, $z = -3.815$, $p < .001$) SonderpädagogInnen aus der Versuchsgruppe deutlich leichter, während für die Kontrollgruppe keine signifikanten Unterschiede festgestellt werden können.

5.4.8 Zusammenhänge zwischen Schwierigkeiten bei der Anwendung der Tests und der außeruniversitären Fortbildung

Hypothese 8

H0: Es gibt keine Unterschiede zwischen der Teilnahme an einer außeruniversitären Fortbildung zur Testdiagnostik und Schwierigkeiten bei der Anwendung von Intelligenztests.

H1: Es gibt einen Unterschied zwischen der Teilnahme an einer außeruniversitären Fortbildung zur Testdiagnostik und Schwierigkeiten bei der Anwendung von Intelligenztests.

Der t-Test vergleicht Q25 (*Haben Sie an einer außeruniversitären Fortbildung zu Intelligenztests teilgenommen?*) mit dem Schwierigkeiten-Index. Beim Mittelwertvergleich gaben TeilnehmerInnen von Fortbildungen zu Intelligenztests ($MW = 3.50$, $SD = .540$) nur moderat weniger Schwierigkeiten bei der Anwendung von Intelligenztests an als SonderpädagogInnen, die an keiner entsprechenden Fortbildung teilnahmen ($MW: 3.37$, $SD = .754$).

Der Levene-Test stellt eine signifikante Abweichung der Varianzgleichheit fest ($p < .001$), so dass die Angaben zur Signifikanz für nicht gleiche Varianzen gewertet werden mit dem Welch-Test. Es besteht eine Tendenz zu schwach weniger erlebten Schwierigkeiten bei der Anwendung von Intelligenztests bei SonderpädagogInnen, die an einer Fortbildung teilnahmen ($t(111.472) = 1.719$, $p = .088$).

5.4.9 Unterschiede zwischen Auswertungsfehlern und der Anwendung von Auswertungsprogrammen

Ab hier werden die Hypothesenprüfungen mit Hilfe der Analyse ausgefüllter Formulare durchgeführt, die im Rahmen von Begutachtungen zur Prüfung sonderpädagogischen Förderbedarfs¹¹⁷ angefertigt worden sind im Rahmen einer Intelligenztestung.

Hypothese 9

H0: Die Anzahl gemachter Fehler unterscheidet sich nicht zwischen Auswertungen mit und ohne Computerauswertungen.

H1: Die Anzahl gemachter Fehler unterscheidet sich zwischen Auswertungen mit und ohne Computerauswertungen.

Der Mann-Whitney-Test konnte keinen signifikanten Unterschied feststellen ($U(122, 123) = 7494.00, z = -0.17, p = .986$); mit Computerauswertung: mittlerer Rang = 123.07; ohne Computerauswertung: mittlerer Rang = 122.93)¹¹⁸. Die Anzahl der gefundenen Fehler in den Formularen ist nach Prüfung aller Fälle nicht abhängig von der Nutzung einer Computerauswertung.

Bei ausschließlicher Betrachtung des WISC-IV ist es ebenfalls unerheblich, ob die Computerauswertung genutzt wurde ($U(32, 28) = 345.00, z = -0.581, p = .114$; mit Computerauswertung: mittlerer Rang = 27.28; ohne Computerauswertung: mittlerer Rang = 34.18). Es wurden zwar mehr Fehler ohne Computerauswertung gemacht, doch ist dieser Unterschied nicht signifikant. Gleiches gilt für den SON-R 6-40 ($U(54, 8) = 205.50, z = -0.249, p = .803$; mit Computerauswertung: mittlerer Rang = 31.69; ohne Computerauswertung: mittlerer Rang = 30.19) und für die KABC-II ($U(8, 11) = 25.50, z = -1.587, p = .129$, mit Computerauswertung: mittlerer Rang = 12.31; ohne Computerauswertung: mittlerer Rang = 8.32). Für die anderen Tests liegen zu wenige Fälle vor bzw. es besteht nicht die Möglichkeit der Auswertung mit einem Computerprogramm.

117 Abhängig vom Bundesland kann es auch anders heißen, z. B. Unterstützungsbedarf oder Bildungsangebot.

118 Asymptotische Signifikanz.

5.4.10 Zusammenhänge zwischen Durchführungs- und Auswertungsfehlern und der Komplexität der Tests

Hypothese 10

H0: Es gibt keinen Zusammenhang zwischen der Anzahl gemachter Fehler und der *Komplexität* der Intelligenztests.

H1: Es gibt einen Zusammenhang zwischen der Anzahl gemachter Fehler und der *Komplexität* der Intelligenztests.

Die nach *Komplexität* in fünf Gruppen aufgeteilten Tests werden korreliert mit Hilfe der Spearman-Korrelation für ordinale Daten; eine Korrelation nach Pearson schließt sich aus, da die Kriterien, nach denen die fünf Gruppen aufgeteilt sind, keine gleichen Abstände zwischen den Gruppen zulassen.

Je komplexer die Tests sind, desto mehr Fehler sind vorhanden ($r_{Sp}(242) = .214$; $p = .001$).

Bei genauerer Betrachtung, unterschieden nach Fehlerart, liegen weder für die Abbruchregel ($r_{Sp}(511) = .004$; $p = .980$) noch für die Umkehrregel ($r_{Sp}(25) = -.077$; $p = .704$) Signifikanzen vor, jedoch für die falsch gezählten Punkte bzw. falschen Auswertungen eine negative Korrelation ($r_{Sp}(93) = -.233$; $p = .023$)¹¹⁹.

Während also tatsächlich mehr Fehler gemacht werden, je komplexer der Test ist, sind weniger Fehler feststellbar beim Addieren der Rohwertpunkte bzw. bei der Bewertung der Aufgaben, je komplexer der Test ist, siehe Tabelle 65.

Tabelle 65. Korrelation zwischen Fehler und Komplexität der Tests.

	Addieren/Auswerten	Abbruchregel	Umkehrregel	Gesamtfehler
Korrelationskoeffizient	-.233	.004	-.077	.214
Signifikanz (zweiseitig)	.023	.980	.704	.001
<i>n</i>	95	53	27	244

Nachvollziehbar hängt die *Komplexität* der Tests auch von der *Dimensionalität* ab, da mehrdimensionale Tests *komplex*, *sehr komplex* oder *außerordentlich komplex* sind, eindimensionale Tests *wenig* oder *leicht komplex*. Deshalb bietet sich im Rahmen der Hypothesenprüfung auch ein Vergleich dieser beiden Gruppen an. Da keine Normalverteilung vorliegt bei der Anzahl gemachter Fehler, sondern die Verteilung stark linkssteil ist (siehe Abbildung C), kommt der t-Test nicht in Frage.

¹¹⁹ Für die Kategorie falsch berechnetes Testalter liegen zu wenige Fälle vor.

Der Unterschiede berechnende Mann-Whitney-U-Test ermittelte eine hohe Signifikanz von $U(109, 85) = 3127.00$, $z = -4.05$, $p < .001$ (eindimensional: mittlerer Rang = 83.69; mehrdimensional: mittlerer Rang = 115.21). Mehrdimensionale Tests sind signifikant fehleranfälliger als eindimensionale Tests (da der mittlere Rang bei mehrdimensionalen Tests höher ist).

Eine genauere Analyse¹²⁰ unterschieden nach Fehlerart¹²¹ ermittelt in der Kategorie Abbruchregel keine Signifikanz ($U(4, 38) = 58.00$, $z = -1.076$, $p = .282$), auch wenn der mittlere Rang bei den mehrdimensionalen Tests höher ist (mittlerer Rang 21.97; eindimensional: mittlerer Rang = 17,00). Bei den falsch gezählten Punkten bzw. bei der falschen Auswertung besteht eine Signifikanz von $U(41, 34) = 527.00$, $z = -2.155$, $p = .031$ (mehrdimensional: mittlerer Rang = 33.00; eindimensional: mittlerer Rang = 42.15). Obwohl also insgesamt mehr Fehler bei den mehrdimensionalen Tests gemacht worden sind, wurden signifikant mehr Punkte falsch gezählt bzw. wurden falsche Auswertungen bei den eindimensionalen Tests vorgenommen.

120 Jeweils mit dem Mann-Whitney-U-Test, asymptotische Signifikanz, zweiseitig.

121 Ohne *falsch berechnetes Alter am Testtag*, da dies einmal zu Beginn eines Tests durchgeführt wird und kein Zusammenhang zur Dimensionalität besteht und ohne *Umkehrregel*, da es diese bei den eindimensionalen Tests nicht gibt; für die Fehlerart *Anfangsitem* liegen zu wenig Fälle vor.

6 Interpretation und Diskussion

Grundlage der erzielten Ergebnisse war unter anderem die Auswertung von Intelligenztest-Formularen unter Berücksichtigung der Durchführungs- und Auswertungsobjektivität, die im Rahmen sonderpädagogischer Begutachtungen durchgeführt worden sind. Der größere Teil der Ergebnisse resultiert jedoch aus den Antworten des Fragebogens für SonderpädagogInnen. Diese Ergebnisse werden zunächst bewertet, anschließend die Ergebnisse aus der Analyse von Testformularen. Am Ende dieses Kapitels sollen Einschränkungen beschrieben werden, die die Ergebnisse relativieren könnten und ein Fazit gezogen werden sowie ein Ausblick auf mögliche Folgeprojekte, die sich aus dieser Arbeit ergeben können.

6.1 Fragebogen

Die Interpretation der Ergebnisse aus der Fragebogenauswertung kann unterteilt werden in Ergebnisse, die sich mit der Anwendung der Tests, mit Bundesländervergleichen, mit dem Alter und Geschlecht und mit der Ausbildung beschäftigen.

Für jeden dieser vier Blöcke werden zusammengefasst Ableitungen und Schlussfolgerungen vorgestellt, die für die Anwendung von Intelligenztests im sonderpädagogischen Kontext interessant sein könnten, evtl. sogar Handlungshinweise vorgeschlagen.

6.1.1 Anwendung

Die ersten drei Hypothesen beschäftigten sich mit der Anwendung und Bedeutung von Intelligenztests und seien kurz erinnert:

- Hypothese 1 fragte nach der Aussagekraft abhängig von der Dimensionalität des Verfahrens.
- Hypothese 2 fragte, ob komplexere Tests seltener angewendet werden.
- Hypothese 3 prüfte, ob es Vorlieben für bestimmte Tests gibt, auch wenn andere vorhanden sind.

6.1.1.1 Interpretation der Ergebnisse zur Anwendung von Intelligenztests

Intelligenztests haben einen hohen Stellenwert im sonderpädagogischen Alltag. Lediglich 3 Prozent der ProbandInnen führten bisher keinen Intelligenztest durch und von 72 Prozent wird in Zukunft erwartet, Intelligenztests durchzuführen. Die Anwendung von Intelligenztests ist keine Randerscheinung und alle in Folge diskutierten Schlussfolgerungen stehen im Zusammenhang mit einem Arbeitsbereich von Bedeutung in der Sonderpädagogik. Selbst bei einer seltenen Anwendung von Intelligenztests kann eine hohe Bedeutung angenommen werden, da die Gefahr einer Stigmatisierung durch die falsche Anwendung der Tests als bekannt unterstellt werden darf.

Die erste Hypothese erfragte die empfundene Aussagekraft von mehr- bzw. eindimensionalen Tests. Signifikant höher wird die Aussagekraft mehrdimensionaler Tests eingeschätzt. Auch wenn das Ergebnis absehbar war, diente sie späteren Argumentationen. Die Motivation für diese Forschung ist u. a. eine Verbesserung bei der Anwendung von Intelligenztests. Es kann nicht ausgeschlossen werden, dass in der Praxis weniger aussagekräftige Tests angewendet werden, obwohl diese kaum pädagogische Schlussfolgerungen oder Ableitungen für Fördermaßnahmen zulassen, begründet mit der einfacheren Anwendbarkeit der eindimensionalen Tests und dem geringeren Zeitaufwand.

Sollten mehrdimensionale Tests nicht angewendet werden, obwohl sie als aussagekräftiger eingeschätzt werden, würde auf einer soliden Grundlage diskutiert werden können, welche Bedingungen die Anwendung aussagekräftigerer Tests verhindern (z. B. Zeitdruck, ungenügende Vorbereitungszeit, Mängel in der Ausbildung usw.). Die solide Grundlage dafür ist diese Hypothesenprüfung und es kann nun ausgeschlossen werden, dass mehrdimensionale Tests als weniger aussagekräftig eingeschätzt werden, z. B. indem Drittvariablen Einfluss auf ein unerwartetes Ergebnis nehmen könnten, z. B. die Einschätzung, dass eindimensionale Tests aussagekräftiger sind, weil die wenigen Ergebnisse besser interpretierbar sind im Gegensatz zu den vielen Ergebnissen eines mehrdimensionalen Tests, dessen kontextuale Würdigung überfordern könnte.

Es ist also festzustellen, dass die ProbandInnen die höhere Aussagekraft mehrdimensionaler Tests wie KABC-II oder WISC-IV honorieren.

In der zweiten Hypothese wurde nach Unterschieden zwischen Anwendungshäufigkeit und Komplexität der Tests gefragt. Die mehrdimensionalen Tests sind auch die komplexeren Tests, also die Tests mit mehr Durchführungsregeln. Entsprechend dem Ergebnis der ersten Hypothese wäre zu erwarten, dass komplexere Tests auf Grund ihrer höheren Aussagekraft häufiger durchgeführt werden.

Bis auf wenige Ausnahmen werden weniger komplexe Tests häufiger als komplexere Tests durchgeführt. Die KABC-II hat in diesem Zusammenhang eine besondere und unerwartete Stellung. Trotz der Neigung zur Anwendung

vermeintlich einfacherer Tests wird ausgerechnet der mit Abstand komplexeste Test KABC-II ebenfalls häufig angewendet. Die KABC-II (*außerordentlich komplex*) wird signifikant häufiger angewendet als alle anderen Tests mit weniger Regeln, bis auf die Tests der CFT-Reihe (*wenig komplex*), hier liegt kein Unterschied in der Anwendungshäufigkeit vor.

Positiv formuliert bedeutet dies, dass der aussagekräftigste Test nicht seltener angewendet wird wie die am wenigsten aussagekräftigen Tests aus der CFT-Reihe, negativ formuliert aber auch nicht häufiger.

Weniger plausible Erklärungen dafür könnten sein, dass die *wenig komplexen* Tests als Zweittest einer komplexen Testbatterie hinzugefügt worden sind, z. B. im Rahmen eines *Cross-battery-assessment* (Renner & Mickley, 2015b). Bei bestimmten Fragestellungen empfehlen die Autoren, neben einem Basistest (z. B. KABC-II) die Anwendung weiterer Subtests aus Testbatterien. Wären z. B. Hinweise zur *fluiden* Intelligenz von Interesse, könnte ein Test der CFT-Reihe zu einer Testung mit einem mehrdimensionalen Test, der u. a. die *fluide* Intelligenz misst, hinzugefügt werden. Eine weitere, ebenfalls weniger plausible Erklärung könnte sein, dass ein Test der CFT-Reihe als Einstieg genutzt worden ist zur Entscheidungsfindung für eine ausführliche Testung. So wurden in Berlin und Brandenburg alle SonderpädagogInnen, die im gemeinsamen Unterricht arbeiten mit den Tests der CFT-Reihe ausgestattet und in der Anwendung geschult¹²². Die mit dem CFT ermittelten Testergebnisse dienen als Grundlage für eine Meldung gegenüber den diagnostischen Teams, die es in diesen beiden Bundesländern seit einigen Jahren gibt (Land Brandenburg, 2013; Senat Berlin, 2012). Die MitarbeiterInnen der diagnostischen Teams ermitteln dann gegebenenfalls fundiertere Ergebnisse aus komplexeren Tests. Diese Vorgehensweise ist jedoch nur in diesen beiden Bundesländern bekannt.

Eine wahrscheinlichere Erklärung dafür, dass die KABC-II trotz der größeren Aussagekraft zwar nicht häufiger, aber auch nicht seltener als die Tests der CFT-Reihe angewendet wird, könnte aus der Zusammensetzung der Stichprobe resultieren. Die Stichprobe besteht aus SonderpädagogInnen, die zur Teilnahme an der Befragung eingeladen worden sind, nachdem sie an einer Schulung zu Intelligenztests teilgenommen haben. In diesen Schulungen wurde überproportional häufig die KABC-II referiert, was die prominente Stellung der KABC-II innerhalb dieser Studie erklären könnte.

Für die Prüfung der Möglichkeit, ob die häufige Anwendung der KABC-II mit einer selektiven Auswahl der Stichprobe zusammenhängt, war der Vergleich der Versuchsgruppe (ProbandInnen, die an Fortbildungen zum Thema teilgenommen haben) mit der Kontrollgruppe interessant (ProbandInnen, die nicht an Fortbildungen zu Intelligenztests teilgenommen haben), da Selektions-

122 Stand: Anfang 2019.

effekte für die Kontrollgruppe nicht anzunehmen sind. Obwohl die mittleren Ränge andeuten, dass auch in der Kontrollgruppe am häufigsten die Tests der CFT-Reihe (jeweils 1. Rang) und die KABC-II (jeweils 2. Rang) angewendet werden, sind die Unterschiede in der Anwendung der KABC-II mit den anderen Kategorien nicht signifikant, und tatsächlich wird die KABC-II signifikant seltener in der Kontrollgruppe als in der Versuchsgruppe genutzt. Auch dieses Ergebnis kann sowohl positiv als auch negativ interpretiert werden: Die KABC-II als mehrdimensionaler und somit aussagekräftiger Test wird weder in der Kontroll- noch in der Versuchsgruppe seltener als vermeintlich einfache Tests durchgeführt. Somit wird mit der Anwendung der KABC-II dessen Aussagekraft durch eine häufige Anwendung respektiert. Andererseits werden aber die wenig aussagekräftigen Tests der CFT-Reihe auch nicht signifikant seltener als die KABC-II angewendet¹²³.

Insgesamt deuten die Ergebnisse eine Präferenz für die Anwendung von vermeintlich einfachen Tests an, sowohl für die Versuchsgruppe, vor allem aber für die Kontrollgruppe. Zwei Ergebnisse weichen von diesem Ergebnis ab:

1. In der Versuchsgruppe gibt es deutliche Hinweise für eine Bevorzugung der KABC-II gegenüber Tests mit weniger Regeln und einer geringeren Aussagekraft. Dies wird mit der selektiven Auswahl der Stichprobe assoziiert, kann aber auch als Beleg für die Nützlichkeit der KABC-II im sonderpädagogischen Kontext interpretiert werden.
2. In der Versuchsgruppe werden signifikant häufiger *sehr komplexe* Tests gegenüber *komplexen* Tests angewendet. Es werden also signifikant häufiger der WISC-IV und der WPPSI-III angewendet als die K-ABC und der WNV, obwohl WISC-IV und WPPSI-III in der Anwendung komplexer sind. Dieses Ergebnis kann damit begründet werden, dass die K-ABC sehr veraltet ist und der WNV wenig verbreitet.

Die dritte Hypothese erfragte Unterschiede zwischen der Verfügbarkeit und der Vorliebe von Intelligenztests. In der Regel befinden sich in den Testschränken, Mediatheken oder Ausleihen für eine Region oder Schule mehrere Tests. Unter der Bedingung, dass wenigstens zwei bestimmte Tests vorhanden sind, ergab die Prüfung der verschiedenen Kombinationsmöglichkeiten, dass die neue KABC-II signifikant häufiger durchgeführt wird als die anderen Tests mit Ausnahme des SON-R 6–40 (keine Signifikanz), während die veraltete K-ABC signifikant seltener angewendet wird als die anderen Tests. Dies ist nicht selbstverständlich, da Bekanntes und Bewährtes evtl. gerne beibehalten werden könnte.

123 Ohne Berücksichtigung der Bonferroni-Korrektur gibt es sogar eine Tendenz zu einer häufigeren Anwendung der CFT-Tests im Vergleich zur KABC-II.

Obwohl der WPPSI-III mehrdimensional ist und auch nicht übermäßig veraltet, scheint dieser Test unbeliebt zu sein. Entweder wird er seltener angewendet als andere Tests (KABC-II; CFT1/CFT1-R; CFT20-R; SON-R 2½-7; SON-R 6-40; IDS) oder es gab keine Signifikanz. Dies bedeutet immerhin, dass es z. B. keinen signifikanten Unterschied gibt in der Häufigkeit der Anwendung zwischen K-ABC und WPPSI-III, obwohl der Flynn-Effekt und die veralteten Stimuli der K-ABC einem fachlichen Vergleich mit dem WPPSI-III nicht standhalten würden.

Relativ häufig wird der SON-R 6-40 durchgeführt, auch wenn andere Tests vorhanden sind. Er wird signifikant häufiger oder tendenziell häufiger durchgeführt als die anderen Tests, bis auf die KABC-II (keine Signifikanz). SON-R 6-40 und KABC-II sind zwar aktuell und teststatistisch solide Verfahren, doch misst der SON-R 6-40 lediglich einen kleinen Teil der Intelligenz und die KABC-II deutlich mehr Intelligenz-Aspekte. Gemessen an der Anzahl von Signifikanzen würden folgende Tests besonders präferiert werden unter der Bedingung, dass wenigstens ein weiterer Test zur Verfügung steht:

1. KABC-II (9/1/0)¹²⁴
2. SON-R 6-40 (7/3/0)
3. SON-R 2½-7 (4/5/1)
4. IDS (4/4/2)
5. CFT1/CFT1-R (3/3/4)
6. SON-R 5½-7 (2/6/2)
7. WISC-IV und CFT20-R (2/5/3)
8. WNV (1/7/2)
9. WPPSI-III (0/2/8)
10. K-ABC (0/1/9)

Aus dieser Aufstellung wird deutlich, dass bei gleichzeitigem Vorhandensein von wenigstens einem weiteren Test KABC-II und SON-R 6-40 bevorzugt angewendet werden und die Tests aus der Wechsler-Reihe, die ansonsten in den psychologischen Beratungsstellen ein hohes Renommee haben, kaum eine Rolle spielen. Daraus kann abgeleitet werden, dass die Tests aus der Wechsler-Reihe im sonderpädagogischen Kontext eine untergeordnete Rolle spielen.

Unter der Bedingung, dass mehrere Tests zur Verfügung stehen,¹²⁵ konnten keine Signifikanzen ermittelt werden. Dies bedeutet u. a., dass der veraltete

124 Die erste Zahl in der Klammer gibt die Anzahl der positiven Signifikanzen wieder (Test wird häufiger durchgeführt als ein anderer), die dritte Zahl die Anzahl negativer Signifikanzen (Test wird seltener als andere durchgeführt); mittlere Zahl in Klammer: Anzahl nicht vorhandener Signifikanzen.

SON-R 5½–17 nicht signifikant seltener als die aktuelle und aussagekräftige KABC-II genutzt worden ist. Aus fachlicher Sicht wäre ein anderes Ergebnis wünschenswerter.

Stehen sowohl die eindimensionalen und die mehrdimensionalen Tests zur Verfügung¹²⁶, wird lediglich der WPPSI-III signifikant seltener angewendet als die KABC-II, SON-R 6–40 und der SON-R 2½–7. Auch für diese Bedingung ist die Betrachtung auf nicht ermittelte Korrelationen interessanter: Aussagekräftige Tests wie die KABC-II oder der WISC-IV werden nicht signifikant häufiger angewendet als die wenig aussagekräftigen Tests der CFT-Reihe und der SON-R 5½–17.

Der SON-R 5½–17 wird nicht einmal signifikant seltener angewendet als die Nachfolgeversion SON-R 6–40, wenn beide Versionen zur Verfügung stehen¹²⁷. Selbst unter Auslassung der konservativen Bonferroni-Korrektur wäre weder eine Signifikanz noch eine Tendenz vorhanden.

Getrennt nach Versuchs- und Kontrollgruppe sind die Unterschiede groß. In der Kontrollgruppe konnten keine Signifikanzen für die K-ABC festgestellt werden, d. h., in jeder geprüften Kombination¹²⁸ wurde die K-ABC nicht signifikant seltener angewendet. Obwohl die aussagekräftige KABC-II vorhanden ist, wird sie nicht gegenüber deutlich weniger aussagekräftigen Tests wie CFT1/ CFT1-R und CFT20-R oder sehr veralteten Tests wie SON-R 5½–17 und K-ABC bevorzugt.

Da die Kontrollgruppe eher die Gesamtheit der SonderpädagogInnen repräsentiert als die Versuchsgruppe, ist zu befürchten, dass die unterschiedliche Nützlichkeit der verschiedenen Verfahren in der sonderpädagogischen Diagnostik nicht genügend honoriert wird, denn die sehr veraltete K-ABC z. B. wird nicht seltener angewendet wie aktuellere Verfahren. Die Nachfolgeversion der K-ABC ist die KABC-II. Sie ist neu normiert, die Stimuli sind aktueller und kindgerechter und aus der Zeit gefallene Items wie bei der K-ABC vorhanden, reduzieren nicht die Aussagekraft. Ein Ergebnis zugunsten der aktuellen KABC-II wäre wünschenswert, ist jedoch im Gegensatz zu der Versuchsgruppe für die Kontrollgruppe nicht erkennbar.

Es ist festzustellen, dass unabhängig von Aussagekraft, Dimensionalität und Komplexität und auch unabhängig vom Alter der Tests und somit dem Einfluss

125 Um die Fallzahlen nicht zu gering zu halten (nur in zwei Fällen standen alle untersuchten 11 Tests zur Verfügung) sind K-ABC, WPPSI-III und WNV in dieser Bedingung ausgeschlossen worden.

126 Zur Wahrung ausreichender Fallzahlen ohne Einbezug der K-ABC.

127 Vergleich unter der Bedingung, dass die neuesten und die ältesten Tests zur Verfügung stehen.

128 Zur Wahrung ausreichender Fallzahlen sind WNV, WPPSI-III und IDS ausgenommen worden.

des Flynn-Effekts auf die Ergebnisse bei veralteten Tests mit dessen teils überholten Stimuli, keine Präferenz zugunsten aktuellerer und nützlicherer Tests ermittelt werden konnte für die Kontrollgruppe. Im Gegenteil: einzig der eindimensionale SON-R 6–40 wurde signifikant häufiger angewendet im Vergleich mit dem aussagekräftigeren WISC-IV.

Die einfache und logisch nachvollziehbare Forderung, dass aktuelle und aussagekräftige Tests genutzt werden sollten, veraltete und wenig aussagekräftige Tests hingegen nicht, findet in den Ergebnissen dieser Studie keine Entsprechung.

6.1.1.2 Bedeutung der Ergebnisse für die Sonderpädagogik

Die ersten drei Hypothesen fragten nach der Anwendung und der Einschätzung der Aussagekraft von Intelligenztests. Komprimiert werden Schlussfolgerungen für die Anwendung von Intelligenztests im sonderpädagogischen Kontext zur Diskussion gestellt.

- Obwohl mehrdimensionale Intelligenztests als aussagekräftiger eingeschätzt werden, resultiert daraus nicht folgerichtig eine höhere Nutzung im Vergleich zu eindimensionalen Tests.
- Die KABC-II erhält hingegen eine herausragende Stellung, obwohl dieser Test am komplexesten ist. Er wird besonders in der Versuchsgruppe bevorzugt genutzt. Die vermehrte Beschäftigung mit der Anwendung von Intelligenztests geht offensichtlich einher mit der richtigen Konsequenz, diesen für die Sonderpädagogik besonders geeigneten Test zu präferieren, aber auch einher mit der Konsequenz, bei eindimensionalen Tests zumindest aktuelle zu verwenden.
- ProbandInnen der Kontrollgruppe, die sich vermeintlich weniger mit der Anwendung von Intelligenztests auseinandergesetzt haben, diese aber dennoch nutzen, unterscheiden weniger nach Aktualität und Aussagekraft der Tests. Ob dies tatsächlich mit einer geringeren Auseinandersetzung mit der Bedeutung von Intelligenztests zusammenhängt, oder ob strukturelle Bedingungen dazu führen (z. B. Zeitnot), bliebe zu diskutieren und gegebenenfalls zu ändern. In der Regel wurde angegeben, dass mehrere Tests zur Verfügung stehen, so dass eine mangelnde Auswahl an Tests nicht die Ursache sein kann.
- Testverfahren aus der Wechsler-Reihe (WISC-IV, WNV, WPPSI-III) spielen in der Sonderpädagogik eine untergeordnete Rolle.
- Es wäre zu überlegen, ob die Anwendung sehr veralteter Tests wie SON-R 5½–17 und K-ABC ausgeschlossen werden sollte, da sie durch aktuelle Versionen bereits vor Jahren ersetzt worden sind. Derzeit gibt es deutliche Hinweise, dass diese Tests noch in der Anwendung sind.

Aus den Ergebnissen dieser Studie resultieren Empfehlungen für die Anwendungspraxis. Im Rahmen einer sonderpädagogischen Begutachtung wäre es wünschenswert, die aktuellsten und aussagekräftigsten Tests anzuwenden, um fundierte Aussagen über das intellektuelle Potential inkl. der Analyse von Stärken und Schwächen vornehmen zu können. Diese Empfehlung könnte unterstrichen werden, indem das Gegenteil verdeutlicht wird: die Anwendung von wenig aussagekräftigen Tests, die kaum Defizite und Ressourcen erkennen können, somit auch nicht geeignet sind, Förderziele oder -maßnahmen abzuleiten.

So ist es nicht möglich, Schwächen in der *auditiven Merkfähigkeit* festzustellen, wenn die eindimensionalen Tests durchgeführt werden, die ausschließlich auf visuellen Stimuli basieren. Hinweise, ob Kinder besser über den auditiven oder den visuellen Kanal Unterrichtsinhalte wahrnehmen können, sind somit nicht möglich. An den Anschaffungskosten kann diese Anwendungspraxis nicht nur liegen, denn der verbreitete eindimensionale SON-R 6–40 kostet mehr als jeder der mehrdimensionalen Tests, wenn auch die Tests der CFT-Reihe günstig sind.

Es gibt ein deutliches Bewusstsein darüber, welche Tests zu Recht als aussagekräftig eingeschätzt werden. Dieses Bewusstsein erhält leider keine Entsprechung in der Anwendungspraxis. Die häufigere Anwendung mehrdimensionaler Tests wird empfohlen, die Bedingungen zur Umsetzung dieser Praxis, die den Kindern gerechter wird, soll an späterer Stelle konkretisiert werden. Zusammengefasst kann festgehalten werden: bestünde der Weg zu einer Handlungsänderung in den zwei Schritten, zunächst ein Bewusstsein für die Sinnhaftigkeit der Änderung zu entwickeln und im zweiten Schritt die Umsetzung der Änderung, so kann der erste Schritt als vorliegend attestiert werden. Darüber hinaus wird empfohlen, die K-ABC und den SON-R 5½–17 weder formell noch informell zu nutzen.

6.1.2 Vergleiche zwischen den Bundesländern

Die vierte Forschungsfrage erwartete Unterschiede in der Häufigkeit von Durchführungsfehlern und Beeinträchtigungen in der Testsituation zwischen den Bundesländern und wurde mit Hilfe von mehreren Hypothesen geprüft:

- Hypothese 4.1: Abhängig vom Bundesland stehen unterschiedliche Tests zur Verfügung.
- Hypothese 4.2: Abhängig vom Bundesland werden unterschiedlich die Durchführungsobjektivität gefährdende Veränderungen vorgenommen.
- Hypothese 4.3: Abhängig vom Bundesland liegen unterschiedliche Beeinträchtigungen wie fehlende oder unvollständige Testmaterialien vor.
- Hypothese 4.4: Abhängig vom Bundesland liegen unterschiedliche Freiheiten vor zu entscheiden, ob ein Intelligenztest durchgeführt werden soll.

- Hypothese 4.5: Abhängig vom Bundesland liegen unterschiedliche Schwierigkeiten im Umgang mit Durchführungsregeln vor.
- Hypothese 4.6: Abhängig vom Bundesland wird die Anwendung von Intelligenztests als schwierig bewertet.
- Hypothese 4.7: Abhängig vom Bundesland wird die zur Verfügung stehende Zeit für die Anwendung als zu kurz bewertet.

6.1.2.1 Interpretation der Ergebnisse zu den Bundesländervergleichen

Die Anwendung von Intelligenztests im Bundesländervergleich bestätigen die Annahme, dass Kinder abhängig vom jeweiligen Bundesland unter nicht einheitlichen Bedingungen getestet werden. Für alle elf primär untersuchten Tests konnten signifikante Unterschiede bezüglich der Verfügbarkeit gefunden werden, bei acht der Tests war $p < .001$. Es stehen somit abhängig vom Bundesland unterschiedliche Tests zur Verfügung. Diese Unterschiede allein erhalten jedoch erst eine Relevanz, wenn veraltete Tests zur Verfügung stehen, daneben aber keine aktuellen. Ansonsten hängt es vom Bundesland ab, ob ein Kind im Rahmen der Feststellung sonderpädagogischen Unterstützungsbedarfs angemessen oder unangemessen mit Intelligenztests überprüft werden würde. Bundesländer, die nach eigener Einschätzung der SonderpädagogInnen (siehe Hypothese 1) die aussagekräftige KABC-II anwenden (die zudem Stärken/Schwächen Analysen zulässt und somit Hinweise auf pädagogische Maßnahmen anbietet), würden eher Kindern mit Unterstützungsbedarf gerecht werden als SonderpädagogInnen aus Bundesländern, die häufiger eindimensionale Tests ohne die Möglichkeit von Ableitungen durchführen und zudem evtl. noch auf Grund des Flynn-Effekts veraltete Normtabellen und antiquierte Grafiken nutzen.

Ein exemplarischer Vergleich zwischen KABC-II (aussagekräftig, aktuell, mehrdimensional) und SON-R 5½–17 (veraltete ausländische Normen, überholte Stimuli) verdeutlicht anschaulich die Unterschiede. In Nordrhein-Westfalen z. B. ist der SON-R 5½–17 signifikant häufiger als im Durchschnitt der sieben untersuchten Bundesländer¹²⁹ vorhanden, die KABC-II jedoch signifikant seltener. In Schleswig-Holstein können 94 Prozent (signifikant mehr) der ProbandInnen auf die KABC-II zugreifen, etwas südlicher in Hamburg 38 Prozent (signifikant weniger).

Folgende Hinweise zeichnen sich für die untersuchten Bundesländer¹³⁰ ab:

129 Es wurden Bundesländer mit akzeptablen Fallzahlen untersucht: Baden-Württemberg (N = 130), Hamburg (N = 29), Hessen (N = 109), Niedersachsen (N = 143), Nordrhein-Westfalen (N = 465), Rheinland-Pfalz (N = 51) und Schleswig-Holstein (N = 31).

130 Für Hamburg, Rheinland-Pfalz und Schleswig-Holstein soll auf eine Bewertung der Ergebnisse auf Grund der niedrigeren Fallzahlen verzichtet werden.

- Positiv sticht Hessen hervor: kein Test ist signifikant seltener vorhanden, aber immerhin sieben Tests signifikant häufiger, darunter die mehrdimensionalen Tests KABC-II und WISC-IV. Durch die große Auswahl an Tests besteht die Möglichkeit, entsprechend der Fragestellung und den Besonderheiten eines Kinds den passgenauen Test zu wählen.
- Negativ sticht Nordrhein-Westfalen hervor: Lediglich der SON-R 5½–17 und der SON-R 6–40 sind häufiger vorhanden, sieben weitere Tests signifikant seltener. Dies bedeutet, dass SonderpädagogInnen in NRW seltener auf aussagekräftige Tests zugreifen können.
- Ebenfalls negativ sticht Niedersachsen hervor: bis auf den CFT20-R sind Tests seltener vorhanden (fünf Tests) oder es liegt keine Signifikanz vor (fünf Tests).
- In Baden-Württemberg sind die aussagekräftigeren Tests KABC-II, WPPSI-III, SON-R 2½–7 und IDS signifikant häufiger vorhanden. Hier fällt auf, dass weniger Wert auf die Tests der CFT-Reihe und auf den SON-R 6–40 gelegt wird, die jeweils seltener vorhanden sind.

Sollten aussagekräftigere Tests weniger vorhanden sein, muss dies noch nicht besorgen, denn es muss nicht einhergehen mit der häufigeren Anwendung der am häufigsten zur Verfügung stehenden Tests. Es ist übliche Praxis, veraltete Tests im Testschrank zu belassen¹³¹. Deshalb kann dieser Befund besser im Kontext der ersten drei Hypothesen interpretiert werden.

Diese ermittelten eine Tendenz, eher einfach durchzuführende Tests anzuwenden, die allerdings auch weniger aussagekräftig sind. Diese Tendenz, verbunden mit Befunden, die Hinweise geben auf eine geringere Verfügbarkeit aussagekräftigerer Tests, bzw. eine bessere Verfügbarkeit eindimensionaler Tests, könnte den Effekt kumulieren, weniger aussagekräftige Tests anzuwenden. Dieser Effekt ist zu befürchten in Nordrhein-Westfalen und Niedersachsen, aber weniger zu befürchten in Hessen.

Hypothese 4.2 fragte nach den die Durchführungsobjektivität gefährdenden Veränderungen. Abhängig vom Bundesland gibt es signifikante Unterschiede bezüglich Veränderungen der Durchführungszeiten. Bevor die genaueren Befunde bewertet werden, soll erwähnt sein, dass zur Beantwortung dieser Frage bei den ProbandInnen ein Bewusstsein über die nicht korrekte Anwendung der Durchführungsregeln unterstellt werden kann. Jede Antwort auf die Frage, ob Durchführungszeiten verändert bzw. weggelassen werden, die nicht mit *nie* beantwortet worden ist, kann als Ergebnis an sich bewertet werden. In Ausnahmefällen gestatten manche Testmanuale eine Abweichung der Durchführungszeiten, sofern diese begründet werden kann und im Testbericht und in der

131 Vermutlich aus Scheu davor, ehemals sehr teure Tests einfach wegzuworfen.

Interpretation entsprechend gewürdigt wird¹³² (z.B. bei Kindern mit körperlich-motorischen Beeinträchtigungen). Im günstigsten Fall könnte angenommen werden, dass die bewusste Abweichung vom standardisierten Vorgehen wenigstens im Testbericht entsprechend gewürdigt wird und das Vertrauens- bzw. Konfidenzintervall vorsichtiger bestimmt wird (z.B. das 90 statt dem 95 prozentigen Vertrauens-/Konfidenzintervall).

Doch ein Blick auf die deskriptivstatistische Auswertung der entsprechenden Fragen (Abbildung 14) gibt einen Hinweis, dass nur ca. vier von fünf Probandinnen niemals die Durchführungszeiten weggelassen und ca. drei von fünf ProbandInnen niemals die Durchführungszeiten verändert haben. Unabhängig also von Unterschieden diesbezüglich zwischen den Bundesländern wäre zumindest eine sorgfältigere Würdigung des standardisierten Vorgehens wünschenswert. Erwähnt sei in diesem Zusammenhang auch, dass es zwar keinen signifikanten Unterschied zwischen den Bundesländern bezüglich der Frage nach einem unerlaubt gegebenen Feedback während der Testung gegenüber dem Kind gibt, jedoch sieben von zehn ProbandInnen diese Abweichung vorgenommen haben, sogar 5 Prozent *oft*. Die Zahlen dürften jedoch höher sein, da auch unbemerkte nonverbale Rückmeldungen (Mimik, Gestik, Reaktionen) zu den Feedbacks gehören¹³³.

Die Hypothese 4.3 gibt einen Hinweis auf unterschiedliche Beeinträchtigungen während der Testanwendung in den Bundesländern. Die Unterschiede zwischen den Bundesländern mit ausreichend hohen Fallzahlen (Baden-Württemberg, Hamburg, Hessen, Niedersachsen, Nordrhein-Westfalen, Rheinland-Pfalz, Schleswig-Holstein) sind signifikant (jeweils $p < .001$). Es kommt unterschiedlich häufig vor, dass Intelligenztests nicht zur Verfügung stehen, die Testmaterialien unvollständig sind oder Formulare bzw. Arbeitsbögen fehlen. Um aus diesen Ergebnissen Rückschlüsse für die Gestaltung der Arbeitsbedingungen bezüglich der Anwendung von Intelligenztests vornehmen zu können, sollen die Unterschiede skizziert und mit den unterschiedlichen Arbeitsbedingungen in den Bundesländern in Verbindung gebracht werden.

Es fällt auf, dass die höchsten Beeinträchtigungen in Schleswig-Holstein vorhanden sind, die niedrigsten in Hamburg und Baden-Württemberg. Im Ver-

132 Selbst in diesem Fall wäre jedoch die Objektivität verletzt, da die Abweichung vom standardisierten Vorgehen ein Vergleich zwischen Testergebnis und Normstichprobe zweifelhaft werden ließe. Zumindest im Falle eines moderaten Umgangs mit den vorgeschriebenen Durchführungszeiten könnten die Vertrauens- bzw. Konfidenzintervalle ebenfalls niedriger bestimmt werden.

133 Es könnte eingewendet werden, dass diese unbemerkten Feedbacks über die nonverbale Kommunikation auch in den Testungen der Normstichprobe vorkommen könnten und somit ausgeglichen werden, doch ist zu erwarten, dass die TesterInnen der Normstichprobe besonders geschult sind und die Bedeutung von nonverbalen Feedbacks entsprechend honoriert haben.

gleich zwischen zwei Bundesländern fällt auf, dass in sieben Vergleichen Beeinträchtigungen in Niedersachsen höher sind, in fünf Fällen in Nordrhein-Westfalen und in vier Fällen in Rheinland-Pfalz und Schleswig-Holstein. Keine signifikanten Beeinträchtigungen konnten im Vergleich mit anderen Bundesländern für Hamburg und Baden-Württemberg ermittelt werden.

Zusammengefasst zeichnet sich die Tendenz ab, dass Beeinträchtigungen in Schleswig-Holstein, Niedersachsen, Nordrhein-Westfalen und Rheinland-Pfalz eher höher, in Baden-Württemberg und Hamburg eher geringer vorhanden sind. Hessische SonderpädagogInnen gaben zumindest an, dass häufiger Testformulare fehlen.

Im Rahmen der Umsetzung der Inklusion sind Förderschulen aufgelöst worden, die in anderen Regelschulen arbeitenden SonderpädagogInnen haben einen schlechteren Zugang zu Testverfahren. Dies könnte ein Erklärungsansatz für die jeweils höchsten Beeinträchtigungen in Schleswig-Holstein sein. Die Umsetzung der Inklusion mit der damit verbundenen Auflösung von Förderschulen und der damit verbundenen Auflösung von Testschranken in den eigenen Räumen liegt allerdings auch in Hamburg vor.

Die geringsten Beeinträchtigungen in Hamburg könnten damit erklärt werden, dass eine Spezialisierung bei der Anwendung von Intelligenztests durch die ReBBZ (Regionale Bildungs- und Beratungszentren) vorliegt. Es könnte auch sein, dass die kürzeren Wege eines Stadtstaates dazu führen, schneller an die Tests zu kommen. Dem gegenüber stünde allerdings, dass in Baden-Württemberg ebenfalls weniger Beeinträchtigungen vorhanden sind. Allerdings ist die Inklusionsquote in Baden-Württemberg mit 34 Prozent deutlich niedriger als in Hamburg mit 63 Prozent¹³⁴ (Lange, 2017). Daraus resultiert eine höhere Dichte an Förderschulen bzw. Sonderschulen¹³⁵ mit der besseren Verfügbarkeit von Testverfahren. Zusammengefasst gibt es Hinweise, dass eine Spezialisierung auf die Anwendung von Intelligenztests und die bessere Verfügbarkeit zu weniger Beeinträchtigungen bei der Anwendung von Intelligenztests führen.

In Niedersachsen können SonderpädagogInnen sowohl in der Versuchs- als auch in der Kontrollgruppe signifikant häufiger selbst entscheiden, ob sie Intelligenztests anwenden oder nicht, diese Freiheit könnte mit zu den vermehrt auftretenden Schwierigkeiten niedersächsischer Lehrkräfte führen. Es ist anzunehmen, dass Tests seltener angewendet werden, wenn dessen Anwendung selbst beschlossen wird und nicht zum institutionalisierten Ablauf gehört. Aus einer selteneren Anwendung würde dementsprechend weniger Erfahrungswissen und Routine entwickelt werden können.

134 Schuljahr 2015/16.

135 „Förderschule“ in Baden-Württemberg ist die Bezeichnung für Sonderschulen für Kinder mit dem Unterstützungsbedarf *Lernen*.

Hypothese 4.5 erfragte Problematiken im Umgang mit den Umkehr- und Abbruchregeln und beim Ausrechnen des Testalters am Testtag. Im Vergleich zwischen allen Bundesländern konnten keine Unterschiede festgestellt werden, auch wenn Problematiken beschrieben worden sind (siehe Tabelle 19). Im Vergleich zwischen den Bundesländern mit einer ausreichend hohen Fallzahl ergaben sich keine Signifikanzen, denen eine Bedeutung beigemessen wird¹³⁶.

Der Schwierigkeiten-Index setzt sich aus mehreren Items zusammen. Hypothese 4.6 fragte nach Unterschieden zwischen empfundenen Schwierigkeiten bei der Anwendung von Intelligenztests und dem Bundesland. Im Vergleich bewerten SonderpädagogInnen aus Niedersachsen und Schleswig-Holstein die Anwendung von Intelligenztests als am schwierigsten, SonderpädagogInnen aus Baden-Württemberg und Hessen als am wenigsten schwierig, doch insgesamt liegen die Mittelwerte dicht beieinander. Einen signifikanten Unterschied gibt es zwischen Baden-Württemberg (weniger Schwierigkeiten) und Niedersachsen (mehr Schwierigkeiten). Interessant im Zusammenhang dieser Hypothese ist der Vergleich zwischen der Versuchs- und Kontrollgruppe. SonderpädagogInnen, die noch niemals an einer Fortbildung zu Intelligenztests teilgenommen haben, empfinden deren Anwendung als tendenziell schwieriger, jedoch gibt es im Vergleich zwischen den Bundesländern mit einer ausreichenden Fallzahl keine signifikanten Ergebnisse. Die bis an diese Stelle diskutierten Ergebnisse deuten an, dass die Anwendung von Intelligenztests in Baden-Württemberg mit weniger und in Niedersachsen mit mehr Problematiken und Schwierigkeiten verbunden ist. Im Zusammenhang mit der Hypothese 4.4 (kann selbst entscheiden, einen Intelligenztest durchzuführen) könnte eine Schlussfolgerung lauten, dass die größere Freiheit, selbst zu entscheiden, Intelligenztests anzuwenden, mit größeren Schwierigkeiten verbunden ist.

Die Hypothese 4.7 fragte nach Unterschieden zwischen den Bundesländern und der zur Verfügung stehenden Zeit für die Anwendung von Intelligenztests. Dazu sollten Aussagen zum Zeitmanagement getroffen werden. Es gehört zu den wichtigen Bedingungen bei der Anwendung der Tests, diese ohne Zeitdruck durchführen und mit entsprechender Vorbereitungszeit lernen zu können. Bevor aus den Ergebnissen Hinweise zu den diesbezüglichen Arbeitsbedingungen in den verschiedenen Bundesländern getroffen werden können, soll ein Blick auf die deskriptivstatistische Auswertung verdeutlichen, dass das Zeitmanagement als ein tatsächliches Problem eingeschätzt wird. Die Frage, ob die Tests in der Freizeit vorbereitet werden, wurde auf einer fünfstufigen Skala von völlig richtig bis völlig falsch deutlich bejaht mit einem Mittelwert von 1,41 ($SD = 0.73$). Auch den anderen Fragen bezüglich der zur Verfügung stehenden Zeit

136 Lediglich hessische SonderpädagogInnen hatten signifikant mehr Schwierigkeiten beim Ausrechnen des Testalters im Vergleich zu rheinland-pfälzischen SonderpädagogInnen.

wurde zugestimmt. Es wird zu wenig Vorbereitungszeit während der Arbeitszeit ($M = 2.18$; $SD = 1.12$), zu wenig Zeit für die Durchführung eines sonderpädagogischen Gutachtens ($M = 2.38$; $SD = 1.21$) und zu wenig Zeit für die Anwendung eines Intelligenztests ($M = 2.71$; $SD = 1.21$) attestiert. Es muss festgestellt werden, dass in diesem Zusammenhang die Arbeitsbedingungen als ungünstig von den ProbandInnen eingeschätzt werden.

Bestehende Unterschiede zwischen den Bundesländern können Hinweise darauf geben, wo diese ungünstigen Arbeitsbedingungen besonders stark ausgeprägt sind und ob in diesen Bundesländern weitere Schwierigkeiten und Problematiken vermehrt auftreten. Tatsächlich kann dies für Niedersachsen und Nordrhein-Westfalen festgestellt werden. Zunächst sei festgehalten, dass sowohl unter Einbezug der Gesamtstichprobe als auch unter Einbezug der Bundesländer mit ausreichend hohen Fallzahlen signifikante Unterschiede zwischen den Bundesländern vorliegen.

Während in Hessen die mittleren Ränge hoch sind, gleichbedeutend mit weniger beschriebenen Problematiken, sind in Hamburg die mittleren Ränge in drei von vier Items niedrig, in Niedersachsen und Nordrhein-Westfalen in zwei von vier Items zu Zeitproblematiken niedrig (= mehr Schwierigkeiten). Im Vergleich zwischen zwei Bundesländern liegen elf signifikante Ergebnisse vor: in fünf Vergleichen lagen signifikant höhere Beeinträchtigungen bei niedersächsischen, in vier Fällen bei nordrhein-westfälischen SonderpädagogInnen vor. In der Kontrollgruppe liegt ein signifikanter Unterschied zu der Frage vor, ob *heutzutage weniger Zeit für die Durchführung sonderpädagogischer Gutachten* vorhanden ist. Hier gaben niedersächsische SonderpädagogInnen eine höhere Belastung im Vergleich zu denen aus Rheinland-Pfalz an.

Der Vollständigkeit halber sei erwähnt, dass niedersächsische SonderpädagogInnen zumindest der Frage nach zu wenig zur Verfügung stehenden Vorbereitungszeit zum Lernen eines Tests im Vergleich mit den anderen Bundesländern am zweitwenigsten zustimmten.

6.1.2.2 Bedeutung der Ergebnisse für die Sonderpädagogik

- Intelligenztests werden unterschiedlich in den Bundesländern angewendet. Es kann vom Bundesland abhängen, mit welchem Test und ob unter günstigeren oder ungünstigeren Bedingungen die Kinder auf Intelligenz getestet werden.
- Für Nordrhein-Westfalen und Niedersachsen gibt es Hinweise, dass weniger aussagekräftige Tests verwendet werden, teils gar veraltete. Für diese beiden Bundesländer gibt es zudem Hinweise, dass die Anwendung von Intelligenztests mit mehr Schwierigkeiten und Problematiken verbunden ist.
- Für Hessen und Baden-Württemberg gibt es Hinweise, dass nicht nur aussagekräftigere Tests angewendet werden, sondern die Auswahl zwischen

Tests größer ist. So besteht eher die Möglichkeit, den passenden Test für die unterschiedlichen Fragestellungen zu wählen. Für Baden-Württemberg gibt es zudem Hinweise, dass weniger Schwierigkeiten und Problematiken vorhanden sind.

- Es gibt Hinweise, dass eine Spezialisierung bei der Anwendung von Intelligenztests zu weniger Beeinträchtigungen während der Testsituation führen wie das Fehlen von Materialien oder unvollständige Materialien. Dieser Effekt liegt in Hamburg vor.
- Es gibt Hinweise, dass eine höhere Dichte an Förder- bzw. Sonderschulen wie in Baden-Württemberg mit der einhergehenden besseren Verfügbarkeit von Testverfahren zu weniger Beeinträchtigungen führt.
- Es gibt Hinweise, dass die Freiheit, darüber selbst zu entscheiden, Intelligenztests anzuwenden oder nicht, zu mehr Schwierigkeiten in der Anwendung führen.

Eine bundesweit einheitliche Regelung, wie mit Intelligenztests im sonderpädagogischen Kontext umgegangen wird, ist nicht zu erwarten. Es widerspräche der Kulturhoheit der Länder, wäre aber auch gar nicht notwendig. Es wäre bereits hilfreich, wenn es orientierungsgebende Regelungen gäbe, die im besten Fall für jedes Bundesland verbindlich wären. Auch bei vorliegender Kulturhoheit der Länder gibt es die Möglichkeit, über die Kultusministerkonferenzen (KMK) Standards festzulegen und zu koordinieren.

Derzeit hängt es vom Zufall, respektive vom Bundesland ab, ob und wie ein Kind im Rahmen einer Gutachtenerstellung getestet wird und wie die Rahmenbedingungen der SonderpädagogInnen gestaltet sind, in denen die Tests angewendet werden. Sowohl bei einem Umzug eines Kinds als auch beim Wechsel einer SonderpädagogIn in ein anderes Bundesland können die Bedingungen sehr unterschiedlich sein.

Standards zur Anwendung von Intelligenztests könnten beinhalten, Tests von der Anwendung auszuschließen (Tests mit veralteten Normierungen und Stimuli wie die K-ABC), die Nutzung eines Tests, für den es eine Folgeversion gibt, auf zwei Jahre zu begrenzen, Vorschläge für angemessene Tests entsprechend der Unterstützungsbedarfe zu erstellen oder eine Auswahl von verfügbaren Tests vorzuschreiben, damit Tests entsprechend den Bedürfnissen der Kinder gewählt werden können. Es wäre zudem hilfreich, wenn der Zeitrahmen zur Verfassung der Gutachten auf mehrere Monate ausgedehnt wird, damit die Tests für eine Probe- und Lernphase ausreichend lange ausgeliehen werden können. Die gleichzeitige Anfertigung der Gutachten innerhalb weniger Wochen in einer Region könnte dazu führen, dass viele SonderpädagogInnen die Tests nur kurz leihen und dementsprechend nur unter Zeitdruck anwenden können.

6.1.3 Alter und Geschlecht

Aus der Forschungsfrage 5 zum Alter und 6 zum Geschlecht resultieren folgende Hypothesen:

- Hypothese 5.1: Mit zunehmendem Alter der TesterInnen werden weniger Schwierigkeiten bei der Anwendung von Intelligenztests erwartet.
- Hypothese 5.2: Mit zunehmendem Alter der TesterInnen werden seltener aktuelle Tests angewendet.
- Hypothese 6: Es wird erwartet, dass sich das Geschlecht nicht auf Schwierigkeiten bei der Anwendung von Intelligenztests auswirkt.

6.1.3.1 Problematiken im Zusammenhang mit Alter und Geschlecht

Hintergrund der Prüfung des Zusammenhangs zwischen Alter und empfundenen Schwierigkeiten ist der Gedanke, dass weniger beschriebene Schwierigkeiten im Umgang mit Intelligenztests durch ältere TestanwenderInnen zu der Schlussfolgerung führen könnten, dass Erfahrungswissen zu weniger Problematiken führt. Infolgedessen wäre dann zu überlegen, wie dieses Erfahrungswissen forciert werden könnte, z. B. durch eine Spezialisierung der SonderpädagogInnen (wenige SonderpädagogInnen führen häufig Tests durch, nicht viele SonderpädagogInnen selten). Tatsächlich ist ein entsprechender Effekt zu beobachten. Das Ergebnis ist jedoch nur schwach ausgeprägt und lediglich als Tendenz erkennbar.¹³⁷ In der Kontrollgruppe sind keine signifikanten Zusammenhänge feststellbar.

Auch bei der Prüfung des Zusammenhangs zwischen Alter und Testanwendung müssen die Ergebnisse vorsichtig interpretiert werden. Die Befürchtung, dass ältere TestanwenderInnen bevorzugt veraltete und somit zwar vertraute, aber nicht mehr angemessene Tests anwenden, hat sich zwar bestätigt, doch führen ältere TestanwenderInnen generell häufiger Tests durch als jüngere, auch aktuelle und aussagekräftige Tests. Dies gilt nicht für die KABC-II, aber signifikant für den WISC-IV und WPPSI-III. Ein interessantes Nebenergebnis ist, dass ältere ProbandInnen häufiger die Wechsler-Tests anwenden, obwohl diese eine untergeordnete Rolle in der Sonderpädagogik zu spielen scheinen.

Es wurde angenommen, dass es keinen Unterschied zwischen Geschlecht und empfundenen Schwierigkeiten bei der Anwendung der Tests gibt. Auf

¹³⁷ Es sei an dieser Stelle erwähnt, dass grundsätzlich zweigerichtet und konservativ-streng geprüft wurde zur Vermeidung von Typ-2 Fehlern. Eine gerichtete Prüfung wäre entsprechend der Hypothesen-Formulierung legitim und würde zu einem signifikanten Ergebnis führen.

zweifelhafte Begriffe wie *sehr signifikant* oder *höchst signifikant* soll im Rahmen dieser Arbeit verzichtet werden, dennoch kann bei Betrachtung des p-Werts von $< .001$ entgegen der Annahme der Alternativhypothese deutlich formuliert werden, dass Männer weniger Schwierigkeiten bei der Anwendung von Intelligenztests beschreiben. Für die Kontrollgruppe werden die Schwierigkeiten tendenziell von Männern geringer beschrieben.

Es kann ein Unterschied sein, ob weniger Schwierigkeiten vorhanden sind oder weniger Schwierigkeiten empfunden werden. In diesem Zusammenhang muss offenbleiben, ob Männer ein geringeres Empfinden gegenüber tatsächlichen Schwierigkeiten haben oder ob bei Männern tatsächlich weniger Schwierigkeiten vorliegen.

6.1.3.2 Bedeutung der Ergebnisse für die Sonderpädagogik

Es gibt moderate Hinweise, dass Erfahrungswissen zu weniger Schwierigkeiten bei der Anwendung von Intelligenztests führt. Dies bekräftigt die Überlegung nach einer Spezialisierung bei der Anwendung von Intelligenztests.

Eine Spezialisierung weniger SonderpädagogInnen könnte dazu führen, dass „Diagnostikansprechpartner“ (Müller, 2009, S. 182) in Fragen der Diagnostik besonders geschult sind und komplexere Tests durchführen. Es wäre z. B. möglich, dass nicht mehr alle innerhalb eines Kollegiums ein Gutachten schreiben und dementsprechend selten testen, sondern dass von der Erstellung eines Gutachtens wenige Lehrkräfte ausgenommen sind. Da diese keine Gutachten mehr schreiben müssen und somit Arbeitszeit gewonnen haben, wird diese für die Anwendung der aufwändigeren Testverfahren genutzt, für deren häufigere Anwendung Routine entwickelt werden könnte. Die gewonnenen Testergebnisse würden in Form von Textbausteinen dem Gutachten beigelegt.

6.1.4 Ausbildung

Drei Hypothesen prüften Zusammenhänge zwischen Bildungserfahrungen und dem Ausmaß an erlebten Schwierigkeiten:

- Hypothese 7.1: Das Ausmaß an Schwierigkeiten bei der Anwendung von Intelligenztests hängt vom Ausmaß der in der universitären Ausbildung besuchten Seminare zur Testdiagnostik ab.
- Hypothese 7.2: Das Ausmaß an Schwierigkeiten bei der Anwendung von Intelligenztests hängt vom Ausmaß der in der universitären Ausbildung referierten Inhalte zur Testdiagnostik ab.

- Hypothese 8: Es wird angenommen, dass TeilnehmerInnen an einer außer-universitären Fortbildung zur Testdiagnostik weniger Schwierigkeiten bei der Anwendung von Testverfahren beschreiben.

6.1.4.1 Auswirkungen der Ausbildung auf Problematiken

Der vermutete Zusammenhang ist eindeutig bei der Frage, ob den ProbandInnen die Anwendung von Intelligenztests leichtfällt: je mehr Seminare bzw. Vorlesungen zum Thema besucht worden sind, desto leichter wird die Anwendung der Tests eingeschätzt, sowohl in der Gesamt-, Kontroll- und Versuchsgruppe.

Keine signifikanten Zusammenhänge können zwischen der Anzahl belegter Seminare bzw. Vorlesungen zum Thema und die Durchführungsobjektivität gefährdenden Veränderungen wie dem Verändern der Durchführungszeiten oder dem unerlaubten Geben von Feedbacks gefunden werden. Es ist jedoch fraglich, ob die Bedeutung des standardisierten Vorgehens während einer Testanwendung explicit an der Universität referiert, oder ob die Bedeutung erst in der Praxis bewusst wurde. Anders verhält es sich mit den grundlegenden Anwendungsregeln. Je mehr Seminare bzw. Vorlesungen besucht worden sind, desto signifikant weniger Schwierigkeiten werden bei der Anwendung der Umkehr- und Abbruchregeln beschrieben, aber auch weniger Schwierigkeiten bei dem Ausrechnen des Alters am Testtag.

Auch in der Kontrollgruppe werden tendenziell mehr Schwierigkeiten bei der Anwendung der Umkehrregel bei weniger belegten Seminaren beschrieben.

Das ungewöhnlichste Ergebnis dieser Arbeit soll nicht verschwiegen werden: in der Kontrollgruppe werden weniger unerlaubte Veränderungen bei den Durchführungszeiten vorgenommen, je weniger Seminare besucht worden sind. Dieses Ergebnis wird allerdings als Zufallsbefund gewertet.

In der Hypothese 7.2 wird gezielter nach Unterschieden zwischen in der Ausbildung referierten Inhalten und beschriebenen Problematiken bei der Anwendung der Tests gefragt. Wurden die fünf Konstrukte *Standardabweichung*, *Durchführungsobjektivität*, *Vertrauens-/Konfidenzintervall*, *Messungenauigkeit* und die *Gaußsche Kurve der Normalverteilung* an der Universität referiert, werden jeweils weniger Schwierigkeiten bei der Anwendung der Tests beschrieben. Diese Unterschiede sind signifikant bei den Konstrukten *Standardabweichung*, *Durchführungsobjektivität* und *Vertrauens-/Konfidenzintervall*, tendenziell signifikant bei dem Konstrukt *Messungenauigkeit/Messfehler*. Bei einer Gruppierung der fünf Konstrukte zu einer gemeinsamen Variablen sind die Ergebnisse ebenfalls eindeutig. Signifikant weniger Schwierigkeiten werden zudem beschrieben, wenn an der Universität Intelligenztests ausprobiert worden sind.

Der *Schwierigkeiten-Index* besteht auch aus Items, die nicht unmittelbar mit der universitären Ausbildung in Verbindung stehen (z.B. Störungen während

der Testsituation). Auch wenn es dennoch eindeutige Belege gibt, dass im Rahmen der universitären Ausbildung referierte Konstrukte zur Anwendung von normierten Testverfahren zu weniger Schwierigkeiten in der praktischen Anwendung führen, hat eine detailliertere Prüfung des Zusammenhangs zwischen eindeutig universitären Inhalten und Problematiken bei der Anwendung der Tests ebenfalls deutliche Ergebnisse ergeben.

Alle signifikanten Unterschiede ermitteln ausnahmslos weniger Problematiken, wenn Inhalte zum Thema an der Universität gelehrt worden sind. Wurde z. B. die Bedeutung der *Durchführungsobjektivität* nicht referiert, haben die ProbandInnen signifikant mehr Schwierigkeiten im Umgang mit der *Umkehr- und Abbruchregel*, dem *Ausrechnen des Alters*, und die Anwendung von Tests fällt ihnen tendenziell schwerer. Ähnliches gilt, wenn die Bedeutung des *Vertrauens- bzw. Konfidenzintervalls* nicht gelehrt wurde, zudem fällt den ProbandInnen die Anwendung der Tests leichter, wenn Intelligenztests an der *Uni ausprobiert* worden sind.

Es fällt auf, dass teilweise signifikant weniger Problematiken beschrieben werden, wenn Inhalte referiert worden sind, die auf den ersten Blick inhaltlich mit einer konkreten Problematik wenig zu tun haben. Wurde z. B. die Bedeutung des *Konfidenzintervalls* gelehrt, wurden signifikant seltener unerlaubte Feedbacks während der Testsituation gegeben. In diesen Fällen könnte es sich einerseits um Zufallsbefunde handeln, es könnte aber auch angenommen werden, dass eine generelle Beschäftigung mit der Thematik im Rahmen der universitären Ausbildung mit einem Anwenden der Tests nach den Regeln der Kunst einhergeht.

Es ist möglich, dass die Befunde durch die Auswahl der Stichprobe eingeschränkt werden. Die meisten ProbandInnen haben an einer Fortbildung zur Testdiagnostik teilgenommen. Für diese Gruppe konnten umfangreiche Signifikanzen festgestellt werden, für die Kontrollgruppe – ProbandInnen, die noch nie an einer Fortbildung zu standardisierten Verfahren teilgenommen hatten – konnten hingegen keine Signifikanzen festgestellt werden, in der Kontrollgruppe sind lediglich zwei Tendenzen zu weniger Schwierigkeiten belegt.

Es ist denkbar, dass bei der Gesamtstichprobe die geringer beschriebenen Problematiken bei der Anwendung der Tests auf die Inhalte der Fortbildung zum Thema zurückzuführen sind, aber nicht mehr genau differenziert werden konnte, ob diese Inhalte an der Universität oder in der Fortbildung gelehrt worden sind.

Möglich ist allerdings auch, dass bedingt durch die hohe Fallzahl der Gesamtstichprobe Signifikanzen im Gegensatz zu den deutlich geringeren Fallzahlen bei der Kontrollgruppe genauer erkannt werden konnten.

Die Hypothese 8 fragte nach einem Zusammenhang zwischen der Teilnahme an einer außeruniversitären Fortbildung zur Testdiagnostik und Schwierigkeiten bei der Anwendung von Intelligenztests. Ergebnisoffen wurde auch hier

zweiseitig geprüft mit dem Ergebnis, dass ProbandInnen, die an einer Fortbildung zum Thema teilnahmen, tendenziell weniger Schwierigkeiten beschrieben.

Mögliche Gründe für die geringe Ausprägung könnten neben der evtl. geringen Wirksamkeit der Fortbildungen oder der geringe Umfang der Inhalte im Rahmen einer ein- bzw. eineinhalbtägigen Veranstaltung auch damit erklärt werden, dass besonders verunsicherte und hilfeschuchende SonderpädagogInnen an einer entsprechenden Fortbildung teilnahmen.

6.1.4.2 Bedeutung der Ergebnisse für die Sonderpädagogik

- Es gibt eindeutige Hinweise, dass die Anwendung von Intelligenztests leichter fällt, je umfangreicher die universitäre Ausbildung war.
- Es gibt ebenfalls eindeutige Hinweise für weniger Schwierigkeiten bei der Anwendung von Durchführungsregeln, wenn die universitäre Ausbildung umfangreicher war.
- Es gibt zumindest für die Gesamtgruppe eindeutige Hinweise, dass die Behandlung an der Universität von zur Testanwendung gehörenden Konstrukten wie *Durchführungsobjektivität* oder *Vertrauens-/Konfidenzintervall* zu weniger Schwierigkeiten bei dessen Anwendungen führt.
- Es gibt moderate Hinweise, dass Fortbildungen zum Thema zu weniger Schwierigkeiten im Umgang mit Intelligenztests führen.
- Zusammengefasst kann die Nützlichkeit einer (universitären) Ausbildung zum Thema belegt werden bzw. die mit einer Reduzierung von entsprechenden Ausbildungsinhalten verbundenen zu erwartenden Schwierigkeiten.

Bestrebungen, universitäre Inhalte zum Thema zu reduzieren zugunsten anderer Inhalte, können als kontraproduktiv angenommen werden. Die Ergebnisse dieser Studie sind eindeutig und belegen die Nützlichkeit universitärer Inhalte zur Testdiagnostik bezüglich der Anwendungssicherheit. Frühere Auseinandersetzungen zwischen VertreterInnen einer Status- bzw. Förderdiagnostik (siehe Eberwein, 1996; Kobi, 1977; Bundschuh, 1985, 2007; Eberwein & Knauer, 1998, Eggert, 1997; Schlee, 2008) könnten zu einer Reduzierung der referierten Inhalte zur Testdiagnostik an den Universitäten geführt haben. Dies ist allerdings wenig nützlich, wenn von den SonderpädagogInnen in der Praxis erwartet wird, Testverfahren anzuwenden.

6.2 Formularanalyse

Im zweiten Teil dieser Arbeit wurden Intelligenztestformulare a posteriori auf Richtigkeit überprüft. Die Intelligenztests wurden ausschließlich im Rahmen einer Begutachtung zur Feststellung sonderpädagogischen Unterstützungsbedarfs durchgeführt. Neben der Darstellung der beiden Hypothesen 9 (ein PC-Auswertungsprogramm erhöht die Auswertungsobjektivität) und 10 (komplexere Tests sind fehleranfälliger), die zu der Fehleranalyse der Testformulare gehören, sollen auch einige interessante deskriptivstatistische Ergebnisse diskutiert werden, da sie Hinweise auf die Anwendungspraxis der Tests zulassen.

Die Auswertung des Fragebogens basiert auf Einschätzungen der ProbandInnen. Diese sind subjektiv und können Verzerrungen unterliegen. Die Antwort z.B. auf die Frage, ob die Anwendung von Intelligenztests leicht falle, kann auch davon abhängen, ob man die Anwendung von Intelligenztests mehr oder weniger nützlich findet und entsprechend mehr oder weniger gut vorbereitet ist. Eine oberflächliche Behandlung mit den Testmanualen, resultierend aus einer vielleicht sogar ablehnenden Haltung gegenüber der Statusdiagnostik könnte zur Folge haben, dass die Anwendung der Tests durch eine weniger intensive Auseinandersetzung eher leicht fällt.

Konkrete Fragen zu Schwierigkeiten bei der Anwendung der Durchführungsregeln wie der *Abbruch-* oder *Umkehrregel* könnten ebenfalls von persönlichen Voraussetzungen beeinflusst sein. Es ist möglich, dass ein Mangel nicht als solcher eingeschätzt wird, wenn es kein Bewusstsein für den Mangel gibt. So ist vorgeschrieben, dass bei der Berechnung des Anfangsitems des Nachfolgetests des SON-R 5½–17 geänderte Regeln angewendet werden: Das Anfangsitem wird durch den Rohwert eines Durchgangs minus 2 beim SON-R 6–40 bestimmt, während beim SON-R 5½–17 das Anfangsitem bestimmt wird durch den Rohwert eines Durchgangs minus 1. Gerade bei Nachfolgeversionen kann es vorkommen, dass die Regeln übernommen werden, obwohl sie sich teilweise änderten. Bei den Antworten des Fragebogens wäre es möglich, Fragen zu Schwierigkeiten bei der Anwendung der Durchführungsregeln als gering einzuschätzen, obwohl die Durchführungsregeln in der Praxis aus Unkenntnis fehlerhaft angewendet werden.

Die Analyse von Testformularen hat also die Aufgabe, typische Schwierigkeiten bei der tatsächlichen Anwendung von Tests festzustellen und die Schwierigkeiten differenziert darzustellen, um somit entsprechende Konsequenzen ziehen zu können. Die Analyse soll auch das Verhältnis zwischen subjektiv geprägter Einschätzung zu Schwierigkeiten und tatsächlich objektiv feststellbaren Schwierigkeiten erkennen.

6.2.1 Analyse von Intelligenztestformularen

In rund drei von fünf Formularen konnten Fehler entdeckt werden. Obwohl die Gesamtfehlerquote von ca. 60 Prozent irritierend hoch scheint, ist dies im Kontext früherer Studien ein erfreuliches Ergebnis, in denen teils Fehlerquoten von 100 Prozent beschrieben worden sind (siehe Alfonso et al., 1998). Dies ist auch ein erfreuliches Ergebnis, da weniger Fehler zu weniger möglichen Fehlurteilen führen könnten (siehe Lipsius et al., 2008). Es sollte auch nicht vergessen werden, dass zur *Klassischen Testtheorie* der grundlegende Gedanke gehört, dass Messfehler vorkommen können und diese mit der Verwendung des *Konfidenzintervalls* aufgefangen werden sollen. In einer Studie von Alfonso et al. (1998) war die Methode ähnlich dieser Arbeit. Nachträglich wurden Testformulare auf Fehler überprüft; neben der Fehlerquote von 100 Prozent konnten bei den 60 untersuchten Formularen durchschnittlich 7,8 Fehler entdeckt werden. Im Rahmen dieser Untersuchung konnten bei den 248 Formularen insgesamt 367 Fehler entdeckt werden, im Schnitt 1,48 Fehler/Formular unter Einbezug der Formulare ohne Fehler, bzw. 2,43 Fehler im Durchschnitt bei den ausschließlich fehlerhaften Formularen.

Obwohl dieser Befund als positiv bewertet wird, muss einschränkend erwähnt werden, dass außerordentlich konservativ ausgezählt wurde. Wenn die Wahrscheinlichkeit sehr groß für einen Fehler war, wurden Fehler nur dann als solche gewertet, wenn sie eindeutig nachweisbar waren. Es ist z. B. sehr unwahrscheinlich, dass TestanwenderInnen nachträglich vor dem Anfangsitem bei nicht durchgeführten Items eine Markierung als positiv beantwortet vornehmen, aber auch nicht gänzlich ausgeschlossen, denn es ist legitim, einen eigenen Notizenstil während der Testsituation zu nutzen, sei er noch so unwahrscheinlich. In diesem Fall wurden die vor dem Anfangsitem vorgenommen Markierungen (z. B. „1“ für richtig) nicht als Fehler gewertet.

Diese sehr vorsichtige Auswertung bedeutet allerdings auch, dass die gefundenen Fehler eindeutig welche sind, und auch wenn insgesamt weniger Fehler wie in vergleichbaren Studien attestiert werden, können diese teils gravierende Auswirkungen auf die Testergebnisse und die daraus resultierenden Schlussfolgerungen haben. Sollten z. B. Bezuschussungen nach dem §35a¹³⁸ einzig von dem Gesamt-IQ abhängig gemacht werden (evtl. gar ohne Einbezug des *Vertrauens-/Konfidenzintervalls*), könnte eine grobe Abweichung durch Fehler bei der *Durchführungs-* und/oder *Auswertungsobjektivität* sogar zu finanziellen Nachteilen führen.

Es wäre übertrieben, im Rahmen dieser Arbeit von häufig auftretenden groben Fehlern zu sprechen, aber eine Randerscheinung sind sie auch nicht. In

138 § 35a: Eingliederungshilfe für seelisch behinderte Kinder und Jugendliche.

29 Fällen wurden mehr als drei Fehler gemacht. In einem Test sind von elf Subtests elf falsch durchgeführt worden. In mehreren Subtests bewirkten die Fehler Abweichungen von einer Standardabweichung oder mehr; in einem Test lag der Gesamtwert bei IQ 90 statt bei dem falsch berechneten Gesamtwert von IQ 67. In einem anderen Test wurde das Testalter um ein Jahr falsch bestimmt, was eine Verschiebung von mehreren Normtabellen bewirkte; in einem weiteren Test sind von sechs Subtests fünf falsch durchgeführt und zudem noch das Alter um zwei Monate falsch berechnet worden.

Am häufigsten sind Punkte falsch zusammengezählt worden, danach machten die richtige Anwendung der Abbruchregel, die Bestimmung des Anfangsitems und die Anwendung der Umkehrregel die meisten Probleme.

Zusammengenommen werden diese Befunde als weiterer Beleg für eine Intensivierung der Ausbildung gewertet, sei es durch das Studium, sei es durch mehr zur Verfügung gestellte Zeit im beruflichen Kontext, um sich intensiver auf die Testungen vorzubereiten und um die Tests auszuprobieren.

Obwohl durch die Berücksichtigung von erheblich mehr Regeln mehr Fehler bei den mehrdimensionalen Tests möglich sind, wurden bei ausschließlicher Betrachtung der fehlerhaften Formulare im Schnitt mehr Gesamtfehler bei den eindimensionalen Tests (2,82) gemacht (2,21 im Durchschnitt bei den mehrdimensionalen Tests). Dieser scheinbare Widerspruch könnte damit zusammenhängen, dass die Anwendung der komplizierteren Tests einhergeht mit einer Einstellung, sich ausführlich damit zu beschäftigen, während die vermeintlich weniger komplizierten Tests dazu verleiten, deren einfachere Anwendung gleichzusetzen mit einer unzureichenden Vorbereitung. Es wäre möglich, mehrdimensionale Tests aus Überzeugung, eindimensionale Tests eher durchzuführen, weil ein Intelligenztestergebnis erwartet und die Anwendung eines einfachen Tests weniger ernst genommen wird. Diese Vermutung wird durch das Ergebnis bestärkt, dass 54 Prozent der eindimensionalen Tests fehlerfrei waren. Entweder wurden die eindimensionalen Tests weitgehend fehlerfrei durchgeführt oder überproportional häufig fehlerhaft.

Vorherige Hypothesenprüfungen konnten belegen, dass es entsprechend der Bildungshoheit der Bundesländer keinen einheitlichen Umgang in der Anwendung von Intelligenztests gibt. Diese Unterschiedlichkeit auf unterschiedliche Regelungen in den Bundesländern zurückzuführen, würde jedoch mit Sicherheit zu kurz greifen. Es ist möglich, dass die Anwendung von Intelligenztests mit den mehr oder weniger unterschiedlichen Regelungen der verschiedenen Schulämter zusammenhängt. Dies in Gänze zu prüfen, wäre ein aufwändiges Projekt im Rahmen einer weiteren Untersuchung. Zumindest bei den sechs Schulämtern, die sich an dieser Dissertation beteiligten, kann von einem unterschiedlichen Umgang bei der Anwendung von Intelligenztests gesprochen werden. Entweder wurden die ein- oder die mehrdimensionalen Tests präferiert. Im Vergleich zwischen den Schulämtern gibt es signifikante Unterschiede be-

züglich der Fehlerhäufigkeiten. Schulamt 6 machte signifikant weniger Fehler, sowohl bei den ein- als auch bei den mehrdimensionalen Tests im Vergleich zu den anderen Schulämtern. Schulamt 4 hat zumindest nicht signifikant weniger oder mehr Fehler bei den eindimensionalen Tests gemacht. Beim Schulamt 6 handelt es sich um ein den Schulämtern vergleichbares Regionales Bildungs- und Beratungszentrum (ReBBZ), in welchem SonderpädagogInnen häufig Tests durchführen, also eine Routine in der Anwendung entwickeln konnten. Erneut zeichnet sich ein Hinweis auf die nicht nur vermuteten, sondern tatsächlichen positiven Auswirkungen einer Spezialisierung bei der Anwendung der Tests ab.

Im Vergleich zwischen den Tests beeindruckt, dass mehr Fehler bei dem einfach zu lernenden SON-R 6–40 gefunden worden sind (durchschnittlich 1,55 Fehler) als bei der KABC-II (durchschnittlich 1,21 Fehler). Insgesamt besteht der SON-R 6–40 aus so wenigen Regeln wie manche der 18 Subtests der KABC-II. Bei der Analyse der Fehler soll kurz erwähnt werden, dass nicht von einer Verzerrung der Stichprobe ausgegangen werden kann. Während die Vermutung im Zusammenhang mit der Analyse der Daten aus dem Fragebogen naheliegend ist, dass die Ergebnisse durch die selektive Zusammensetzung der Stichprobe beeinflusst worden sind (viele ProbandInnen sind ehemalige Seminar- teilnehmerInnen), kann dies für die Analyse der Fehlerhäufigkeiten in den Testformularen nicht angenommen werden. Dies bedeutet, die prominente Stellung der KABC-II in der Sonderpädagogik wird auch deutlich, wenn es keine Hinweise auf eine Beeinflussung der Stichprobe zugunsten der KABC-II gibt.

Aus den Ergebnissen können auch einige konkrete Hinweise für Schulungsmaßnahmen abgeleitet werden:

- eine Auseinandersetzung mit der Bestimmung der Rohwerte für die IDS,¹³⁹
- eine Auseinandersetzung mit den Abbruchregeln und den Umkehrregeln des WISC-IV,¹⁴⁰
- eine Auseinandersetzung mit der Bestimmung des Anfangsitems für den SON-R 6–40.

Die Hypothese 9 prüfte, ob es einen Unterschied zwischen der Anzahl gemachter Fehler und der Nutzung eines Auswertungsprogramms gibt. Dieser Unterschied konnte nicht nachgewiesen werden. Aus diesem Ergebnis kann nur bedingt abgeleitet werden, dass die Nutzung eines Computerprogramms keinen Einfluss auf die Auswertungsobjektivität hat, denn für diese Hypothesenprüfung müssen methodische Mängel eingeräumt werden. Es wurde nicht bedacht,

139 Besonders viele Fehler wurden beim Subtest *Aufmerksamkeit selektiv* gezählt; diesen Subtest gibt es in einer modifizierten Variante auch in der inzwischen erschienenen IDS-2.

140 Im inzwischen erschienen Nachfolgetest WISC-V gibt es ebenfalls Umkehr- und Abbruchregeln in ähnlicher Form (teilweise sogar vereinfacht).

dass vor der Nutzung eines Computerprogramms die Tests richtig durchgeführt werden und die Rohwerte manuell richtig bestimmt werden müssen. Die Prüfung der Intelligenztestformulare auf mögliche Fehler ist unabhängig von der Nutzung eines Computerprogramms. Erst nach der Prüfung der meisten dieser Arbeit zugrunde liegenden Fragen kommt ein Computerprogramm zum Einsatz, spielt also für diese Arbeit keine wesentliche Rolle mehr.

Bejahten die ProbandInnen die Frage nach der Nutzung eines Computerprogramms, dann nur für die Tests, bei denen in das Computerprogramm die bereits gezählten Rohwerte eingegeben werden, ab hier hörten die meisten der Prüfungen für diese Arbeit auf. Es war im geringen Umfang möglich, die Summe aller per Hand berechneten Gesamtergebnisse mit denen zu vergleichen, die per Computer ausgerechnet worden sind. Bei zu vielen Fällen sind allerdings bereits vor Berechnung der Gesamtwerte so viele Fehler in den Formularen gemacht worden, dass die Berechnung eines Gesamtergebnisses auf falschen Rohwerten basiert hätte, so dass sehr hypothetisch-abstrakte Vergleiche mit falschen Rohwertbestimmungen vorgenommen worden wären.

Vorsichtige Hinweise werden für SON-R 6–40, WISC-IV und KABC-II beschrieben: für jeden dieser Tests konnten keine Signifikanzen ermittelt werden für die wenigen Fälle, wo mehrere ProbandInnen per Hand und nachvollziehbar Gesamtergebnisse ermittelten mit den Fällen, in denen Gesamtergebnisse per Computerprogramm ausgewertet worden sind. Vorausgesetzt, die Computerprogramme berechnen zu 100 Prozent¹⁴¹ korrekt, wäre die Bestätigung der Nullhypothese ein Beleg für die Sorgfalt bei der Berechnung der Gesamtergebnisse.

Die zehnte und somit letzte Hypothese prüfte den Zusammenhang zwischen tatsächlichen (und nicht auf Einschätzungen basierenden) Durchführungs- bzw. Auswertungsfehlern und der Komplexität eines Tests. Dieser Zusammenhang ist signifikant und auch vorhanden, wenn eindimensionale Tests mit mehrdimensionalen verglichen werden und gilt insbesondere für die Anwendung der *Abbruchregel* und der *Umkehrregel*. Für das Addieren der (Rohwert-)Punkte gilt dies nicht: es werden weniger Punkte falsch gezählt, je weniger komplex ein Test ist, was als erneuter Beleg dafür gewertet wird, dass die Anwendung von aufwändigeren Tests einhergeht mit einer intensiveren Auseinandersetzung.

Würde diese Vermutung zutreffen, muss dennoch festgestellt werden, dass eine noch intensivere Auseinandersetzung mit komplexeren bzw. mehrdimensionalen Tests angemessen ist. Selbst wenn die Anwendung aufwändigerer Tests mit einer angemessenen Auseinandersetzung mit den Tests einhergehen würde, die Anzahl real gezählter Fehler nimmt nichtsdestotrotz signifikant mit der Komplexität der Tests zu.

141 Wofür es Gegenbeispiele bei Computerauswertungen aus der Vergangenheit gibt.

Eigentlich ist es eine logische Konsequenz, die nicht beunruhigen sollte: je mehr Fehler möglich sind, desto mehr Fehler werden gemacht. Wichtig ist die Beurteilung, ob die quantitativ vorhandenen Fehler unabhängig von der Komplexität vertretbar sind und mit der Anwendung eines *Vertrauens-* bzw. *Konfidenzintervalls* hinreichend aufgefangen und relativiert werden können. Im Vergleich mit ähnlichen Untersuchungen (Lipsius et al., 2008) ist das Ergebnis erfreulich, ändert aber nichts an insgesamt 367 gefunden Fehlern in 151 von 248 Formularen, die zu falschen Ergebnissen führten, die in Subtests und sogar in Gesamtwerten über eine Standardabweichung hinaus vom tatsächlichen Ergebnis abwichen.

Es stellt sich erneut die Frage, ob die Anzahl vorhandener Fehler durch eine Spezialisierung mit der einhergehenden Entwicklung einer Routine vor allem bei mehrdimensionalen bzw. komplexen Tests verringert werden könnte.

6.2.2 Zusammenfassung und Bedeutung der Ergebnisse

- Im Vergleich zu früheren Studien ist die Fehlerquote geringer. Dies kann als Beleg für die Qualität und Professionalisierung bei der Anwendung von Intelligenztests im sonderpädagogischen Kontext gewertet werden.
- Es gab weniger ausgewertete Testformulare, die fehlerfrei waren als fehlerhafte. Dies wird als Beleg gewertet, die Anwendung von Intelligenztests noch professioneller gestalten zu müssen. Es gibt deutliche Hinweise, dass eine Spezialisierung zu weniger Fehlern führt. Eine Spezialisierung hätte zur Folge, dass wenige SonderpädagogInnen häufig vor allem die komplexen Tests durchführen sollten, um eine Routine entwickeln zu können.
- Aus der Analyse der Fehlerarten und Häufigkeiten können Empfehlungen für die Schulämter vorgenommen werden: Bezogen auf die in dieser Untersuchung geprüften Aspekte im Bereich der Durchführungsobjektivität könnte Schulamt 4 erwägen, eher eindimensionale Tests durchzuführen, da die Ergebnisse kaum von Fehlern beeinträchtigt sind, bei den mehrdimensionalen Tests werden hingegen überproportional häufig Fehler gemacht. Mehrdimensionale Tests wären zwar weniger aussagekräftig, die Ergebnisse dafür aber auch weniger von Fehlern beeinflusst. Schulamt 6 hingegen könnte empfohlen werden, die aussagekräftigeren mehrdimensionalen Tests durchzuführen, da – obwohl durch die höhere Anzahl von Durchführungsregeln fehleranfälliger – im Durchschnitt unter Einbezug aller Formulare kaum weniger Fehler festzustellen sind.
- In wenigen Fällen führten die Fehler in der Testanwendung oder -auswertung zu Ergebnissen, die über eine Standardabweichung vom korrekten Testergebnis abwichen, auch bei Gesamtergebnissen. Da alle Tests im Rahmen einer Begutachtung zur Erkennung sonderpädagogischen Förderbedarfs

durchgeführt worden sind, können negative Auswirkungen auf die Schlussfolgerungen im Rahmen der sonderpädagogischen Begutachtung nicht ausgeschlossen werden.

- Die meisten Fehler werden bei der Bestimmung der Rohwertpunkte, bei der Anwendung der Abbruchregel und bei der Anwendung der Umkehrregel gemacht, so dass die vertiefende Auseinandersetzung mit diesen Konstrukten sinnvoll wäre.
- Auffällig häufig wird beim SON-R 6–40 das Anfangsitem falsch bestimmt.
- Es gibt erneut Hinweise, dass die KABC-II ein Mittel der Wahl in der Sonderpädagogik ist und die Auseinandersetzung mit dem Testmanual ernster genommen wird, als bei deutlich einfacher zu lernenden Tests. Obwohl die KABC-II die mit Abstand meisten Regeln hat, konnten im Durchschnitt weniger Fehler als bei dem SON-R 6–40 gefunden werden, obwohl der Test insgesamt so viele Regeln hat wie mancher der 18 Subtests der KABC-II.
- Es fällt auf, dass entweder viele Fehler gemacht worden sind oder keine bis wenige. Es gibt Hinweise, dass die Tests, unabhängig von Dimensionalität oder Komplexität, entweder weitgehend beherrscht werden oder ungenügend.
- Bei der Anwendung der IDS-2 wird empfohlen, die Regeln zur Zählweise für die Bestimmung der Rohwerte für den Subtest *Zwei Merkmale durchstreichen* ausreichend zu internalisieren, da in der Vorgängerversion IDS der sehr ähnliche Subtest *Aufmerksamkeit selektiv* fehleranfällig war.
- Bei der Anwendung des WISC-V wird empfohlen, die Abbruch- und Umkehrregeln genügend zu internalisieren, da in der Vorgängerversion WISC-IV diese fehleranfällig waren und in beiden Versionen ähnlich angewandt werden.
- Mehrdimensionale bzw. komplexere Tests sind signifikant fehleranfälliger, was erneut für eine Spezialisierung zur Entwicklung von Routine in der Anwendung spricht.

Auch die Analyse dieser Ergebnisse lässt die Schlussfolgerung zu, dass ein häufigeres Testen als Mittel der Wahl im Sinne einer angemessenen Wahrung der Durchführungs- und Auswertungsobjektivität günstig wäre. Werden die Tests selten angewendet, steigt nicht nur die Gefahr einer fehlerhaften Anwendung, es bindet auch unnötig zeitliche Ressourcen, denn zu einer seltenen Durchführung ohne die Möglichkeit, Routine entwickeln zu können, gehört jedes Mal eine zeitintensive Vorbereitung.

Würden TestexpertInnen häufig testen, könnte sich dies auf die wenigen (über-)komplexen Tests wie KABC-II, WISC-IV, WPPSI-III oder IDS bzw. deren Nachfolgeversionen beschränken. Es würde also nicht eine Kernkompetenz in der Sonderpädagogik wegfallen. Sowohl die Anwendung eindimensionaler Tests, die Interpretation von allen Testverfahren (auch der komplexen) und die

Anwendung weiterer Bausteine der Diagnostik wie z. B. die Beobachtung gehörten weiter zur diagnostischen Expertise.

Schlussfolgerungen wie den bis hierhin beschriebenen beschäftigen sich mit der Struktur der Arbeit in der Sonderpädagogik, z. B. der Verbesserung der Testanwendungen durch eine Spezialisierung, oder die Verlängerung des Zeitraums für eine Gutachtenerstellung. Es könnte allerdings auch hinterfragt werden, ob die Konstruktion der Testverfahren angemessen ist. Insbesondere komplexe Tests wie KABC-II oder die Wechsler-Tests sind inklusive aller Regeln ins Deutsche adaptiert worden. Es ist nicht anzunehmen, dass genügend berücksichtigt wurde, dass in Deutschland *special education teachers* im Gegensatz zu denen aus anderen Ländern auch Intelligenztests anwenden. Es wäre wünschenswert, wenn bei der Adaption die Arbeitsbedingungen in der Sonderpädagogik durch eine weniger komplexe Konstruktion berücksichtigt werden würden. Es ist fraglich, ob ein Test wie die KABC-II alleine für die Testsituation mit dem Kind und ohne Berücksichtigung der allgemeinen Regeln aus ca. 580 Regeln bestehen muss. Jeder Subtest hat andere Regeln und es wäre durchaus möglich, die Regeln für jeden Subtest eines Tests gleich oder ähnlich zu gestalten.

Eine Perspektive könnte die Konstruktion eines Intelligenztests speziell für den sonderpädagogischen Einsatzbereich sein, dessen Anzahl an Regeln den Arbeitsbedingungen in der Sonderpädagogik entspricht, dessen Konstruktion den speziellen Bedürfnisse der Kinder entspricht (sehr kindgerechte Gestaltung der Items zur Erhöhung der *Compliance*; unbewertete Einführungs-Items; viele einfache Items zur Vermeidung von Bodeneffekten usw.) und dessen spezielle Normstichproben für die Bereiche *Geistige Entwicklung*, *Lernen* etc. den besonderen Kindergruppen in der Sonderpädagogik entsprechen.

6.3 Methodenkritik und Einschränkungen der Untersuchung

Die Aussagekraft der Ergebnisse wird durch einige Aspekte beeinträchtigt, die im Folgenden beschrieben werden.

Die Stichprobe resultiert überwiegend aus ehemaligen TeilnehmerInnen von Diagnostikseminaren. Vorhandene Datensätze wurden genutzt für die Rekrutierung von ProbandInnen für die Beantwortung des Fragebogens, dem wichtigsten Pfeiler dieser Untersuchung. Daraus resultiert eine Stichprobe, die nicht zwingend die Gesamtheit der SonderpädagogInnen repräsentiert. So ist es möglich, dass die Stichprobe aus Personen besteht, die sich bereits vermehrt mit der Anwendung von Intelligenztests beschäftigt haben, es ist auch möglich, dass die ProbandInnen besonders verunsichert in der Anwendung der Tests sind und sich daraus die Motivation an der Teilnahme an einem Seminar zum Thema ableitete. Möglich sind auch Beeinflussungen durch den Autor dieser Arbeit, der vielen ProbandInnen als Referenten bekannt ist. Es ist menschlich,

eine Haltung zu einer bekannten Person zu entwickeln und sich durch Sympathie oder Antipathie beeinflussen zu lassen.

Es ist anzunehmen, dass zumindest die Versuchsgruppe nicht repräsentativ für die Gesamtheit der SonderpädagogInnen steht. Über diese Annahme hinaus muss kritisch eingewendet werden, dass Vermutungen über die Zusammensetzung der Versuchsgruppe kaum vorgenommen werden können. Es ist also nicht auszuschließen, dass Dritt- bzw. Störvariablen die Ergebnisse beeinflussen. Es wäre z.B. möglich, dass überproportional häufig ProbandInnen an der Befragung teilgenommen haben, die technikaffin genug sind, einen E-Mail Anschluss zu konfigurieren, da Anfragen zur Teilnahme an der Studie ausschließlich elektronisch versandt worden sind¹⁴². In diesem Fall wäre es möglich, dass eine positive Einstellung zur Technik und die Anwendung komplexer Tests sich gegenseitig beeinflussen und somit die geringe Teilnahme wenig technikaffiner ProbandInnen die Ergebnisse verzerren. Es kann also festgehalten werden, dass die Studienergebnisse durch eine selektive Auswahl der ProbandInnen beeinflusst wurde und es kann festgehalten werden, dass darüber hinaus kaum Hinweise präsentiert werden können, welche Eigenschaften die ProbandInnen der Versuchsgruppe repräsentieren.

Obwohl diese Effekte als gering eingeschätzt werden, können sie nicht gänzlich ausgeschlossen werden. Zur Erhöhung der Repräsentativität wurden Gewichtungen vorgenommen und es gab Vergleiche mit einer Kontrollgruppe: ProbandInnen, die noch nie an einer Fortbildung zum Thema teilnahmen und also weder von entsprechenden Inhalten noch vom Referenten beeinflusst waren. Es sollte auch bedacht werden, dass bei rund 1 100 ProbandInnen bei einer Grundgesamtheit von ca. 68 000 in Deutschland arbeitenden SonderpädagogInnen (Destatis, 2017) es sich um eine Stichprobe nicht nur im Promille-, sondern im Prozentbereich handelt. Doch während die Stichprobe mit einer Fallzahl im vierstelligen Bereich beeindrucken könnte, träfe dies auf die Kontrollgruppe nicht zu.

Hinzu kommen Belege im Bereich der Wahlforschung, dass eine kleine repräsentative Gruppe genauer ist als eine große, aber weniger repräsentative Gruppe (Bandilla, 1999, S. 18).

Eine weitere Einschränkung liegt im konservativen Vorgehen im Rahmen der Signifikanzprüfungen. Die grundsätzlich zweiseitige Prüfung, obwohl eine einseitige Prüfung an der einen oder anderen Stelle begründbar wäre, und die grundsätzliche Verwendung korrigierter Signifikanz erhöht die Gefahr, Nullhypothesen zu Unrecht zu bestätigen. Diese Arbeit verfolgte bereits im Ansatz

142 Tatsächlich nutzten SeminarteilnehmerInnen oftmals die Accounts der PartnerInnen oder der Schulen für die Anmeldungen zu den Seminaren, in einigen Fällen wurde ausschließlich über den Postweg angemeldet.

das Ziel, Schlussfolgerungen aus den Ergebnissen ableiten zu können und gegebenenfalls Empfehlungen auszusprechen für den Umgang mit den Tests, für die Anschaffungspraxis, für die universitäre Ausbildung oder für den strukturellen Rahmen, in denen die Tests durchgeführt werden.

Die Ergebnisse dieser Arbeit resultieren aus einem konservativen Vorgehen und versuchen zu Unrecht abgeleitete Empfehlungen zu verhindern, die dann eher anmaßend als hilfreich wären. Die Annahme eines guten Rats für eine Verbesserung ist verbunden mit dem Eingeständnis, verbesserungswürdig gehandelt zu haben. Daraus resultiert die Verpflichtung, Ergebnisse eindeutig belegen zu können, erhöht zwangsläufig aber die Gefahr, signifikante Ergebnisse nicht entdeckt zu haben. Als Beispiel sei die Bonferroni-Korrektur genannt, die teils in der Wissenschaft abgelehnt wird, da sie zu konservativ sei (Hemmerich, 2015b), für diese Arbeit aber bewusst angewendet worden ist. Der konsequente Versuch, Typ-1 Fehler zu vermeiden, erhöht die Gefahr von Typ-2 Fehlern. Ein Blick auf die Ergebnisse dieser Arbeit lässt schnell erkennen, dass manche Signifikanzen und Tendenzen bei einer weniger vorsichtigen Vorgehensweise erkennbar wären.

Es muss eingeräumt werden, dass die Berechnung von Effektstärken eine sinnvolle Ergänzung an einigen Stellen sein könnte, z.B. bei der Nutzung des Chi-Quadrat-Tests, bei dem aus hohen Fallzahlen Signifikanzen bereits bei sehr geringen Unterschieden berechnet werden können und in diesen Fällen die Effektstärken evtl. Signifikanzen relativieren könnten. Die Bedeutung von Effektstärken wurde erst am Ende dieser Arbeit erkannt und blieb im Sinne der Arbeitsökonomie unberücksichtigt, wird zumindest als Lernprozess verstanden für zukünftige Forschungsarbeiten.

Einige Hypothesen beschäftigten sich mit Unterschieden in der Anwendung der Tests abhängig vom Bundesland. Bei genauerer Betrachtung ist eine Analyse der Anwendungspraxis in den Bundesländern nur schwer möglich, da der Rahmen im Umgang mit den Tests nicht einheitlich geregelt scheint in den Bundesländern, sondern abhängig vom Schulamt ist, den Einstellungen der MitarbeiterInnen der Schulämter zur Testdiagnostik, evtl. sogar von den Einstellungen der jeweiligen Schulen oder Förderzentren in den zu einem Schulamt gehörenden Regionen, aber auch von politischen Konstellationen, die zudem wechseln können. Die Inklusionsquote ist steigend und in einigen Bundesländern ein Politikum und der Erhalt oder die Schließung von Förderschulen ideologisch beeinflusst. Eine Plakatparole der *Alternative für Deutschland* (AfD) lautete im Europawahlkampf 2019 „Schließung der Förderschulen verhindern“. Die Anwendung von Intelligenztests ist nicht unbeeinflusst von politischen Rahmenbedingungen. Ein inklusiv beschultes Kind hat in der Regel einen sonderpädagogischen Unterstützungsbedarf attestiert bekommen. Das Attest kann stark von Intelligenztestergebnissen beeinflusst sein. Die Ergebnisse können also zu einem Anstieg der Quote von Kindern mit Unterstützungsbedarf beitra-

gen oder nicht und sind beeinflusst von einem gesellschaftlichen Klima. SonderpädagogInnen könnten sich dem Verdacht von politischer Seite ausgesetzt fühlen, leichtfertig sonderpädagogischen Unterstützungsbedarf zu attestieren, um die Anzahl der Betreuungsstunden im inklusiven Unterricht zu erhöhen und somit dem Verdacht, Intelligenztestergebnisse zu instrumentalisieren.

Politik als teilweise kurzlebige Geschäft zu bezeichnen wäre wenig gewagt. Dementsprechend ist eine richtungsweisende Orientierung in der Anwendung von Intelligenztests in den Bundesländern kaum erkennbar und es bleibt bei der richtigen Feststellung von Gaus und Drieschner, die dem Bildungssystem absprechen, einer internen Logik zu folgen und es als „labil chaotisch“ bezeichnen (2014, S. 29). Die gefundenen Unterschiede und Zusammenhänge im Vergleich zwischen den Bundesländern sollten also mit Vorsicht wahrgenommen werden, da es zweifelhaft ist, dass die Anwendung von Intelligenztests in den Bundesländern kongruenten Bedingungen folgen. Verbindliche Standards für die Anwendung wären sinnvoll, vor allem in Anbetracht der Tatsache, dass eine der wichtigsten Säulen in sonderpädagogischen Gutachten zur Erkennung von Förderbedarf die Würdigung von Intelligenztestergebnissen ist. Positive Beispiele stellen hier entsprechende Handreichungen in den Bundesländern Berlin und Brandenburg dar, die es bereits seit Jahren gibt, und die verbindlich und kontinuierlich angepasst an aktuelle Entwicklungen die Anwendung von Intelligenztests nicht nur in der Sonderpädagogik, sondern sogar in den unterschiedlichen Unterstützungsbedarfen regeln (Land Brandenburg, 2013; Senat Berlin, 2012). Es wäre interessant, Unterschiede und Zusammenhänge zwischen diesen beiden Bundesländern und den Bundesländern mit wenig geregelten Vorgaben zu untersuchen.

Oben beschriebene Einschränkungen in der Aussagekraft dieser Arbeit resultieren aus Problematiken, die im Zusammenhang mit den Untersuchungsbedingungen stehen. Eine weitere Einschränkung soll abschließend durch das Ausbleiben eines Forschungszweigs beschrieben werden. Diese Arbeit sollte ursprünglich auf drei Säulen stehen: die Befragung durch den Fragebogen, die Untersuchung von Intelligenztestformularen auf Durchführungs- und Auswertungsfehler und drittens auf Videoanalysen von Testsituationen. Diese nicht durchgeführte dritte Säule sollte SonderpädagogInnen bei der Anwendung von Intelligenztests filmen, die im Rahmen sonderpädagogischer Begutachtungen durchgeführt worden wären. Es ist möglich, dass die Einschätzungen bei der Beantwortung des Fragebogens subjektiv geprägt sind. Es ist auch möglich, dass Angaben zu Durchführungs- und Auswertungsschwierigkeiten durch ein mangelndes Bewusstsein über entsprechende Problematiken beeinflusst sind. Auf die Frage, ob während der Testsituation unerlaubte Rückmeldungen gegeben werden, könnte mit „Nein“ beantwortet werden, obwohl unbewusst über die nonverbale Kommunikation Rückmeldungen gegeben werden könnten. Eine Videoanalyse von realen Testsituationen wäre hilfreich und würde mögliche

subjektiv geprägte Einschätzungen objektivieren. So wären Rückschlüsse möglich im Sinne einer Sensibilisierung gegenüber den Auswirkungen nonverbaler Kommunikation über sinnvolle Schulungsinhalte.

Die notwendigen Datenschutzbestimmungen waren jedoch so aufwändig, dass der für die Umsetzung dieses Forschungszweigs notwendige zeitliche Rahmen diese Arbeit gesprengt hätte. Es wäre z. B. notwendig gewesen, Verträge mit den Datenschutzbeauftragten der jeweiligen Bundesländer bzw. Regionen mit jeweils anderen Bestimmungen abzuschließen, für jedes Kind einen Vertrag mit den Schulleitungen und einen Vertrag mit den Eltern. Neben dem Angebot individueller Rückmeldungen sollten monetäre Anreize SonderpädagogInnen ermutigen, sich an der Studie zu beteiligen. Dem stand gegenüber, dass die meist verbeamteten Lehrkräfte keine Zusatzeinkünfte während der Arbeitszeit erzielen dürfen. Es gab zusammengefasst so viele Hürden, dass ohne je eine Videoaufnahme verwirklicht zu haben, so viel Zeit und Mühen durch Telefonate, Akquise, Beratungen und Vertragsprüfungen investiert worden sind wie für das Schreiben mehrerer Kapitel dieser Arbeit. Abschließend betrachtet kann ohne Wertung und nicht kritisierend festgestellt werden, dass der Datenschutz den Forscherdrang ausgebremst hat. Bedingt durch den rechtlichen und zeitlichen Aufwand könne eine Videoanalyse an sich Forschungsgegenstand im Rahmen eines vermutlich umfangreichen Forschungsdesigns sein.

Die vorliegenden Befunde dieser Arbeit wären abgerundet gewesen bei einer Umsetzung des Videoprojekts und deshalb soll die Nichtverwirklichung als Einschränkung beschrieben werden.

6.4 Fazit und Ausblick

Ziel dieser Arbeit war es, Alltagsbeobachtungen, resultierend aus vielfach durchgeführten Fortbildungen zu standardisierten normierten Testverfahren, empirisch zu belegen. Von SonderpädagogInnen beklagte strukturelle Mängel bei der Anwendung von Intelligenztests wurden ebenso festgestellt wie beobachtete Mängel in der Durchführungs- und Auswertungsobjektivität. Unterschiede in der Anwendungspraxis führen dazu, dass die Auswahl der Testverfahren eher von der Region und weniger von den Fragestellungen und Besonderheiten der Kinder abhängen. Im Vergleich zu ähnlichen Untersuchungen ist die Anzahl gemachter Fehler bei der Anwendung der Tests gesunken, was für die Qualität der Anwendung von Intelligenztests im sonderpädagogischen Kontext spricht. Es werden allerdings eher einfach durchzuführende Tests mit Ausnahme der KABC-II präferiert, obwohl aussagekräftigere Tests zur Verfügung stehen.

Abgeleitet aus den Ergebnissen sind mehrere Schlussfolgerungen möglich. Es gibt eindeutige Belege, dass die mehrdimensionalen Tests geeigneter und

aussagekräftiger, somit nützlicher für den sonderpädagogischen Kontext sind. Die logische Konsequenz ist die häufigere Anwendung dieser Tests.

Es gibt Hinweise, dass eine Spezialisierung zu einer verbesserten Anwendung der Tests führt. Daraus kann die Empfehlung resultieren, dass zumindest die komplexen bzw. mehrdimensionalen Tests von wenigen SonderpädagogInnen häufig und nicht von vielen selten durchgeführt werden. Dies betrifft einige wenige aufwändige Tests und berührt und reduziert nicht die generelle diagnostische Expertise von SonderpädagogInnen.

Es gibt deutliche Hinweise, dass sich das Ausmaß universitärer Inhalte zur Testdiagnostik positiv auf die Anwendung auswirkt. Daraus kann die Empfehlung zu einer umfangreicheren Ausbildung, zumindest nicht zu einem Abbau der Inhalte im Rahmen der universitären Ausbildung resultieren.

Auch ein persönliches Fazit soll gezogen werden. Durch die vor Beginn selbst vielfach durchgeführten Intelligenztests und durch das Referieren derer Anwendung im Rahmen von Fortbildungen, bestand bereits vor Beginn dieser Arbeit eine gefestigte Haltung gegenüber Intelligenztests. Diese Haltung hat sich modifiziert. Es ist erstaunlich, wie viel im Rahmen einer umfangreichen Studie an Wissen hinzukommt. Die intensive Auseinandersetzung zur Thematik führte dazu, die Anwendung von Intelligenztests und dessen Ergebnisse noch kritischer zu betrachten. Dies kann begründet werden sowohl mit der Instrumentalisierung der Ergebnisse für zweifelhafte Zwecke (Intelligenztests als Kriegsmittel; zur Verbesserung der menschlichen Rasse im Rahmen eugenischer Bestrebungen; als Grundlage für rassistische Argumentationslinien). Es kann aber auch mit vielfachen methodischen Mängeln begründet werden, die vermutlich deutlicher zu Tage treten, je intensiver sich mit einem Thema beschäftigt wird. Weder das Konstrukt Intelligenz ist je bewiesen noch die Normalverteilung des nicht bewiesenen Konstrukts Intelligenz; die Begründung für die Validität wird durch Vergleiche mit anderen Tests seit jeher belegt, doch war bereits der erste Test in dieser Begründungskette methodisch fragwürdig und ein Grundgedanke der Klassischen Testtheorie ist die Fehlerhaftigkeit der Tests.

Die Nützlichkeit von Intelligenztests soll nach wie vor nicht in Frage gestellt werden, die Interpretation von Testergebnissen wird aber mit mehr Respekt und Vorsicht vorgenommen werden. Im Rahmen der Anwendung von Intelligenztests im sonderpädagogischen Kontext wäre zu wünschen, dass Intelligenztestergebnisse nicht die Grundlage eines sonderpädagogischen Gutachtens, sondern eine Ergänzung sind, insbesondere für die Unterstützungsbedarfe *Lernen* und *Geistige Entwicklung*.

Abgeleitet aus den Befunden dieser Arbeit resultieren Forschungsvorhaben für die Zukunft.

Die bereits geplante Videoanalyse von Testsituationen könnte weitere Hinweise auf Schwierigkeiten bei der Anwendung der Tests ergeben. Daraus könn-

ten sich konkrete Schulungsmaßnahmen ableiten lassen und in Bezug zu den Befunden dieser Arbeit gesetzt werden.

Ein Forschungsvorhaben könnte die Realisierung von Handreichungen für die Anwendung von Intelligenztests, welche im Rahmen sonderpädagogischer Begutachtungen durchgeführt werden, für eine Region oder ein Schulumt sein, auch unter Berücksichtigung bereits bestehender Handreichungen. Die Umsetzung der Empfehlungen könnte begleitet werden mit einer Evaluation bezüglich der Qualität der sonderpädagogischen Gutachten.

Interessant wäre zudem die Erfassung zu Einstellungen von SonderpädagogInnen zur Intelligenzdiagnostik. Daraus ließen sich Schlussfolgerungen für die Anwendung der Tests ziehen. Läge z. B. ein geringer Glaube an die Nützlichkeit vor, wären die positiven Aspekte der Intelligenzdiagnostik hervorzuheben, z. B. Hinweise für die Unterrichtsgestaltung bei erkannten Defiziten oder Ressourcen. Resultieren aus den Ergebnissen Hinweise auf Mängel zu testtheoretischen Konstrukten, können diese deutlicher im Rahmen des Studiums hervorgehoben werden. Daten für ein solches Projekt sind aus ökonomischen Gründen bereits mit den Daten für diese Arbeit miterhoben worden.

Die u. a. in dieser Arbeit gefundenen Schwierigkeiten im Umgang mit Intelligenztests könnten zu Empfehlungen bezüglich besserer Schulungen im Umgang mit den Durchführungsregeln oder zu einer Verbesserung der strukturellen Rahmenbedingungen führen, um die Schwierigkeiten zu verringern. Es scheint logisch, dass bei vermehrten Fehlern bei den Umkehr- und Abbruchregeln die Anwendung dieser Regeln genauer beachtet werden sollte. Ansätze dieser Art setzen bei den TestanwenderInnen an.

Es wäre aber auch möglich, die Anwendungsregeln weniger komplex zu gestalten. Dieser Ansatz setzt bei der Konstruktion der Intelligenztests an.

Tests wie die KABC-II oder WISC-IV gelten im Ursprungsland USA als *Level C* Tests, dürfen dort nur von besonders befähigten Personen durchgeführt werden, z. B. speziell geschulten PsychologInnen. Die Anwendung von Intelligenztests ist zwar ein wichtiger, aber meist selten umgesetzter Arbeitsbereich im sonderpädagogischen Kontext.

Ein lohnendes Projekt wäre die Konstruktion einer Intelligenztestbatterie, bei der sich die Regeln in jedem Subtest gleichen und die sich auf das Notwendigste beschränken. Solch ein Test würde den besonderen Arbeitsbedingungen in der Sonderpädagogik Rechnung tragen und kann dennoch aussagekräftig sein.

Ansätze hierzu bietet der SON-R 6–40. Jeder der vier Subtests unterliegt den gleichen Regeln. Durch das adaptive Testsystem (das Testergebnis eines Durchgangs bestimmt das Anfangsitem des nächsten Durchgangs, angepasst an die Fähigkeit des Kinds) werden fehleranfällige Regeln wie die Umkehrregel verhindert. Der SON-R 6–40 ist zwar nur bedingt aussagekräftig, aber es sollte möglich sein, statt vier Subtests eine Vielzahl von Subtests zu konstruieren bei

gleichzeitiger Beibehaltung der wenigen, sich in jedem Subtest wiederholenden Regeln. Ein weiterer Ansatz in die beschriebene Richtung kann für die Nachfolgeversion des WISC-IV (Petermann & Petermann, 2007) festgestellt werden, da die aus dem WISC-IV bekannten Regeln sich nun im Sinne der Anwendungsfreundlichkeit beim WISC-V (Wechsler, 2017) vereinfacht und reduziert haben.

Literatur

- Alfonso, V. C., Johnson, A., Patinella, L. & Rader, D. E. (1998). Common WISC-III examiner errors: Evidence from graduate students in training. *Psychology in the Schools*, 35, S. 119–125.
- Amelang, M. & Bartussek, D. (1990). *Differentielle Psychologie und Persönlichkeitsforschung* (3. überarbeitete und erweiterte Aufl.). Stuttgart: Kohlhammer.
- Amelang, M. & Zielinski, W. (1994). *Psychologische Diagnostik und Intervention*. Berlin: Springer.
- Amthauer, R. (1953). *Intelligenz-Struktur-Test*. Göttingen: Hogrefe.
- Amthauer, R. (1973). *I-S-T 70: Intelligenz-Struktur-Test* (3., erw. Aufl.). Göttingen: Hogrefe.
- Avenarius, H. (1990). *Anwendung diagnostischer Testverfahren in der Schule: ein Rechtsgutachten*. Weinheim: Beltz.
- Bandilla, W. (1999). WWW-Umfragen – Eine alternative Datenerhebungstechnik für die empirische Sozialforschung? In Batinic, B., Werner, A., Gräf, L. & Bandilla, W. (Hrsg.), *Online research: Methoden, Anwendungen und Ergebnisse* (S. 9–20). Göttingen: Hogrefe.
- Baudson, T. G. (2011). Pygmalion in der Schule. *MinD-Magazin* 82, S. 8–10. Abrufbar unter: <https://www.uni-due.de/imperia/md/content/dia/mindmag82-tgb.pdf> [6. 8. 2017].
- Baudson, T. G. (2012). Der Aufbau der Intelligenz: Die CHC-Theorie als Strukturmodell kognitiver Fähigkeiten. *MinD-Magazin. Die offizielle Zeitschrift von Mensa in Deutschland*, 91, S. 8–10.
- Bayerisches Staatsministerium für Unterricht und Kultus (2014). *Inklusion durch eine Vielfalt schulischer Angebote in Bayern*. Abrufbar unter: <https://bc.pressmatrix.com/de/profiles/66f86c543d18/editions/7a602ff9fe241e3e5208/pages>. [26. 07. 2019].
- Bayerisches Staatsministerium für Unterricht und Kultus (2018). *Der beste Bildungsweg für mein Kind mit sonderpädagogischem Förderbedarf – Informationen zur Einschulung*. Abrufbar unter: <https://bc.pressmatrix.com/de/profiles/66f86c543d18/editions/8bc1567bb1259df6ba18/pages> [26. 7. 2019].
- Beauducel, A. & Leue, A. (2014). *Psychologische Diagnostik*. Göttingen: Hogrefe.
- Behörde für Schule und Berufsbildung (Hamburg) (2014). *Liste der Testverfahren zur Diagnostik bei Förderbedarf Lernen, Sprache sowie emotionale und soziale Entwicklung*. Abrufbar unter: <https://www.hamburg.de/diagnostikverfahren/nofl/4341042/sd-testverfahren-final/> [4. 7. 19].
- Behörde für Schule und Berufsbildung (Hamburg) (2016) *Diagnosebogen Lernen*. Abrufbar unter: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwjkiNvQv_XjAhVbQ0EAHZDpBKUQFjAAegQIAhAC&url=https%3A%2F%2Fwww.hamburg.de%2Fcontentblob%2F4557324%2F7cd8aa65d504d2797b25466f34049b32%2Fdata%2Fdiagnosebogen-lernen-dl-aktuell.docx&usg=AOvVaw10qqC-qASb-_Z5Y4utdnYB [16. 7. 2019].
- Behörde für Schule und Berufsbildung (Hamburg) (2017). *Leitfaden – prozessorientierte Diagnostik*. Abrufbar unter: <https://www.hamburg.de/contentblob/4341038/3077d8d8e596c04c20599db96d1da781/data/sd-leitfaden-klaerung-final.pdf;jsessionid=DE61B9C0F83FC4D15A9C7479DE34846A.liveWorker2> [2. 7. 2019].
- Berger, M. (2014). Josefine Kramer – Ihr Leben und Wirken. *Zeitschrift für Heilpädagogik*, 2, S. 24–27.

- Bezirksregierung Münster (2017). *Handreichung für die Schulen der Sekundarstufen in der schulfachlichen Aufsicht der Bezirksregierung Münster*. Abrufbar unter: https://www.bezreg-muenster.de/zentralablage/dokumente/schule_und_bildung/inklusion/inklusionsordner/Inklusionsordner_Kapitel-7_AOSF_Handreichung_sekundarstufen.pdf [10. 8. 2019].
- Bildungsministerium Rheinland-Pfalz (2017). *Handreichung zur Feststellung des sonderpädagogischen Förderbedarfs*. Abrufbar unter: https://egs.bildung-rp.de/fileadmin/user_upload/egs.bildung-rp.de/FoeGu/Feststellungsverfahren_sonderpaedagogischer_Foerderbedarf_2_2017.pdf. [10. 8. 2019].
- Bildungsserver Saarland (2019). *Verfahren zur Feststellung eines sonderpädagogischen Förderbedarfs und zur Aufnahme in eine Förderschule bzw. zur gemeinsamen Unterrichtung in einer Schule der Regelform*. Abrufbar unter: https://www.saarland.de/dokumente/thema_bildung/VerfahrenFeststellungSonderpaedagogischerFoerderbedarf.pdf [10. 8. 2019].
- Billig, M. (1979). *Psychology, Racism & Fascism*. A. F. & R. Publications.
- Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Année Psychologique*, 11, S. 191–244.
- Bobertag, O. (1928). *Über Intelligenzprüfungen nach der Methode von Binet und Simon*. (3. Aufl.). Leipzig: Barth.
- Boring, E. G. (1923). Intelligence as the tests test it. *The New Republic*, 6, S. 35–37.
- Bortz, J. *Statistik* (4. Aufl.). Berlin: Springer.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler* (4. Aufl.). Berlin: Springer.
- Bourdieu, P., Beister, H. & Schwibs, B. (1993). *Soziologische Fragen* (5. Aufl.). Frankfurt am Main: Suhrkamp
- Brocke, B. & Beauducel, A. (2001). Intelligenz als Konstrukt. In E. Stern & J. Guthke (Hrsg.), *Perspektiven der Intelligenzforschung* (S. 13–42). Lengerich: Pabst.
- Brosius, F. (2017). *SPSS 24 für Dummies*. Weinheim: Wiley-VCH.
- Brügelmann, H. (2006). Sind Noten nützlich – und nötig? In H. Bartnitzky (Hrsg.), *Beiträge zur Reform der Grundschule: Band 121. Pädagogische Leistungskultur, Materialien für Klasse 3 und 4* (S. 17–46). Frankfurt am Main: Grundschulverb.
- Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3. aktualisierte Aufl.). München: Pearson Studium.
- Bühner, M., Ziegler, M., Bohnes, B. & Lauterbach, K. (2006). Übungseffekte in den TAP Untertests Test Go/Nogo und Geteilte Aufmerksamkeit sowie dem Aufmerksamkeits-Belastungstest (d2). *Zeitschrift für Neuropsychologie*, 17, S. 191–199.
- Bundschuh, K. (1985). *Dimensionen der Förderdiagnostik bei Kindern mit Lern-, Verhaltens- und Entwicklungsproblemen*. München: Reinhardt.
- Bundschuh, K. (2007). *Förderdiagnostik konkret: theoretische und praktische Implikationen für die Förderschwerpunkte Lernen, geistige, emotionale und soziale Entwicklung*. Bad Heilbrunn: Julius Klinkhardt.
- Bundschuh, K. (2010). *Einführung in die sonderpädagogische Diagnostik* (7. Aufl.). München: E. Reinhardt.
- Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 103, S. 276–279.
- Carroll, J. B. (1993). *Human Cognitive Abilities: A survey of factor-analytical studies*. Cambridge: Cambridge University Press.
- Castello, A. & Nestler, J. (2003). Praxis psychologischer Testdiagnostik an Erziehungsberatungsstellen. *Verhaltenstherapie & psychosoziale Praxis*, 35 (3), S. 555–566.
- Cattell, R. B. (1936). Is national intelligence declining? *The Eugenic Review*, 3/1936, S. 181.
- Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York: Yonkers-on-Hudson.

- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54, S. 1–22.
- Cattell, R. B. (1997). *Open Letter to the APA*. Abrufbar unter: <http://www.cattell.net/devon/openletter.htm>. [15. 8. 2019].
- Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *Journal of General Psychology*, 130 (3), S. 290–304.
- Chwallek, G. (2005). Tödlicher Intelligenzquotient. Abrufbar unter: <https://www.stern.de/panorama/stern-crime/todeskandidat-daryl-atkins-toedlicher-Intelligenzquotient-3547898.html> [15. 8. 2019].
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3), 297–334.
- Dalbert, C. (2013). *Gerechtigkeit in der Schule*. Berlin: Springer.
- Daseking, M., Petermann, F. & Waldmann, H.-C. (2008). Der allgemeine Fähigkeitsindex (AFI) – eine Alternative zum Gesamt-Intelligenzquotienten (G-IQ) des HAWIK-IV? *Diagnostica*, 54 (4), S. 211–220.
- Deary, I. J., Strand, S., Smith, P. & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, 35 (1), S. 13–21.
- Deimann, P. & Kastner-Koller, U. (2008). Testbesprechung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 40 (3), S. 161–165.
- Destatis (2015). *Bevölkerungsstand*. Abrufbar unter: <https://www.statistikportal.de/de/bevoelkerung/flaeche-und-bevoelkerung> [26. 6. 18].
- Destatis (2017). Fachserie 11 Reihe 1. *Bildung und Kultur. Allgemeinbildende Schulen. Schuljahr 2016/2017*. Abrufbar unter: https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Schulen/AllgemeinbildendeSchulen2110100177004.pdf;jsessionid=E60E0086401A40187B4A3376E225361A.InternetLive1?__blob=publicationFile [28. 3. 18].
- Deutsches Institut für Normung (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Dubois, B. & Burns, J. A. (1975). An analysis of the meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, 35, S. 869–884.
- Eberwein H. (1996). Förderdiagnostik als Lernprozeßdiagnostik. *Behinderte in Familie, Schule und Gesellschaft*, 19 (1), S. 5–14.
- Eberwein, H. & Knauer, S. (Hrsg.) (1998). *Handbuch Lernprozesse verstehen. Wege einer neuen (sonder-)pädagogischen Diagnostik*. Weinheim: Beltz.
- Eggert, D. (1997). *Von den Stärken ausgehen...: Individuelle Entwicklungspläne in der Lernförderungsdiagnostik. Ein Plädoyer für andere Denkgewohnheiten und eine veränderte Praxis*. Dortmund: Borgmann.
- Eser, K.-H. (2007). *Teilhabe-Barriere „Sprachlich-Symbolische Enthinderung“: Über den defizitären Umgang mit defizitären Lernprozessen*. Abrufbar unter: [http://www.sankt-nikolaus.de/web/st_nikolaus.nsf/gfx/26FB86E1E9A71896C12572D60043D8D0/\\$file/2007_05_07_eser_teilhabe_barriere.pdf](http://www.sankt-nikolaus.de/web/st_nikolaus.nsf/gfx/26FB86E1E9A71896C12572D60043D8D0/$file/2007_05_07_eser_teilhabe_barriere.pdf) [26. 01. 2016].
- Esser, G & Wyschkon, A. (2012). *BUEVA–II. Basisdiagnostik umschriebener Entwicklungsstörungen im Vorschulalter – Version II*. Göttingen: Hogrefe.
- Evers, A. (2001a). The Revised Dutch Ratings System for Test Quality. *International Journal of Testing*, 1, S. 155–182.
- Evers, A. (2001b). Improving Test Quality in the Netherlands: Results of 18 years of Test Ratings. *International Journal of Testing*, 1, S. 137–153.

- EVuP (Erste Verordnung für unterstützende Pädagogik) (2013). (Ohne Titel). Abrufbar unter: https://www.transparenz.bremen.de/sixcms/detail.php?gsid=bremen2014_tp.c.87628.de&template=00_html_to_pdf_d [2. 6. 2019].
- Finzsch, N. (1999). Wissenschaftlicher Rassismus in den Vereinigten Staaten – 1850 bis 1930. In Kaupen-Haas, H & Saller, C., *Wissenschaftlicher Rassismus: Analysen einer Kontinuität in den Human- und Naturwissenschaften* (S. 84–111). Frankfurt am Main: Campus Verlag.
- Flanagan, D. P., McGrew, K. S. & Ortiz, S. O. (2000). *The Wechsler Intelligence Scales and Gf-Gc theory: A contemporary approach to interpretation*. Needham Heights, MA: Allyn & Bacon.
- Flanagan, D. P., Ortiz, S. O. & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3. Aufl.). Hoboken, NJ: Wiley.
- Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, 95, S. 29–51.
- Funke, J. & Vaterrodt, B. (2009). *Was ist Intelligenz?* (3. aktualisierte Aufl.). München: C. H. Beck.
- Funke, J. (2006). Alfred Binet (1857–1911) und der erste Intelligenztest der Welt. In Lamberthi, G. (Hrsg.), *Intelligenz auf dem Prüfstand – 100 Jahre Psychometrie* (S. 23–40). Göttingen: Vandenhoeck & Ruprecht.
- Galton, F. (1869). *Hereditary genius*. New York: Macmillan.
- Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Book.
- Gaus, D. & Drieschner, E. (2014). Grundlegung einer Theorie- und Forschungsperspektive auf strukturelle Kopplungen des Bildungssystems. In: Drieschner, E. & Gaus, D. (Hrsg.): *Das Bildungssystem und seine strukturellen Kopplungen – Umweltbeziehungen des Bildungssystems aus historischer, systematischer und empirischer Perspektive* (S. 17–55). Berlin: Springer.
- Gehring, U. W. & Weins, C. (2009). *Grundkurs Statistik für Politologen und Soziologen* (5. Aufl.). Wiesbaden: Verlag für Sozialwissenschaften.
- Geißler, L. (2008). *Eine empirische Analyse der Typik der Intelligenz-Strukturen bei ADHS-Kindern und Jugendlichen mittels des HAWIK-III im Vergleich mit der Norm-Stichprobe*. (Unv. Diss.). Technische Universität Chemnitz: Institut für Psychologie.
- Goleman, D. (2012). *Emotional Intelligence: 10th Anniversary Edition*. New York: Random House Publishing Group.
- Gottfredson, L. (1994). Mainstream Science on Intelligence: An Editorial with 52 Signatories, History and Bibliography. Nachdruck in: *Intelligence*, 1/1997, S. 13–23.
- Gottfredson, L. S. (2002). g: Highly general and highly practical. In Sternberg, R. J. & Grigorenko, E. L. (Hrsg.), *The general factor of intelligence: How general is it?* (S. 331–380). Mahwah, NJ: Erlbaum.
- Grob, A. & Hagmann-von Arx, P. (2011). Replik auf die Rezension von. [sic]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 43, S. 246–249.
- Grob, A., Meyer, C. S. & Hagmann-von-Arx, P. (2009). *Intelligence and development scales: IDS; Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren*. Bern: Huber.
- Grob, A., Reimann, G., Gut, J. & Frischknecht, M.-C. (2013). *IDS-P. Intelligence and Development Scales. Intelligenz- und Entwicklungsskalen für das Vorschulalter*. Bern: Huber.
- Groffmann, K. J. (1983). Die Entwicklung der Intelligenzmessung. In K. J. Groffmann & Michel, L. (Hrsg.), *Enzyklopädie der Psychologie Themenbereich B, Methodologie und Methoden, Serie II Psychologische Diagnostik, Band 2 Intelligenz- und Leistungsdiagnostik* (S. 2–76). Göttingen: Hogrefe.

- Gruber, N. & Tausch, A. (2015). TBS-TK Rezension: „CFT 20-R mit WS/ZF-R. GrundIntelligenztest Skala 2-Revision (CFT 20-R) mit Wortschatztest und Zahlenfolgentest – Revision (WS/ZF-R)“. *Report Psychologie* 10, S. 403–404.
- Guilford, J.P. (1967) *Crystallized intelligences: The nature of human intelligence*. New York: McGraw-Hill.
- Guilford, J.P. (1977). *Way beyond the IQ: Guide to improving intelligence and creativity*. New York: McGraw-Hill.
- Hain, J. (2018). *Varianzanalyse ANOVA*. Lehrstuhl für Mathematik-Statistik. Universität Würzburg. Abrufbar unter: https://www.uni-wuerzburg.de/fileadmin/10040800/user_upload/hain/SPSS/ANOVA.pdf [2. 7. 2018].
- Haller, M. & Niggeschmidt, M. (Hrsg.) (2012). *Der Mythos vom Niedergang der Intelligenz: Von Galton zu Sarrazin: Die Denkmuster und Denkfehler der Eugenik*. Berlin: Springer.
- Hartig, J., Frey, A. & Jude, N. (2007). Validität. In: Moosbrugger, H. & Kelava, A. (Hrsg.): *Testtheorie und Fragebogenkonstruktion*. Heidelberg: Springer.
- Hausknecht, J.P., Halpert, J.A., Di Paolo, N.T. & Gerrard, M.M.O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92 (2), S. 373–385.
- Hebenstreit, G.K. (2000). *Die Fehleranfälligkeit der Auswertung beim Adaptiven Intelligenz Diagnostikum (Kubinger & Wurst, 1991)*. Unveröffentlichte Diplomarbeit, Universität Wien.
- Helmke, A. (2007). *Unterrichtsqualität. Erfassen, bewerten, verbessern*. Seelze: Kallmeyer.
- Hemmerich, W. (2015a). *Cronbachs Alpha: Auswerten und berichten*. Abrufbar unter: <https://statistikguru.de/spss/reliabilitaetsanalyse/auswerten-und-berichten-2.html> [23. 6. 2019].
- Hemmerich, W. (2015b). *Bonferroni-Korrektur*. Abrufbar unter: <https://statistikguru.de/lexikon/bonferroni-korrektur.html>. [29. 6. 2018].
- Herrnstein, R. & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.
- Hofmans, J., Theuns, P., Baekelandt, S., Mairesse, O., Schillewaert, N. & Cools, W. (2007). Bias and changes in perceived intensity of verbal qualifiers affected by scale orientation. *Survey Research Methods* 1, S. 97–108.
- Hogrefe (2017). *Testkatalog*. Abrufbar unter: <https://www.testzentrale.de/shop/tests/Intelligenztests.html?cat=2544> [24. 7. 2017].
- Holert, T. (2004). Intelligenz. In Bröckling, H.; Krasmann, S. & Lemke, T. (Hrsg.), *Glossar der Gegenwart* (S. 125–131). Frankfurt am Main: Suhrkamp.
- Holling, H., Preckel, F. & Vock, M. (2004). *Intelligenzdiagnostik*. Göttingen: Hogrefe.
- Horn, J.L. (1965). *Fluid and crystallized intelligence: A factor analytic study of the structure among primary mental abilities*. Thesis. University of Illinois.
- Huber, C. (1999). *Der Einsatz des Intelligenztests im Spannungsfeld traditioneller und gegenwärtiger Diagnostik*. Unveröffentlichte Examensarbeit am Seminar für Sondererziehung und Rehabilitation der Körperbehinderten. Heilpädagogische Fakultät der Universität zu Köln.
- Huber, C. (2000). Sonderpädagogische Diagnostik im Spannungsfeld traditioneller und gegenwärtiger Sichtweisen. Ergebnisse und kritische Reflexion einer Praxisuntersuchung an Schulen für Körperbehinderte. *Zeitschrift für Heilpädagogik*, 10, S. 411–416.
- Intelltheory. (2013a) *John L. Horn. American Psychologist*. Abrufbar unter: <https://www.intelltheory.com/horn.shtml> [15. 8. 2019].
- Intelltheory. (2013b) *John B. Carroll. American Educational Psychologist*. Abrufbar unter: <https://www.intelltheory.com/horn.shtml> [15. 8. 2019].
- Irblich, D. (2010). Neuere Testverfahren. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 59, S. 316–325.

- Irblich, D. (2015). Neuere Testverfahren. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 64, S. 635–648.
- Jäger, A. O. & Althoff, K. (1983). *Der Wilde-Intelligenz-Test (WIT) – Ein Strukturdiagnostikum*. Göttingen: Hogrefe.
- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, 23, S. 195–225.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, 35 (1), S. 21–35.
- Jäger, A. O. (1986). Validität von Intelligenztests. *Diagnostica*, 32, S. 272–289.
- Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997a). Berliner Intelligenzstruktur – Test. Form 4. In Sarges, W. & Wottawa, H. (Hrsg.): *Handbuch wirtschaftspsychologischer Testverfahren* (S. 95–101). Lengerich: Pabst Science Publishers.
- Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997b). *Berliner Intelligenzstruktur-Test: BIS-Test; Form 4; mit einem separat verwendbaren Kurztest der allgemeinen Intelligenz und der Verarbeitungskapazität; Handanweisung*. Hogrefe: Verlag für Psychologie.
- Jantzen, W. (2003). A. N. Leont'ev und das Problem der Raumzeit in den psychischen Prozessen. Eine methodologische Rekonstruktion. In Jantzen, W. & Siebert, B. (Hrsg.), „Ein Diamant schleift den anderen“ – Evald Vasilevič Il'enkov und die Tätigkeitstheorie (S. 400–462). Berlin: Lehmanns.
- Jantzen, W. (2011). Der wissenschaftliche Weg von Vygotsky. Frei gehaltener Vortrag bei den 6. Görlitzer Heilpädagogischen Tagen/Fachtagung: „Begriffe, Praxen, Perspektiven – kulturhistorische Ideen für inklusives Handeln“ (13.–15.05.2011). Transkriptionen und Durchsicht durch Christmann, S.; Jödecke, M. & Reif, B.
- Jantzen, W. (Hrsg.) (2004). *Gehirn, Geschichte und Gesellschaft: die Neuropsychologie Alexander R. Lurijas (1902–1977)*. Berlin: Lehmanns Media.
- Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
- Joél, T. (2017). Das Dilemma der Intelligenzdiagnostik in der Sonderpädagogik – erläutert anhand der neuen KABC-II. *Zeitschrift für Heilpädagogik*, 68, S. 12–21.
- Joél, T. (2018). Intelligenzdiagnostik mit geflüchteten Kindern und Jugendlichen. *Zeitschrift für Heilpädagogik*, 69, S. 196–206.
- Johnson, W. & Bouchard Jr., T. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, 33 (4), S. 393–416.
- Kahl, R. (2006). Schlechte Noten für die Noten. *Pädagogik*, 7-8 (06), S. 96. Abrufbar unter: <http://www.reinhardkahl.de/dokumente/pdf/PS%207%208%20noten.pdf> [24.7.2019].
- Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement. A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, S. 361–372.
- Karg Fachportal Hochbegabung. (2017). SON-R 2 ½–7 – Snijders-Oomen Non-Verbaler Intelligenztest 2½–7. Abrufbar unter: <https://www.fachportal-hochbegabung.de/Intelligenztests/son-r-2-7-snijders-oomen-non-verbaler-Intelligenztest-2-7-revidierte-fassung/> [17.8.2019].
- Kastner-Koller, U. & Deimann, P. (2012). *WET. Wiener Entwicklungstest. Ein Verfahren zur Erfassung des allgemeinen Entwicklungsstandes bei Kindern von 3 bis 6 Jahren* (3., überarb. u. erw. Aufl.). Göttingen: Hogrefe.
- Kaufman, A. & Kaufman, N. (2004). *Kaufman Assessment Battery for Children – Second Edition*. Bloomington: NCS Pearson.
- Kaupen-Haas, H & Saller, C. (1999). *Wissenschaftlicher Rassismus: Analysen einer Kontinuität in den Human- und Naturwissenschaften*. Frankfurt am Main: Campus Verlag.

- Kendall, M. G. (1962). *Rank Correlation Methods*. London: Griffin.
- Kern, H., Boldt-Mußmann, G., Dietmann, T., Heuel, K. J., Heuel, J., Jacob, E., Matthias, B. & Winter-Witschurke, C. (2017). *Leitfaden zur Feststellung sonderpädagogischen Förderbedarfs an Berliner Schulen*. Abrufbar unter: https://www.berlin.de/sen/bildung/schule/.../leitfaden_foerderbedarf-2017.pdf [29. 7. 19].
- Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau*, 57 (4), S. 243–253.
- Keune, S. & Frohnenberg, C. (2004). *Nachteilsausgleich für behinderte Prüfungsteilnehmer/innen. Handbuch mit Fallbeispielen und Erläuterungen für die Prüfungspraxis*. Bielefeld: Bertelsmann.
- Knebel, L. & Marquardt, P. (2012). Vom Versuch, die Ungleichwertigkeit von Menschen zu beweisen. In Haller, M. & Niggeschmidt, M. (Hrsg.), *Der Mythos vom Niedergang der Intelligenz: Von Galton zu Sarrazin: Die Denkmuster und Denkfehler der Eugenik* (S. 87–126). Wiesbaden: Springer-Verlag.
- Kobi, E. (1977). Einweisungsdiagnostik – Förderdiagnostik: eine schematische Gegenüberstellung. *Vierteljahresschrift für Heilpädagogik und ihre Nachbargebiete*, 46, S. 115–123.
- Koch, H., Kastner-Koller, U. & Deimann, P. (2011). Testbesprechung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 43 (2), S. 108–113.
- Kottmann, B. (2006). *Selektion in der Sonderschule: das Verfahren zur Feststellung von sonderpädagogischem Förderbedarf als Gegenstand empirischer Forschung*. Bad Heilbrunn: Julius Klinkhardt.
- Kramer, Jochen (2009). *Metaanalytische Studien zu Intelligenz und Berufsleistung in Deutschland*. Unv. Diss., Philosophische Fakultät der Rheinischen Friedrich-Wilhelms-Universität zu Bonn.
- Kramer, Josefine (1972). *Intelligenztest. Mit einer Einführung in Theorie und Praxis der Intelligenzprüfung* (4. revidierte Aufl.). Solothurn: Antonius.
- Krebs, D. & Hoffmeyer-Zlotnik, J. H. P. (2010). Positive first or negative first? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6 (3), S. 118–127.
- Krosnick, J. A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, S. 201–219.
- Krosnick, J. A. & Fabrigar L. R. (1997). Designing rating scales for effective measurement in surveys. In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N. & Trewin, D. (Hrsg.), *Survey measurement and process quality* (S. 141–164). New York: John Wiley & Sons.
- Krosnick, J. A. & Presser S. (2010). Question and Questionnaire Design. In Peter V. Marsden und James D. Wright (Hrsg.), *Handbook of Survey Research* (S. 264–313). Bingley, UK: Emerald.
- Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, 47, S. 583–621.
- Kubinger, K. D. & Holocher-Ertl, S. (2014). *Adaptives Intelligenz Diagnostikum – Version 3.1 (AID 3)*. Göttingen: Hogrefe.
- Kubinger, K. D. (2009a). *Adaptives Intelligenz Diagnostikum – Version 2.2 (AID 2) samt AID 2-Türkisch*. Göttingen: Hogrefe.
- Kubinger, K. D. (2009b). *Psychologische Diagnostik. Theorie und Praxis psychologischen Diagnostizierens* (2., überarb. und erw. Aufl.). Göttingen: Hogrefe.
- Kühl, S. (1999). Die soziale Konstruktion von Wissenschaftlichkeit und Unwissenschaftlichkeit in der internationalen eugenischen Bewegung. In Kaupen-Haas, H & Saller, C., *Wissenschaftlicher Rassismus: Analysen einer Kontinuität in den Human- und Naturwissenschaften* (S. 111–121). Frankfurt am Main: Campus Verlag.

- Kühl, S. (2014). *Die Internationale der Rassisten: Aufstieg und Niedergang der internationalen eugenischen Bewegung im 20. Jahrhundert* (2. aktualisierte Aufl.). Frankfurt am Main: Campus Verlag.
- Kultusministerium Sachsen-Anhalt. (o. J.). *Handreichung zur sonderpädagogischen Förderung in Sachsen-Anhalt*. Abrufbar unter: https://www.sachsen-anhalt.de/fileadmin/Bibliothek/Politik_und_Verwaltung/MK/MK/Textdokumente/Publikationen/Bildung/handreichung_sonderpaedagogische_foerderung.pdf [10. 8. 2019].
- Kultusministerkonferenz (1997). *Eingliederung von Berechtigten nach dem Bundesvertriebenengesetz in Schule und Berufsbildung*. Abrufbar unter: https://www.kmk.org/fileadmin/Dateien/pdf/ZAB/Hochschulzugang_Beschluesse_der_KMK/BVFG.pdf [1. 7. 2019].
- Kultusministerkonferenz (2019). *Aufgaben der Kultusministerkonferenz*. Abrufbar unter: <https://www.kmk.org/kmk/aufgaben.html>. [01.07.2019].
- Kuschel, A., Kamp-Becker, I. & Ständer, D. (2017). TBS-TK Rezension: „Kaufman Assessment Battery for Children-2 (KABC-II)“. *report psychologie* 5/2017, S. 211–2012.
- LAG (Landesarbeitsgemeinschaft Baden-Württemberg – Gemeinsam leben – gemeinsam lernen e. V.). (2016). *Inklusion macht Schule. „Ich kenne meine Rechte“ – Das neue Schulgesetz in Baden-Württemberg*. Abrufbar unter: <https://www.lag-bw.de/PDF2015/Elternratgeber.pdf> [2. 7. 2019].
- Lamberti, G. (2006). *Intelligenz auf dem Prüfstand: 100 Jahre Psychometrie*. Göttingen: Vandenhoeck & Ruprecht.
- Land Brandenburg (Ministerium für Bildung, Jugend und Sport) (Hrsg.) (2013). *Handreichung zur Durchführung des sonderpädagogischen Feststellungsverfahrens*. Presse- und Öffentlichkeitsarbeit Referat 32.
- Lange, V. (2017). *Ländervergleich. Inklusive Bildung in Deutschland*. Friedrich Ebert Stiftung. Abrufbar unter: <http://library.fes.de/pdf-files/studienfoerderung/13493.pdf> [18. 8. 2019].
- Legg, S. & Hutter, M. (2007). A collection of definitions of intelligence. *Frontiers in Artificial Intelligence and Applications*, 157, S. 17–24.
- Leibniz-Zentrum für Psychologische Information und Dokumentation (2017). *Internationale Richtlinien für die Testanwendung*. Abrufbar unter https://www.zpid.de/pub/tests/itc_richtlinien.pdf [03.05.2017].
- Lipsius, M., Petermann, F. & Daseking, M. (2008). Wie beeinflussen Testleiter die HAWIK-IV-Befunde? *Kindheit und Entwicklung*, 17 (2), S. 107–117.
- Loehlin, J. C., Lindzey, G. & Spuhler, J. N. (1975) *Race differences in intelligence*. San Francisco: Freeman.
- Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley.
- Luria, A. R. (1932). *The nature of human conflicts*. New York: Liveright.
- Luria, A. R. (1966). *Human brain and psychological processes*. New York: Harper and Row.
- Lurija¹⁴³, A. R. (1970). The functional organization of the brain. *Scientific American*, 222, S. 66–78.
- Maltby, J., Day, L. & Macaskill, A. (2011). *Differentielle Psychologie, Persönlichkeit und Intelligenz* (2. aktualisierte Aufl.). München: Pearson.
- Mayo, E. (1930). The human effect of mechanization. *Papers and Proceedings of the 42th Annual Meeting of the American Economic Association*, Vol. XX, Nr. 1, S. 156–176.
- Mayo, E. (1933). *Humans Problems of an Industrial Civilization*. New York: McGraw-Hill.
- MBJS (Land Brandenburg: Ministerium für Bildung, Jugend und Sport, Hrsg.). (2018). *Handreichung zur Durchführung des sonderpädagogischen Feststellungsverfahrens*. Abrufbar un-

143 Lurija/Luria wird nicht einheitlich geschrieben.

- ter: https://mbjs.brandenburg.de/media_fast/6288/final_handreichung_2018.pdf [30.7.19].
- McCrae, R. R. & Costa, P. T. (1989). Reinterpreting the Myers-Briggs Type Indicator from the Perspective of the Five-Factor Model of Personality. *Journal of Personality*, S. 17–40.
- McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, 33, S. 337–346.
- Mecklenburg, J. (2002). *Was tun gegen rechts*. Berlin: Espresso-Verlag.
- Melchers, P. & Melchers, M. (2015). *Kaufman Assessment Battery for Children – II (KABC-II). Deutsche Bearbeitung der Kaufman Assessment Battery for Children – Second Edition von Alan und Nadeen Kaufman*. Frankfurt: Pearson Assessment.
- Melchers, P. & Preuß, U. (2009). *Kaufman Assessment Battery for Children* (deutsche Version) (8. unveränd. Aufl.). Frankfurt/M.: Pearson Assessment.
- Menold, N. & Bogner, K. (2015). *Gestaltung von Ratingskalen in Fragebögen (Version 1.1). (GESIS SurveyGuidelines)*. Mannheim: GESIS – Leibniz-Institut für Sozialwissenschaften. Abrufbar unter: https://doi.org/10.15465/gesis-sg_015 [18.8.19].
- Merton, R. K. (1948). The self-fulfilling prophecy. *The Antioch Review*, S. 193–210.
- Mickley, M. & Renner, G. (2010). Intelligenztheorie für die Praxis: Auswahl, Anwendung und Interpretation deutschsprachiger Testverfahren für Kinder und Jugendliche auf Grundlage der CHC Theorie. *Klinische Diagnostik u. Evaluation*, 3, S. 447–466.
- Mickley, M. (2013). Nutzen, Schaden und Qualität standardisierter Entwicklungs- und Intelligenzdiagnostik. *Kinder- und Jugendmedizin*, 3, S. 1–5.
- Mickley, M. (2015). Besprechung der deutschen Bearbeitung der „Wechsler Nonverbal Scale of Ability“ (WNV). *Frühförderung interdisziplinär*, 34, S. 105–113.
- Mienert, M. & Pitcher, S. (2011). *Pädagogische Psychologie – Theorie und Praxis des lebenslangen Lernens – Lehrbuch*. Berlin: Springer-Verlag.
- Ministerium für Bildung und Kultur (Saarland). (2019). *Sonderpädagogische Förderung in Regel- und Förderschulen*. Abrufbar unter: <https://www.saarland.de/216957.htm> [10.8.2019].
- Ministerium für Bildung, Wissenschaft und Kultur (Mecklenburg-Vorpommern) (Hrsg.) (2015). *Standards der Diagnostik*. Abrufbar unter: https://www.bildung-mv.de/export/sites/bildungserver/downloads/Handbuch_Diagnostik_2015.pdf [9.8.2019].
- Ministerium für Bildung, Wissenschaft und Kultur Thüringen (Hrsg.) (2013). *Handreichung für den Gemeinsamen Unterricht*. Abrufbar unter: https://www.thueringen.de/mam/th2/schulaemter/handreichung_gu.pdf [11.8.2019].
- Moosbrugger, H. & Höfling, V. (2007). Standards für psychologisches Testen. In H. Moosbrugger, H & A. Kelava, A. (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 193–212). Heidelberg: Springer.
- Moosbrugger, H. & Kelava, A. (2007). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In: Moosbrugger, H. & Kelava, A. (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7–26). Heidelberg: Springer.
- Müller, C.M. (2009). Schulinterne Diagnostikansprechpartner. Eine Praxiskonzeption zur Umsetzung systematischer Diagnostik im Schulalltag. *Zeitschrift für Heilpädagogik*, 5, S. 180–187.
- Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5. Aufl.). Upper Saddle River, NJ: Prentice Hall.
- Myers, D. G. (2008). *Psychologie* (3. Aufl.). Berlin: Springer-Verlag.
- Naescher, S. (2009). *SON-R 2½–7 Snijders-Oomen Non-verbaler Intelligenztest 2½–7 – Revidierte Fassung (PSYNDEX Tests Review)*. Abrufbar unter: <https://www.zpid.de/retrieval/PSYNDEXTests.php?id=9003441> [3.8.2017].

- Naescher, S. (2010). *IDS – Intelligenz- und Entwicklungsskalen für Kinder von 5–10 Jahren* (PSYINDEX Tests Review). Abrufbar unter: <https://www.zpid.de/retrieval/PSYINDEXTests.php?id=9006118> [4. 8. 2017].
- Norden, I. (1953). *Anleitung zur Intelligenzprüfung nach Binet-Bobertag „Binetarium“*. Göttingen: Hogrefe.
- Norden, I. (1956). *Binetarium. Hilfsmittel zur Intelligenzprüfung nach Binet-Bobertag* (12. Aufl.). Göttingen: Hogrefe.
- O’Muircheartaigh, C., Krosnick, J.A. & Helic, A. (1999). Middle alternatives, acquiescence, and the quality of questionnaire data. Paper presented at the annual meeting of the American Association for Public Opinion Research. St. Petersburg, Florida.
- Pearsonassessment (2019). *Qualifications*. Abrufbar unter: <https://www.pearsonassessments.com/professional-assessments/ordering/how-to-order/qualifications.html> [20. 8. 2019].
- Petermann, F. & Petermann, U. (Hrsg.) (2007). *Hamburg-Wechsler-Intelligenztest für Kinder-HAWIK-IV*. Bern: Huber.
- Petermann, F. (2009). *WPPSI-III. Wechsler Preschool and Primary Scale of Intelligence – third edition. Deutschsprachige Adaptation nach D. Wechsler*. Frankfurt: Pearson Assessment.
- Petermann, F. (2014). *Wechsler Nonverbal Scale of Ability (WNV)*. Frankfurt: Pearson.
- Pilshofer, B. (2001). *Wie erstelle ich einen Fragebogen*. Wissenschaftsladen Graz Institut für Wissens- und Forschungsvermittlung. Abrufbar unter: https://www.ph-ludwigsburg.de/fileadmin/subsites/2d-sprt-t-01/user_files/Hofmann/SS08/erstellungvonfragebogen.pdf [17. 1. 2017].
- Ploetz, A. (1895). *Die Tüchtigkeit unserer Rasse und der Schutz der Schwachen: ein Versuch über Rassenhygiene und ihr Verhältnis zu den humanen Idealen, besonders zum Socialismus. Grundlinien einer Rassen-Hygiene, 1. Theil*. Berlin: S.Fischer.
- Porteus, S.D. (1965). *Porteus Maze Tests: Fifty years application*. Palo Alto: Pacivic Books.
- Poseschill, M. (1996). *Praktische Statistik*. Weinheim: Psychologie Verlags Union.
- Preckel, F. & Vock, M. (2004). *Intelligenzdiagnostik*. Göttingen: Hogrefe.
- Raab-Steiner, E. & Benesch, M. (2015). *Der Fragebogen. Von der Forschungsidee zur SPSS-Auswertung* (4. aktualisierte und überarbeitete Aufl.). Wien: Facultas Universitätsverlag.
- Raven, J.C., Raven, J. & Court, J.H. (2010). *CPM – Coloured Progressive Matrices. Deutsche Bearbeitung und Normierung Bulheller, S. & Häcker, H.* Frankfurt a.M.: Pearson Assessment.
- Renner, G. & Fricke, T. (2001). *Der Hamburg-Wechsler-Intelligenztest für Kinder – dritte Auflage (HAWIK-III)*. *Report Psychologie*, 26, S. 460–477.
- Renner, G. & Mickley, M. (2015a). Berücksichtigen deutschsprachige Intelligenztests die besonderen Anforderungen von Kindern mit Behinderungen? *Praxis der Kinderpsychologie und Kinderpsychiatrie*, 64, S. 88–103.
- Renner, G. & Mickley, M. (2015b). Intelligenzdiagnostik im Vorschulalter. *CHC – theoretisch fundierte Untersuchungsplanung und Cross-battery-assessment*. *Frühförderung interdisziplinär*, 34, S. 67–83.
- Renner, G. (2010). Testbesprechung. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 42 (3), S. 177–182.
- Renner, G. (2014). Neue diagnostische Verfahren für die Sonderpädagogik. *CFT 1-R. Grundintelligenztest Skala 1 – Revision. Sonderpädagogische Förderung heute*, 59, S. 107–112.
- Renner, G. (2015). Neue diagnostische Verfahren für die Sonderpädagogik. *Sonderpädagogische Förderung*, 4, S. 439–444.
- Renner, G., Schmid, S., Irblich, D. & Krampen, G. (2012). Psychometrische Eigenschaften der „Kaufman-Assessment Battery for Children.“ (K-ABC) bei 5- und 6-jährigen Kindern: Re-

- liabilität und Validität in einer klinischen Stichprobe. *Frühförderung Interdisziplinär*, 31, S. 197–206.
- Ricken, G., Fritz, A., Schuck, K. D., Preuß, U. (Hrsg.) (2007). *HAWIVA-III. Hannover-Wechsler-Intelligenztest für das Vorschulalter – III*. Bern: Huber.
- Riepl, W. (2013). *Methodenberatung: Welcher statistische Test passt zu meiner Fragestellung und meinen Daten?* Abrufbar unter: <https://statistik-dresden.de/archives/6026> [16.10.2019].
- Rindermann, H., Flores-Mendoza, A. & Mansur-Alves, M. (2010). Reciprocal effects between fluid and crystallized intelligence and their dependence on parents' socioeconomic status and education. *Learning and Individual Differences*, 20, S. 544–548.
- Roethlisberger, F. J., Dickson & W. J. (1939). *Management and the Worker*. Cambridge, Mass.: Harvard University Press.
- Rohrman, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, 1978, S. 222–245.
- Rollett, B. & Preckel, F. (2011). TBS-TK Rezension: „K-ABC: Kaufman – Assessment Battery for Children.“ *Psychologische Rundschau*, 63, S. 139–141.
- Rosenthal, R. & Fode, K. L. (1963). „The Effect of Experimenter Bias on the Performance of the Albino Rat“. *Behavioral Science* 8, S. 183–189.
- Rosenthal, R. & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart & Winston.
- Sächsisches Staatsinstitut für Bildung und Schulentwicklung. (2005). *Material- und Methodensammlung zur Förderdiagnostik*. Abrufbar unter: https://www.schule.sachsen.de/download/download_smk/material_foerderdiagnostik_teil2.pdf [11. 8.2019].
- Salgado, J., 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 (6), S. 1068–1081.
- Saris, W. E. & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Sarrazin, T. (2010). *Deutschland schafft sich ab: Wie wir unser Land aufs Spiel setzen*. München: DVA.
- Sattler, J. M. & Dumont, R. (2004). *Assessment of children: WISC-IV and WPPSI-III supplement*. San Diego: Jerome Sattler, Publisher.
- SBA-VO. (2016). *Verordnung des Kultusministeriums über die Feststellung und Erfüllung des Anspruchs auf ein sonderpädagogisches Bildungsangebot*. Abrufbar unter: <http://www.lag-avmb-bw.de/Themenfelder/Inklusion/Verordnung-uber-sonderpadagogische-Bildungsangebote---SBA-VO---SM-BW-2016-.pdf> [2. 7.2019].
- Schaefer, R. I., Goos, M. & Goeppert, S. (2017). *Online-Lehrbuch Medizinische Psychologie*. http://www.medpsych.uni-freiburg.de/OL/body_testgutekriterien.html. [3. 11.2017].
- Schermelleh-Engel, K. & Werner, C. (2007). Methoden der Reliabilitätsbestimmung. In: Moosbrugger, H. & Kelava, A. (Hrsg.): *Testtheorie und Fragebogenkonstruktion* (S. 114–133). Heidelberg: Springer.
- Schlee, J. (2008). 30 Jahre „Förderdiagnostik“ – eine kritische Bilanz. *Zeitschrift für Heilpädagogik*, 4, S. 122–131.
- Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. *Psychological Bulletin*, 124, S. 262–274.
- Schmidt-Atzert, L. & Amelang, M. (2012). *Psychologische Diagnostik (Lehrbuch mit Online-Materialien)* (5. Aufl.). Berlin: Springer Science & Business Media.
- Schmukle, S. & Schulze, R. (2016). TBS-TK Rezension: „WISC-IV. Wechsler Intelligence Scale for Children – Forth Edition“, *Psychologische Rundschau*, 67, S. 158–160.

- Schneider, W. J. & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment. Theories, tests, and issues*. (3. Aufl.) (S. 99–144). New York: Guilford Press.
- Schröter, G. (1981). *Zensuren? Zensuren! Allgemeine und fachspezifische Probleme: Grund-erkenntnisse und neue Forschungsergebnisse für Lehrer, Eltern und interessierte Schüler* (3. erw. Aufl.). Schneider Baltmannsweiler: Pädagogischer Verlag Burgbücherei Schneider.
- Schroth, J. (2013). SON-R 6–40 – Non-verbaler Intelligenztest (PSYNDEX Tests Review). Abrufbar unter: <https://www.zpid.de/retrieval/PSYNDEXTests.php?id=9006517> [4. 8. 2017].
- Schroth, J. (2015). WNV (D) – Wechsler Nonverbal Scale of Ability – deutsche Bearbeitung (PSYNDEX Tests Review). Abrufbar unter: <https://www.zpid.de/retrieval/PSYNDEXTests.php?id=9006789> [10. 11. 2017].
- Schuenger, J. M. & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology*, 45 (2), S. 294–302.
- Schulam für die Stadt Dortmund (Hrsg.) (2010). *Leitfaden Sonderpädagogische Förderung im Gemeinsamen Unterricht und in der Förderschule*. Abrufbar unter: http://www.gsbnet.de/pdf/Leitfaden_Sonderpaedagogische_Foerderung_2010.pdf [10. 8. 2019].
- Schulämter für Bochum und Herne (Hrsg.) (o. J.). *Leitfaden zum Feststellungsverfahren gemäß AO-SF und zum Gemeinsamen Lernen*. Abrufbar unter: [https://www.bochum.de/C12571A3001D56CE/vwContentByKey/W2A3FFMR369BOCMDE/\\$FILE/AO_SF_Leitfaden.pdf](https://www.bochum.de/C12571A3001D56CE/vwContentByKey/W2A3FFMR369BOCMDE/$FILE/AO_SF_Leitfaden.pdf) [10. 8. 2019].
- Schumann, B. (2013). *Inklusive Bildung braucht inklusive Diagnostik*. Abrufbar unter: <https://bildungsklick.de/schule/detail/inklusive-bildung-braucht-inklusive-diagnostik> [15. 8. 19].
- Seidler-Brandler, U. (2002). Kramer-Test. In Brähler, E., Holling, H., Leutner, D. & Petermann, F. (Hrsg.): *Brickenkamp. Handbuch psychologischer und pädagogischer Tests* (3., vollst. überarb. u. erw. Aufl.) (S. 180–183). Göttingen: Hogrefe.
- Senat Berlin, Senatsverwaltung für Bildung, Jugend und Wissenschaft (Hrsg.) (2012). *Leitfaden zur Feststellung sonderpädagogischen Förderbedarfs an Berliner Schulen*. Ohne Verlag.
- Serpell, R. (1979). How specific are perceptual skills? A cross-cultural study of pattern reproduction. *British Journal of Psychology*, 70, S. 365–380.
- Sesin, C.-P. (2012). Sarrazins dubiose US Quellen. In Haller, M. & Niggeschmidt, M. (Hrsg.): *Der Mythos vom Niedergang der Intelligenz: Von Galton zu Sarrazin: Die Denkmuster und Denkfehler der Eugenik* (S. 27–48). Wiesbaden: Springer.
- Shuey, A. M. (1966). *The Testing of Negro Intelligence* (2., erw. Aufl.). New York: Social Science Press.
- Slate, J. R. & Jones, C. H. (1990). Identifying students' errors in administering the WAIS-R. *Psychology in the Schools*, 27, S. 83–87.
- Slate, J. R., Jones, C. H., Coulter, C. & Covert, T. L. (1992). Practitioners' administration and scoring of the WISC-R: Evidence that we do err. *Journal of School Psychology*, 30, S. 77–82.
- Snijders, J. T., Tellegen, P. J. & Laros, J. A. (1997). *Snijders-Oomen non-verbaler Intelligenztest SON-R 5½–17: Manual*. Göttingen: Hogrefe.
- Spearman, C. E. (1904). „General Intelligence“ objectively determined and measured. *American Journal of Psychology*, 15, S. 201–293.
- Speck, O. (2003). Die Ökonomisierung des Lebenswertes als Gefährdung behinderten Lebens. In Dederich, M. (Hrsg.), *Bioethik und Behinderung* (S. 104–139). Leipzig: Klinkhardt.
- Speckemeier, C. (2011). *Intelligenz und Wissen: Schlüsselkompetenzen im Kontext personal-politischer Eignungsdiagnostik; Validierungsstudie des Wissenstests START-W für Berufseinsteiger*. Dissertation Fachbereich Erziehungswissenschaft und Psychologie, Freie Universität Berlin. Abrufbar unter: <http://www.diss.fu-berlin.de/diss/servlets/MCRFileNode>

- Servlet/FUDISS_derivate_00000009612/Dissertation.pdf;jsessionid=FF7E2F1D8098A11B D3F3E69AB58C4319?hosts= [31.1.2015].
- Staatsministerium für Kultus Sachsen. (2005). *Handbuch zur Förderdiagnostik. Handlungs- und Arbeitsgrundlage zum Verfahren zur Feststellung des Sonderpädagogischen Förderbedarfs*. Abrufbar unter: https://www.schule.sachsen.de/download/download_bildung/foer-diagnostik.pdf [11.8.2019].
- Staud, E. & Staud, M. (2011). *Sonderpädagogik: Erkenntnisse der Hirnforschung und ihre Bedeutung für die Körperbehindertenpädagogik* (2. Aufl.). Norderstedt: BoD – Books on Demand.
- Stern, W. (1912). Die psychologischen Methoden der Intelligenzprüfung. In F. Schumann (Hrsg.), *Bericht über den 5. Kongress für Experimentelle Psychologie* (S. 1–109). Leipzig: Barth.
- Stern, W. (1914). *The psychological methods of testing intelligence*. Baltimore: Warwick & York.
- Sternberg, R. J. (1987). Intelligence. In Gregory, R. L. (Hrsg.), *The Oxford companion to the mind* (S. 375–383). New York: Oxford University Press.
- Sternberg, R. J. (1985). *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge: Cambridge University Press.
- Sternberg, R. J. (1986). *Intelligence applied*. San Diego: Harcourt Brace Jovanovich.
- Sternberg, R. J. & Detterman, D. K. (Hrsg.) (1986). *What is Intelligence?* Norwood, USA: Ablex.
- Strehle, W. (1961). Ein Binet-Test für Blinde. *Der Blindenfreund*, 81, S. 105–129.
- Süddeutsche Zeitung. (2017). *Förderschüler wider Willen*. Abrufbar unter: <http://www.sueddeutsche.de/bildung/schule-foerderschueler-wider-willen-1.3408588> [3.5.2017].
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2012). *SON-R 6–40 Non-verbaler Intelligenztest. 4. Überarbeitung des Sniijders-Oomen non-verbale Intelligenztests*. Technisches Manual. Göttingen: Hogrefe.
- Tellegen, P. J., Laros, J. A. & Petermann, F. (2007). *SON-R 2½–7. Non-verbaler Intelligenztest*. Göttingen: Hogrefe.
- Testzentrale (2017). *SON-R 5½–17*. Abrufbar unter: <https://www.testzentrale.de/shop/non-verbaler-Intelligenztest-49009.html> [3.8.2017].
- Tewes, U. (1983). *HAWIK-R*. Bern: Huber.
- Tewes, U., Rossmann, P. & Schallberger, U. H. (Hrsg.) (1999). *Hamburg-Wechsler-Intelligenztest für Kinder – Version III*. Bern: Huber.
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.
- Tiedemann, F. (1836). On the Brain of the Negro, Compared with that of the European and the Orang-Outang. *Philosophical Transactions of London*, 126, S. 497–527.
- Toepoel, V. (2008). *A Closer Look at Web Questionnaire Design*. Tilburg: Tilburg University Press.
- Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, 133 (5), S. 859–883.
- Tourangeau, R., Couper, M. P. & Conrad, F. G. (2004). Spacing, position and order. Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, 68, S. 368–393.
- Tourangeau, R., Rips, L. J. & Rasinski, K. A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.
- Universität Zürich. (2016). *Kruskal-Wallis-Test*. Abrufbar unter: https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/kruskal.html [18.8.19].
- Universität Zürich. (2018). *Friedman-Test*. Abrufbar unter: https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/friedman.html [19.10.2019].

- Van Dijk, H. & Tellegen, P. J. (2004). *NIO Nederlandse Intelligentietest voor Onderwijsniveau: Handleiding*. Amsterdam: Boom Test Uitgevers.
- Velden, M. (2013). *Hirntod einer Idee: Die Erblichkeit der Intelligenz*. Göttingen: Vandenhoeck & Ruprecht.
- Vernooij, M. A. (2013). *Sonderpädagogische Begutachtung. Thüringer Diagnostikkonzept zur Qualitätssicherung*. Abrufbar unter: https://ngu.jena.de/wp-content/uploads/2015/09/Sonderpädagogische_Begutachtung_Vernooij.pdf [11. 8. 2019].
- Walter-Busch, E. (1989). *Das Auge der Firma*. Stuttgart: Enke.
- Wechsler, D. & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability*. San Antonio, TX: Harcourt Assessment.
- Wechsler, D. (1939). *Wechsler-Bellevue Intelligence Scale*. New York: Psychological Corporation.
- Wechsler, D. (1956). *Die Messung der Intelligenz Erwachsener*. Bern: Huber.
- Wechsler, D. (1958). *The measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins.
- Wechsler, D. (1991). *Wechsler Intelligence Scale for Children – Third Edition*. San Antonio: Harcourt Assessment.
- Wechsler, D. (2003). *The Wechsler Intelligence Scale for Children – Fourth Edition*. Technical and interpretive manual. San Antonio: Harcourt Assessment.
- Wechsler, D. (2017). *WISC-V Durchführungs- und Auswertungsmanual. Deutsche Fassung der WISC-V Wechsler Intelligence Scale for Children – Fifth Edition*. Bearbeiter der deutschen Fassung F. Petermann. Frankfurt: Pearson.
- Weiss, R. H. & Osterland, J. (2013). *CFT 1-R. Grundintelligenztest Skala 1-Revision*. Göttingen: Hogrefe.
- Weiss, R. H. (2006). *CFT 20-R. Grundintelligenztest Skala 2-Revision*. Göttingen: Hogrefe.
- Werning, R. & Lichtblau, M. (2012). Sonderpädagogische Diagnostik. In Werning, R., Balgo, R., Palmowski, W. & Sassenroth, M. *Sonderpädagogik: Lernen, Verhalten, Sprache, Bewegung und Wahrnehmung* (S. 229–260). München: Oldenbourg Verlag.
- Willerman, L., Schultz, R., Rutledge, J.N. & Bigler, E. (1991). In vivo brain size and intelligence. *Intelligence*, 15, S. 223–228.
- Wolf, J. (1999). *SON-R 5½–17 – Snijders-Oomen Non-Verbaler Intelligenztest*. PSYINDEX Tests Review. Abrufbar unter: <https://www.zpid.de/retrieval/PSYINDEXTests.php?id=9003771> [26. 05. 2018].
- Woodcock, R. W., McGrew, K. S. & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.
- Zaborowski, K. U., Meier, M. & Breidenstein, G. (2011). *Leistungsbewertung und Unterricht: Ethnographische Studien zur Bewertungspraxis in Gymnasium und Sekundarschule* (1. Aufl.). Berlin: Springer-Verlag.
- Ziegler, M., Schmidt-Atzert, L., Bühner, M. & Krumm, S. (2007). Fakability of Different Measurement Methods for Achievement Motivation: Questionnaire, Semi-projective, and Objective. *Psychology Science Quarterly*, 49 (4), S. 291–307.
- Zimbardo, P. G. (1992). *Psychologie* (5. Aufl.). Berlin: Springer-Verlag.