TECHNISCHE UNIVERSITÄT
CHEMNITZ

UNTERSUCHUNGEN ZUR PRÄDIKTIVEN VALIDITÄT VON KONZENTRATIONSTESTS

EIN CHRONOMETRISCHER ANSATZ ZUR ÜBERPRÜFUNG DER ROLLE VON ITEMSCHWIERIGKEIT, TESTLÄNGE UND TESTDIVERSIFIKATION

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt der

Fakultät für Human - und Sozialwissenschaften

der Technischen Universität Chemnitz

von

Dipl.-Psych. Frank Schumann

geboren am 06.04.1976 in Zwickau

Chemnitz

2015

Tag der Einreichung: 02.11.2015

Tag der Verteidigung: 06.06.2016

Erstgutachterin: Prof. Dr. Anja Strobel

Zweitgutachter: Prof. em. Dr. Karl Westhoff

Inhaltsverzeichnis

Zusammenfassung **XI**

1 Einführung und Ziele **1**

2 Diagnostik von Konzentrationsfähigkeit **10**

2.1 Historische Einordnung 10

2.2 Kognitive Modellierung 24

2.3 Psychometrische Modellierung 39

3 Prädiktive Validität von Konzentrationstests **57**

3.1 Reliabilität, Konstruktvalidität, Kriterienvalidität 57

3.2 Konstruktions- und Validierungsstrategien 62

3.3 Ableitung der Fragestellung 71

4 Beschreibung der Fragebögen und Tests **81**

5 Empirischer Teil **93**

5.1 Studie 1 - Itemschwierigkeit 93

5.1.1 Methode 93

5.1.2 Ergebnisse 95

5.1.3 Diskussion 99

5.2 Studie 2 - Testverlängerung und Testverlauf 99
5.2.1 Methode 99
5.2.2 Ergebnisse 101
5.2.3 Diskussion 105
5.3 Studie 3 - Testdiversifikation 106
5.3.1 Methode 106
5.3.2 Ergebnisse 109
5.3.3 Diskussion 113
5.4 Studie 4 - Validität in realer Auswahlsituation (I-S-T 2000 R) 114
5.4.1 Methode 114
5.4.2 Ergebnisse 117
5.4.3 Diskussion 121
5.5 Studie 5 - Validität in realer Auswahlsituation (WIT-2) 122
5.5.1 Methode 122
5.5.2 Ergebnisse 124
5.5.3 Diskussion 126

6 Diskussion **128**
6.1 Sind schwierige Tests besser als leichte Tests? 135
6.2 Sind lange Tests besser als kurze Tests? 142
6.3 Sind Testbatterien besser als Einzeltests? 150
6.4 Sind Tests auch unter „realen“ Bedingungen valide? 155
6.5 Validität unter realen Bedingungen - Generalisierung 157

7 Theoretische Implikationen **160**

8 Praktische Konsequenzen **165**

9 Literaturverzeichnis **170**

Anhang **XIII**

Tabellenverzeichnis

5.1 Ergebnisse der Studie 1: Korrelationen zwischen standardisierten Leistungstests und Schulnoten 96
5.2 Ergebnisse der Studie 1: Korrelationen zwischen Reaktionszeit im SMACT und Schulnoten als Funktion von niedriger versus hoher Itemschwierigkeit und Itemhomogenität 97
5.3 Ergebnisse der Studie 2: Korrelationen zwischen klassischen Leistungstests und Schulnoten 102
5.4 Ergebnisse der Studie 2: Korrelationen zwischen Reaktionszeit im SMACT und Schulnoten als Funktion von Testlänge (kurz versus lang) und Testverlauf (Abschnitt 1-4) 103
5.5 Ergebnisse der Studie 3: Korrelationen zwischen klassischen Leistungstests und Schulnoten 109
5.6 Ergebnisse der Studie 3: Korrelationen zwischen klassischen Leistungstests und Schulnoten 110
5.7 Ergebnisse der Studie 3: Korrelationen zwischen Leistungsmotivation und Schulnoten 111
5.8 Ergebnisse der Studie 4: Korrelationen zwischen klassischen Leistungstests und Schulnoten im Kontext einer realen Bewerberauslesesituation an der Hochschule Mittweida 118

5.9 Ergebnisse der Studie 5: Korrelationen zwischen klassischen Leistungstests und Schulnoten im Kontext einer realen Personalauslesesituation 125

Abbildungsverzeichnis

2.1 Drei verschiedene Testvarianten von nach dem Bourdon-Prinzip konstruierten Durchstreich-Konzentrationstests. 16
2.2 Drei Beispiele von Rechenkonzentrationstests, die nach dem Bourdon-Prinzip konstruiert wurden. 23
4.1 Beispielitems im Komplexen Konzentrationstest (KKT). 83
4.2 Beispielitems aus dem Serialen Mentalen Additions- und Komparationstest (SMACT). 84
4.3 Beispielitems im Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II). 86
5.1 Darstellung der Korrelationen der aggregierten Parameter des I-S-T 2000 R als Vergleich der Experimentalsituation (Studie 3) und der realen Bewerbersituation (Studie 4). 119
5.2 Darstellung der Korrelationen der Subtests des I-S-T 2000 R als Vergleich zwischen Experimentalsituation (Studie 3) und realer Bewerbersituation (Studie 4). 120

Abkürzungen und Definitionen

Allgemein

Hz	Hertz
cm	Zentimeter
ms	Millisekunde
RT	mittlere Reaktionszeit (ms)
RSI	Reaktions-Stimulus-Intervall

Leistungskennwerte

GZ	Gesamtzahl bearbeiteter Zeichen
F%	Fehlerprozentwert (in Konzentrationstests)
KL	Konzentrations-Leistung
KG	Konzentrations-Genauigkeit
KH	Konzentrations-Homogenität

Statistische Abkürzungen

M	Mittelwert
SD	Standardabweichung
r	Korrelationskoeffizient
R	multipler Korrelationskoeffizient
R^2	Bestimmtheitsmaß
p	Irrtumswahrscheinlichkeit

Danksagung

Zum Gelingen der vorliegenden Arbeit haben mehrere Personen entscheidend beigetragen, denen ich an dieser Stelle herzlich danken will. Frau Prof. Dr. Strobel danke ich ganz herzlich für die unkomplizierte Betreuung als externer Doktorand und für die wertvollen Anregungen und Hinweise sowie für die Übernahme des Gutachtens zur Dissertation. Besonders bedanke ich mich bei Herrn Prof. Dr. Karl Westhoff, der mir mit seinem umfangreichen Wissen und seiner langjährigen Erfahrung ein hochgeschätzter Ansprechpartner gewesen ist und mir von Anbeginn beratend zur Seite gestanden hat. Seine Überlegungen stellen das Fundament der vorliegenden Arbeit dar. Insbesondere seine Arbeiten zu den Facetten von Leistungskennwerten in Konzentrationstests (Geschwindigkeit, Genauigkeit, und Leistungsschwankung) bilden die theoretische Grundlage für die vorliegende Dissertation. Herrn Prof. Dr. Westhoff danke ich zudem auch für die Übernahme des Gutachtens zur Dissertation.

Darüber hinaus danke ich Herrn Dr. Hagen C. Flehmig, der stets hilfsbereit war und zahlreiche wertvolle Hinweise zu methodischen Fragen und aktuellen Entwicklungen in der Testdiagnostik gegeben hat. Ebenfalls gebührt Herrn Dr. Michael Steinborn besonderer Dank, der ebenfalls viele Hinweise gegeben hat, die insbesondere deshalb sehr nützlich waren, weil er als Kognitionspsychologe einen sehr detaillierten Überblick über Informationsverarbeitungsprozesse und deren Modellierung besitzt und ich von diesem Wissen sehr profitieren konnte. Von allen genannten Personen konnte ich sehr viel lernen und ich bin sehr dankbar über die mir gewährte Unterstützung

und Freundschaft.

Frau Prof. Dr. Ulla Meister von der Hochschule Mittweida danke ich herzlich dafür, dass sie mich mit sanftem Druck stets zum Weitermachen motivierte und mir auch die nötigen Rahmenbedingungen schuf, die ich für die Fertigstellung der hier vorliegenden Arbeit benötigte. Besonders dankbar bin ich meiner Frau, Bianca Schumann, die mir über den gesamten Zeitraum der Datenerhebung und des Schreibens der Dissertation den Rücken frei hielt und währenddessen allein die Betreuung unseres Sohnes übernahm.

Zusammenfassung

In der hier vorliegenden Arbeit wurde die Validität von Aufmerksamkeits- und Konzentrationstests untersucht. Im Vordergrund stand dabei die Frage nach dem Einfluss verschiedener kritischer Variablen auf die prädiktive Validität in diesen Tests, insbesondere der Itemschwierigkeit und Itemhomogenität, der Testlänge bzw. des Testverlaufs, der Testdiversifikation und der Validität im Kontext einer echten Personalauslese. In insgesamt fünf Studien wurden die genannten Variablen systematisch variiert und auf ihre prädiktive Validität zur (retrograden und konkurrenten) Vorhersage von schulischen und akademischen Leistungen (Realschule, Abitur, Vordiplom/Bachelor) hin analysiert. Aufgrund der studentischen (d. h. relativ leistungshomogenen) Stichprobe bestand die Erwartung, dass die Korrelationen etwas unterschätzt werden. Da die Validität in dieser Arbeit jedoch „vergleichend" für bestimmte Tests bzw. experimentelle Bedingungen bestimmt wurde, sollte dies keine Rolle spielen. In Studie 1 (N = 106) wurde zunächst untersucht, wie schwierig die Items in einem Rechenkonzentrationstest sein sollten, um gute Vorhersagen zu gewährleisten. Dazu wurden leichte und schwierigere Items vergleichend auf ihre Korrelation zum Kriterium hin untersucht. Im Ergebnis waren sowohl leichte als auch schwierigere Testvarianten ungefähr gleich prädiktiv (r = .30** bis .37**). In Studie 2 (N = 103) wurde die Rolle der Testlänge untersucht, wobei die prädiktive Validität von Kurzversion und Langversion in einem Rechenkonzentrationstest vergleichend untersucht wurde. Im Ergebnis zeigte sich, dass die Kurzversion valider war als die Langversion und dass die Validität in der

Langversion im Verlauf abnimmt. In Studie 3 (N = 388) stand der Aspekt der Testdiversifikation im Vordergrund, wobei untersucht wurde, ob Intelligenz besser mit einem einzelnen Matrizentest (Wiener Matrizen-Test, WMT) oder mit einer Testbatterie (Intelligenz-Struktur-Test, I-S-T 2000 R) erfasst werden sollte, um gute prädiktive Validität zu gewährleisten. Die Ergebnisse sprechen klar für den Matrizentest, welcher ungefähr gleich valide war wie die Testbatterie, aber dafür testökonomischer ist. In den Studien 4 (N = 105) und 5 (N =97) wurde die prädiktive Validität zur Vorhersage von Schulleistungen im Kontext einer realen Personalauswahlsituation untersucht. Während die großen Testbatterien, Wilde-Intelligenz-Test 2 (WIT-2) und Intelligenz-Struktur-Test 2000R (I-S-T 2000 R), nur mäßig gut vorhersagen konnten, war der Komplexe Konzentrationstest (KKT), insbesondere der KKT-Rechentest ein hervorragender Prädiktor für schulische und akademische Leistungen. Auf Basis dieser Befunde wurden schließlich Empfehlungen und Anwendungshilfen für den strategischen Einsatz von Testinstrumenten in der diagnostischen Berufspraxis ausgesprochen.

Abstract

In the present work, the predictive validity of psychometric tests of attention and concentration was investigated. The focus was on the effect of several critical experimental variables on predictive validity, particularly item difficulty and item homogeneity, test length and performance fluctuation across tests, diversity of tests and the validity in the context of real personnel assessment (selection and classification) situations. In five studies, these variables were systematically varied and analyzed for their validity (backward and concurrent) to predicting school and academic performance (secondary school, high school, and bachelor). Due to the use of student-based samples (and the restricted range of variance, respectively), it was expected that the height of correlations will be somewhat underestimated in the present experiments. Since the predictive validity was comparably assessed for different psychometric tests or experimental conditions, limitations in sample variance would not hamper the interpretation of our results. In study 1 (N = 106) it was examined whether the predictive validity would be better for easy or difficult items in the self-paced mental addition and comparison task. The correlation obtained for these experimental conditions were then comparatively evaluated with respect to their correlation to school achievement. As a result, both easy and difficult items were equally predictive (r = .30** .37** up). In study 2 (N = 103), the role of test length was examined, comparing the predictive validity of a short and a long version of the self-paced mental addition and comparison task. As a result, the short version yielded slightly better correlations to school grades than the long version,

indicating that the test length of the short version was already sufficient in length than that of the long version. In study 3 (N = 388) aspects of test diversification was examined, comparing a rather short matrices test (Wiener Matrices-Test, WMT) with a rather long test battery (Intelligence-Structure-Test 2000R), both being constructed to measure general cognitive ability. The results indicate that predictive validity was rather similar, which would render the WMT as the better choice, because this instrument is more economic with respect to time constraints. In study 4 (N = 105) and study 5 (N = 97), the predictive validity of psychometric test devices were examined in the context of a real personnel selection situation, which might be more stressful and more motivating for participants as compared to the typical situation in the psychological laboratory. Whereas the psychometric test batteries, Wilde-Intelligenz-Test 2 (WIT-2) and Intelligence-Structure-Test 2000R (I-S-T 2000 R), were only modestly effective, the Complex Concentration Test (KKT) appeared to being an excellent predictor of school and academic achievement. Based on these findings, several recommendations were provided for practitioners in the field of school and personnel psychology.

1 Einführung und Ziele

Aufmerksamkeits- und Konzentrationstests gelten als wichtige Verfahren zur Messung der allgemeinen kognitiven Leistungsfähigkeit und werden sowohl für akademische Forschungsfragen als auch Fragestellungen im Kontext anwendungsorientierter psychologischer Diagnostik eingesetzt. Damit sind sie unentbehrliche Instrumente in zahlreichen Feldern der Eignungsdiagnostik. Im Bereich der Schulpsychologie können Konzentrationstests ein Bestandteil einer umfangreicheren Diagnostik sein und dabei helfen, das Leistungspotential von Schülern abzuschätzen. Derzeit gibt es circa 1500 praktizierende Schulpsychologen in der Bundesrepublik Deutschland, die etwa 120000 diagnostische Untersuchungen jährlich durchführen. Dies zeigt, wie wichtig es ist, dass qualitativ hochwertige Tests zur Verfügung stehen, um eine valide Diagnostik gewährleisten zu können. Konzentrationstests sind aber auch für zahlreiche Felder innerhalb der neuropsychologischen Diagnostik sehr relevant. Obwohl die Instrumente in diesem Bereich oft zur Diagnostik spezifischer kognitiver Defizite ausgerichtet sind, können Konzentrationstests einen Beitrag leisten, um die allgemeine kognitive Leistungsfähigkeit bzw. Funktionsfähigkeit von Patienten abzuschätzen (vgl. Sturm & Willmes, 2001; Sturm, Willmes, Orgass, & Hartje, 1997). Die Testbatterie zur Aufmerksamkeitsprüfung (TAP) von Sturm und Zimmermann (2000) besteht zum Beispiel aus verschiedenen Untertests, die gemäß der mehrdimensionalen Auffassung von Aufmerksamkeitskomponenten konstruiert wurde. Aber auch in weiteren Feldern berufspraktischer Diagnostik können Konzentrationstests hilfreich sein, so zum Beispiel

zur Beurteilung der Verkehrstüchtigkeit oder zur Auswahl von Bewerbern innerhalb eines Personalauswahlverfahrens in großen Organisationen (Westhoff & Hagemeister, 2005; Schmidt-Atzert & Deter, 1993).

Diese Dissertation verfolgt das Ziel, den Wissensstand im Bereich der angewandten Testdiagnostik mit aktuellen kognitionspsychologischen Befunden anzureichern und auch zu aktualisieren, so dass im Ergebnis eine Reihe von praktischen Empfehlungen für die Konstruktion von zukünftigen Testverfahren, insbesondere von Konzentrationstests, gegeben werden können. Um dieses Ziel zu erreichen, wurde die wissenschaftliche Literatur in den genannten Bereichen umfangreich gesichtet und ausgewertet. Die Daten von fünf experimentellen Studien sollen einen empirischen Beitrag leisten. Im Vordergrund stehen sowohl die diagnostischen Konstrukte „Aufmerksamkeit“ und „Konzentration“ als auch deren Operationalisierung in messbare Tests und schließlich deren prädiktive Validität bei der Vorhersage von bestimmten Außenkriterien (z. B. Schulleistungen). Ein fachkundiger Leser wird sich an dieser Stelle fragen müssen: Ist diese Sache nicht längst empirisch geklärt? Tatsächlich gibt es bereits zahlreiche Studien zu dieser Fragestellung. Ein umfangreicher Körper an empirischen Befunden legt unter anderem nahe, dass sich Leistungen in der Schule (Schulnoten) am besten durch Leistungen in Intelligenztests vorhersagen lassen. Viele Autoren haben sich jedoch gefragt, ob und inwiefern Testleistungen für spezifischere Fähigkeiten, insbesondere in Aufmerksamkeits- und Konzentrationstests, einen inkrementellen Beitrag zur prädiktiven Validität leisten können. Insofern sind auch noch viele Fragen offen. Die vorliegende Arbeit soll einen Beitrag zur Beantwortung einiger dieser offenen Fragen leisten. Dabei liegt der Schwerpunkt nicht auf der Anwendung von handelsüblichen Testverfahren, sondern auf der experimentellen Analyse kritischer Variablen mit potentiellem Einfluss auf die Validität eines Tests.

Aus theoretischen Modellen über Testreliabilität und -validität und auch aus

Modellen über sich akkumulierende mentale Ermüdung bzw. psychische Sättigung wurden Hypothesen über den Einfluss von Testverlauf, Testlänge und Itemschwierigkeit auf die prädiktive Validität eines Tests zur Vorhersage von Schulleistungen abgeleitet und empirisch überprüft (Smit & Van der Ven, 1995; Van Breukelen et al., 1996; Westhoff & Kluck, 1984). Unter praktisch arbeitenden Diagnostikern wird gern über diese Faktoren philosophiert und man kommt schnell zu ad-hoc-Hypothesen, ohne diese jemals zu überprüfen. In der Literatur findet sich zum Beispiel oft die These, dass die prädiktive Validität (hinsichtlich eines bestimmten Kriteriums) besser wird, wenn Probanden unter abträglichen oder widrigen Bedingungen getestet werden (z. B. unter Stress, Ermüdung etc.), jedoch findet sich kaum empirische Unterstützung für eine solche Behauptung (s. Moosbrugger & Goldhammer, 2006; Westhoff & Hagemeister, 2005). Außerdem liest man nicht minder selten, dass sich die prädiktive Validität eines Tests mit zunehmender Testlänge (d. h. Anzahl der Items) monoton erhöht, was sich aus den Annahmen zur klassischen Testtheorie direkt ableitet. Auch die These, dass die prädiktive Validität mit zunehmender Itemschwierigkeit steigt, ist bislang weitgehend ungeprüft, zumindest gibt es keine systematisch angelegten Studien. Miller und Ulrich (2013) nehmen an, dass selbst die Reliabilität eines Tests von einer Vielzahl von Einflussfaktoren abhängt und daher immer einer empirischen Überprüfung im individuellen Fall bedarf. Die vorliegenden Analysen schließen sich der Schlussfolgerung der Autoren an, indem sie zeigen, dass die Realität komplex ist und dass ein tiefgehendes Verständnis kognitionspsychologischer Mechanismen für die Konstruktion von (auf Reaktionszeiten basierenden) psychometrischen Leistungstests unbedingt erforderlich ist.

Eine Analyse des Testverlaufs von Konzentrationstests diente der Beantwortung der (zunächst praktischen) Frage, ob sich die prädiktive Validität während der andauernden Bearbeitung des Tests ändert. Führt Testermüdung zu einer Verbesserung der

prädiktiven Validität oder zu einer Verschlechterung derselben? Aus theoretischer Sicht stellt sich die Frage, welches Prozessmodell gelten sollte, um den Einfluss von Ermüdungsprozessen auf individuelle Differenzen in der Testleistung beschreiben zu können. Einerseits lässt sich leicht vorhersagen, dass Probanden mit zunehmender Ermüdung besser differenziert werden können, sofern angenommen wird, dass die leistungsfähigen Probanden robuster gegenüber den rasch eintretenden Ermüdungs- und Sättigungsprozessen sind. Andererseits lässt sich aber auch plausibel vorhersagen, dass Leistungsunterschiede zwischen Probanden durch Ermüdung nivelliert werden, sofern angenommen wird, dass ermüdungsbedingte Leistungseinbußen asymmetrisch stärker für die leistungsstärkeren Probanden ausfallen, verglichen zu den leistungsschwächeren Probanden einer Stichprobe. Die diagnostische Forschung hat dafür keine Antworten vorrätig, weshalb man sich im Forschungsfeld der grundlagenorientierten Daueraufmerksamkeitsforschung umsehen muss, um Antworten zu finden. Langner und Kollegen (Langner, Eickhoff, & Steinborn, 2011; Langner, Steinborn, Chatterjee, Sturm, & Willmes, 2010) führen eine Reihe von aufgabenspezifischen und situationsspezifischen Faktoren an, welche bei der Aufstellung von Hypothesen über ermüdungsbedingte Leistungsunterschiede eine wichtige Rolle spielen. Wenn man von Ermüdung spricht, muss man natürlich auch unterscheiden, ob es sich um durch „time-on-task“ induzierte Ermüdung handelt (Steinborn, Flehmig, Westhoff, & Langner, 2010) oder um z. B. durch Schlafdeprivation induzierte Änderungen kognitiver Leistungsfähigkeit (Bratzke, Rolke, Steinborn, & Ulrich, 2009).

Eine Analyse von Effekten der Testlänge dient der Beantwortung der Frage, ob sich eine Akkumulation von Items verbessernd hinsichtlich der prädiktiven Validität auswirkt, wie es aus der klassischen Testtheorie ableitbar ist, oder ob sich das zu messende Merkmal während langer Testungen ändert, wie es aus einem kognitiv-energetischen Modell heraus ableitbar wäre (z. B. Hockey, 1997; Langner & Eickhoff,

2013; Sanders, 1998). Daraus ergibt sich eine ganz praktische Frage: Wie lang sollte ein Testverfahren sein, damit es noch reliabel genug ist, um valide sein zu können, aber dennoch ökonomisch ist? Die Analyse der Itemschwierigkeit dient schließlich der Beantwortung der Frage nach der optimalen Schwierigkeit in Konzentrationstests einer bestimmten Aufgabenform. Führt eine Erhöhung der Itemschwierigkeit (einhergehend mit einer höheren Arbeitsgedächtnisbelastung) in einem psychometrischen Test zur Messung von Konzentrationsfähigkeit zu einer Verbesserung der prädiktiven Validität hinsichtlich eines bestimmten Kriteriums (z. B. Schulnoten)? Im Unterschied zu obengenannten Hypothesen über Ermüdungseinflüsse lassen sich Hypothesen über Testlängeneffekte nur in eine Richtung formulieren: Je länger der Test, desto höher sollten die Gütekriterien Reliabilität und Validität ausfallen (vgl. Miller & Ulrich, 2013; Steinborn, Flehmig, Westhoff, & Langner, 2008; Westhoff & Hagemeister, 1992), was einen Vergleich von gleichem Aufgabenmaterial bei jeweils kurzer und langer Testform impliziert. Interessant ist hierbei jedoch vor allem die Frage, wann ein Test genügend lang ist, sodass er hinreichend reliabel und valide ist und welche Rolle die Wahl der Leistungskennwerte (Geschwindigkeit, Genauigkeit, und Variabilität) dabei spielt (vgl. Flehmig, Steinborn, Langner, Scholz, & Westhoff, 2007; Steinborn et al., 2010; Van Breukelen et al., 1996; Westhoff & Kluck, 1984).

Eine Analyse der Itemschwierigkeit in Konzentrationstests dient weiterhin der Beantwortung der Frage, welcher Grad an Schwierigkeit optimale Werte bezüglich Reliabilität und Validität eines Tests hervorbringt. Damit zusammenhängend stellt sich auch die Frage, mit welchem Aufgabenmaterial sich Schwierigkeit am besten experimentell manipulieren lässt. Klassische Verfahrensweisen implizieren eine Erhöhung der Anzahl der Reiz- oder Reaktionsalternativen als Indikator der Itemschwierigkeit (vgl. Westhoff & Graubner, 2003; Westhoff & Kluck, 1984), während aktuelle Arbeiten eine Variierung der Problemgröße in Additionsaufgaben empfehlen (vgl. Steinborn

et al., 2008; Steinborn et al., 2010), wobei nicht nur die Problemgröße (ein Kennwert für die Schwierigkeit von Additionsaufgaben) sondern auch die Variationsbreite der Schwierigkeit innerhalb eines Aufgabensets von Bedeutung ist. In der hier vorliegenden Arbeit wurde daher überprüft, ob eine simultane Variation der Itemschwierigkeit und der Itemhomogenität zu einer Verbesserung der Gütekriterien Reliabilität und Validität führt. Da anzunehmen ist, dass mit zunehmender Itemschwierigkeit die Korrelationen zum Intelligenzkonstrukt zunehmen, war es außerdem von großer Bedeutung, ob jegliche Effekte auf die Validität inkrementeller Natur sind (d. h. sie leisten einen zusätzlichen Beitrag zum Intelligenztest), oder ob diese lediglich aufgrund einer höheren Redundanz zum Intelligenzkonstrukt entstehen (Campbell & Fiske, 1959). Eine Besonderheit der hier vorliegenden Arbeit ist jedoch, dass die Einflüsse der Itemschwierigkeit um sogenannte Mixkosteneffekte bereinigt werden können, da hier leichte und schwierigere Items jeweils in homogenen (d. h. in geblockten) und heterogenen (d. h. in gemixten) Testeinheiten überprüft wurden (Los, 1999a, 1999b; Los, Knol, & Boers, 2001).

Wenn man einen Test mit dem Ziel verlängern will, dessen Gütekriterien damit zu verbessern, dann stellt sich die Frage, ob man dies mit gleichartigen Items oder mit andersartigen Items erreichen möchte. Im ersten Fall handelt es sich um eine klassische Variante der Testverlängerung (mit homogenen Items), im zweiten Fall handelt es sich stets auch um „Testverlängerung durch Testdiversifikation“. Aus theoretischer Sicht ist eine Methodentriangulation diagnostischer Konstrukte durchaus wünschenswert, sofern sich dadurch eine robustere Erfassung des zu messenden Merkmals ergibt, und sofern die Kosten (d. h. der zeitliche und materielle Umfang) einer derartigen diagnostischen Strategie nicht überhandnehmen. Die Kosten müssen noch im Verhältnis zum erreichten Nutzen stehen. So klar diese Aussage nämlich ist, so schwierig ist es tatsächlich einzuschätzen, ob ein bestimmter Zuwachs an prädiktiver Validität durch

einen bestimmten Kostenfaktor noch gerechtfertigt werden kann. Für den Fall der Reliabilität haben Miller und Ulrich (2013) eine entsprechende Strategie vorgeschlagen, indem sie die funktionale Beziehung zwischen Itemmenge und Reliabilität durch eine psychometrische Funktion beschreiben, wobei die optimale Itemmenge durch Bestimmung des „Knickpunkts“ der gegen eine Asymptote verlaufenden Funktion bestimmt wird. In der vorliegenden Arbeit wird dieser Gedanke auf Validitätsaspekte generalisiert. Aufbauend auf den Arbeiten und Gedanken von Westhoff und Graubner (2003) wurde geprüft, ob die Validität einer Testbatterie besser durch „Diversifikation von Tests eines Konstrukts“ oder durch „spezifische Wahl eines Konstrukts“ optimiert werden kann (vgl. Flehmig et al., 2007; Schmidt-Atzert, Bühner, & Enders, 2006).

Schließlich wurde auch der Versuch unternommen, die prädiktive Validität von Konzentrationstests unter realen Bedingungen zu bestimmen. Dazu wurde eine diagnostische Strategie innerhalb einer echten Eignungsfeststellung verwirklicht. Es stellte sich hier die Frage, ob eine echte Prüfungssituation zu einer Differenzierung oder zu einer Nivellierung von individuellen Differenzen führt. Im ersteren Fall sollten sich die Gütekriterien verbessern, während sie sich im zweiten Fall verschlechtern sollten. Auch ein komplexes Ergebnismuster wurde in Erwägung gezogen. So zum Beispiel wäre es denkbar, dass bestimmte kognitive Fähigkeiten (z. B. einfache Wahlreaktionen) unter realen Bedingungen besser differenziert werden können, während andere (z. B. komplexere Denkleistungen) sogar nivelliert werden. Da die Literatur keine Antworten (bezüglich psychometrischer Gütekriterien) bereithält, war es eine empirische Frage, die theoretisch motiviert war. Ausgehend von kognitiv-energetischen Modellen, wie sie von Sanders (1998, Kapitel 9), Hockey et al. (Hockey, 1997; Hockey & Earle, 2006), und auch von Langner (Langner & Eickhoff, 2013; Langner et al., 2011) formuliert wurden, wurden Hypothesen über den Effekt von realem Prüfungsstress auf individuelle Leistungsunterschiede formuliert. Eine solche Fragestellung impliziert

einen Vergleich von Testleistungen in lebensechter und artifizieller Situation. Dazu wurden zwei Datensätze verglichen, welche einander identische Testleistungen enthalten, sodass ermittelt werden kann, ob sich Testleistungen in realer versus artifizieller Situation unterscheiden, und falls ja, ob sich spezifische Ergebnismuster herausstellen.

Um die oben genannten Ziele verwirklichen zu können, wurden fünf Studien durchgeführt, in denen jeweils eine große Stichprobe (von 97 bis 388 Probanden) hinsichtlich korrelativer Beziehungen zwischen Tests untersucht wurde. Während die Hauptfragestellungen mittels experimentalpsychologischer Methoden untersucht wurden, wie sie von Ulrich, Miller, und Kollegen (Miller & Ulrich, 2013; Ulrich & Miller, 2008; Ulrich & Miller, 1994) im Rahmen des chronometrischen Ansatzes beschrieben werden, so wurden auch Beziehungen zwischen handelsüblichen Testverfahren ausgewertet. Die prädiktive Validität von bekannten Konzentrationstests (z. B. dem Test d2) wurde in diesem Sinne „en passant“ ermittelt. Für die Hauptfragestellungen (d. h. Rolle von Testverlauf, Testlänge, und Itemschwierigkeit die prädiktive Validität von Konzentrationstests) kamen jedoch Varianten des Serialen Mentalen Additions- und Komparationstests (SMACT) zum Einsatz, da sich diese Aufgabenform hervorragend für chronometrische Fragestellungen eignet (dazu siehe Methodenteil). Ein weiteres (jedoch nachgeordnetes) Ziel bestand darin, berichtete Zusammenhänge zwischen objektiv (durch Tests) und subjektiv (durch Selbstbericht im Alltag) ermittelten Leistungen auf ihre Wiederholbarkeit hin zu überprüfen. Insbesondere ging es hierbei um selbstberichtete kognitive Fehlleistungen, die im alltäglichen Leben von Personen auftreten können, und mit Instrumenten wie dem CFQ (Broadbent, Cooper, Fitzgerald, & Parkes, 1982) oder dem SIML (Reason & Lucas, 1984) gemessen werden können.

Schlussendlich sollten die Ergebnisse der erwähnten Analysen dazu dienen, einen praktischen Beitrag zur psychologischen Konzentrationsdiagnostik zu leisten. Deshalb war das erklärte Ziel der vorliegenden Studien, konkrete Hinweise und Ver-

fahrensregeln für den in der beruflichen Praxis tätigen psychologischen Diagnostiker geben zu können. Dieses Ziel motiviert sich zunächst aus der Selbstdefinition einer an der Schnittstelle zwischen Grundlagenforschung und Praxis liegenden Disziplin, wird doch die psychologische Diagnostik als jenes Fach ausgewiesen, welches anwendbare Technologien für praktische Berufsfelder im Bereich der Personalauslese oder im schulpsychologischen Dienst liefert. Entsprechend sollen auf Basis der vorliegenden Analysen allgemeine Empfehlungen über die Wahl von Testverfahren und auch über die optimale Position in der Reihenfolge subsequent zu administrierender Testverfahren gegeben werden können. Der hauptsächliche Adressat der vorliegenden Forschungsergebnisse ist jedoch nicht der in praktischen Berufsfeldern arbeitende Psychologe, sondern es ist der praktische Technologien herstellende Testentwickler, an welchen sich die Empfehlungen richten. Für zukünftige Testentwicklungen im Bereich der Konzentrationsdiagnostik soll hier eine Empfehlung über folgende Problemfelder gegeben werden können: Wie lang sollte (oder muss) ein Konzentrationstest konstruiert sein, damit seine psychometrischen Gütekriterien optimal sind? Wie sollte man mit Verlaufsmerkmalen umgehen? Sollte man ermüdungsbedingte Veränderungen vermeiden oder sollte man sie explizit nutzbar machen? Wie schwierig sollten die Items eines Konzentrationstests sein? Führt höhere Itemschwierigkeit etwa zu Redundanz mit verwandten Konstrukten (z. B. Arbeitsgedächtnis)? Wie viel Redundanz darf man diesbezüglich zulassen?

2 Diagnostik von Konzentrationsfähigkeit

2.1 Historische Einordnung

Die ersten Versuche, die mentale Leistungsfähigkeit von Personen systematisch zu erfassen, gab es bereits Mitte des 19. Jahrhunderts. Zu nennen ist hier insbesondere das Institut für Psychologie von Wilhelm Wundt in Leipzig, der auch das erste „Chronoskop“ zur Erfassung von individuellen Reaktionszeiten entwickelte. Dieses Gerät war ein hochpräzises Instrument zur Messung der Reizdarbietungs- und Reaktionszeit, welches Reaktionen mit einer Genauigkeit von ca. 10 Millisekunden erfassen konnte. Chronoskope waren im Prinzip hochpräzise Uhrwerke, die durch Gewichte angetrieben wurden, und durch einen elektromagnetischen Mechanismus gestartet und gestoppt wurden. In einem typischen Experiment musste der Proband eine Telegraphentaste gedrückt halten (womit er den Haltestrom über dem Chronoskop geschlossen hielt). Am Zeitpunkt der Reizdarbietung wurde dieser Stromkreis unterbrochen und das Chronoskop begann zu laufen (d. h. es wurde Zeit registriert). Reagierte der Proband auf den Reiz durch Loslassen der (bislang gedrückten) Telegraphentaste, dann wurde der Stromkreis wieder geschlossen und das Chronoskop angehalten. Die bis dahin verstrichene Zeit - die Reaktionszeit des Probanden auf

den Reiz - konnte auf diese Weise erfasst werden. Heute beschreibt der Terminus „mentale Chronometrie“ einen Forschungsansatz innerhalb der Kognitionspsychologie. Vornehmlich geht es um die Fragen, (1) wie Informationsverarbeitungsprozesse organisiert sind und (2) wie man auf Prozesse und Mechanismen, die einer kognitiven Denktätigkeit zugrunde liegen, schließen kann (Ulrich, Mattes, & Miller, 1999). Erster Punkt beinhaltet Fragen zur Struktur der Verarbeitung, während sich zweiter Punkt auf die Logik zulässiger Schlussfolgerungen bezieht (vgl. Miller & Ulrich, 2003; Miller & Ulrich, 2013; Ulrich & Miller, 2008; Ulrich & Miller, 1994).

Klassische Ansätze

Eine zentrale Stellung innerhalb des chronometrischen Ansatzes nimmt die explizite Manipulation experimenteller Variablen sowie deren Effekt auf die Reaktionszeit der Probanden ein. Das Ziel der kognitionspsychologisch orientierten Forschung (wie sie von Wilhelm Wundt und Franciscus Cornelis Donders begründet wurde) besteht im Kern darin, Gesetzmäßigkeiten im allgemeinen menschlichen Leistungsverhalten aufzudecken und zu erklären. Das Ziel der diagnostisch orientierten Forschung (wie sie von James McKeen Cattell und Francis Galton begründet wurde) besteht hingegen darin, individuelle Unterschiede in der Informationsverarbeitung zu erklären und für Erfassungstechnologien nutzbar zu machen. Das letztere Ziel, die Entwicklung von psychometrischen Technologien, ist jedoch vornehmlich eine Leistung von Alfred Binet und Théodore Simon, die den ersten (nach heutigen Maßstäben) psychometrischen Test zur Vorhersage von Schulleistungen entwickelten. Schaut man sich die frühen Arbeiten näher an, wie sie in der Literatur zum Ende des 19. und des frühen 20. Jahrhunderts beschrieben wurden, so kann man zwischen jenen zur „Apperzeption“ als auch zu „Fluktuationen“ der Aufmerksamkeit unterscheiden. Erstere behandelten dabei vornehmlich den Aspekt der selektiven Aufmerksamkeit und Handlungssteuerung, während zweite Studien eher Aspekte der Aufrechterhaltung

der Aufmerksamkeit sowie deren Bedingungen behandelten (vgl. Langner & Eickhoff, 2013; Steinborn et al., 2010; Steinborn & Langner, 2011; Van Breukelen et al., 1996; Van der Ven, Smit, & Jansen, 1989).

Man kann es auch so sagen: Im Fokus der Apperzeptionsstudien stand stets der komputationale Aspekt kognitiver Prozesse (von der Reizaufnahme bis hin zur Reaktionsauswahl und motorischen Aktivierung); im Fokus der Fluktuationsstudien stand jedoch immer der energetische Aspekt kognitiver Prozesse, insbesondere waren dies Fragen nach den Mechanismen der Erreichung und Aufrechterhaltung eines angemessenen Grundniveaus für eine kontinuierlich hohe Arbeitseffektivität. Flehmig (2010) hat unlängst für eine Integration beider Traditionen (d. h. für eine zwar theoriegeleitete, aber technologisch orientierte Konzentrationsdiagnostik) plädiert und eine Berücksichtigung sowohl komputationaler als auch energetischer Mechanismen als Determinanten der Aufgabenleistung gefordert (vgl. Flehmig, 2010; Flehmig et al., 2007; Flehmig, Steinborn, Langner, & Westhoff, 2007a). In ähnlicher Weise wurde auch von Krumm (2008) argumentiert, welcher aber eher auf die Frage nach der Konstruktvalidität eines Konzentrationstests abzielt, insbesondere auf die Unterscheidung von Tests zur (1) Messung von mentaler Geschwindigkeit und jenen zur (2) Messung von Konzentration (vgl. Krumm, 2008; Krumm, Schmidt-Atzert, & Eschert, 2008; Krumm, Schmidt-Atzert, Michalczyk, & Danthiir, 2008; Krumm, Schmidt-Atzert, Schmidt, Zenses, & Stenzel, 2012). Tatsächlich wurden Tests zur Messung elementarer kognitiver Fähigkeiten seit Beginn des 20. Jahrhunderts von einigen Autoren jeweils als „Speedtest“, von anderen wiederum als „Konzentrationstests“ bezeichnet, obgleich das zugrundeliegende Aufgabenmaterial oft sehr ähnlich war. Die Bezeichnung „Speedtest” wurde vornehmlich von englischen und amerikanischen Forschern verwendet, während sich in der kontinentaleuropäischen Tradition der Begriff „Konzentrationstest“ etablierte (vgl. Smit & Van der Ven, 1995; Van Breukelen et al., 1996).

Die meisten Arbeiten im Bereich der Konzentrationsforschung des frühen 20. Jahrhunderts vertraten eine kognitiv-energetische Position die sich in entsprechenden Fragestellungen manifestierte: Wie lange können Probanden einfache Aufgaben ohne Leistungseinbußen bearbeiten und welche wichtigen Einflussfaktoren gibt es? Wie wirken sich Pausen auf die Leistung bei fortdauernder Bearbeitung aus und wie sollten Pausenstrukturen geschaffen sein, um optimale Arbeitseffektivität über lange Zeiträume aufrechterhalten zu können? Wie ist der Zustand „konzentriert sein“ überhaupt zu beschreiben und wie ist er von dem Zustand „abgelenkt sein“ zu differenzieren? Der Zustand „Konzentration“ wurde tatsächlich häufig gleichgesetzt mit „eine bestimmte Repräsentation im Bewusstsein zu haben“, während der Zustand „abgelenkt sein“ die Situation beschrieb, in der eine bestimmte aufgabenrelevante Repräsentation durch eine aufgabenirrelevante ersetzt wurde. Ein mathematisches Modell zur Beschreibung von Konzentrationsleistungen (Smit & Van der Ven, 1995; Van der Ven et al., 1989) basiert tatsächlich auf der theoretischen Annahme, dass der Zustand eines Probanden während der Bearbeitung von Konzentrationstests kontinuierlich zwischen „konzentriert“ und „abgelenkt“ wechselt, wobei die Häufigkeitsrelation beider Zustände ein Maß für Konzentration bzw. die Konzentrationsfähigkeit eines Probanden ist.

Besonders wichtig sind diesbezüglich die Arbeiten von Emil Kraepelin und seinen Kollegen (Kraepelin, 1902, 1903). Kraepelin konnte in seinen Studien zur Arbeitskurve zeigen, dass die von einer Person für die Bearbeitung fortlaufender (relativ leichter) Additionsaufgaben benötigte Zeit in verschiedene Faktoren unterteilt werden kann. In der Standardversion waren die Probanden angehalten, anderthalb Stunden lang auf einem Papierbogen dargebotene Aufgaben kontinuierlich zu bearbeiten. Die Menge korrekt bearbeiteter Items wurde in fünfminütigen Intervallen in einem Ergebnisplot aufgetragen. Aus diesen Leistungsdaten ließen sich dann verschiedene Kennwerte ermitteln, die, jeder für sich, auch bezüglich ihrer Validität hin untersucht

wurden. Natürlich geschah dies nicht nach den heute üblichen Maßstäben, sondern die Bewertung und folgerichtige Einschätzung von Probanden verlangte ein hohes Maß an Erfahrung vom durchführenden Diagnostiker. Üblicherweise wurden Parameter wie die mittlere Geschwindigkeit der Bearbeitung (d. h. die Anzahl bearbeiteter Items pro Zeit), die Genauigkeit (d. h. der prozentuale Anteil an inkorrekten Lösungen) sowie der Verlauf des Leistungsverhaltens ermittelt und standen der diagnostischen Auswertung zur Verfügung. Kraepelin versuchte auch, sein Verfahren zu validieren, jedoch geschah dies nicht durch korrelative Analysen innerhalb einer Normalpopulation an Probanden, sondern durch Gruppenvergleiche zwischen gesunden Probanden und Patienten der psychiatrischen Klinik. Das Forschungsprogramm von Emil Kraepelin glich demzufolge also eher einer heutigen neuropsychologischen Forschungsstrategie, welche das Ziel verfolgt, spezifische Beeinträchtigungen in kognitiven Funktionsbereichen von Patienten (verglichen zu gesunden Probanden) durch Testadministration mit der Arbeitskurve festzustellen.

Kraepelin gilt heute als einer der Pioniere der Konzentrationsforschung und seine Arbeitskurve wird heute als Prototyp aller Konzentrationstestverfahren angesehen (vgl. Brandstätter, 1995). Dabei sollte nicht vergessen werden, dass Kraepelin selbst niemals eine psychometrische Normierung vollzogen (und womöglich nie angestrebt) hat. Dies blieb nachfolgenden Forschern überlassen. Nennenswert sind an dieser Stelle die Arbeiten von Richard Pauli, welcher die erste psychometrisch normierte Version der Arbeitskurve nach Kraepelin entwickelte und validierte (Pauli, 1938). In der Folge gab es eine ganze Reihe von Validierungsstudien zum „Pauli-Test“ (wie die Arbeitskurve fortan genannt wurde), die bereits darauf angelegt waren, das allgemeine Fähigkeitsniveau von Personen im Rahmen von Personalselektion und -kategorisierung zu bestimmen. Zahlreiche Studien wurden unter anderem vom psychologischen Dienst der deutschen Wehrmacht mit dem Ziel der Personalselektion

und Personalkategorisierung von Rekruten und Offiziersanwärtern durchgeführt. Die Zeitperiode zwischen 1910 und 1945 kann also durchaus als die Geburtsstunde der Psychologischen Diagnostik bezeichnet werden, da sich die Psychologie als Wissenschaft von der erkenntnisorientierten Forschung hin zur praktischen Anwendung im Dienste der drei „Kulturaufgaben“ (1) Personalauslese, (2) Berufsberatung und (3) Training, weiterentwickelte (vgl. van Drunen, 1993). Insgesamt gesehen hat das sich aus der Arbeitskurve ableitende Aufgaben- und Konstruktionsprinzip zahlreiche Weiterentwicklungen stimuliert, was häufig mit der großen Flexibilität in der Generierung von Items erklärt wird (vgl. Steinborn et al., 2010). Zu nennen wären hier unter anderem der Konzentrations-Leistungs-Test (KLT) von Düker (1949) sowie der Revisionstest von Marschner (1971).

Klassische Testprinzipien

Eine ganz andere Materialform, welche für diagnostische Fragestellungen häufige Verwendung findet, sind die Durchstreichtests, in denen Personen Zielreize durchstreichen und Distraktorreize ignorieren müssen. Das erste standardmäßig eingesetzte Testverfahren, das nach diesem Prinzip konstruiert wurde, war der Bourdon-Test (1895). Der Test ist nach dem französischen Psychologen Benjamin Bourdon benannt, welcher verschiedene Testformen für seine Durchstreichverfahren verwendete, um Konzentrationsfähigkeit bei Kindern für schuldiagnostische Zwecke zu untersuchen. Ein häufig eingesetzter Test, der auf dem gleichen Aufgabenprinzip basiert, ist der d2 Aufmerksamkeits-Belastungstest von Brickenkamp (2002). Der Test besteht aus Zeilen der Buchstaben p und d, die unten und oben jeweils mit ein oder zwei Strichen versehen sind. Alle d mit zwei Strichen sind Zielreize und daher zu markieren, alle anderen Zeichen sind auszulassen. Tatsächlich gibt es zahlreiche weitere Tests, welche nach dem genau gleichen Prinzip konstruiert sind und mehr oder weniger vergleichbar hinsichtlich der psychometrischen Gütekriterien sind (z. B. Schmidt-Atzert

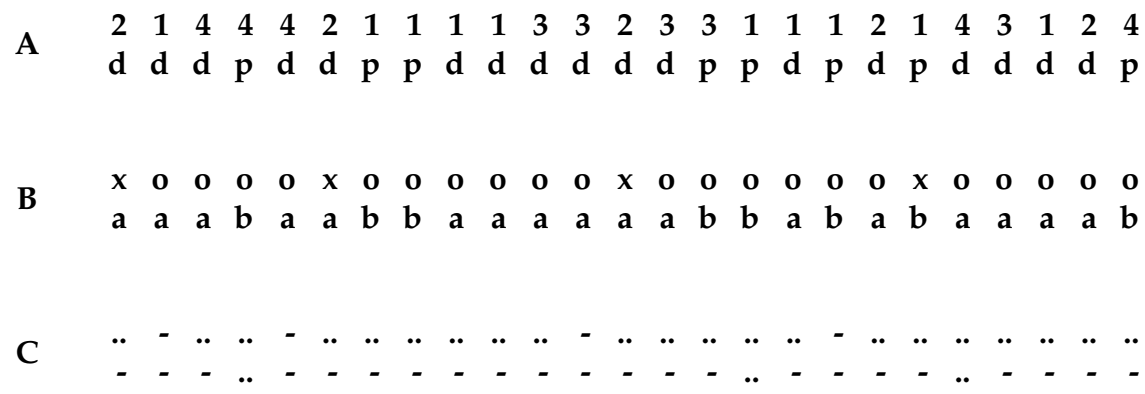

Abbildung 2.1 – Zielreize müssen markiert werden, alle anderen Reize müssen ausgelassen werden. Panel A stellt eine alternative Variante des Test d2 dar, alle d-2-Kombinationen müssen markiert werden. Panel B stellt eine auf Buchstaben basierende Testvariante dar, a-x Kombinationen sind Zielreize, alle anderen Kombinationen sind Distraktoren. In Panel C müssen alle „gleichen" Kombinationen (Strich-Strich, Punkte-Punkte) markiert werden, alle unterschiedlichen Kombinationen (Strich-Punkte, Punkte-Strich) sind Distraktoren.

& Bühner, 1997). In Abbildung 2.1 sind verschiedene Varianten entsprechend dem Bourdon-Prinzip konstruierte Durchstreich-Konzentrationstests abgebildet. Gemäß einiger Autoren sind Durchstreichverfahren nach dem Bourdon-Prinzip sehr anfällig gegenüber der Möglichkeit von Testverfälschung, da nicht alle Zeichen markiert werden müssen. Es ist zum Beispiel gezeigt worden, dass Probanden dazu instruiert werden können, schneller und etwas oberflächlicher zu arbeiten, um auf diese Weise eine größere Gesamtmenge pro Zeit zu bearbeiten. Dabei ist der Anstieg der Gesamtmenge relativ größer als der Anstieg der Fehler, was zu einem besseren Testergebnis führt. Dieses Validitätsproblem kann durch die Anwendung des vollständigen Markierungsprinzips gelöst werden, indem alle dargebotenen Items zu markieren sind. Der Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II) von Moosbrugger und Goldhammer (2005) ist nach diesem Prinzip konstruiert worden.

Neben Rechen- und Durchstreichverfahren sind auch andere Aufgabenformen als Konzentrationstest verwendet worden, wie zum Beispiel Kategorisierungs- und

Sortierverfahren. Allen Testverfahren gemeinsam ist jedoch, dass sie meistens sehr einfache und homogene Items enthalten, die von jeder Person gelöst werden können und möglichst wenig Spielraum für Testübungseffekte zulassen. Gemäß Westhoff und Hagemeister (2002) sind Konzentrationstests dennoch übungsabhängig. Jedoch sind sie im Vergleich zu Tests zum Arbeitsgedächtnis deutlich weniger anfällig für Verbesserungen. So konnten z. B. Compton und Logan (1991) zeigen, dass mit zunehmender Größe der verwendeten Alphabetarithmetik-Aufgaben die Reaktionszeiten kontinuierlich anstiegen. Die Reaktionszeiten verringerten sich nach Übung in Abhängigkeit von der Aufgabengröße. Entsprechend fiel der Übungseffekt für schwierigere Aufgabenformen höher aus als für leichtere Aufgaben. Trotz dieser zahlreichen Möglichkeiten der Testkonstruktion messen diese Tests ein ähnliches Konstrukt, das als Konzentration bezeichnet werden kann. Um möglichst gute Testleistungen in Durchstreichtests zu erreichen, müssen Probanden ihre Aufmerksamkeit auf die Zielreize fokussieren und dabei irrelevante Reize ignorieren. Bei nach dem vollständigen Markierungsprinzip konstruierten Tests müssen Probanden jedes Item aufmerksam verarbeiten und nach einer intern repräsentierten Regel beantworten. Kritisch für die Konzentrationsleistung ist aber die aufgabenunabhängige Fähigkeit, sich fortdauernd für die Bearbeitung zu energetisieren und diesen Zustand über die Testbearbeitungsdauer aufrechtzuerhalten. Das andauernde Aufrechterhalten einer optimalen Fokussierung in Konzentrationstests wird vor allem durch die Verwendung homogener Aufgabenformen erschwert, da diese nicht intrinsisch motivierend sind und bei langdauernder Bearbeitung zu psychischer Sättigung führen. Das in einer konkreten Testsituation verwendete Aufgabenmaterial ist also nur ein Mittel zum Zweck der Messung einer generellen Fähigkeit, die als Konzentration bezeichnet wird. Die bekanntesten im Handel üblichen Verfahren werden oft nach ihrer Aufgabenform klassifiziert, wobei man drei Arten unterscheidet: Rechentests, Durchstreichtests und Sortierverfahren.

Westhoff und Kluck (1984) haben eine vom Aufgabenmaterial unabhängige Einteilung von Konzentrationstests vorgenommen, aus der sich auch praktische Hinweise für die Konstruktion von Tests ableiten lassen. Den Autoren zufolge gibt es drei Facetten, die als die wesentlichen Determinanten von Konzentrationsleistungen angesehen werden können: (1) die Komplexität der Reizgrundlage, (2) die Komplexität der geforderten Reaktionsalternativen und (3) die Geübtheit einer Testperson mit dem spezifischem Aufgabenmaterial. Aus einer Kombination der Ausprägungsgrade dieser drei Facetten kann die individuelle Leistung der Probanden vorhergesagt werden. Westhoff und Kluck (1984) zufolge ist eine weitere wesentliche Leistungsanforderung in Konzentrationstests, dass diese in einem selbstbestimmten Arbeitstempo (engl. self-paced) durchgeführt werden. Da Probanden grundsätzlich instruiert sind, möglichst schnell und dabei fehlerarm zu arbeiten, ist es wichtig, dass diese die Balance zwischen Geschwindigkeit und Genauigkeit selbst einstellen. Die Leistungsanforderung in Konzentrationstests besteht also nicht nur darin, die richtigen Reaktionen auf der Basis der vorgegebenen Regel zu tätigen, sondern auch darin, die eigene optimale Arbeitsgeschwindigkeit zu finden und über die Dauer der Testbearbeitung beizubehalten. Diese Ansicht wird auch von anderen Autoren geteilt und es ist weitgehend akzeptiert, dass die individuelle Konzentrationsfähigkeit nur im Bereich des persönlichen Tempos erfasst werden kann. Würde man die gleiche Aufgabenform dagegen taktgebunden (engl. forced-paced) darbieten, so würde man ein anderes Konstrukt messen, je nachdem, ob ein Proband durch die Taktung unterfordert oder überfordert würde. Eine optimale und gut vorhersehbare Taktung würde einer Person die Möglichkeit eröffnen, sich zeitlich auf den Moment der Reizdarbietung vorzubereiten (vgl. Steinborn & Langner, 2012).

Es ist tatsächlich empirisch belegt worden, dass sich die faktorielle Validität des Bearbeitungsmodus leicht ändert, wenn Probanden eine gleiche Aufgabenform in

unterschiedlichem Modus bearbeiten. Krumm, Schmidt-Atzert und Eschert (2008) haben diesbezüglich Aufgabenformen in jeweils unterschiedlichem Pacing-Modus (self-paced versus forced-paced) und unterschiedlicher Darbietungsweise (Einzeldarbietung versus Listendarbietung) administriert. Gleicher Pacing-Modus führte bei unterschiedlichem Aufgabenmaterial zu einer höheren Interkorrelation, was belegt, dass ein ähnliches Konstrukt (unabhängig vom Aufgabenmaterial) gemessen wird. Gemäß Flehmig (2010) ist es bei der Bearbeitung von Konzentrationstests also entscheidend, dass man sich selbst immer weiter antreiben muss und möglichst schnell eine optimale interne Taktrate finden muss, um gute Konzentrationsleistungen zu zeigen (vgl. Rabbitt, 1969; Rabbitt & Rogers, 1965). Dies impliziert, dass individuelle Verhaltenstendenzen von Personen (z. B. Impulsivität, Motivation, etc.) zu Leistungsunterschieden im Test führen können, die nichts mit mentaler Informationsverarbeitung zu tun haben. Ob dies ein Vorteil oder Nachteil ist, ist Ansichtssache. Einige Autoren argumentieren, dass Konzentrationstests daher einen Aspekt der Arbeitsweise und damit eine Facette der Persönlichkeit messen, was unter Umständen von diagnostischer Bedeutung sein kann. Andere Autoren haben wiederum versucht zu zeigen, dass die individuelle Arbeitsweise den wahren Leistungswert verunreinigt und damit ein Messartefakt darstellt. Der von Moosbrugger und Goldhammer (2005) konstruierte FAKT-II versucht dieses Problem durch eine adaptive Leistungsmessung (in der Testform E umgesetzt) zu lösen. Die Aufgabe in diesem computerisierten Test ist es, Items nach einer vorgegebenen Regel zu markieren, wobei die Darbietungsdauer der Items adaptiv variiert wird. Arbeitet eine Person über viele Durchgänge fehlerfrei, dann wird die Darbietungsdauer gekürzt, bei vielen Fehlern jedoch verlängert.

Da Konzentrationstests keine domainspezifischen Fähigkeiten messen sollen, ist es wichtig, dass eine hohe Vertrautheit mit dem verwendeten Aufgabenmaterial besteht. Aufseiten der Testperson werden also die Bekanntheit der Aufgabenform

und die zur Lösung der Aufgabe erforderlichen mentalen Operationen vorausgesetzt. Für die Konstruktion eines Konzentrationstests sind also all jene Aufgabenformen besonders geeignet, welche hoch überlernt und von jeder gesunden Person bearbeitet werden können. Aus diesem Grunde dienen die klassischen Kulturtechniken oft als Grundlage für die Konstruktion von Aufgabenmaterial. Zahlreiche Verfahren sind als Rechentest realisiert und erfordern fortlaufendes mentales Addieren, andere Aufgaben simulieren den Prozess des Lesens und erfordern die reihenweise Bearbeitung von Zeichen. Rasch (1980) vertrat dabei die Ansicht, dass der Prozess des Lesens in hohem Maße dem kognitiven Mechanismus entspricht, der in Durchstreichtests realisiert ist. Aus diesem Grunde sind das Rasch-Modell der Lesegeschwindigkeit und das Rasch-Modell für Speedtests einander äquivalent. Obgleich die Nutzung von Kulturtechniken als Aufgabenmaterial die Testfairness erhöhen soll, so ist gelegentlich auch angeführt worden, dass dieser Umstand die Testfairness beeinträchtigen könnte, da bestimmte Berufsgruppen dadurch bevorteilt werden könnten. So könnte es sein, dass ein Bankangestellter, dem man den täglichen Umgang mit Zahlen unterstellen darf, einen Vorteil in Rechenkonzentrationstests hat, während ein Deutschlehrer besonders gute Leistungen in Durchstreichverfahren zeigt. Tatsächlich ist wiederholt gezeigt worden, dass Bankangestellte besonders gute Leistungen in psychometrischen Tests zeigen, weshalb sich im angloamerikanischen Bereich auch das Konstrukt „clerical speed“ bzw. „clerical aptitudes“ etabliert hat.

Obwohl ein gewisser Konsens darüber besteht, dass die Aufgaben in Konzentrationstests relativ leicht sein sollten, sind immer wieder Versuche mit komplexerem Aufgabenmaterial unternommen worden. Für eine höhere Aufgabenkomplexität sprechen mehrere Gründe. Einige Autoren haben angeführt, dass sich die psychometrischen Gütekriterien Reliabilität und Validität mit zunehmender Komplexität verbessern und daher komplexe Testvarianten zu bevorzugen seien (z. B. Schweizer

& Moosbrugger, 2004; Schweizer, Moosbrugger, & Goldhammer, 2005; Westhoff & Graubner, 2003; Westhoff & Kluck, 1984). Dies mag bis zu einem bestimmten Grad an Komplexität wohl auch zutreffen. Andere Autoren haben angeführt, dass mit zunehmender Aufgabenkomplexität ein breiteres Spektrum an Kriteriumsvalidität erfasst wird (vgl. Cronbach, 1975). Dies ist womöglich durch die damit einhergehende höhere Korrelation zu dem Konstrukt Intelligenz zu erklären. Dabei kommt es natürlich sehr darauf an, welche Art von Komplexität in einer Aufgabenform umgesetzt wurde. Die Taxonomie von Westhoff und Kluck (1984) basiert auf einer Variierung der Anzahl der Reiz- und Reaktionsalternativen als Mittel der Komplexitätserhöhung. Düker (1949) hat argumentiert, dass beim Bearbeiten von Rechenkonzentrationstests unterschiedliche Einzeltätigkeiten koordiniert werden müssen, wie Auffassen, Behalten oder Vergleichen. Ob Leistungsunterschiede mit zunehmender Komplexität deutlicher werden oder gar nivelliert werden, ist eine empirische Frage, deren Beantwortung sehr wahrscheinlich von der Art der experimentellen Manipulation abhängt. Es gilt jedoch als gut belegt, dass individuelle Leistungsunterschiede immer deutlicher hervortreten, wenn die Aufgabenkomplexität erhöht wird (vgl. Jensen, 2006). Bei der Konstruktion eines Tests ist es jedoch immer eine empirische Frage, bei welchem Grad die Itemschwierigkeit optimal ist, damit die psychometrischen Gütekriterien bestmögliche Werte erreichen. Es macht Sinn, die Antwort auf diese Frage durch experimentelle Variation im Rahmen von Voranalysen herauszufinden.

Ein typischer Konzentrationstest, der komplexes Aufgabenmaterial enthält, ist der von Düker (1949) entwickelte Konzentrations-Leistungstest (KLT). Dieser Test verlangt von den Probanden die Bearbeitung relativ komplexer Rechenaufgaben über eine Zeitspanne von ca. 20 Minuten. Dabei besteht jede Aufgabe aus zwei Zeilen. In jeder Zeile werden drei einstellige Zahlen addiert bzw. subtrahiert, die Zwischenergebnisse müssen kurzfristig behalten werden. In Abhängigkeit davon, ob das Ergebnis der

oberen oder der unteren Zeile größer ist, werden die Zwischenergebnisse der beiden Zeilen addiert oder subtrahiert. Es soll die Fähigkeit zur Konzentration gemessen werden, die hier allerdings als die Fähigkeit verstanden wird, kognitive Subroutinen effizient zu sequenzieren und zu koordinieren. Während im KLT ein sogenanntes Produktionsparadigma realisiert wurde, so verwendet der Revisionstest von Marschner (1971) ein typisches Verifikationsparadigma. In einem solchen Paradigma werden komplette Additionsterme (mit Ergebnis) dargeboten und die Probanden sind aufgefordert, die Richtigkeit des Terms durch eine entsprechende Markierung bzw. durch Tastendruck zu verifizieren. Im Vergleichsparadigma werden wiederum zwei dargebotene Zahlenterme hinsichtlich ihrer Größe verglichen. Im Serialen Mentalen Additions- und Vergleichstest (engl. serial mental addition and comparison task, SMACT) wird ein Additionsterm lateral dargeboten und ein Ergebniswert auf der gegenüberliegenden Seite dargeboten (vgl. Restle, 1970; Steinborn et al., 2010). Die Probanden müssen entscheiden, ob das dargebotene Ergebnis größer oder kleiner als der mental errechnete Zahlenwert ist, und dies mit einem Tastendruck (linke versus rechte Taste) indizieren.

Es ist natürlich auch möglich, das Rechenprinzip und das Durchstreichprinzip miteinander zu kombinieren, denn beide Konstruktionsweisen schließen sich nicht gegenseitig aus. Letztlich zählt nur, ob ein bestimmter Test den psychometrischen Anforderungen standhält und valide Vorhersagen machen kann. In der Abbildung 2.2 auf der nächsten Seite sind drei Varianten eines auf Rechenzeichen basierenden Tests nach dem Durchstreichverfahren abgebildet.

Insgesamt gesehen ist zu beobachten, dass sich heute zwar viele handelsübliche Testverfahren zur Messung von Konzentrationsfähigkeit auf das Prinzip des fortlaufenden Addierens beziehen, dass aber fast alle modernen Verfahren (1) eine wesentlich geringere Testlänge aufweisen, (2) die Itemschwierigkeit höher ist und (3) neben der Papierversion fast immer eine computerisierte Version erhältlich ist. Bislang

A

2 1 4 4 4 2 1 1 1 1 3 3 2 3 3 1 1 1 2 1 4 3 1 2 4
1 4 7 2 7 2 2 3 4 7 3 6 4 6 4 6 6 1 3 3 2 5 5 2 4

B

1 2 3 4 5 6 7 1 2 3 4 1 2 3 4 5 6 7 1 2 3 4 1 2 3
2 3 4 5 6 7 9 2 3 4 6 2 3 4 5 6 7 9 2 3 4 6 2 3 4

C

3 2 4 1 5 2 2 3 3 2 4 1 0 2 2 5 1 2 1 0 1 2 3 4 4
2 3 1 4 0 3 3 2 4 3 1 4 5 3 4 0 4 3 5 5 4 3 2 1 2

Abbildung 2.2 – Zielreize sind zu markieren, alle anderen Reize sind auszulassen. In Panel A (Additionsaufgabe) sind alle Zahlenkombinationen zu markieren, deren Summe 6 beträgt. In Panel B (Subtraktionsaufgabe) sind jene Zahlenkombinationen zu markieren, deren Differenz 2 beträgt. In Panel C (Abweichungsaufgabe) sind alle Zahlenkombinationen zu markieren, deren Summe ungleich 5 ist.

ist jedoch nicht hinreichend geklärt, ob die Testlänge lediglich aus ökonomischen Gesichtspunkten heraus abgenommen hat oder ob sich kürzere Versionen als diagnostisch trennschärfer erwiesen haben. Zahlreiche Autoren haben sich diesbezüglich geäußert, jedoch gibt es kaum empirische Arbeiten, welche den Effekt der Testlänge systematisch untersucht haben. Ebenso ungeklärt ist es, ob eine höhere Itemschwierigkeit in Konzentrationstests zu einer Verbesserung psychometrischer Gütekriterien (d. h. Reliabilität, Validität) führt. Zwar kann man den Trend beobachten, dass die Itemschwierigkeit im Laufe des 20. Jahrhunderts angestiegen ist, jedoch gibt es kaum vergleichende Untersuchungen zur prädiktiven Validität von leichten versus schwierigen Items in Konzentrationstests. Diese Lücken sollten mit der vorliegenden Arbeit geschlossen werden. Die Frage, ob Papierversionen und computerisierte Versionen von Konzentrationstests ein ähnliches Konstrukt messen, ist jüngst von Krumm und Kollegen (Krumm, Schmidt-Atzert, & Eschert, 2008; Krumm, Schmidt-Atzert, et al., 2012) untersucht worden. Im Ergebnis dieser Forschungsbemühungen konnte für eine Reihe handelsüblicher Verfahren faktorielle Validität nachgewiesen werden, was den

computerisierten Versionen hinreichende Güte gegenüber den Papier- und Bleistift-Versionen bescheinigt. Allerdings haben computerisierte Tests den großen Vorteil, dass die Reaktionszeiten separat ausgewertet werden können, was wiederum Analysen der Reaktionszeitverteilung ermöglicht. Dieser Vorteil computerisierter Versionen sollte in der aktuellen Arbeit genutzt werden.

2.2 Kognitive Modellierung

Wie kann man das Leistungsverhalten von Probanden in Konzentrationstests im Allgemeinen beschreiben? Welche kognitiven Mechanismen liegen der Leistung während der Bearbeitung solcher Aufgaben zugrunde? Um diese Frage beantworten zu können, muss man sich zunächst einmal darüber im Klaren sein, dass „sich konzentrieren“ eine aufgabenunabhängige Leistung ist. Man kann schließlich bei jeder denkbaren Aufgabe „konzentriert“ oder auch „unkonzentriert“ (d. h. abgelenkt) sein. Daher haben zahlreiche Autoren auch versucht, das Konstrukt Konzentration aufgabenunabhängig über ein sogenanntes Zwei-Zustands-Modell (engl. two-state mixture model) zu beschreiben (vgl. Miller, 2006; Steinborn et al., 2010; Van Breukelen et al., 1996). Es wird angenommen, dass sich Probanden während der Bearbeitung von serialen Wahlreaktionen zunächst in einem Zustand hoher Konzentriertheit befinden (d. h. sie reagieren schnell), aber gelegentlich in einen Zustand der Unkonzentriertheit wechseln (d. h. sie reagieren langsam), und dass die Konzentrationsfähigkeit eines Probanden an dem Verhältnis beider Zustände festzumachen ist. Weiterhin wird angenommen, dass mit zunehmender Dauer der Bearbeitung nicht etwa die mentale Verarbeitungsgeschwindigkeit abnimmt, sondern lediglich der relative Anteil an Zuständen der Unkonzentriertheit zunimmt (d. h. Aufmerksamkeitslapsus tritt häufiger auf). Dafür gibt es schon sehr frühe empirische Belege. Schon von Voss (1899) stellte fest, dass nach einer fortdauernden Periode der Bearbeitung einfacher Additionen die relative

Häufigkeit langsamer Reaktionen ansteigt, ohne dass jedoch Änderungen bei den schnellsten Reaktionen zu verzeichnen sind. Er machte dafür wiederkehrend auftretende Fluktuationen der Aufmerksamkeit verantwortlich, die sich durch Pausen und Übung vermindern lassen.

Zwei-Zustandsmodelle

Es gibt noch andere sehr frühe Belege, die für die oben genannte Ansicht sprechen. So zum Beispiel hat Oehrn (1895) in sehr ähnlicher Weise bereits vermutet, dass eine durch Lapsus bedingte erhöhte Reaktionszeitvariabilität während der Testbearbeitung ein Indikator für eine verschlechterte Arbeitseffizienz ist. Auch er ging davon aus, dass die kontinuierliche Reaktionsleistung von Probanden hauptsächlich durch zwei bestimmende Faktoren beeinflusst wird, durch vorausgehende Übung und durch aufgabeninduzierte mentale Ermüdung. Durch die Verwendung von einfachen und homogenen Aufgaben wie fortlaufendes Addieren oder Varianten von Zuordnungsaufgaben konnte er zeigen, dass kurze Erholungspausen die Effekte von aufgabeninduzierter Ermüdung auf die Reaktionszeit verringern. Die Effekte der Übung dagegen bleiben erhalten. Im angloamerikanischen Raum wurden ähnliche Ergebnisse aus Studien über gelegentlich auftretende „mentale Blockierungen“ während fortlaufender Aufgabenbearbeitung berichtet. Die meisten der heute noch zitierten Arbeiten sind von Bills (1931, 1937) durchgeführt worden, der zahlreiche Aufgabenformen (wie fortlaufendes Addieren, Kategorisieren, Sortieren, Benennen usw.) verwendete. Seine Analysen der Einzelreaktionen ergaben, dass sich nach ungefähr 5 Minuten ununterbrochener Arbeit sogenannte mentale Blockierungen beobachten ließen, welche definiert waren als sehr langsame Reaktionen verglichen zum individuellen Mittelwert der Reaktionen eines Probanden (z. B. Blocks = alle Reaktionszeiten, die länger waren als das Doppelte des Medians der Reaktionszeiten). Die wichtigsten Ergebnisse dieser Studien lassen sich so zusammenfassen, dass (1) Testlänge (engl. time-on-task) zu

einer Häufung von Blockierungen führt, nicht jedoch zu einer generellen Verlangsamung der Reaktionen, dass (2) Kurzpausen diesen Effekt kompensieren und dass (3) aufgabenspezifische Übung (Übung am gleichen Material) präventiv gegenüber Blockierungen ist, während unspezifische Übung (Übung an ähnlichem Material) keine derartigen Effekte hat (Van Breukelen et al., 1996).

Zusammenfassend kann man sagen, dass es mehrere Strömungen gibt, von denen aus versucht wurde, Konzentrationsfähigkeit zu erfassen. Allen gemeinsam ist jedoch, dass ein Zwei-Zustands-Modell angenommen wird: Während konzentrierter Tätigkeit wechseln Probanden stetig zwischen einem Zustand hoher Konzentration und Ablenkung, wobei die Bezeichnungen terminologisch leicht variieren. Bezüglich der zugrundeliegenden Mechanismen gibt es jedoch verschiedene theoretische Annahmen. Kraepelin (1902, 1903) vermutete, dass sich während fortlaufenden Addierens eine Art „Widerstand gegen die Fortführung der Aufgabe“ entwickelt, der von den Probanden durch „Willenskraft“ überwunden werden muss. In ähnlicher Weise haben auch Karsten (1928) und Lewin (1928) argumentiert, dass sich während der repetitiven Ausführung einer Handlung „psychische Sättigung“ einstellt, die subjektiv als eine ausgeprägte Abneigung gegenüber jedweder Fortführung der Handlung empfunden wird. Schulz-Hardt und Kollegen haben eine Reihe von Experimenten durchgeführt, in denen Probanden Varianten eines Rechenkonzentrationstests (KLT, Düker, 1949) unter verschiedenen experimentellen Bedingungen bearbeiten sollten. Von besonderem Interesse war dabei jedoch das subjektive Erleben (d. h. der gemessene Grad an „psychischer Sättigung “). Psychische Sättigung konnte als Funktion zweier Faktoren weiter spezifiziert werden: Demzufolge war die subjektiv erlebte Sättigung größer bei (1) geringer Peripherisierbarkeit der Aufgabe (d. h. bei größerer Auslastung des Arbeitsgedächtnisses) und (2) bei hoher persönlicher Relevanz der Aufgabe für die Probanden (Mojzisch & Schulz-Hardt, 2007; S. Schulz-Hardt, Meinken, Rott, & Frey,

2001; S. Schulz-Hardt, Rott, Meinken, & Frey, 2001). In anderen Worten führt hohe Belastung und hohe Bedeutung der eigenen Leistung für die Probanden in einer Situation zu höheren Graden an Sättigung.

Die oben erläuterte moderne Konzeptualisierung des Konstrukts der psychischen Sättigung als Ergebnis kognitiv anstrengender Arbeitstätigkeit ist im Wesentlichen durch die Theorie der reaktiven Hemmung von Hull (1943) beeinflusst, welche die Position vertritt, dass die Ausführung jeder fortlaufend durchgeführten Entscheidungstätigkeit ab einer bestimmten Anzahl an Wiederholungen ein „reaktives Hemmungspotential“ auf nachfolgende Ausführungen derselben Tätigkeit auslöst. Dabei nimmt Hull konkret an: Wenn eine Person einmal wiederholt auf einen Reiz mit einer bestimmten Handlung aufgrund einer Entscheidungsregel reagiert (d. h. eine Art von Klassifikationstätigkeit ausführt), so vergrößert sich zunächst das Reaktionspotential. Diese Vorhersage entspricht in etwa dem Reaktionsvorteil bei Aufgabenwiederholungen (verglichen zu einem Aufgabenwechsel) im Aufgabenwechselparadigma (vgl. Kiesel et al., 2010). Wenn eine Person jedoch immer wieder dieselbe Aufgabe ausführt, dann wird aus dem Vorteil zunehmend ein Nachteil. Gemäß Hull (1943) vergrößert sich mit zunehmenden Ausführungen derselben Handlung ein „primär negativer Antrieb“. Dieser negative Antrieb hemmt in Abhängigkeit seines Ausmaßes das Reaktionspotential für jede nachfolgende Handlung, sofern sie unmittelbar nach der vorangehenden ausgeführt wird. Vergeht jedoch eine bestimmte Zeit bis zur nächsten Handlungsausführung, so löst sich der negative Antrieb spontan auf. Der Zeitbedarf bis zum Erlöschen des negativen Antriebspotentials ist abhängig von der jeweiligen Art und Länge der Tätigkeit, erstreckt sich aber im Bereich zwischen zwei Sekunden und mehreren Minuten. Die Theorie der reaktiven Hemmung macht unter anderem auch Aussagen über Effekte der Testlänge, Effekte kurzer und langer Pausen sowie über die Effekte von Abwechslung bei monotonen Arbeitstätigkeiten (Lewin, 1928).

Akkumulation von Overload

Während die oben genannten Modelle den subjektiven Aspekt eines zunehmenden „Widerwillens“ gegen die Fortführung der Aufgabe betonten, der überwunden werden muss, so gibt es eine andere Modellklasse, die einen kumulativen Refraktäreffekt als Resultat fortwährender Bearbeitung gleicher Aufgabenformen annimmt (vgl. Pashler, 1994, für eine Übersicht; Ulrich & Miller, 2008). Insbesondere die Arbeiten von Arthur Bills (1931, 1937) waren diesbezüglich sehr einflussreich und die meisten der heute debattierten Zwei-Zustands-Modelle (vgl. Miller, 2006; Van Breukelen et al., 1996) basieren auf den von ihm vorgelegten Grundgedanken. So vermutete Bills (1937), dass die bei fortwährenden serialen Reaktionstätigkeiten auftretenden mentalen Blockierungen durch eine akkumulierte Überladung bestimmter kognitiver Systemkomponenten entstehen. Bills nahm konkret an, dass insbesondere die zentrale Stufe der Informationsverarbeitung (Stufe der Reaktionswahl) betroffen ist, während die Stufen der Informationsaufnahme (perzeptuelle Stufe) und jene der motorischen Ausführung (motorische Stufe) unbeeinträchtigt bleiben sollten. Ob sich der postulierte kumulative Refraktäreffekt tatsächlich selektiv auf nur eine Verarbeitungsstufe beschränkt, ist empirisch bislang ungeklärt. Auf jeden Fall sind, gemäß dieser Sichtweise, mentale Blockierungen anzusehen als kurzzeitig auftretende, vom kognitiven System erzwungene Erholungspausen, welche durch einen akkumulierten Refraktäreffekt der zentralen (für Klassifikationen, d. h. Reiz-Reaktions-Zuweisungen verantwortlichen) Systemkomponente entstehen. Diese Annahme wird durch den Befund gestützt, dass die relative Häufigkeit von Blockierungen maximal ist, wenn das Zeitintervall zwischen Durchgängen (Reaktions-Stimulus-Intervall, RSI) kurz ist, jedoch abnimmt, wenn das RSI verlängert wird (Jentzsch & Leuthold, 2006; Soetens, Boer, & Hueting, 1985).

Basierend auf den eben erläuterten Modellen über sich akkumulierende mentale Ermüdung (Dodge, 1917; Kraepelin, 1903), psychische Sättigung (Karsten, 1928; Lewin,

1928) bzw. kumulative kognitive Überladung (Bills, 1931; Weaver, 1942) und daraus resultierender Ablenkung haben zahlreiche Autoren formale Modelle für Reaktionszeiten in Konzentrationstests vorgestellt, die im Folgenden als Inhibitionsmodelle zusammenfasst werden (z. B. Smit & Van der Ven, 1995; Van Breukelen et al., 1996; Van der Ven et al., 1989). Diesbezüglich wird postuliert, dass sich das arithmetische Mittel der Reaktionszeit aus zwei separaten Prozesskomponenten formiert, dem Anteil an abgegebenen Reaktionen, welche aus einem Zustand mentaler Konzentration heraus abgegeben wurden, und jenem Anteil an Reaktionen, die im Zustand mentaler Distraktion abgegeben wurden. Der Geltungsbereich dieser Annahmen ist explizit für die Domäne routinierter und in einem selbstbestimmten Arbeitsrhythmus (self-paced) ausgeführten Aufgaben definiert. Damit fallen praktisch alle handelsüblichen Konzentrationstests in den Geltungsbereich dieser Modellklasse (Steinborn et al., 2010). Das theoretische Grundgerüst des Inhibitionsmodells für Konzentration ist die vorangehend erörterte Theorie der reaktiven Hemmung (Hull, 1943). Diese Theorie sagt aus, dass sich bei fortlaufender monotoner Bearbeitung eine Art negativer Antrieb (Hemmung) ausbildet, der wiederum einen Gegengradienten zur aktuellen Zielrepräsentation bildet. Dieser Gegengradient kann als eine „Distraktionstendenz“ bezeichnet werden und steigt als Funktion der Anzahl konzentriert bearbeiteter Items an. Er führt bei Erreichung eines bestimmten kritischen Werts unmittelbar zu Ablenkung, was sich in extrem langsamen Reaktionen (mentale Blockierungen) äußert. Die Wahrscheinlichkeit, in einen Zustand der Abgelenktheit zu driften, erhöht sich monoton als Funktion konzentriert bearbeiteter Items, vermindert sich jedoch durch Kurzpausen, mit größer werdendem Reaktions-Stimulus-Intervall (RSI) etc.

Alle Variationen des Inhibitionsmodells für Konzentrationsleistungen nehmen an, dass die mittlere Reaktionszeit einer Person die Summe zweier Prozesse darstellt, nämlich des Anteils der Reaktionen, die im Zustand der Konzentration ausgeführt

wurden (dieser Anteil entspricht der mentalen Geschwindigkeit einer Person), und des Anteils an Reaktionen, welche im Zustand von Distraktion ausgeführt wurden. Dieser Sichtweise folgend wäre das (in fast allen Studien verwendete) arithmetische Mittel der Reaktionszeiten kein geeignetes Maß für die Konzentrationsleistung. Es können jedoch verschiedene Leistungskennwerte direkt aus dem Inhibitionsmodell abgeleitet werden: Die „minimale Reaktionszeit“ kann als Schätzer für die mentale Geschwindigkeit einer Person interpretiert werden. Dieser Kennwert kann durch Zusammenfassung der schnellen Perzentile einer individuellen Reaktionszeitverteilung berechnet werden. Die „ablenkungsfreie Reaktionszeit“ kann als Schätzer für die Effektivität einer Person bei fortlaufender, konzentrierter Tätigkeit interpretiert werden. Dieser Kennwert kann als das geometrische Mittel der individuellen (logarithmierten) Reaktionszeiten berechnet werden. Die Distraktionstendenz kann aus dem Verhältnis der schnellsten Perzentile zu denen der langsamsten Perzentile geschätzt werden. Fehlreaktionen spielen im Inhibitionsmodell keine Rolle. Van Breukelen et al. (1996) führen an, dass der Fehleranteil maximal 12 Prozent betragen sollte und dass Personen mit einem höheren Fehleranteil auf instruktionswidriges Verhalten schließen lassen. Entsprechend dieser Ausführungen ist festzuhalten, dass die Gesamtleistung einer Person aus der Sicht des Inhibitionsmodells durch zwei Fähigkeiten bestimmt wird: Geschwindigkeit der Verarbeitung (fluide Intelligenz) und Ablenkungstendenz (Konzentration). Beide Konstrukte sollten, gemäß dem Inhibitionsmodell, durch sorgfältige Analyse voneinander separiert werden, weshalb empfohlen wird, respektive Kennwerte separat auf ihre prädiktive Validität hin zu prüfen (Schmiedek, Oberauer, Wilhelm, Suss, & Wittmann, 2007; Steinborn et al., 2008).

Leider macht das Inhibitionsmodell keine konkreten Vorhersagen bezüglich der Itemschwierigkeit bzw. der Verarbeitungsstufe, auf dem die Schwierigkeit von Items experimentell manipuliert wird. Das ist jedoch von einigen Autoren als besonders

wichtig empfunden worden, da eine Schwierigkeitsmanipulation jeweils auf perzeptueller, zentraler oder aber motorischer Verarbeitungsstufe geschehen kann (vgl. Pieters, 1983; Sternberg, 2001; Ulrich & Miller, 2008) und sich für den Testkonstrukteur die Frage stellt, welche Schwierigkeit auf welcher Stufe am besten in einem Testverfahren zu realisieren ist (vgl. Flehmig et al., 2007; Westhoff & Graubner, 2003). Dieser Gedanke ruft nach einer systematischen Überprüfung von Tests mit unterschiedlichen Schwierigkeitsniveaus. Vertreter des Inhibitionsmodells jedoch fordern möglichst leichte und homogene Items, da unterschiedlich schwierige (d. h. leichte und schwierigere) Items das Problem der asymmetrischen Übungssensitivität mit sich bringen würden, unabhängig davon, auf welcher Verarbeitungsstufe die Schwierigkeitsmanipulation vorgenommen wurde (vgl. Hagemeister, 2007; Pashler & Baylis, 1991). Gemäß Logan und Kollegen (Compton & Logan, 1991; Logan, 2002) geschieht dies deshalb, weil mit zunehmender Anzahl an Testwiederholungen eine Veränderung der Informationsverarbeitung von komputationaler (kontrollierter) Verarbeitung hin zu auf Gedächtnisabruf basierender (automatischer) Verarbeitung stattfindet. Da der Anteil an komputationaler Verarbeitung bei schwierigen Items größer ist, fällt der Lerneffekt für schwieriges Aufgabenmaterial größer aus. Hagemeister (2006) hat in einer umfangreichen empirischen Arbeit zeigen können, dass asymmetrisches prozedurales Lernen unterschiedlich schwieriger Items tatsächlich eine große Herausforderung für Testkonstrukteure ist, dies aber dadurch gelöst werden kann, indem eine große Anzahl an Items verwendet wird. Je mehr Items vorhanden sind, desto schwieriger sind diese zu lernen, was sich präventiv gegenüber gedächtnisbasierten (automatisierten) Bearbeitungsstrategien auswirken sollte. Hier zeigt sich der exklusive Vorteil von numerischem Aufgabenmaterial, da zahlenmäßig sehr viele Items für ein bestimmtes Schwierigkeitsniveau generierbar sind (vgl. Ashcraft, 1992; Steinborn et al., 2008). Somit stellt sich die Frage, wie man Itemschwierigkeit innerhalb einer theoretisch

motivierten, jedoch praktisch orientierten Konzentrationsdiagnostik behandeln sollte.

Integration kognitiver und energetischer Mechanismen

Das kognitiv-energetische Modell (KEM) von Sanders (vgl. Sanders, 1983; Sanders, 1998, Kap. 9) ist zwar kein Modell, das sich exklusiv auf die Domäne von Konzentrationsleistungen spezifiziert hat, jedoch ist sein Geltungsbereich geeignet, konkrete Vorhersagen hinsichtlich zahlreicher (für die Konzentrationsdiagnostik relevanter) experimenteller Variablen zu machen. Das Modell integriert komputationale und energetische Ansätze der Informationsverarbeitung zu einem interaktiven System. Demgemäß ist die Leistung von Probanden nicht allein durch komputationale Parameter bestimmt, sondern auch von der Motivationslage oder vom aktuellen Zustand bzw. der Befindlichkeit einer Person abhängig. Grundsätzlich ist das KEM ein Modell für Verarbeitungsstufen, welches aber energetische Aspekte mit berücksichtigt. Der Terminus Leistung ist dabei entsprechend definiert als die Effizienz der Verarbeitung von Reizen, deren Klassifikation gemäß einer Zuordnungsregel sowie die Übersetzung in eine entsprechende Reaktionsalternative. Auf komputationaler Ebene unterscheidet das KEM vier Verarbeitungsstufen, die jeweils in Verbindung mit dem energetischen System stehen: (1) Reizvorverarbeitung, (2) Merkmalserkennung, (3) Reaktionswahl und (4) motorische Adjustierung. Auf der energetischen Ebene unterscheidet das KEM zwei Bottom-up Level Systeme, (a) phasisches Arousal und (b) tonische Aktivierung sowie (c) ein Top-down gesteuertes Effort-System. Dabei nimmt das Modell auf der Metaregulationsebene einen mit Evaluation bezeichneten Überwachungsmechanismus an.

Obwohl das kognitiv-energetische Modell (KEM) nicht explizit für die Vorhersage von individuellen Differenzen in der Aufgabenleistung konstruiert wurde, so lassen sich dennoch Hypothesen bezüglich individueller Differenzen in der Konzentrationsfähigkeit aufstellen. Tatsächlich hat das Modell große Beachtung in der

neuropsychologischen Forschung zum Aufmerksamkeits-Defizit-Hyperaktivitäts--Syndrom (ADHS) gefunden. Sergeant und Kollegen (Sergeant, 2000, 2005; Sergeant & van der Meere, 1990) haben dazu ein umfangreiches Forschungsprogramm durchgeführt, in dem die Aufgabenleistung von Personen mit klinisch-diagnostizierter ADHS Symptomatik zu der Leistung von gesunden Probanden (Kontrollgruppe) verglichen wurden. Im Ergebnis kommen die Autoren zu dem Schluss, dass Personen mit ADHS nicht etwa durch eine defizitäre Verarbeitungsweise auf einer bestimmten Stufe charakterisiert sind, sondern eher durch eine ineffiziente Regulation der energetischen Systeme „Arousal“ und „Aktivierung“ durch das supervisorische Überwachungssystem („Evaluation“). Demzufolge sind die Personen aus dieser Gruppe (verglichen zu Personen einer Kontrollgruppe) schlechter in der Lage, die Konzentration während kontinuierlicher (insbesondere langwährender) Tätigkeit zu regulieren, was sich in einer höheren Variabilität der Reaktionsleistung äußert. Gemäß Sergeant und Kollegen gibt es also keine bestimmte „kognitive Funktion“, die mit einer bestimmten „Aufgabenform“ gemessen werden kann, welche Personen mit ADHS charakterisiert. Vielmehr haben diese Personen Probleme mit der energetischen Zustandsregulation während kognitiv anstrengender Aufgaben. Konsistent mit dieser Sichtweise sind auch Befunde, welche Assoziationen zwischen ADHS und Reaktionszeitvariabilität in kontinuierlichen Aufgabenformen nachgewiesen haben (z. B. de Zeeuw et al., 2008; Ridderinkhof, Scheres, Oosterlaan, & Sergeant, 2005).

Das kognitiv-energetische Modell (KEM) scheint also durchaus geeignet zu sein, individuelle Differenzen in der Konzentrationsleistung zu beschreiben. Flehmig (2010) hat die Ansicht vertreten, dass mit dem Modell auch individuelle Differenzen in der Konzentrationsleistung normaler Personen beschrieben werden können, wenngleich die Unterschiede auch etwas geringer ausfallen sollten, als bei der Analyse von Gruppenunterschieden (d. h. Patienten versus Kontrollprobanden). Dem Autor folgend

sind Konzentrationstests hervorragend dafür geeignet, Prozesse der Selbstregulation kognitiv-energetischer Zustände zu erfassen, da Personen bei dieser Testform Aufgaben in selbstbestimmter Geschwindigkeit (self-paced) bearbeiten müssen. Außerdem lassen sich anhand des Modells auch Fragen nach der Lokalisierung von Verarbeitungsstufen mit Bezug auf individuelle Differenzen beantworten. Diesbezüglich haben Rammsayer und Stahl einige Arbeiten vorgelegt, welche zum Beispiel zeigen, dass die Persönlichkeitseigenschaft Extraversion mit einem (in Ruhe bzw. unter geringer Stimulation) verminderten Arousal assoziiert ist, jedoch bei Stimulation mit einer stärkeren motorischen Aktivierbarkeit assoziiert ist (z. B. Rammsayer & Stahl, 2004; Stahl & Rammsayer, 2004; Stahl & Rammsayer, 2008). Flehmig (2010) hat in seiner Arbeit verschiedene Möglichkeiten erläutert, wie sich die Interaktion von Verarbeitungsstufen und energetischen Faktoren im Rahmen des KEM für diagnostische Zwecke (z. B. Testentwicklung) nutzbar machen lassen könnten. Ein von ihm vorgeschlagener, stärker theoriegeleiteter Ansatz psychometrischer Testentwicklung wäre wünschenswert. Wünschenswert wäre womöglich auch eine stärkere gemeinsame Berücksichtigung von kognitiven Theorien und psychometrischen Modellen der Konzentrationsleistung.

So zum Beispiel besteht immer noch Unklarheit darüber, welche Rolle verschiedene experimentelle Faktoren in klassischen Konzentrationstests hinsichtlich psychometrischer Gütekriterien (Reliabilität und Validität) besitzen. Eine Testverlängerung führt bekanntlich bei allen Personen zu Ermüdung (bzw. Sättigung, Motivationszerfall) und zu einem Abfall der Testleistung. Unklar ist jedoch bis heute, ob eine solche experimentelle Manipulation zu einer Vergrößerung oder Verkleinerung individueller Differenzen führt (Langner & Eickhoff, 2013). Wie lang sollte ein Test eigentlich sein, damit er bestmöglich valide, aber immer noch hinreichend ökonomisch ist? Die Antwort kann nur empirisch ermittelt werden und hängt sehr wahrscheinlich auch davon ab, welche Zielgruppe (Patienten, Bewerber in einem Personalverfahren etc.)

mit dem Verfahren diagnostiziert werden soll (s. Stuss, Meiran, Guzman, Lafleche, & Willmer, 1996; Van Breukelen et al., 1996; Westhoff & Hagemeister, 1992). Ebenso ungeklärt ist die Frage nach der optimalen Schwierigkeit eines Konzentrationstests sowie der Verarbeitungsstufe, auf welcher die Schwierigkeit hergestellt wird. Verschiedene Autoren haben diesbezüglich argumentiert, dass Konzentrationstests besonders leicht sein sollten (z. B. Van Breukelen et al., 1996; Van der Ven et al., 1989; Westhoff & Kluck, 1984), aber ist dieser Vorschlag auch optimal hinsichtlich der Testvalidität? Die Antwort auf diese Frage kann nur empirisch herausgefunden werden. Forster und Lavie (2007) haben zum Beispiel beobachtet, dass eine Erhöhung der Itemschwierigkeit auf perzeptueller Stufe zu einer Verminderung (nicht Vergrößerung) der individuellen Differenzen führte. Ebenso verringerte sich die Kriteriumsvalidität (zu selbstberichteter Fehlerneigung im Alltag).

Wie die soeben erläuterten Befunde vermuten lassen, spielen motivationale Einflüsse bei der Informationsverarbeitung eine große Rolle. In einer Reihe von Studien wurde gezeigt, dass sich eine schon vor der Testung bestehende motivationale Einstellung auch während der Bearbeitung von Tests ändern kann. Hierfür wurde der Einfluss untersucht, den Testbearbeitungszeit (Time-On-Task, TOT) oder Schlafmangel auf Parameter der Aufgabenleistungen haben. Obwohl der Terminus Motivation vielfältig theoretisch abgeleitet und praktisch gebraucht wird, soll er hier (im Rahmen des KEM) als „energetische Variable“ und „auf die Aufgabe bezogen“ verwendet werden. Ein hohes Maß an Motivation ist also gegeben, wenn eine Person gewillt ist, der Aufgabeninstruktion zu folgen, sich bestmöglich anzustrengen und über eine lange Testdauer hinweg durchzuhalten. Als Zustandsvariable kann Aufgabenmotivation mittels dafür entwickelter Fragebogenverfahren gemessen werden. Ein bekanntes Verfahren ist hierfür der DSSQ (Dundee Stress State Questionnaire) von Matthews und Kollegen, welcher die drei grundlegenden Dimensionen „task enga-

gement“ (TE, Aufgabenmotivation), „distress“ (DI, Nervosität) und „worry“ (WO, Besorgnis) in Leistungssituationen erfasst (Matthews et al., 2002). Aufgabenspezifische Motivationseffekte werden üblicherweise als Vorher-Nachher-Messung realisiert. Dabei haben experimentelle Variablen unterschiedliche Effekte auf die subjektive (mittels DSSQ gemessene) Aufgabenmotivation bzw. den Zustand einer Person. Langner, Steinborn und Kollegen haben zum Beispiel zeigen können, dass sich über die Dauer der Bearbeitung einer Einfachreaktionsaufgabe eine Verminderung der DSSQ-Dimension TE („task engagement“) beobachten lässt, sich jedoch kaum eine Verminderung der Dimensionen DI („distress“) und WO („worry“) zeigt (Langner et al., 2010).

Innerhalb einer kognitiv-energetischen Sichtweise ist aber weiterhin noch zu unterscheiden, ob sich die Ausprägung von Motivation auf die allgemeine (auf die Aufgabe ausgerichtete) Energetisierung bezieht oder ob sie selektiv (durch Instruktion) auf bestimmte Aspekte der Leistungssituation ausgerichtet wird. Man kann verallgemeinern, dass sich allgemeine (d. h. unselektive) Aufgabenmotivation durch Instruktion nur kurzfristig erreichen lässt. Falkenstein und Kollegen (Falkenstein, Hoormann, Hohnsbein, & Kleinsorge, 2003) haben in einer bemerkenswerten Studie zeigen können, dass die Reaktionen der Probanden schneller werden, wenn gelegentlich (d. h. in 20 Prozent aller Durchgänge) ein Hinweis gegeben wird, sich „jetzt anzustrengen“ (sogenannter „effort-cue“). Dies äußerte sich auch in elektrophysiologischen (EEG) Markern, welche als Indikatoren der vorbereitenden Mobilisierung von kognitiven Ressourcen gelten. Die Autoren führen aber an, dass diese Art der experimentell induzierten Ressourcenmobilisierung nur effektiv sein kann, wenn der Effort-Cue selten gegeben wird. Würde man einen solchen Hinweis häufig administrieren, dann wäre keine derartige Steigerung der Aufgabenleistung zu erwarten. Während die Methode des gelegentlichen Hinweisreizes eine Möglichkeit darstellt, die Aufgabenleistung von Testpersonen allgemein zu steigern, so gibt es auch zahl-

reiche Techniken zur selektiven Ausrichtung der individuellen Arbeitsweise. Eine Möglichkeit besteht in einer Veränderung der Balance zwischen Geschwindigkeit und Genauigkeit von Testpersonen, wobei man hier zwischen Mikro- und Makrostrategien unterscheidet (Sanders, 1998, S. 39).

Zu den Mikrostrategien zählen experimentelle Manipulationen, die zu einer kurzfristigen Dynamik des Arbeitsverhaltens führen. Zum Beispiel wird angenommen, dass Rückmeldungen über Fehler (engl. error feedback) in Reaktionszeitaufgaben kurzfristig zu vorsichtigerem Reaktionsverhalten von Testpersonen führen. Es soll an dieser Stelle jedoch angemerkt werden, dass Testpersonen auch ohne jedwede Fehlerrückmeldung auf ihre eigenen Fehler reagieren, indem sie im nachfolgenden Durchgang langsamer sind. Dieses als Fehlerverlangsamung (engl. posterror-slowing) bezeichnete Phänomen tritt besonders stark in Testsituationen auf, in denen subsequente Durchgänge unmittelbar nacheinander administriert werden (engl. self-paced), nimmt jedoch in seiner Stärke ab, wenn das Reaktions-Stimulus-Intervall (RSI) vergrößert wird (s. Jentzsch & Dudschig, 2009; Jentzsch & Leuthold, 2006). Außerdem ist die Fehlerverlangsamung größer, wenn aufgabenbedingt weniger Fehler gemacht werden, d. h. wenn Fehler selten sind (z. B. Jentzsch & Leuthold, 2006; Notebaert et al., 2009). Innerhalb einer experimentellen Situation zeigen jene Testpersonen eine stärker ausgeprägte Fehlerverlangsamung, deren Arbeitsweise besonders akkurat ist (z. B. Steinborn, Flehmig, Bratzke, & Schroter, 2012). Zu den Makrostrategien gehören experimentelle Manipulationen, welche die Arbeitsweise von Testpersonen gezielt beeinflussen. So können Versuchspersonen zum Beispiel instruiert werden, entweder besonders fehlerarm oder besonders schnell zu arbeiten. Jentzsch und Leuthold (2006) haben überdies gezeigt, dass Mikro- und Makroebene der Geschwindigkeits-Genauigkeits-Regulation interagieren. Unter der Instruktion, besonders genau zu arbeiten (verglichen zur Instruktion, besonders schnell zu arbeiten), zeigen Versuchs-

personen auch eine stärkere Reaktivität nach Fehlern, d. h. sie neigen zu stärker ausgeprägtem post-error slowing.

Insgesamt gesehen kann festgestellt werden, dass sich die im Test erfassbare Konzentrationsleistung von Probanden nur durch Berücksichtigung sowohl komputationaler als auch energetischer Komponenten beschreiben lässt. Alle Modelle gehen von der restriktiven Definition von Konzentrationstest aus, derzufolge ein ebensolcher Test in (1) selbstgewählter (self-paced) Arbeitsgeschwindigkeit durchgeführt wird und (2) die Testbearbeitung über eine bestimmte Zeitdauer erfolgt. Davon abzugrenzen sind also alle Verfahren, in denen die Items zeitlich voneinander durch einen Reaktions-Stimulus-Intervall (RSI) separiert sind und in denen die Präsentation der Items durch ein Warnsignal angekündigt wird. In einer solchen Situation hätten die Testpersonen die Möglichkeit, das Target (auf das zu reagieren ist) zeitlich zu antizipieren (s. Steinborn & Langner, 2011; Steinborn & Langner, 2012). Eine zeitliche Separierung von Durchgängen durch die Einfügung eines mehr oder weniger langen RSI würde außerdem die Möglichkeit eröffnen, sich kurzfristig zu erholen. Gemäß Jensen (2006, S. 180) wäre jede experimentelle Manipulation, welche die Aufmerksamkeit der Testperson extern optimiert (d. h. sicherstellt, dass die Testperson während der Bearbeitung konzentriert ist), dazu geeignet, das Konstrukt „mentale Geschwindigkeit“ zu messen. Jede experimentelle Manipulation, welche eine Erschwernis der Arbeitsleistung bewirken soll, ist demgemäß jedoch geeignet, das Konstrukt Konzentration zu messen. Dieser Aspekt wurde auch von Schmidt-Atzert und Kollegen betont, denen zufolge Konzentration als die aktive und andauernde Fokussierung der Aufmerksamkeit auf eine nicht intrinsisch motivierende Tätigkeit verstanden werden kann (s. Schmidt-Atzert et al., 2006; Schmidt-Atzert, Büttner, & Bühner, 2004).

2.3 Psychometrische Modellierung

Ein ganz wichtiges Merkmal von Konzentrationstests ist, dass von den Testpersonen eine aktive Reaktionstätigkeit abverlangt wird. Dieses Prinzip einer aktiven Zuwendung auf die fortdauernde Reaktionstätigkeit ist das eigentliche Grundprinzip eines jeden Konzentrationstests und alle kognitionspsychologischen Modelle zur Beschreibung von Konzentrationsleistungen basieren darauf (Pieters, 1985; Van Breukelen et al., 1996; Van der Ven et al., 1989; Westhoff & Kluck, 1984). Das Modell kumulativer Überladung nimmt dabei an, dass sich während der Testbearbeitung ein stetig ansteigender Widerstand gegen die weitere Fortführung der Tätigkeit ergibt. Dieser Gradient steigt bis zu dem Punkt an, an dem die Fortführung der Reaktionstätigkeit kurzfristig unterbrochen ist, was sich in extrem langen Reaktionszeiten (sogenannten mentalen Blockierungen) äußert. Bills hat dieses Phänomen in zahlreichen Experimenten nachgewiesen (s. Bills, 1931). Dem Modell gemäß steigt die Wahrscheinlichkeit für eine mentale Blockierung kontinuierlich an, bis diese tatsächlich eintritt. Zahlreiche Autoren haben dieses Phänomen als eine erzwungene Kurzpause interpretiert, welche die kognitive Überladung abbaut und danach das Weiterarbeiten ermöglicht. Das kognitiv-energetische Modell (s. Sanders, 1983; 1998, Kap. 9) steht mit dem Überladungsmodell nicht im Gegensatz, jedoch erweitert es den theoretischen Mechanismus um zwei Aspekte, (1) die Lokalisierung von Blockierungen auf Verarbeitungsstufen (perzeptuell, zentral, motorisch) und (2) die Möglichkeit proaktiver Steuerung (Selbstregulation) von kognitiver Beanspruchung. Damit ist gemeint, dass Testpersonen in Konzentrationstests das Auftreten von Blockierungen vorausschauend vermeiden können, indem sie ihre Arbeitsgeschwindigkeit optimieren und somit kumulative Überladung verhindern.

Kognitive Korrelate und kognitive Komponenten

Während jedoch die kognitionspsychologische Analyse von Reaktionsleistungen sowohl eine Überprüfung von Hypothesen über Effekte experimenteller Faktoren erfordert, als auch deren Modellierung, so erfordert die psychometrische Analyse eine Überprüfung von Testverfahren hinsichtlich ihrer Testgütekriterien Reliabilität und Validität. Beide Forschungstraditionen sind derzeitig nur ungenügend miteinander verbunden und aktuelle Entwicklungen in der kognitionspsychologischen Welt werden in der testdiagnostischen Welt nur sehr verzögert wahrgenommen. In der Tat hat man den Eindruck, dass die meisten handelsüblichen Konzentrationstests weniger aus theoretischen Überlegungen heraus entstanden sind, sondern aus eher intuitiven Überlegungen und Ideen heraus. Zahlreiche Autoren haben es daher für sehr wünschenswert erachtet, dass kognitionspsychologische und diagnostische Forschung zu Konzentrationsleistungen in der Zukunft stärker miteinander verzahnt werden sollten. Dies betrifft mindestens zwei Aspekte: (1) die Terminologie theoretischer Modelle und (2) die methodische Herangehensweise bei empirischen Studien. Während zum Beispiel diagnostische Studien Begriffe wie Aufmerksamkeit, Konzentration oder Arbeitsgedächtnis als Etikett für bestimmte durch Testverfahren erfassbare Konstrukte auffassen, so werden selbige in kognitionspsychologischen Studien kontextsensitiv behandelt. Der Begriff Aufmerksamkeit ist dort kein eigenständiges Konstrukt, dessen Ausprägung (bei Personen und in Situationen) „gemessen“ werden kann, sondern ein Oberbegriff für alle beteiligten Prozesse, welche bei der Reizaufnahme eine Rolle spielen. Zweitens verwenden diagnostische Studien in der Regel auf der Korrelationsanalyse basierende Verfahren (z. B. Faktorenanalyse), um Zusammenhänge zwischen Tests zu ermitteln, ohne jedoch experimentelle Faktoren, die kritisch sein können, zu berücksichtigen.

Schweizer (2005) hat dieses Problem aufgegriffen und einen stärkeren Bezug

auf kognitionspsychologische Grundlagen in der diagnostischen Forschung gefordert. In der Literatur zu individuellen Unterschieden in der kognitiven Verarbeitungsgeschwindigkeit (engl. mental speed) hat sich eine starke Verzahnung experimenteller und korrelativer Forschung abgezeichnet, die Schweizer als (1) den Ansatz kognitiver Korrelate und (2) den Ansatz kognitiver Komponenten bezeichnet. Zu den Untersuchungstechniken des ersteren Ansatzes gehören im Wesentlichen korrelative Analysemethoden. Es wird nach Zusammenhängen zwischen Tests gefragt, wobei das große Ziel einer solchen Strategie die Strukturmodellierung von Testvarianten sowie deren Zuordnung zu latenten Konstrukten ist. Psychometrische Tests zur Erfassung zweier diagnostischer Konstrukte, z. B. Aufmerksamkeit und Arbeitsgedächtnis, sollten zum Beispiel jeweils hoch untereinander korrelieren, zwischen Repräsentanten beider Konstrukte sollte die Korrelation allerdings nur niedrig ausfallen. Schmidt-Atzert, Bühner und Enders (2006) haben sich zum Beispiel gefragt, ob die verschiedenen im Handel erhältlichen Testverfahren zur Erfassung von Konzentration überhaupt dasselbe Konstrukt erfassen. Dazu wurde eine große Stichprobe mit den entsprechenden Verfahren getestet und deren Testleistung dann miteinander korreliert. Da alle verwendeten Tests ausgewiesen waren, die Konstrukte Aufmerksamkeit und Konzentration zu messen, erwarteten die Autoren eine entsprechend hohe Korrelation der Testleistungen. Nach Durchführung einer explorativen Faktorenanalyse zeigte sich dann zwar auch ein gemeinsamer Faktor, der als „Konzentration“ bezeichnet wurde, jedoch gab es auch viele kleinere Faktoren, die andere Prozesse messen (z. B. Gedächtnis, Koordination). Konzentrationstests messen also im Prinzip schon ein sehr ähnliches Konstrukt, je nach Aufgabenmaterial und Antwortmodus jedoch auch verschiedene Konstrukte.

Die Frage nach der faktoriellen Validität von Konzentrationstests hat Krumm (2008) in seiner Dissertation weiter verfolgt, indem er den Einfluss von verschiedenen Randbedingungen untersuchte. Krumm, Schmidt-Atzert und Eschert (2008) haben

zum Beispiel untersucht, inwiefern bestimmte Aufgabencharakteristiken die Interkorrelationen zwischen verschiedenen Tests beeinflussen. Die 110 Probanden der Stichprobe bearbeiteten dabei Testvarianten jeweils in selbstbestimmter (self-paced) oder fremdbestimmter (computer-paced) Arbeitsgeschwindigkeit und jeweils in Einzeldarbietung versus Reihendarbietung. Insbesondere der Arbeitsmodus (self-paced versus computer-paced) spielte eine bedeutsame Rolle, denn unterschiedliche Tests mit gleichem Arbeitsmodus korrelierten höher miteinander als gleiche Tests mit unterschiedlichem Arbeitsmodus. In einer weiteren Studie haben Krumm et al. (2008) untersucht, ob sich Tests, welche entweder explizit als Konzentrationstest oder als Intelligenztest (zur Erfassung der mentalen Informationsgeschwindigkeit) ausgewiesen sind, ein ähnliches Konstrukt erfassen bzw. ob sich diese Konstrukte mittels Strukturgleichungsmodellierung differenzieren lassen. Dazu wurden den 199 Probanden der Stichprobe insgesamt 24 verschiedene Testverfahren dargeboten, die sich jeweils konzeptuell als Konzentrationstest oder Speedtest klassifizieren ließen. Die Ergebnisse dieser Studie zeigen jedoch, dass diese Testverfahren ein sehr ähnliches Konstrukt messen, denn die Korrelation der Tests waren sowohl innerhalb als auch außerhalb der beiden Kategorien ähnlich hoch gelagert. Gemäß Krumm et al. (2008) liegt das einerseits daran, dass innerhalb von Testkonstrukteuren ein gewisser Grad an terminologischer Unschärfe besteht, andererseits aber auch eine tatsächlich existierende konzeptuelle Überlappung der Konstrukte Konzentration und mentale Verarbeitungsgeschwindigkeit besteht.

Die Studie von Moosbrugger, Goldhammer und Schweizer (2006) ist an dieser Stelle noch nennenswert, da die Autoren eine Unterscheidung bzw. Kategorisierung von Testverfahren nach Verarbeitungsstufen vorgenommen haben. Allerdings sind keine experimentellen Variablen manipuliert worden, sondern es sind lediglich Tests verwendet worden, welche die Autoren auf entweder der perzeptuellen oder der

zentralen Stufe lokalisierten. Die Autoren haben die 232 Probanden der Stichprobe mit insgesamt elf Testverfahren untersucht, welche entweder als Kategorie wahrnehmende Aufmerksamkeit (perceptual attention) oder handlungsorientierte Aufmerksamkeit (executive attention) klassifiziert wurden. In ersterer Kategorie befanden sich typische visuelle Durchstreichtests, während sich in der zweiten Kategorie eher Tests vom Typ der Wahlreaktionsaufgabe befanden. Basierend auf ihren Ergebnissen der Strukturgleichungsmodellierung postulierten die Autoren ein Zwei-Faktorenmodell mit den Faktoren perzeptuelle Aufmerksamkeit (Wahrnehmen) und exekutive Aufmerksamkeit (Handeln). Im Prinzip hat diese Studie gezeigt, dass Tests zur Erfassung von diagnostischen Konstrukten wie Aufmerksamkeit, Konzentration und Arbeitsgedächtnis sich in eine nomologische Struktur bringen lassen, die sich in zwei Hauptklassen unterscheidet, visuelle Scan-und-Check-Verfahren und Wahlreaktionsverfahren (vgl. Van Breukelen et al., 1996). Diese Studie bestätigt empirisch eine bereits von Westhoff und Kollegen vorgeschlagene Differenzierung von Aufmerksamkeit und Konzentration. Danach soll Aufmerksamkeit als kognitive Tätigkeit betrachtet werden, die sich auf das Wahrnehmen bezieht, während Konzentration als eine auf das Handeln bezogene kognitive Tätigkeit betrachtet werden soll. Eine Limitation der Studie besteht darin, dass lediglich die Reaktionszeiten nicht aber die Fehler als Leistungskennwerte untersucht worden sind.

Innerhalb des zweiten Ansatzes (kognitiver Komponenten) spielen experimentelle Methoden und kognitionspsychologische Theorien eine wichtige Rolle. Hier werden experimentelle Faktoren systematisch variiert (z. B. Itemschwierigkeit, Aufgabenlänge etc.), die hinsichtlich der Testgütekriterien eines Verfahrens als besonders kritisch gelten. Das Ziel einer solchen Strategie ist zunächst die Lokalisierung von beteiligten kognitiven Prozessen als Determinanten individueller Differenzen. Die definite Zielstellung besteht jedoch in der Bereitstellung von Wissen zur optimalen Konstruk-

tion von Testverfahren und auch in der Optimierung von Testverfahren bezüglich ihrer psychometrischen Gütekriterien Reliabilität und Validität. So ist zum Beispiel mehrfach gezeigt worden, dass der Zusammenhang zwischen mentaler Verarbeitungsgeschwindigkeit (erfasst mittels einer Wahlreaktionsaufgabe) und Intelligenz (erfasst mittels eines Matrizentests) nicht einfach per se existiert, sondern durch die Komplexität der Reaktionen in der Wahlreaktionsaufgabe moduliert wird. Dazu wurde das Hick-Hyman-Paradigma verwendet (Hick, 1952; Hyman, 1953), in welchem die Wahlreaktionszeit als Funktion der Anzahl der Reaktionsalternativen ermittelt wird. In diesem Paradigma müssen Probanden auf dargebotene Reize mit einer Wahlreaktion antworten, wobei die Wahlmöglichkeiten variiert werden. Dabei muss nach Erscheinen von entweder zwei, vier oder acht möglichen Reizen eine dem Reiz zugeordnete Taste gedrückt werden. Dabei zeigte sich, dass die Reaktionen als Funktion der Anzahl der Alternativen immer langsamer werden, vor allem aber, dass die Korrelation zwischen Reaktionszeit und Intelligenz als Funktion dieser Alternativen ansteigt. Einige Autoren haben auch die Steigung dieser Hick-Hyman-Funktion mit Intelligenz in Zusammenhang gesetzt, und obgleich die Korrelationen nicht allzu hoch ausfielen (r =.20 bis .50), ist dieser Befund mehrfach repliziert worden (s. Beauducel & Brocke, 1993; Jensen, 2006, S. 158-162).

Wenn tatsächlich ein spezifischer korrelativer Zusammenhang zwischen Wahlreaktionszeit und Intelligenztestleistung besteht, der durch die Anzahl der Reaktionsalternativen moduliert wird, dann liegt es nahe, auch einen allgemeinen Zusammenhang zwischen Aufgabenkomplexität in chronometrischen Paradigmen und Intelligenz anzunehmen. Innerhalb des Ansatzes kognitiver Komponenten wäre es also zu erwarten, dass die psychometrischen Gütekriterien von Konzentrationstests (Reliabilität und Validität) auch durch eine Manipulation der Aufgabenkomplexität optimiert werden könnten. Westhoff und Kluck (1984) haben diesbezüglich bereits einen theoretischen

Ansatz vorgelegt, indem sie Konzentrationsleistungen im Test als Funktion der Komplexität des Reizmaterials, der Komplexität der Reaktionen und der Erfahrung (bzw. Testübung) beschreiben. Westhoff und Graubner (2003) sind noch weiter gegangen und haben den theoretischen Ansatz empirisch überprüft, indem sie sowohl Aufgabenform (figurales, numerisches, verbales Material) als auch Aufgabenkomplexität (in fünf Stufen) systematisch auf ihre prädiktive Validität hin untersucht haben. Die Komplexität wurde dabei jedoch auf unterschiedlichen Stufen variiert. Bei figuralem Material bestand die Schwierigkeit darin, die perzeptuelle Ähnlichkeit von Target und Distraktor zu erhöhen, bei numerischem Material wurde die Kettenlänge von Additionsaufgaben erhöht und beim Wortumkehrtest wurde die Häufigkeit des lexikalischen Gebrauchs als Index der Schwierigkeit verwendet. Dies zeigt, dass es viele Möglichkeiten gibt, die Komplexität der Aufgabe zu variieren. Während eine Erhöhung der Target-Distraktor-Ähnlichkeit vornehmlich Anforderungen an die Wahrnehmung stellt, so wird durch eine Manipulation der Kettenlänge die Arbeitsgedächtnisanforderung erhöht.

Effekte experimenteller Variablen auf Testgüte

Da eine Erhöhung der Aufgabenkomplexität sowohl die Gütekriterien als auch die praktische Anwendungsbreite (das Spektrum der prädiktiven Validität) von Konzentrationstests erhöhen könnte, ist dieser Ansatz für die Testkonstruktion und psychologische Diagnostik sehr wichtig. Mit einer Erhöhung der Aufgabenkomplexität sind aber mehrere Probleme verbunden: (1) die Konstruktvalidität von Konzentrationstests verändert sich, indem entsprechende Tests einen stärkeren Anteil am Arbeitsgedächtnis messen. Ein solcher Test würde also nicht mehr allein Konzentrationsfähigkeit messen, sondern ein Mischkonstrukt (s. Schmidt-Atzert et al., 2006; Schmidt-Atzert et al., 2004). Sofern eine solche Manipulation die prädiktive Validität erhöht, kann dies in praktischen diagnostischen Kontexten auch ein Vorteil sein. (2) Die Übungsabhängigkeit von Tests erhöht sich mit zunehmender Komplexität. Gemäß Logan und Kollegen

werden Items in Additionsaufgaben bei mehrfach wiederholter Bearbeitung zunehmend nicht mehr durch mentales Rechnen sondern durch puren Gedächtnisabruf gelöst. Schwierigere Items zeigen dabei eine stärkere Lernkurve als leichtere Items (s. Compton & Logan, 1991; Logan, 1992; Logan, 2002). Um diesem Problem zu begegnen, haben Hagemeister und Kollegen (Hagemeister, 2003, 2007; Hagemeister, Scholz, & Westhoff, 2002) verschiedene Techniken vorgeschlagen. In einer Serie von Experimenten wurde überprüft, ob sich Geübtheit an bestimmten Indikatoren des Testverhaltens erkennen lässt, z. B. dem Verhältnis der Reaktionszeit von leichten und schwierigen Items oder ihrer Lernkurve über den Test. Eine recht brauchbare Möglichkeit besteht darin, möglichst viele Items zu verwenden (s. Steinborn et al., 2008; Steinborn et al., 2010), sodass ein auf gedächtnisbasierter Abruf unmöglich gemacht wird. (3) Eine Manipulation der Aufgabenkomplexität muss auf Basis theoretischer Modelle geschehen, welche die psychometrischen Gütekriterien (Reliabilität und Validität) systematisch als Funktion der Aufgabenschwierigkeit in Beziehung setzt.

Ein theoretisches Modell der Aufgabenkomplexität in Konzentrationstests wurde von Düker (1930, 1949) vorgestellt und auf die Anforderungen der psychologischen Testdiagnostik zugeschnitten. Dabei definierte er Konzentration als die „Fähigkeit zur Anspannung zum Zwecke der Koordination von Teilhandlungen“. Eine wichtige Rolle spielt dabei der Begriff der Koordination, den der Autor beschreibt als „das zu einer Gesamttätigkeit geordnete Zusammenwirken der Einzeltätigkeiten, die zur Erreichung eines bestimmten Zweckes erforderlich sind“. Soll eine Rechenaufgabe, für die die Konzeption hauptsächlich gilt, gelöst werden, müssen beispielsweise folgende Einzeltätigkeiten koordiniert werden: Auffassen, Behalten, Reproduzieren, Vergleichen, Abstrahieren und Antworten. Die Aufgabe kann nicht (bzw. nur fehlerhaft) gelöst werden, wenn eine dieser Einzeltätigkeiten gar nicht ausgeführt wird oder wenn deren Abfolge nicht korrekt koordiniert wird. Der von Düker und Lienert (1965) entwickelte

Konzentrations-Leistungs-Test (KLT) basiert auf dieser theoretischen Konzeption und ist heute noch in revidierter Fassung im Handel erhältlich (Düker, Lienert, Lukesch, & Mayrhofer, 2002). Während es zahlreiche Validierungsstudien gibt, um die psychometrischen Gütekriterien des KLT zu bestimmen, gibt es aber kaum Studien, die eine systematische Variation bestimmter Testeigenschaften des KLT vorgenommen haben, um deren Einfluss auf die psychometrischen Eigenschaften festzustellen. Wir wissen also nicht, ob der KLT hinsichtlich experimenteller Variablen wie Itemschwierigkeit, Itemhomogenität und Testlänge optimal gestaltet ist. Die systematische Variation kritischer Variablen wird aber von vielen Autoren gefordert, da sonst jegliche Testkonstruktionsforschung lediglich Produktforschung bleibt (vgl. Jensen, 2006, S. 155-187; Matthews et al., 2002; Miller & Ulrich, 2013).

Um jedoch generalisierbaren Erkenntnisgewinn zu erbringen, ist es nötig, Testvarianten mit systematischer Variation kritischer Faktoren auf ihre Reliabilität und Validität hin zu überprüfen. Das kognitiv-energetische Modell (KEM) von Sanders (1998, Kap. 9) ermöglicht eine theoretisch motivierte Herangehensweise bei der Manipulation von kritischen Aufgabenparameter. Innerhalb dieses Modells kann z. B. Itemschwierigkeit auf verschiedenen Verarbeitungsstufen (perzeptuell, zentral, motorisch) erfolgen, und deren Effekt auf psychometrische Gütekriterien geprüft werden. Kann zum Beispiel gezeigt werden, dass eine bestimmte Aufgabenanordnung in einem Test eine höhere Retest-Reliabilität oder gar prädiktive Validität aufweist als eine herkömmliche Anordnung, dann wäre erstere in der Zukunft zu bevorzugen. Tatsächlich gibt es zahlreiche Studien, welche die prädiktive Validität verschiedener Tests oder kognitiver Paradigmen hinsichtlich eines Kriteriums (z. B. Schulnoten) überprüft haben (z. B. Colom, Escorial, Shih, & Privado, 2007; Krumm, Ziegler, & Bühner, 2008; Rindermann & Neubauer, 2004; Spinath et al, 2006; Westhoff & Graubner, 2003). Ein Problem dabei ist jedoch, dass die oftmals verwendeten Testverfahren nicht vergleich-

bar sind, da eine unterschiedliche Anzahl an Items in den Tests verwendet wurde, was wiederum Unterschiede in der zugrundeliegenden Reliabilität hervorrufen könnte. Miller und Ulrich (2013) haben in der Tat darauf hingewiesen, dass faktorenanalytische Ergebnisse schwer zu interpretieren sind, wenn sich die Tests hinsichtlich der Anzahl der Items unterscheiden. Gemäß den Autoren wären jegliche Strukturgleichungsmodelle nur interpretierbar, wenn die formal (d. h. extern, durch Anzahl der Items) hergestellte psychometrische Reliabilität für alle Testverfahren vergleichbar ist.

Gemäß Miller und Ulrich (2013) können die Gütekriterien verschiedener Testverfahren eigentlich nur vergleichend untersucht werden, wenn die Anzahl der Items gleich ist. Sonst hätte ein Test mit vielen Items einen Vorteil gegenüber einem Test mit nur wenigen Items. Dies folgt direkt aus den Annahmen der klassischen Testtheorie, gemäß dieser eine Messung immer genauer wird, je mehr Messereignisse vorliegend sind, aus denen ein Gesamtwert gebildet wird. Das bedeutet ganz einfach gesagt: Die Reliabilität eines Testverfahrens zur Messung von Konzentration vergrößert sich mit zunehmender Anzahl an Items, aus denen der Test besteht. Daraus folgt, dass eine Langversion eines Tests immer reliabler und prädiktiver sein sollte als die Kurzversion desselben Tests. Hagemeister (1994) konnte zum Beispiel zeigen, dass der Fehlerprozentwert im d2 Aufmerksamkeits-Belastungstest (Brickenkamp, 2002) nur ungenügend reliabel im Retest ist, dass jedoch eine Testverlängerung zu einer substantiellen Erhöhung der Reliabilität dieses Kennwerts führte. Dagegen war der Kennwert für Geschwindigkeit bereits in der Normalversion hoch reliabel und dementsprechend gab es auch keine nennenswerte Verbesserung der Reliabilität dieses Kennwerts. Dies zeigt wiederum, dass eine vergleichende Analyse der Gütekriterien verschiedener Durchstreichtests nur sinnvoll erfolgen kann, wenn die Anzahl der Items dabei relativiert wird. Gemäß Miller und Ulrich (2013) kann eine solche Relativierung nur empirisch erfolgen, indem die Reliabilität als Funktion der Anzahl von Items abgetra-

gen wird. Die Autoren zeigen in ihrer Arbeit, dass bestehende Korrekturformeln zu Reliabilitätsschätzungen bei Testverlängerung nicht bedenkenlos angewendet werden können, da experimentelle Faktoren eine große Rolle spielen.

Einige Autoren haben außerdem vermutet, dass sich die zugrundeliegenden Prozesse ändern könnten, wenn der Test extrem verlängert wird, da aufgabeninduzierte Ermüdung eintreten kann. Kraepelin (1902, 1903) und auch Pauli (1938) haben sogar explizit darauf abgezielt, den kognitiven Zustand der Probanden während der Testbearbeitung zu verändern. Aus einer testtheoretischen Perspektive stellt sich hier jedoch die Frage, ob aufgabeninduzierte Ermüdung (bzw. psychische Sättigung) zu einer Verbesserung oder gar einer Verschlechterung der Gütekriterien eines Tests führt. Führt aufgabeninduzierte Ermüdung zu einer Differenzierung der Probanden im Test oder gar zu einer Nivellierung? Aus Sicht der klassischen Testtheorie sollte sich jede Testverlängerung immer nur positiv auswirken (d. h. Probanden können besser differenziert werden, je länger der Test ist), da mehr Beobachtungen registriert werden, aus denen ein Mittelwert gebildet wird. Eine Änderung der Prozesse über den Testverlauf hinweg (z. B. das Hinzukommen von Ermüdung, Sättigung, Erfahrung, Übung, etc.) ist in den Annahmen der klassischen Testtheorie nicht enthalten. Da sich aber insbesondere die frühe Forschung mit energetischen Fragestellungen beschäftigt hat (vgl. Bills, 1937; Brandstätter, 1995), ist diese Annahme interessant. Pauli (1938) hat den Effekt lange andauernder Testbearbeitung als besonders kritisch für die diagnostische Eignungsfeststellung erachtet. Mierke (1955) vertrat ebenfalls die Ansicht, dass sich Konzentration im Test nur erfassen lässt, wenn eine entsprechende Grundsituation hergestellt wird, in der diese Fähigkeit auch evoziert wird.

Die Induktion von aufgabenspezifischer Ermüdung hat eine lange Tradition in der ergonomischen Aufmerksamkeits- und Vigilanzforschung, in der kognitive Zustände als kritisch für das Verhalten in bestimmten Situationen angesehen werden (vgl.

Finomore, Matthews, Shaw, & Warm, 2009; Langner & Eickhoff, 2013). Die Idee, einen Zustand aufgabeninduzierter Ermüdung im Probanden herzustellen, entspricht dem transaktionalen Ansatz, der sich insbesondere in der verkehrspsychologischen Forschung zu Fehlerverhalten und Unfällen etabliert hat (vgl. Gulian, Matthews, Glendon, Davies, & Debney, 1989; Matthews et al., 2002). Der Punkt ist, dass Fehlerverhalten oft nur in einer kritischen Situation gezeigt wird und diese Situation daher im Test induziert werden müsste, um Fehlerverhalten psychometrisch zu messen. Aus diesem Grund basieren zahlreiche Tests im Bereich der Fahrsimulation auf der Idee kritischer Ereignisse. Bei der Bearbeitung eines Konzentrationstests besteht jedoch der Effekt von Time-on-Task (TOT) nicht (oder nur selten) in einer Erhöhung der Fehlerrate, sondern vielmehr in einer Erhöhung der Reaktionszeitvariabilität (Langner et al., 2010; Steinborn et al., 2010). Zahlreiche Arbeiten haben sich mit den zugrundeliegenden Prozessen dieser Leistungsschwankungen beschäftigt. So ging z. B. Bills (1931) davon aus, dass die bei kontinuierlichen Tätigkeiten auftretenden mentalen Blockierungen durch eine akkumulierte Überladung des kognitiven Systems entstehen, die sich durch kurze Pausen schnell wieder abbaut (vgl. Bertelson & Joffe, 1963; Van Breukelen et al., 1996). Diese Auffassung ist durch zahlreiche Befunde unterstützt, da gezeigt werden konnte, dass TOT-induzierte Ermüdung vornehmlich die langsamen Reaktionen beeinflussen, während die schnellen Reaktionen einer Reaktionszeitverteilung unbeeinflusst bleiben (Sanders & Hoogenboom, 1970; Steinborn et al., 2010).

Während einige Autoren sich vornehmlich mit der psychometrischen Modellierung von Reaktionszeitverteilungen beschäftigt haben, so haben andere Autoren versucht, die Leistungsschwankung durch einen geeigneten Kennwert auszudrücken, der diagnostisch genutzt werden kann. Flehmig (2010) hat in seiner Dissertation die Frage untersucht, ob Kennwerte der Leistungsschwankung in Konzentrationstests psychometrisch reliabel und valide sind. Dabei untersuchte er die psychometrischen

Gütekriterien des Variationskoeffizienten gegenüber der Standardabweichung als klassisches Variabilitätsmaß. Es zeigte sich jedoch, dass die Reliabilität der Kennwerte stark abhängig vom Aufgabenmaterial war. In auf mentaler Addition basierenden Tests war die Retest-Reliabilität des Variationskoeffizienten zufriedenstellend hoch, während die Retest-Reliabilität in Scan-und-Check-Tests ungenügend war. Außerdem war die Reliabilität in allen untersuchten Tests geringer als die Reliabilität der gemittelten Reaktionszeiten. Immerhin steigt die Retest-Reliabilität des Variationskoeffizienten an, wenn ein Konzentrationstest über lange Zeit bearbeitet werden muss. In einer Studie von Steinborn et al. (2010) kam eine Testversion des SMACT zum Einsatz, in der die Probanden aufgefordert sind, einen dargebotenen Additionsterm zu berechnen und dann mit einer ebenfalls dargebotenen Zahl hinsichtlich ihrer Größe zu vergleichen. Die Autoren haben gezeigt, dass sich die aufgabeninduzierten Ermüdungseffekte beim fortlaufenden mentalen Addieren vornehmlich im Variationskoeffizienten (und weniger in den mittleren Reaktionszeiten) manifestieren. Dies kann als ein Beleg dafür gewertet werden, dass dieser Kennwert sensitiv für energetische Zustandsänderungen einer Person während der Bearbeitung von Konzentrationstests ist.

Obwohl der Einsatz von Kennwerten der Variabilität viele neue Möglichkeiten eröffnen würde, ist man von einer generellen Empfehlung eines psychometrischen Kennwert derzeit noch weit entfernt. Dies hat mehrere Gründe: (1) Zunächst sind Kennwerte der intraindividuellen Variabilität oft mit dem Mittelwert der Reaktionszeiten korreliert. Die Höhe hängt dabei von der Wahl des Kennwerts und von der im Test verwendeten Aufgabenform ab. Zusammenhänge mit der Standardabweichung liegen dabei höher (r = .30 bis .70) als Zusammenhänge mit dem Variationskoeffizienten (r = .00 bis .50). Werden relativ komplexe Aufgabenformen verwendet, dann sind diese Zusammenhänge ebenfalls höher, verglichen zu wenig komplexen Aufgabenformen (vgl. Flehmig et al., 2007). Es gibt also einen nicht unerheblichen Grad an

Redundanz zwischen Kennwerten der zentralen Tendenz und jenen der Streuung. (2) Alternative Verfahren der Analyse von Reaktionszeitverteilungen sind in jüngster Zeit immer mehr in Mode gekommen, welche auch eine Extraktion von Kennwerten erlauben und in Konkurrenz mit den klassischen statistischen Kennwerten stehen. Die Methode der Anpassung von Verteilungen erlaubt zum Beispiel die Extraktion von Parametern der Skalierung, Streuung und Schiefe einer Funktion (Bresin, Robinson, Ode, & Leth-Steensen, 2011; Leth-Steensen, Elbaz, & Douglas, 2000; Schmiedek et al., 2007). Lapsus während der Bearbeitung lässt sich hierbei durch den Kennwert der Schiefe ausdrücken und auf psychometrische Gütekriterien hin überprüfen. (3) Ein generell akzeptiertes Modell gibt es derzeit nicht. Zwei-Zustands-Modelle nehmen an, dass die Schiefe der Reaktionszeitverteilung durch zwei voneinander unabhängige Prozesse entsteht, also eine Mischung von Reaktionen sind, die im Zustand der (a) Konzentration und (b) der Abgelenktheit entstanden sind. Im Gegensatz dazu nehmen Prozess-Delay-Modelle an, dass die Schiefe durch Variabilität zentraler Prozesse entsteht. Eine Klärung der Gültigkeit beider Modelle bzw. deren situative Validität steht derzeit noch aus.

Neben der Frage der optimalen Itemschwierigkeit und Testlänge gibt es noch einen dritten wichtigen Aspekt, welcher bei der Konstruktion von Konzentrationstests zu berücksichtigen ist: Welches Aufgabenformat ist eigentlich am besten für die psychometrische Erfassung von Konzentrationsleistung geeignet? Wie bereits erörtert können zahlreiche Aufgabenformen als Konzentrationstest dienen, sofern deren Bearbeitung in selbstbestimmter Geschwindigkeit (self-paced) ausgeführt wird. Westhoff und Graubner (2003) haben hierbei zwischen numerischen, verbalen und figuralen Aufgabenmaterial unterschieden, das jeweils in unterschiedlichen Schwierigkeitsstufen dargeboten wird. Gemäß den Autoren ist es sinnvoll, Konzentrationsfähigkeit mittels einer Batterie diversifizierter Testverfahren zu erfassen, die jeweils verschiedenes Auf-

gabenmaterial enthalten. Gemäß dem Standpunkt diversifizierter Testadministration ist also eine multimethodale Herangehensweise auf der Ebene von Tests sinnvoll, da der Einfluss sowohl von materialspezifischen Erfahrungsunterschieden als auch von Teilleistungsschwächen auf die Testleistung minimiert werden kann. Es ist bereits von vielen Autoren angemerkt worden, dass bestimmte Berufsgruppen (z. B. Büroangestellte, Buchhalter usw.) einen Vorteil bei der Bearbeitung von Scan-und-Check-Konzentrationstests, also klassischen Durchstreichverfahren haben könnten. Nicht umsonst werden diese Verfahren im angloamerikanischen Raum als „clerical tasks“ bezeichnet (Jensen, 2006, S. 100). Zum anderen könnten Personen mit bestimmten Teilleistungsschwächen bei bestimmten Aufgabenformen besonders benachteiligt sein. Denkbar wäre zum Beispiel, dass Personen mit Lese- und Rechtschreibschwäche in Durchstreichverfahren besonders benachteiligt sind oder dass Personen mit Dyskalkulie in Rechenkonzentrationstests benachteiligt sind. Diversifikation von Aufgabenmaterial in Tests betrifft also auch den Aspekt, dass die Testfairness dadurch erhöht werden kann.

Dem Ansatz der Diversifikation steht der Standpunkt optimalen Testmaterials entgegen. Gemäß dieser Idee sollte es immer das Bestreben von Testentwicklern sein, jene Varianten von Aufgabenformen mit optimaler psychometrischer Eignung zu entdecken und für diagnostische Zwecke nutzbar zu machen. In der Intelligenzdiagnostik zeigt sich dieser Gegensatz besonders am Beispiel der Konstruktion von Intelligenztests, die innerhalb des Ansatzes des optimalen Testmediums immer als Matrizentests konzeptualisiert sind, innerhalb des Ansatzes der Testdiversifikation jedoch als Batterie mit zahlreichen Untertests konzeptualisiert sind. Das mit Matrizentests gemessene Konstrukt „schlussfolgerndes Denken“ steht dem allgemeinen Intelligenzkonzept wohl am meisten nahe und ist daher geeignet, im Rahmen einer ökonomischen diagnostischen Strategie als Testmedium zu dienen (vgl. Jensen, 2006,

Kap. 11). Es stellt sich für die Konzentrationsdiagnostik die Frage, wie eine Batterie von Tests aufgebaut sein sollte bzw. aus welchen Testvarianten sie zusammengesetzt sein sollte. Benötigt man überhaupt eine Testbatterie von Konzentrationstests? Schließlich vertreten einige Autoren den Standpunkt, dass Konzentrationstests hinsichtlich ihrer prädiktiven Validität zur Vorhersage verschiedenster Erfolgskriterien kein eigentlich wichtiges Konstrukt messen, das sich mit Intelligenz messen könnte (vgl. Jensen, 2006, Kap. 11). Andere Autoren führen an, dass die Überlegenheit von Intelligenztests vor allem dem Umstand geschuldet sei, dass eben große Testbatterien verwendet werden, welche schon allein durch ihren Umfang und den damit verbundenen Reliabilitätszuwachs einen Vorteil bei der Bewertung ihrer Validität haben (vgl. Cronbach, 1975; Hagemeister, 1994).

Man darf sich also durchaus folgende Frage stellen: Welche Arten von Tests sollten in welcher Zusammensetzung (und in welcher Gewichtung) in einer solchen Batterie enthalten sein? Im Folgenden sollen daher verschiedene Ansätze erläutert werden, aus denen Hinweise für eine systematische Zusammensetzung einer Konzentrationstestbatterie erfolgen könnte. Ausgehend von Westhoff und Graubner (2003) sollte eine Konzentrationsdiagnostik immer mehrere Materialarten verwenden, um erfahrungsspezifische Effekte zu minimieren. Die Autoren schlagen einen taxonomischen Ansatz vor, demzufolge eine Testbatterie aus einer faktoriellen Kombinierung der drei Materialarten (verbal, numerisch, figural) und der Itemschwierigkeit (leicht, mittel, schwer) bestehen könnte. Die Studie von Schmidt-Atzert et al. (2006) bestätigt diesen Vorschlag größtenteils. In zwei Studien (117 und 90 Probanden) wurde mit Hilfe konfirmatorischer Faktorenanalysen (11 und 10 Tests) gezeigt, dass es einen allgemeinen Konzentrationsfaktor gibt, dass aber auch ein Teil der Varianz durch materialspezifische Komponenten (verbal, figural, numerisch) und Merkfähigkeit erklärt wird. Ein anderes Modell orientiert sich an den von Posner und Petersen (1990)

vorgeschlagenen Komponenten der Aufmerksamkeit: (1) Alertnessregulation, (2) orientierende Aufmerksamkeit und (3) exekutive Aufmerksamkeit. Der von Fan et al. (2002) entwickelte ANT (engl. attention network test) beansprucht für sich, diese drei Komponenten der Aufmerksamkeit diagnostisch erfassen zu können. Allerdings hat höchstens die dritte Komponente (exekutive Aufmerksamkeit) etwas mit dem hier interessierenden Konstrukt Konzentration zu tun. Ob jedoch eine theoretische oder einer pragmatische Lösung im Rahmen einer Konzentrationstestbatterie angestrebt wird, bleibt immer eine empirische Frage. Denn schließlich geht es in angewandten Kontexten immer um die Validität bezüglich relevanter Kriterien einschließlich der Ökonomie der verwendeten Testverfahren.

Zusammenfassend ist festzustellen, dass es zur Frage der Testkonstruktion viele offene Probleme gibt, die bisher nicht gelöst worden sind. Die aktuelle Arbeit soll einige Aspekte offener Fragen aufgreifen und mittels experimenteller Methoden beantworten. Im Vordergrund stehen dabei Fragen nach der optimalen Testlänge und der richtigen Itemschwierigkeit eines Tests. Diese Aspekte sind vor allem deshalb wichtig, weil zahlreiche Studien diese überhaupt nicht berücksichtigt haben. Zahlreiche Strukturanalysen von Konzentrationstests basieren auf Verfahren, welche sich in Testlänge und Schwierigkeit unterscheiden, ohne dass dies in den Analysen berücksichtigt ist. Aus dieser Perspektive ist es unklar, ob es überhaupt ein Materialfaktor gibt oder ob dieser in Wirklichkeit ein Schwierigkeitsfaktor ist. So sind Rechenkonzentrationstests (numerisches Material) viel komplexer als Durchstreichtests (oft figurales Material) und es ist unklar, ob faktorenanalytische Studien nicht einfach die Schwierigkeit differenzieren ohne dass dies intendiert ist. Aus diesen methodischen Problemen ergibt sich die Motivation der folgenden Arbeit. Aus psychometrischen Vorstellungen über die Gütekriterien Reliabilität und Validität ergeben sich potentiell überprüfbare Hypothesen über den Einfluss dieser kritischen Variablen auf die prädiktive Validität.

Miller und Ulrich (2013) sind zu der Ansicht gekommen, dass selbst die Reliabilität eines Tests von einer Vielzahl von Einflussfaktoren abhängt und daher immer einer empirischen Überprüfung im individuellen Fall bedarf. Die hier vorliegende Arbeit soll einen Beitrag dazu leisten, den Einfluss dieser experimentellen Faktoren auf die Gütekriterien zu überprüfen.

3 Prädiktive Validität von Konzentrationstests

3.1 Reliabilität, Konstruktvalidität, Kriterienvalidität

Das Konstrukt „Konzentrationsfähigkeit“ stellt eine wichtige Fähigkeitskomponente dar, welche sich innerhalb von taxonomischen Aufmerksamkeitsmodellen im Bereich der Daueraufmerksamkeit einordnen lässt, gemäß einiger Autoren jedoch eine Sonderstellung einnimmt (vgl. Schmidt-Atzert et al., 2004; Westhoff & Kluck, 1984). So hat Sanders (1998, Kap. 9) angeführt, dass zahlreiche Aufmerksamkeitstests lediglich eine passive Überwachung und Detektion seltener Reize erfordern, die oft unter dem Terminus Vigilanz geführt werden. Aufgabenformen, welche jedoch eine aktive Bearbeitung über eine bestimmte Dauer erfordern, werden häufig unter dem Terminus „prolonged work tasks“ geführt (vgl. Healy, Kole, Buck-Gengler, & Bourne, 2004; Van Breukelen et al., 1996). Zahlreiche Studien deuten darauf hin, dass passive und aktive Daueraufmerksamkeit als zwei unterschiedliche Konstrukte angesehen und dementsprechend nicht in einen Topf geworfen werden sollten. Charakteristisch für Konzentrationstests ist dabei nicht allein die Art der Aufgabe bzw. des Paradigmas (es können zahlreiche Aufgabenformen als Materialgrundlage dienen), sondern dass die Geschwindigkeit selbstbestimmt (self-paced) erfolgt. In einem solchen Arbeits-

modus kann kurzfristige Überlastung schnell entstehen, die oft als psychologisches Refraktärphasenphänomen bezeichnet wird. Außerdem muss die Balance zwischen Geschwindigkeit und Genauigkeit regelmäßig überwacht und wenn nötig korrigiert werden. Ein behavioraler Hinweis auf eine solche Korrektur ist das sogenannte „post-error slowing“ Phänomen (vgl. Steinborn et al., 2012). Personen tendieren dazu, nach einem Fehler ihre Arbeitsgeschwindigkeit zu „überdenken“, wobei bislang unklar ist, ob dies durch eine strategisch motivierte Evaluation oder durch einen automatischen Mechanismus entsteht. Obwohl sich Autoren einig sind, dass die Leistung in Konzentrationstests nicht durch Strategien der Bearbeitung beeinflusst sein sollte, so eröffnet die explizite Analyse von Bearbeitungsstrategien gerade in Konzentrationstests neue Forschungsmöglichkeiten.

Schmidt-Atzert, Krumm und Bühner (2008) haben eine für den praktisch arbeitenden Diagnostiker sinnvolle Einordnung von Konzentrationstests innerhalb der Landschaft existierender Leistungstests vorgeschlagen. Demnach messen viele als Aufmerksamkeitstests bezeichnete Verfahren jene Aspekte, die sich als jeweils kritisch für die Wahrnehmung oder für die Ausführung der Reaktion einteilen lassen. Vigilanztests beziehen sich dabei eher auf das Wahrnehmen, während sich als Konzentrationstests bezeichnete Verfahren auf die Weiterverarbeitung der Information beziehen (vgl. Westhoff, 1985; Westhoff & Dewald, 1990; Westhoff & Kluck, 1984). In Tests wie dem Revisionstest (Marschner, 1971), dem KLT (Düker & Lienert, 1965) oder dem SMACT (Steinborn et al., 2010) sind die Aspekte der Reizerkennung schließlich überhaupt nicht kritisch für die Aufgabenleistung, wohl aber ihre effiziente Weiterverarbeitung, die aus mentalem Addieren, Verifizieren bzw. Vergleichen besteht. Die Autoren haben auch Bezug auf unterschiedliche Traditionen der Theoriebildung genommen. Als rationalen Ansatz haben die Autoren die experimentelle bzw. kognitionspsychologische Forschung bezeichnet, als faktorenanalytischen Ansatz bezeichnen

sie die diagnostisch motivierte Forschung zur faktoriellen Validität von Testverfahren. Die Autoren weisen auch darauf hin, dass die einfache Übernahme von theoretischen Konstrukten aus der allgemein-psychologischen Forschung nicht sinnvoll ist, da die Terminologie dort oft sehr kontextsensitiv angewandt wird und höchstens partielle Überlappung mit diagnostischen Konstrukten besteht. Für testdiagnostische Zwecke sind allein die Gütekriterien Reliabilität, Konstruktvalidität (faktorenanalytische Struktur) und Kriterienvalidität relevant, da man wissen muss, wofür ein Testverfahren eingesetzt werden kann und welche anderen verfügbaren Tests eine diagnostische Strategie ergänzen können.

Ein Vorteil von Konzentrationstests gegenüber zahlreichen anderen Testformen ist technischer Natur, dass nämlich in kurzer Zeit sehr viele Reaktionen auf Items erfasst werden können (vgl. Jensen, 2006, Kap. 1; Miller & Ulrich, 2013). Dies führt zu einer hohen Test-Retest-Reliabilität von Konzentrationstestverfahren im Allgemeinen und liefert damit die Grundlage für eine hohe prädiktive Validität bezüglich bestimmter Kriterien (Lienert & Raatz, 1998). Die Reliabilität eines Tests ist bekanntlich ein Maß für die formale Genauigkeit einer Leistungsmessung. Das Gegenteil von Reliabilität ist Unreliabilität, was in der klassischen Testtheorie als Messfehler konzeptualisiert ist. Hochreliable Instrumente sind dementsprechend frei von (oder arm an) Zufallsfehlern, d. h. bei einer wiederholten Testung unter gleichen Rahmenbedingungen würde das genau gleiche Testergebnis wieder erzielt (Lienert & Raatz, 1998). Obwohl das Konstrukt Reliabilität auch die Aspekte Stabilität und Konsistenz umfasst, soll der Begriff hier im engeren Sinne als Messgenauigkeit bei Testwiederholung aufgefasst werden. Zahlreiche Faktoren haben Einfluss auf die Reliabilität, wobei die meisten dieser Faktoren auf theoretischer Ebene thematisiert wurden (Miller & Ulrich, 2013). Die empirische Befundlage zu Einflussfaktoren auf die Reliabilität ist sehr dünn und wenig systematisch. In der Regel wird meist nur mehr oder weniger „gute Reliabilität“ für

ein Testverfahren ausgewiesen, jedoch nicht systematisch untersucht, welche Faktoren die Reliabilität beeinflussen. Miller und Ulrich (2013) haben unlängst kritisiert, dass in den meisten Studien immer nur Reliabilität für einen Test ausgewiesen wird, ohne dies in Beziehung zur Anzahl an verwendeten Items zu setzen. Um die Reliabilität zweier Testverfahren vergleichen zu können, müsste die Anzahl an Items in beiden Tests jedoch auf eine gleiche Anzahl gebracht werden. Formeln zur Hochrechnung der Reliabilität bei Testverlängerung sind für Reaktionszeitparadigmen ungeeignet, da die Veränderungskurve der Reliabilität als Funktion der Itemmenge von spezifischen Faktoren abhängig ist, die in den Formeln nicht berücksichtigt sind (z. B. Übungsanfälligkeit, Itemschwierigkeit etc.).

Miller und Ulrich (2013) haben vorgeschlagen, die Reliabilität (d. h. die Korrelation von Test und Retest) von zu vergleichenden Testverfahren als Funktion der Itemmenge darzustellen, vergleichend für verschiedene Tests (bzw. Aspekte von Tests). Die Fläche unter dieser Funktion ergibt die Gesamtreliabilität einer bestimmten Aufgabenform, unabhängig von der tatsächlich verwendeten Itemmenge eines handelsüblichen Testverfahrens. Demnach ist es ein Kennzeichen besonders geeigneten Aufgabenmaterials, dass diese Funktion besonders im Bereich weniger Items zwischen Tests differenziert und dass eine hohe Anzahl an Items jedes Testverfahren sehr reliabel macht. Da ökonomische Aspekte insbesondere in praktischen Situationen eine große Rolle spielen, ist dieser Aspekt also nicht unerheblich (vgl. Schmidt-Atzert et al., 2008; Westhoff & Hagemeister, 1992). Davon ausgehend, dass Reliabilität die Voraussetzung für eine hohe prädiktive Validität ist, wird klar, dass die Testlänge (d. h. die Itemmenge) eine große Rolle spielt. Es ist anzunehmen, dass längere Konzentrationstests bessere psychometrische Gütekriterien aufweisen als kürzere Tests, da aufgrund der höheren Anzahl an Items bereits eine bessere statistische Schätzung erfolgen kann (vgl. Lienert & Raatz, 1998). Diese Vorhersage wird allein aus den Annahmen der klassi-

schen Testtheorie heraus abgeleitet, empirische Belege dafür gibt es nur sehr wenige. Nennenswert ist der Beitrag von Hagemeister (1994), die den Effekt der Testverlängerung auf die Reliabilität von Geschwindigkeitskennwerten und Fehlerkennwerten in Konzentrationstests untersucht hat. Im Test d2 zeigte sich dabei, dass vierfache Testverlängerung zu einer beachtenswerten Steigerung der Reliabilität von Fehlerkennwerten führte, während sich beim Geschwindigkeitskennwert keine derartigen Effekte zeigten. Dies lag allerdings daran, dass dieser Wert bereits bei der Standardversion des Test d2 sehr reliabel war und aus diesem Grunde nur noch wenig Spielraum für weitere Verbesserungen bestand.

Insgesamt gesehen gibt es derzeit keine allgemein anerkannte Reliabilitätstheorie für die Domäne von kognitiven Leistungstests. Miller und Ulrich (2013) haben einen vielversprechenden Ansatz vorgelegt, der auch Verarbeitungsstufen kognitiver Tests mit berücksichtigt. Das von den Autoren vorgelegte Modell berücksichtigt dabei auch die Stichprobenvarianz und bietet einen Korrekturmechanismus an. Denn es ist bekannt, dass hohe merkmalsrelevante Stichprobenvarianz zu einer Überschätzung von Reliabilitätsschätzungen führt, während niedrige Stichprobenvarianz zu einer Unterschätzung führt. Empirisch gibt es jedoch kaum Studien, welche sich systematisch mit dem Einfluss von experimentellen Variablen in bestimmten Paradigmen beschäftigt haben. Studien zur Validierung von Kurzversionen bestimmter Tests finden sich zwar recht häufig in der Literatur, jedoch wird die Reliabilität nur sehr selten vergleichend zur Langversion berichtet. Insgesamt gesehen wird die aktuelle Praxis der Berichterstattung zur Reliabilität als methodisch sehr unzureichend angesehen (vgl. Green, Chen, Helms, & Henze, 2011). Um sich theoretisch weiterzuentwickeln, benötigt die Testkonstruktionsforschung neue Impulse und einen systematischen Ansatz. Viele Fragen sind bislang ungeklärt. So ist unklar, ob sich durch Testverlängerung das zu messende Konstrukt ändert. Eine systematische Analyse der Reliabilität über den

Testverlauf würde diese Frage beantworten können. Führt Testermüdung zu einer Verbesserung oder Verschlechterung der Reliabilität? Die Daten von Steinborn et al. (2010) geben einen Hinweis darauf, dass es dabei auf den Leistungskennwert ankommt, denn die Reliabilität des Variationskoeffizienten verbesserte sich, während jene des Reaktionszeitmittelwerts konstant blieb. Eine Analyse der Reliabilität verschiedener Grade der Itemschwierigkeit bei gleichem Aufgabenmaterial ist eine weitere zu beantwortende Frage. Sehr wahrscheinlich ist, dass es ein Optimum gibt, das nicht weiter gesteigert werden kann. Schließlich ist es ungeklärt, ob es für die Testkonstruktion besser wäre, homogene Items oder heterogene Items zu verwenden.

3.2 Konstruktions- und Validierungsstrategien

Es gibt zahlreiche Studien, welche die Validität von kognitiven Leistungstests in irgendeiner Form belegen. Dabei sind mehrere Aspekte von Bedeutung, die in unmittelbarem Bezug zum Gütekriterium Validität stehen. So zum Beispiel gibt die Konstruktvalidität an, inwiefern ein Testverfahren ein bestimmtes Konstrukt (z. B. Aufmerksamkeit) überhaupt erfasst. Dies wird in der Regel über die Interkorrelation des Tests mit anderen Tests ermittelt. Tests, welche ein gleiches Merkmal messen, sollten entsprechend hoch miteinander korrelieren, während Tests, die ein unterschiedliches Merkmal messen, nicht bzw. nur gering miteinander korrelieren sollten (Cronbach & Meehl, 1955). Die Konstruktvalidität wird oft mit der sogenannten Multitrait-Multimethod-Analyse (MTMM) ermittelt, die ein heuristisches Verfahren zur Beurteilung der korrelativen Zusammenhänge ähnlicher und unähnlicher diagnostischer Instrumente ist. Dabei werden die Instrumente so angeordnet, dass jeweils die Retest-Reliabilität, die konvergente und die diskriminante Validität abgelesen werden können. Konvergente Validität liegt vor, wenn verschiedene Messmethoden (z. B. unterschiedliche Testverfahren), die das gleiche Konstrukt (z. B. Aufmerksamkeit) repräsentieren, auch hoch

untereinander korrelieren. Diskriminante Validität liegt vor, wenn gleiche Verfahren, die ein unterschiedliches Konstrukt repräsentieren, auch nur moderat miteinander korrelieren. Ein spezieller Fall der Konstruktvalidität ist die faktorielle Validität, welche mit Verfahren der Strukturgleichungsmodellierung erreicht werden kann. Wie bereits dargestellt, haben Schmidt-Atzert et al. (2006) unter Verwendung dieser Methode zeigen können, dass verschiedene im Handel verfügbare Konzentrationstests ein sehr ähnliches Merkmal messen, das mit Konzentration bezeichnet werden kann. Dennoch haben auf korrelativen Methoden basierende quantitative Verfahren auch ihre Grenzen. Schmidt-Atzert, Krumm und Bühner (2008) haben daher darauf hingewiesen, dass nur die simultane Betrachtung von faktorenanalytischen Studien und theoretischen Ansätzen zu einer Weiterentwicklung der Testkonstruktionsforschung führen kann.

Westhoff und Kluck (1984) haben ein Modell vorgeschlagen, welches die Konzentrationsleistung im Test anhand dreier Facetten beschreibt, die bei der Konstruktion von Konzentrationstests maßgeblich sind. Demgemäß bestimmen die Komplexität des Reizmaterials, die Komplexität der Reaktionsweise und die bisher erfolgte Übung mit dem Aufgabenmaterial die Leistung im Test. Aus diesem Modell lassen sich Vorhersagen ableiten, welche eine systematische Testentwicklung ermöglichen. Westhoff und Kollegen (Westhoff, 1985, 1989, 1995; Westhoff & Kluck, 1984) haben dazu eine Reihe von Studien durchgeführt, um die Effekte der Komplexität von Reizmaterial und Reaktionsweisen sowie der Übung zu ermitteln. Aufgrund dieser Studien kamen die Autoren zu dem Schluss, dass Konzentrationstests recht einfach (d. h. wenig komplex) gestaltet sein sollten, da mit zunehmender Komplexität die Gefahr von Übungseffekten besteht. Tatsächlich hat Hagemeister (Hagemeister, 2006, 2007) zeigen können, dass schwierigere Items, verglichen zu leichten Items, eine steilere Lernkurve aufweisen. In jedem Fall lässt es sich oft nicht verhindern, dass sich Testübungseffekte und Testermüdungseffekte irgendwie vermischen (Steinborn et al., 2008, 2010).

Inwiefern die Gefahr der Validitätsbeeinträchtigung durch Testübung tatsächlich ein relevantes Problem ist, bleibt eine offene Frage, da keine Daten vorliegen, welche die Häufigkeit von Testadministration in praktischen Kontexten abschätzen und aufzeigen welche Rolle der zeitliche Abstand zwischen den Testungen dabei spielt. Sofern Konzentrationstests als Indikator in Längsschnittstudien, Trainingsstudien oder im Bereich neuropsychologischer Rehabilitation eingesetzt werden, sind Übungseffekte sehr relevant. Hagemeister (2007) hat diesbezüglich mehrere Methoden auf ihre Effektivität hin untersucht, um potentielle Einflüsse auf die Validität von diagnostischen Strategien zu minimieren. Vielversprechend war dabei die Analyse differentieller Übungseffekte leichter und schwieriger Items, die Verwendung häufiger und seltener Items sowie die Analyse differentieller Testverläufe. Eine wirkliche Kontrolle von Übungseffekten ist derzeit jedoch nicht möglich.

Auf der Basis bestehender faktorenanalytischer Studien und theoretischer Modelle von Aspekten der Aufmerksamkeit haben Schmidt-Atzert, Krumm und Bühner (2008) eine taxonomische Einordnung von Tests vorgenommen. Dieses Modell erweitert den theoretischen Ansatz der Testkonstruktion von Westhoff und Kluck, indem es Aspekte der linearen Strukturmodellierung und der kognitiven Komponenten verbindet bzw. gemeinsam berücksichtigt. Die Autoren verfolgen das übergeordnete Ziel, die im deutschen Sprachraum erhältlichen allgemeinen Leistungstests hinsichtlich ihrer zugrundeliegenden theoretischen Mechanismen zu charakterisieren und auf Basis eines pragmatischen Prinzips zu kategorisieren. Die Autoren betonen ebenfalls, dass eine solche Zielstellung weder allein durch kognitionspsychologische noch durch faktorenanalytische Ansätze erreicht werden kann, sondern eine grundlegende hermeneutische Analyse des Gegenstandes erfordert. Im Ergebnis haben die Autoren ein Modell der Aufmerksamkeitsstruktur (diagnostischer Leistungstests) vorgestellt, das die Komponenten Alertness, Selektion, Daueraufmerksamkeit und fokussierte

(exekutive) Aufmerksamkeit enthält und in das sich alle in deutscher Sprache verfügbaren Testverfahren einordnen lassen. Dieses Modell ist daher vor allem für den praktisch arbeitenden Diagnostiker im Bereich der Schulpsychologie oder der neuropsychologischen Diagnostik und Intervention besonders nützlich, da schnell und unkompliziert abgelesen werden kann, welche Tests für welche Zwecke optimal sind. Schmidt-Atzert, Krumm und Bühner (2008) betonen auch, dass dieses Modell die Grundlage für die Konstruktion einer Testbatterie von Aufmerksamkeitsfacetten dienen kann. Es ist diesbezüglich auch bereits von Westhoff und Kollegen (Westhoff & Graubner, 2003; Westhoff & Hagemeister, 2005) betont worden, dass eine solche Diagnostik mit mehreren Arten von Konzentrationstests eine bessere Vorhersagekraft hinsichtlich der intendierten Kriterien erreichen könnte (vgl. auch Flehmig et al., 2007).

Aufgrund der vorherigen Erörterungen darf man schlussfolgern, dass Konzentrationstests eine wichtige Rolle in der Welt existierender diagnostischer Testanwendungen spielen und außerdem wichtige Instrumente in Grundlagenforschung, Feldforschung und Praxis darstellen (Westhoff & Hagemeister, 2005). Dies setzt natürlich voraus, dass Konzentrationstests auch tatsächlich eine hohe prädiktive Validität bezüglich relevanter Kriterien besitzen oder zumindest einen hohen inkrementellen Beitrag zur Vorhersage von Verhalten in verschiedenen diagnostischen Bereichen liefern. Es stellt sich also die Frage, wie valide allgemeine Leistungstests überhaupt sind, und wie sie sich nützlich in eine diagnostische Gesamtstrategie einpassen. Hier ist vor allem das Gütemerkmal der Kriteriumsvalidität von Bedeutung, das die Beziehung eines Messinstruments zu einem externen Kriterium ausdrückt. In der Regel wird die Kriteriumsvalidität als die Korrelation zwischen dem Leistungswert in einem Test (Prädiktor) und dem Wert auf dem Kriterium (Kriterium) ausgedrückt, weshalb man auch von prädiktiver Validität spricht (vgl. Lienert & Raatz, 1998). Dabei bezieht sich der Terminus prädiktiv auf die statistische Vorhersage in einem Regressionsmodell.

Um die zeitliche Abfolge zu benennen, werden häufig die Bezeichnungen anterograde (vorwärts gerichtete), retrograde (rückwärts gerichtete) und konkurrente (zeitgleiche) Vorhersage verwendet. In dieser Arbeit soll die zeitliche Unterscheidung keine Rolle spielen, da die prädiktive Validität vergleichend für bestimmte Testinstrumente bzw. experimentelle Manipulationen verwendet wird. Es soll daher einem Test eine hohe prädiktive Validität bescheinigt werden, wenn dieser eine hohe statistische Vorhersagekraft besitzt. Empirisch hat sich gezeigt, dass Konzentrationstests (bzw. Aufgabenformen, welche sich als Konzentrationstest klassifizieren lassen) eine hohe prädiktive Validität zur Vorhersage von Schulerfolg besitzen, aber zumindest einen inkrementellen Beitrag innerhalb einer komplexeren diagnostischen Strategie leisten (Schmidt-Atzert & Deter, 1993).

Die Korrelationen zwischen Test als Prädiktor und Schul- bzw. Ausbildungserfolg als Kriterium liegen in der Regel bei moderaten bis mittleren Korrelationen (d. h. zwischen .25 und .50), je nach Test (Westhoff & Hagemeister, 2005). Dabei ist die Höhe dieser Zusammenhänge vergleichbar mit denen von Intelligenztests, die als Schnelligkeitstest konstruiert worden sind. Dies liegt womöglich daran, dass Testverfahren, wie der von Oswald und Roth (1987) eingeführte Zahlenverbindungstest (ZVT), alle relevanten Merkmale mit Konzentrationstests teilen, inklusive der selbstbestimmten Arbeitsgeschwindigkeit. Rindermann und Neubauer (2004) haben die prädiktive Validität des ZVT zur Vorhersage von Schulnoten genauer untersucht. Dazu wurde eine Stichprobe von 271 Probanden mit verschiedenen Leistungstests untersucht. Als Indikator für mentale Geschwindigkeit (mental speed) wurde unter anderem der ZVT (Oswald & Roth, 1987) verwendet, als Indikator für schlussfolgerndes Denken (reasoning) der Raven Test (Raven, 1998) und als Indikator für Kreativität verschiedene entsprechende Tests. Dabei zeigte sich, dass schlussfolgerndes Denken am höchsten mit Schulerfolg korrelierte (r = .50**), dass aber auch für mentale Geschwindigkeit

ein Zusammenhang gefunden werden konnte (r = .35**). Am niedrigsten war der Zusammenhang für Kreativität (r = .15 bis .25**). Aufgrund der Tatsache, dass es auch eine Korrelation zwischen mentaler Geschwindigkeit sowohl mit Kreativität als auch mit Denken gab (r = .31**, r = .33**), zogen die Autoren den Schluss, dass der Effekt von mentaler Geschwindigkeit auf die Effizienz kognitiver Leistungsanforderungen im Alltag eigentlich durch Kreativität und schlussfolgerndes Denken moderiert ist. Demgemäß gibt es keinen direkten, sondern nur einen indirekten Zusammenhang zwischen mentaler Geschwindigkeit und Schulnoten, der durch andere Faktoren moduliert ist.

Aus kognitionspsychologischer Sicht würde sich eine andere Sichtweise ergeben und es wäre vor allem eine andere experimentelle Strategie impliziert. Demgemäß wäre der ZVT ein Konzentrationstest, dessen Komponenten wiederum separiert werden könnten in mentale Geschwindigkeit und Konzentrationsfähigkeit. Gemäß den bereits erörterten Zwei-Zustands-Modellen der Konzentrationsleistung (Smit & Van der Ven, 1995; Van Breukelen et al., 1996) ist die Gesamtleistung eine Mischung aus Reaktionen, die im Zustand hoher Aufmerksamkeit erzeugt wurden und Reaktionen, die im Zustand des Abgelenktseins erzeugt wurden. Durch eine Analyse der Reaktionszeitverteilung können diese Komponenten separiert werden, wobei hier verschiedene Techniken existieren. Eine technisch einfach zu handhabende Möglichkeit besteht in der separaten Verwendung von Kennwerten der zentralen Tendenz (als Indikator für Speed) und Kennwerten der Reaktionszeitvariabilität (als Indikator für Konzentration). Problematisch ist bisher lediglich die Umsetzung der richtigen Operationalisierung für beide Kennwerte, da im Reaktionszeit-Mittelwert „zu viel Konzentration“ steckt und in Kennwerten der Variabilität „zu viel mentale Geschwindigkeit“. Derzeit gibt es eine Debatte darüber, ob die Reaktionszeitvariabilität besser als Standardabweichung oder als Variationskoeffizient berechnet werden sollte (vgl. Flehmig, 2010). Einige Autoren führen sogar an, dass man lediglich die schnellsten 25 Prozent der Reaktionszeitvertei-

lung als Indikator für Geschwindigkeit mitteln sollte und die langsamsten 25 Prozent als Indikator für Konzentration (bzw. Ablenkungstendenz). Andere Autoren haben vorgeschlagen, die gesamte Reaktionszeitverteilung zu analysieren und entsprechende Parameter zu extrahieren (vgl. Miller, 2006). Eine derzeit häufig verwendete Möglichkeit besteht in der Anpassung der Verteilung an eine exponentiell modifizierte Gauß-Funktion, auch als ex-Gauß-Funktion bekannt, welche ungefähr den Eigenschaften der empirischen Reaktionszeitverteilungen entspricht (Schmiedek et al., 2007).

Bisher gibt es wenige Studien, welche die prädiktive Validität von Konzentrationstests zur Vorhersage von Schul- und Ausbildungserfolg im theoretischen Kontext von Zwei-Zustands-Modellen untersucht haben. Eine aktuelle Studie von Steinborn et al. (2008) hat sich dieser Frage jedoch gewidmet. In der Studie wurde eine Stichprobe von 100 Probanden mit einer Version des SMACT getestet. Die Testleistung der Probanden wurde in einem Regressionsmodell als Prädiktor für Schulnoten verwendet, wobei das Reaktionszeitmittel als Indikator für mentale Geschwindigkeit verwendet wurde und der Variationskoeffizient als Indikator für Ablenkungstendenz. Da Kennwerte der zentralen Tendenz in Konzentrationstests sich als gute Prädiktoren für Schulnoten erwiesen haben, war dies auch in dieser Studie zu erwarten. Die kritische Frage war, ob ein Kennwert der Variabilität einen inkrementellen Beitrag zum Gesamtmodell der Vorhersage machen kann und über welchen Kennwert dies am besten erreicht wird. Konkret wurde also gefragt, ob ein Zwei-Prädiktor-Modell (Speed + Konzentration) zu einer besseren Vorhersage führt als ein Ein-Prädiktor-Modell (Speed). Die Ergebnisse der Studie zeigen zunächst, dass der SMACT beachtliche Fähigkeiten zur Vorhersage von Schulnoten hat und dass Kennwerte der zentralen Tendenz die besseren Prädiktoren sind (r = .42**). Der Variationskoeffizient als Kennwert der Reaktionszeitvariabilität zeigte sich dabei zwar als ebenfalls hoch prädiktiv (r = .31**), jedoch war der Wert mit dem Reaktionszeitmittel substantiell korreliert (r = .62**).

Daher haben die Autoren geschlussfolgert, dass der Variationskoeffizient in seiner derzeitigen Operationalisierungen partiell redundant zum Reaktionszeitmittel ist und daher nur eingeschränkt empfohlen werden kann. Eine wichtige Aufgabe wäre es, verschiedene Operationalisierung der Reaktionszeitvariabilität systematisch auf ihre chronometrischen und psychometrischen Eigenschaften hin zu überprüfen. Die weiter oben erörterten Möglichkeiten der Analyse von Reaktionszeitverteilungen wären hier eine Methode, die vielversprechend ist. Leider gibt es bislang noch keine Studien zur Validität entsprechender Kennwerte (vgl. Steinborn et al., 2008, 2010).

Eine andere Möglichkeit besteht darin, dass experimentelle Faktoren (z. B. Testlänge, Itemschwierigkeit, etc.) systematisch manipuliert werden und deren Einfluss auf die Validität entsprechend überprüft wird. Dazu gibt es bislang überhaupt keine empirische Basis. Gemäß Flehmig et al. (2007) liegt dies womöglich daran, dass die diagnostische Forschung darauf angewiesen ist, fertige Paradigmen verwenden zu können, welche in der allgemeinpsychologischen Forschung entwickelt wurden. Dies klingt auch plausibel, jedoch zeigen gerade die Daten von Krumm et al. (2008), dass experimentelle Faktoren (z. B. Einzel- versus Seriendarbietung, self-paced versus computer-paced etc.) einen substantiellen Einfluss auf die Konstruktvalidität von Konzentrationstests haben. Um aber einen geeigneten Startpunkt zu finden, muss man die prädiktive Validität verschiedener kognitiver Paradigmen kennen, die mehr oder weniger als Konzentrationstests bezeichnet werden können. Tatsächlich gibt es einige empirische Belege, in denen Testverfahren verwendet wurden, welche Konzentrationstests ähnlich sind (obgleich nicht unbedingt immer als solche deklariert). Luo, Thompson und Dettermann (2003) haben verschiedene Aufgabenformen auf ihre prädiktive Validität zur Vorhersage von Schulerfolg hin untersucht. In dieser Studie wurden 569 Probanden mit zahlreichen Aufgabenformen getestet, von denen einige, gemäß dem taxonomischen Modell von Schmidt-Atzert, Krumm und Bühner

(2008), die Konstrukte mentale Geschwindigkeit bzw. Konzentration messen. Obwohl Westhoff (1995) auf einen konzeptuellen Unterschied von Konzentration und mental speed hinweist, kommen Schmidt-Atzert et al. aufgrund ihrer empirischen Daten zu dem Schluss, dass sich beide Konstrukte nicht voneinander diskriminieren lassen (siehe auch Krumm et al., 2008). Schulerfolg wurde mit einem dafür konstruierten Wissenstest erfasst. Für relevante Aufgabenformen wie Einfachreaktionsparadigma, Wahlreaktionsparadigma und Kurzzeitgedächtnis fanden sich Korrelationen zu Schulerfolg in einer Höhe zwischen r = .30** und .50**, woraus die Autoren schlussfolgerten, dass mentale Geschwindigkeit die grundlegende Fähigkeit darstellt, welche für den Erfolg in Aspekten schulischer Leistung verantwortlich ist.

Dem aufmerksamen Leser dürfte an dieser Stelle nicht entgangen sein, dass verschiedene Autoren aufgrund ihrer Studienergebnisse zu unterschiedlichen Schlüssen gelangen, was die kognitiven Mechanismen sind, welche der Testvalidität zugrunde liegen. Rindermann und Neubauer (2004) haben argumentiert, dass mentale Geschwindigkeit nur einen indirekten Einfluss ausübt, der durch schlussfolgerndes Denken und Kreativität moderiert ist. Luo, Thompson und Dettermann (2003) haben wiederum argumentiert, dass mentale Geschwindigkeit die grundlegende Fähigkeit ist, welche dem Zusammenhang mit schulischen Erfolgskriterien zugrunde liegt. Aus kognitionspsychologischer Sicht können jedoch solche Aussagen aus den Studien gar nicht gezogen werden, da dies eine experimentelle Manipulation kritischer Variablen erfordern würde. Um die Mechanismen zu isolieren, welche an der Vorhersage von Schulleistungen beteiligt sind, dürfen daher nicht einfach Tests verwendet werden, welche bestimmte Konstrukte repräsentieren (z. B. Test d2 = „Aufmerksamkeit“, ZVT = „mentale Geschwindigkeit“), sondern es müssen die Operationen systematisch untersucht werden, welche diesen Konstrukten zugrunde liegen. Krumm, Lipnevich, Schmidt-Atzert und Bühner (2012) sind diesen Weg gegangen, indem sie die prädiktive

Validität von Aspekten der Koordinationsfähigkeit zur Vorhersage von akademischen Studienleistungen untersucht haben. Es wurde eine Stichprobe von 161 Probanden mit verschiedenen Aufgaben getestet, welche jeweils einer von den drei kognitiven Komponenten (1) „Relationale Integration“, (2) „Kurzzeitgedächtnis“ und (3) „mentale Geschwindigkeit“ zugewiesen werden können. Relationale Integration bezeichnet dabei die Fähigkeit, mentale Repräsentationen funktional zu erzeugen und sie mit anderen in Relation zu setzen. Dieses Konstrukt zeigte sich dabei als besonders hoch prädiktiv in der Vorhersage akademischer Leistungen. Kurzzeitgedächtnis bezeichnet die Fähigkeit, Inhalte kurzfristig behalten zu können. Mentale Geschwindigkeit bezeichnet die Fähigkeit, schnell elementare Informationen verarbeiten zu können. Man kann das Ergebnis von Krumm et al. (2012) auch so interpretieren, dass kognitive Kapazität bei koordinativen Anforderungen ein kritisches Merkmal bei der Vorhersage schulischer bzw. akademischer Leistungen ist.

3.3 Ableitung der Fragestellung

Aus den vorausgegangenen Erörterungen ergeben sich verschiedene Hypothesen, die innerhalb eines chronometrischen Ansatzes überprüft werden sollen. Im Vordergrund steht dabei die Frage der prädiktiven Validität von allgemeinen Leistungstests, konkret die Frage nach der Vorhersage von schulischen und akademischen Leistungen durch ebendiese Tests (vgl. Schmidt-Atzert et al., 2004; Westhoff & Hagemeister, 2005). Zunächst soll erst einmal bzw. noch einmal festgestellt werden, was bereits bekannt ist und was nicht, und wie offene Fragen eventuell beantwortet werden können. Bekannt ist, dass kognitive Leistungstests als Prädiktor geeignet sind, um Schulleistungen als Kriterium vorherzusagen. Dabei besitzen unterschiedliche Aufgabenformen eine mehr oder weniger hohe prädiktive Validität (zwischen r = .20 bis .50), welche durch einige technische Faktoren ebenfalls beeinflusst ist. Bekannt ist ebenfalls, dass

die Konstrukte mentale Geschwindigkeit und Arbeitsgedächtnis die wesentlichen Komponenten sind, welche zu dieser Vorhersage beitragen. Immer wieder haben sich Autoren jedoch gefragt, ob die Konstrukte nicht zu grob aufgefasst werden und ob eine Betrachtung auf der Ebene kognitiver Operationen nicht zu einer besseren Vorhersage führen kann. Unbekannt ist jedoch bislang, inwiefern experimentell manipulierbare Faktoren wie die Itemschwierigkeit, die Itemhomogenität, die Testlänge oder die Pausenstruktur einen kritischen Einfluss ausüben. Westhoff und Graubner (2003) haben den Komplexen Konzentrationstest (KKT) vorgestellt, der drei Materialarten (numerisch, verbal, figural) und insgesamt fünf Grade der Itemschwierigkeit differenziert. Leider haben die Autoren keine systematische Analyse auf der Ebene der Itemschwierigkeit vorgenommen, sondern lediglich den Gesamtleistungswert aus der Summe aller Itemschwierigkeiten berechnet. Diese Lücke soll mit der vorliegenden Arbeit geschlossen werden.

Die Manipulation experimenteller Variablen ist eine Herangehensweise der kognitionspsychologischen (und meist grundlagenorientierten) Forschungstradition. Im Gegensatz zu in der diagnostischen Forschungstradition verbreiteten Nutzung handelsüblicher Testverfahren wird der Einfluss bestimmter kritischer Variablen auf die Gütekriterien eines Tests systematisch untersucht. Experimentelle Variablen sind jedoch keine unbedeutenden Randbedingungen in kognitiven Aufgabenformen, wie gelegentlich von diagnostisch arbeitenden Autoren vermutet (Schmidt-Atzert et al., 2008, S. 64), sondern können ein kritisches Element hinsichtlich der Validität eines Verfahrens sein. Die Studie von Forster und Lavie (2007) belegt diesen Fakt auf eindrucksvolle Weise. In der Studie wurde die Reaktionsleistung in einer Variante des visuellen Suchparadigmas (als Prädiktor) zu selbstberichteter Konzentration im Alltag (als Kriterium) in Zusammenhang gesetzt. Der Selbstbericht wurde dabei mit dem von Broadbent und Kollegen (1982) konstruierten Cognitive Failures Questionnaire (CFQ)

erhoben, einem Instrument, das die Häufigkeit von verschiedenen Arten von Lapsus und Fehlern im Alltag während der letzten sechs Monate erfragt. Forster und Lavie (2007) manipulierten in ihrer Studie die Itemschwierigkeit auf der perzeptuellen Stufe und überprüften dabei den Effekt auf die Prädiktor–Kriterium Korrelation. Was würde man intuitiv erwarten? Man würde natürlich erwarten, dass eine höhere Schwierigkeit zu einer besseren Differenzierung der Probanden führt, was sich in höherer Reliabilität und Validität äußern sollte. Genau das Gegenteil war jedoch der Fall. Eine Erhöhung der perzeptuellen Belastung (perceptual load) nivellierte individuelle Differenzen und verringerte entsprechend die Beziehung der Aufgabenleistung zum Kriterium. Dies zeigt, wie kritisch bestimmte experimentelle Anordnungen sind und dass die Testentwicklung darauf ausgerichtet sein sollte, diese kritischen Variablen zu identifizieren und Testverfahren diesbezüglich zu optimieren. An dieser Stelle soll die vorliegende Arbeit ansetzen und einen Beitrag leisten.

Testlänge, Testverlauf und Itemschwierigkeit

Aus den im Theorieteil erörterten kognitionspsychologischen und psychometrischen Modellen wurden Hypothesen über den Einfluss potentieller experimenteller Variablen auf die prädiktive Validität generiert. Insbesondere ging es dabei um die Frage nach der optimalen Testlänge, der Itemschwierigkeit, und möglicher Veränderungen über den Testverlauf. Solche Hypothesen werden oft diskutiert, aber nur selten überprüft. Tatsächlich kann man in Testhandbüchern oft die Feststellung der Autoren lesen, dass zum Beispiel die prädiktive Validität besser wird, wenn Probanden unter nachteiligen situativen Bedingungen getestet werden, was oft so begründet wird, dass Personen mit hoher Fähigkeitsausprägung dann erst recht viel besser abschneiden als Personen mit entsprechend niedriger Fähigkeitsausprägung. In bestimmten Situationen mag dies womöglich zutreffen, in anderen mag genau das Gegenteil zutreffen, was die empirische Überprüfung solcher Behauptungen im konkreten Fall motiviert. Die

bereits erörterte Studie von Forster und Lavie (2007) ist ein kritisches Beispiel für den Fall einer Nivellierung von individuellen Differenzen durch eine kritische Aufgabenvariable. Die Analyse der Itemschwierigkeit in Konzentrationstests ist dementsprechend eine der kritischen experimentellen Manipulationen, welche in der aktuellen Arbeit auf ihre Effekte auf die Validität eines Tests hin untersucht werden soll. Dazu wird der von Flehmig et al. (2010) und Steinborn et al. (2010) bereits genutzte SMACT verwendet, der hier in jeweils leichter und schwieriger Itemschwierigkeit vorliegt. Dabei ist der Faktor Itemschwierigkeit nochmals in homogener (leichte und schwierige Items sind separiert zwischen Experimentalblöcken) und heterogener (leichte und schwierige Items gemischt innerhalb von Experimentalblöcken) Variante variiert. Es ist nämlich gezeigt worden, dass nicht nur Itemschwierigkeit, sondern auch die Mischung wichtig ist, was als Mixkosteneffekt bezeichnet wird (vgl. Flehmig, 2010; Los, 1996).

Aus den Annahmen der klassischen Testtheorie geht klar hervor, dass sich die Gütekriterien Reliabilität und Validität mit zunehmender Testlänge (d. h. Anzahl der Items im Test) erhöhen sollten. Die Annahme resultiert daher, weil die Schätzung des Mittelwerts der Kennwerte einer Person mit zunehmender Anzahl an Items genauer wird. Da die Reliabilität eine Voraussetzung für die Validität ist, führt Testverlängerung also indirekt auch zu einer Verbesserung der Validität, welche hier von Belang ist. So einleuchtend dies erscheint, ist es doch eine empirische Frage, ob dies auch bei einem bestimmten Test der Fall ist. Hagemeister (1994) hat zeigen können, dass eine Testverlängerung im Test d2 zu einer Verbesserung der Reliabilität des Fehlerkennwerts führt, nicht aber zu einer gleichwertigen Verbesserung des Kennwerts für Geschwindigkeit. Dies resultiere daher, so die Autorin, weil der Kennwert für Geschwindigkeit bereits recht reliabel war und daher bereits im Bereich der Asymptote lag. Es gab also wenig Spielraum für eine Steigerung der Reliabilität in diesem Kennwert. Die Annahmen der klassischen Testtheorie über den Effekt der Testverlängerung auf die Reliabilität

gelten außerdem nur dann, wenn sich das zu messende Merkmal nicht über den Testverlauf verändert, wie das zum Beispiel bei Schlafdeprivationsstudien oft der Fall ist. Dies könnte dann der Fall sein, wenn energetische Zustände durch die Testung selbst betroffen sind und testbedingte Ermüdung oder Motivationszerfall eintritt. In den Testvarianten der frühen Konzentrationsforschung wurde dies tatsächlich auch intendiert, um die Willenseigenschaften einer Person zu messen (vgl. Brandstätter, 1995). Eine Analyse von Effekten der Testlänge auf die Güteeigenschaften des SMACT dient in der vorliegenden Arbeit also der Beantwortung der Frage, ob sich eine Akkumulation von Items verbessernd hinsichtlich der prädiktiven Validität auswirkt oder ob sich das zu messende Merkmal während langer Testungen ändert. Anhand der Ergebnisse soll eine ganz praktische Frage beantwortet werden können. Wie lang sollte ein Testverfahren sein, damit es noch reliabel genug ist, aber auch dem ökonomischen Aspekt Rechnung trägt?

Die Analyse von Testlängeneffekten impliziert auch eine Analyse des Testverlaufs. Tatsächlich waren die Hypothesen früherer Arbeiten oft über den Verlauf gemacht worden. Kraepelin nahm sogar verschiedene Phasen des Verlaufs an, die er als Erwärmungsphase, stabile Leistungsphase und Ermüdungsphase bezeichnete. Allgemeinpsychologische Studien zeigen eindeutig, dass Ermüdung im Verlauf zu längeren und variableren Reaktionen führt, es ist jedoch unklar, ob diese Veränderungen einen Test reliabler oder unreliabler machen. Die wesentlichen Fragen sind: (1) Führt Testermüdung zu einer Verbesserung oder Verschlechterung der psychometrischen Gütekriterien? (2) Wie lassen sich Verlaufsmerkmale diagnostisch nutzbar machen? Eine Analyse des Testverlaufs in einer langen Variante des SMACT diente daher der Beantwortung der Frage, ob sich die Gütekriterien Reliabilität und Validität im Verlauf der Testbearbeitung ändern. Dazu wurde eine Langversion des SMACT verwendet, in der Testverlaufseffekte systematisch untersucht wurden. Dabei wurde außerdem

jeweils eine Kurzversion des SMACT vor und nach der Langversion administriert, wobei zwischen den Testungen jeweils Pausen stattfanden. Um die prädiktive Validität der jeweiligen Konzentrationsleistungen in kurzen und langen Tests abschätzen zu können, wurden schulische Leistungen als Kriterium erhoben. Dass die Schulleistungen dabei retrospektiv erhoben wurden, sollte für unsere Frage wenig problematisch sein, da lediglich die Prädiktor–Kriteriums-Relation von Interesse ist. Streng genommen würde eine retrospektive Erhebung des Kriteriums immer eine Unterschätzung des wahren Zusammenhangs bedeuten, da sich das kognitive Leistungsmerkmal über die Zeit geändert haben könnte. Potentielle Zusammenhänge würden also als eine Mindestschätzung interpretiert werden müssen. Zusätzlich wurde die Konzentration im Alltag als weiteres Kriterium erhoben. Wie in der Studie von Forster und Lavie (2007) kam auch hier der CFQ in der Übersetzung von Klumb (1995) zum Einsatz, welcher als Instrument zur Messung von Aufmerksamkeitslapsus und Fehlertendenzen in alltäglichen Situationen allgemein anerkannt ist.

Testdiversifikation und Testkontext

Die Analyse von Testdiversifikationseffekten war eine weitere Fragestellung der hier vorliegenden Arbeit. In der Testkonstruktionsforschung zum Intelligenzkonstrukt ist es allgemein anerkannt, dass Testbatterien eine genauere Schätzung des Leistungsmerkmals „psychometrische Intelligenz“ ermöglichen. Dabei sind Testbatterien zur Messung von Intelligenz natürlich immer aufgrund einer Theorie der Messung psychometrischer Intelligenz konstruiert. Die wohl bekannteste Theorie ist das Generalfaktormodell (g-model) von Spearman und das Modell mehrerer gemeinsamer Faktoren von Thurstone. Das Konstrukt Konzentrationsfähigkeit ist dabei meistens in einem der Untertest vertreten, denn gemäß dieser globalen Theorien ist Konzentrationsfähigkeit ein Aspekt allgemeiner kognitiver Leistungsfähigkeit. Einige Autoren haben mehr oder weniger explizit die Entwicklung von Konzentrationstestbatterien angeregt bzw.

deren Notwendigkeit für eine umfangreiche Diagnostik unterstrichen, jedoch aus zum Teil unterschiedlichen Gründen. Schmidt-Atzert, Krumm und Bühner (2008) haben argumentiert, dass in einer solchen Testbatterie unterschiedliche Aspekte von Konzentrationsfähigkeit (wie z. B. Wahrnehmen, Merken, Zwischenspeichern, Reagieren) optimal abgebildet werden sollten. Westhoff und Graubner (2003) haben jedoch eher den Aspekt der methodischen Triangulation über die drei Materialarten (numerisch, verbal, figural) und die Diversifikation der Komplexität hervorgehoben. Da bisher keine validierte Testbatterie für Konzentrationstests besteht, wurde die Frage anhand einer Intelligenztestbatterie (dem I-S-T 2000 R) untersucht. Die prädiktive Validität der einzelnen Untertests zur Vorhersage von Schulleistungen und von Konzentration im Alltag wurde systematisch untersucht. Um zu prüfen, ob die Validitätseffekte durch Testdiversifikation oder lediglich durch einfache Testverlängerung entstanden sind, wurde die Validität als Funktion der Anzahl der Untertest betrachtet, analog zu der von Miller und Ulrich (2013) angewandten Methodik.

Die Analyse der prädiktiven Validität unter realen Bedingungen ist ein ganz wichtiger, aber oft vernachlässigter Aspekt der Testforschung. Es stellt sich ja immer die Frage, ob die anhand von studentischen Stichproben gewonnenen Prädiktor–-KriteriumRelationen überhaupt generalisierbar sind. Dies wird grundsätzlich angenommen, jedoch gibt es kaum Studien, welche diese Frage überprüft haben (vgl. Westhoff & Hagemeister, 2005). Die Annahme ist gerechtfertigt, dass Probanden aus studentischen Stichproben, welche für Versuchspersonenstunden zur Testung kommen, weniger unter Druck stehen bzw. weniger gestresst sind als Probanden, welche zum Zwecke der Eignungsbeurteilung in einem Personalverfahren zur Testung kommen. In jedem Fall ist es hier eine Frage der ökologischen Validität, ob ein für ein Personalauswahlverfahren konstruiertes Instrument auch in der als Anwendung vorgesehenen Situation valide ist (vgl. Cronbach, 1975). Daher wurde der Versuch unter-

nommen, die prädiktive Validität von Tests unter realen Bedingungen zu bestimmen. Dazu wurde eine diagnostische Strategie innerhalb einer echten Eignungsfeststellung verwirklicht. Es stellte sich hier die Frage, ob eine echte Prüfungssituation zu einer Differenzierung oder zu einer Nivellierung von individuellen Differenzen führt. Im ersteren Fall sollten sich die Gütekriterien verbessern, während sie sich im zweiten Fall verschlechtern sollten. Auch ein komplexes Ergebnismuster wurde in Erwägung gezogen. So zum Beispiel wäre es denkbar, dass bestimmte kognitive Fähigkeiten (z. B. einfache Wahlreaktionen) unter realen Bedingungen besser differenziert werden können, während andere (z. B. komplexere Denkleistungen) sogar nivelliert werden. Da die Literatur keine Antworten (bezüglich psychologischer Gütekriterien) bereithält, war es eine empirische Frage, die natürlich auch theoretisch motiviert war.

Insgesamt gesehen sollten die Ergebnisse der vorliegenden Arbeit sowohl neue Erkenntnisse als auch einen praktischen Beitrag zur psychologischen Konzentrationsdiagnostik leisten. Daher war es das Ziel, konkrete Hinweise und Verfahrensregeln für den in der beruflichen Praxis tätigen Testpsychologen geben zu können. Dieses Ziel motiviert sich auch schon allein daher, weil die psychologische Diagnostik als jenes Fach ausgewiesen wird, das anwendbare Technologien für praktische Berufsfelder im Bereich der Personalauslese oder im schulpsychologischen Dienst liefert (vgl. Westhoff, 1995). Die vorliegenden Ergebnisse sind jedoch in erste Linie an den wissenschaftlich arbeitenden Testentwickler gerichtet, welcher Technologien für die diagnostische Praxis erstellt, in zweiter Linie natürlich an den in praktischen Berufsfeldern arbeitenden Psychologen. Die Ergebnisse der Studie sollen konkrete Antworten für folgende Fragen geben können: Wie lang sollte ein Konzentrationstest sein, damit seine psychometrischen Gütekriterien hinreichend gut sind? Wie sollte man mit Verlaufsmerkmalen umgehen? Sollte man ermüdungsbedingte Veränderungen vermeiden (z. B. durch Kurzpausenmanagement) oder sollte man sie explizit erzeugen? Wie

schwierig sollten die Items eines Konzentrationstests sein? Wenn ein Test komplexer gestaltet wird, dann steigen auch die Anforderungen an die Koordinationsfähigkeit, die Düker (1949) und Krumm (Krumm, Schmidt-Atzert, Bracht, & Ochs, 2011) als eine spezielle Variante von Arbeitsgedächtnisfähigkeit betrachtet haben. Komplexere Konzentrationstests messen natürlich auch einen Aspekt des Arbeitsgedächtnisses und sollten sogar höher mit psychometrischer Intelligenz korrelieren, im Vergleich zu sehr einfachen Tests. Die Frage ist, ob dies problematisch ist, sofern ein Test reliabel und valide ist. Gemäß Cronbach (1975) ist ein Test so viel wert, wie er vorherzusagen imstande ist.

Konservativität

Für Korrelationshypothesen gilt ebenso wie für Unterschiedshypothesen, dass sehr viele Hypothesen zu einer Akkumulation des Typ-1 Fehler führen kann. Darunter versteht man die Wahrscheinlichkeit, dass die Nullhypothese (H0) zugunsten der Alternativhypothese (H1) irrtümlicherweise verworfen wird. Entsprechend sind in jeder wissenschaftlichen Arbeit geeignete Maßnahmen zu treffen, um das Risiko dieser Fehlerform zu minimieren. Da nicht vorausgesetzt werden kann, dass jeder interessierte Leser mit statistischen Kennwerten vertraut ist, soll die Problemstellung und die getroffenen Maßnahmen zur Prävention kurz erläutert werden. Unter der (idealisierten) Annahme, dass hinsichtlich einer bestimmten Forschungsfrage nur eine einzelne Hypothese geprüft würde, ergibt sich, dass bei alpha = 0,05 die Wahrscheinlichkeit für die fälschliche (falsch-positive) Annahme 5% beträgt. Werden aber nicht nur ein Test sondern zahlreiche Einzeltests durchgeführt, dann kommt es zu einer artifiziellen Erhöhung der Fehlerwahrscheinlichkeit. Wenn man zum Beispiel 20 verschiedene Hypothesen mit alpha = 0,05 prüft, hat man eine gute Chance, dass mindestens eine davon signifikant wird. Dies muss in einer Studie immer berücksichtigt werden. Diese sogenannte Typ-1 Fehlerakkumulierung tritt bekanntlich immer dann auf, wenn eine

Globalhypothese anhand mehrerer Signifikanztests geprüft wird. Bortz (1999, S. 260) schlägt vor, diese mittels der Bonferroni-Methode zu korrigieren.

Eine wichtige Voraussetzung jedoch ist, dass die Hypothesen auch tatsächlich voneinander unabhängig sind. Prinzipielle Orthogonalität ist eine wichtige Voraussetzung. Die Tests müssen unabhängig sein. Denn bei ähnlichen Tests hat man ja auch eine ähnliche Erwartung. Bei gleichen Tests hat man eine gleiche Erwartung. Ist die Unabhängigkeitserwartung verletzt, ist die Bonferroni-Methode gemäß Bortz (1999) kontraindiziert. Das ist in der vorliegenden Arbeit weitgehend der Fall, da hier kanonische Erwartungen vorliegen, die auch als sogenannte Set-Hypothesen bezeichnet werden. Zum einem werden handelsübliche Tests verwendet, die ein ähnliches Konstrukt messen (Aufmerksamkeit, Konzentration, Intelligenz), zum anderen werden konstruierte Tests eingesetzt, welche ebenfalls diese Konstrukte abbilden sollen. Daher ist keine der Hypothesen wirklich unabhängig voneinander weshalb gemäß Bortz (1999) keine Konservativitätskorrektur eingesetzt werden darf. In der folgenden Studie werden daher keine Korrekturen vorgenommen, da diese nicht vorgenommen werden dürfen, weil die zugrunde liegenden Hypothesen nicht voneinander unabhängig sind.

4 Beschreibung der Fragebögen und Tests

Alle Instrumente, die in der hier vorliegenden Dissertationsschrift zur Anwendung kamen, sind im folgenden Abschnitt aufgeführt und beschrieben. Detaillierte Beschreibungen zu bestimmten experimentellen Manipulationen oder anderen Besonderheiten in der Anwendung finden sich im Methodenteil der jeweiligen Studien 1 bis 5.

Komplexer Konzentrationstest (KKT)

Der von Westhoff und Graubner (2003) entwickelte Test soll die Konzentrationsleistung einer Person im Sinne einer kognitiven Teilfähigkeit erfassen. Die dargebotenen Wahlreaktionsaufgaben sind fortlaufend zu bearbeiten, ohne dass es Pausen oder Unterbrechungen gibt. Auch falsche Reaktionen werden nicht rückgemeldet, um die Äquivalenz von Papier-und-Bleistift Version und computerisierter Version zu gewährleisten. Der KKT besteht aus drei Subtests, die jeweils in einem figuralen, verbalen und numerischen Aufgabenformat dargeboten werden. Auf diese Weise soll die von spezifischen Formaten unabhängige Fähigkeit zur Konzentration gemessen werden. Zuerst wird der Pfeiltest (figural) dargeboten, danach wird der Rechentest (numerisch) dargeboten, dem sich der Wortumkehrtest (verbal) anschließt. Die Komplexität der Reizgrundlage steigt in jedem der Untertest progressiv in fünf Stufen an. Jeder Untertest dauert 10 Minuten. Die hier verwendete computerisierte Version lief auf einem

Standardcomputer und die Reaktionen (linke versus rechte Taste) erfolgten auf einer Standardtastatur mit entsprechend markierten Tasten. Im Pfeiltest bestehen die Reize aus unterschiedlichen Pfeilen, welche in eine von vier möglichen Richtungen zeigen. Weitere Arten verschiedener Pfeile kommen mit steigender Komplexität hinzu. Die Probanden sind instruiert, auf den Zielreiz (einfacher Pfeil, zeigt nach rechts oben) durch Drücken der rechten Taste zu reagieren, während alle anderen Reize mit der linken Taste zu beantworten sind. Die Aufgabenkomplexität steigt durch Hinzufügen von Pfeilen mit unterschiedlicher figuraler Komplexität. Im Rechentest bestehen die Reize aus Kettenaufgaben mit Ergebnis, deren Komplexität über die Kettenlänge erhöht wird. Die Probanden sind instruiert, auf richtige Terme mit der rechten und auf falsche Terme mit der linken Taste zu reagieren. Im Wortumkehrtest bestehen die Reize aus deutschen Wörtern in richtiger und umgekehrter Darbietung, auf die jeweils mit rechter oder linker Taste reagiert werden soll. Die Komplexität wird durch die Länge des Wortes erhöht. In der Abbildung 4.1 auf der nächsten Seite sind exemplarisch einige Items aus dem Pfeiltest (A), dem Rechentest (B) und dem Wortumkehrtest (C) dargestellt. Der KKT wurde als computerisierte Version (Westhoff & Scholz, 2007) über das Wiener Testsystem (WTS) dargeboten.

Serial Mental Addition and Comparison Test (SMACT)

Der Test wird in der vorliegenden Arbeit ausschließlich am Computer dargeboten. Das allgemeine Aufgabenprinzip wurde von Restle (1970) erstmalig berichtet, jedoch unter einer kognitionspsychologischen Fragestellung. Verschiedene diagnostische Fragestellungen wurden von Flehmig et al. (Flehmig, 2010; Flehmig et al., 2010) und Steinborn et al. (Steinborn, 2004; Steinborn et al., 2012) untersucht. Die Probanden sind aufgefordert, dargebotene Terme zu addieren und dann mit einem ebenfalls dargebotenen Zahlenwert hinsichtlich der Größe zu vergleichen. Jedes dargebotene Item besteht aus einem Additionsterm und einer Zahl, die räumlich durch einen senkrech-

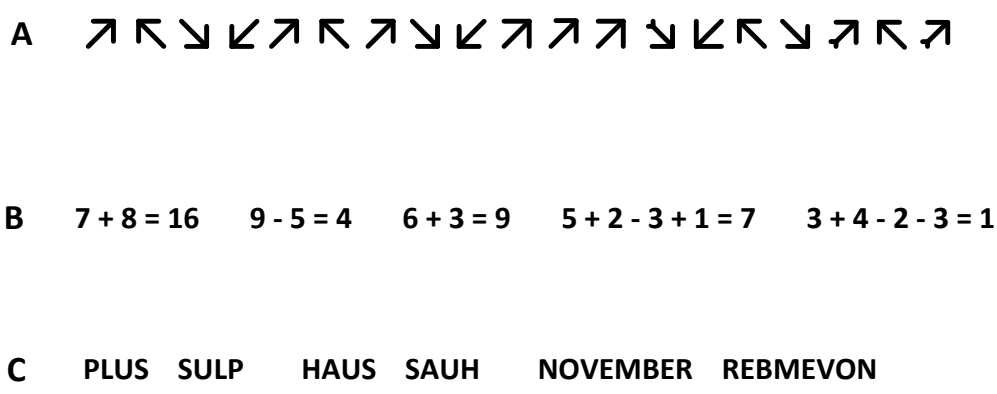

Abbildung 4.1 – Im Pfeiltest bestehen Zielreize aus Pfeilen, die nach oben rechts ausgerichtet sind. Die Itemschwierigkeit wird über die Erkennbarkeit manipuliert. Im Rechentest muss die Richtigkeit einer Kettenaufgabe mit Ergebnis beurteilt werden. Die Itemschwierigkeit wird über die Kettenlänge manipuliert. Im Wortumkehrtest muss festgestellt werden, ob ein umgekehrtes Wort dem normal dargestellten Wort entspricht oder nicht. Die Itemschwierigkeit wird über die Länge des Wortes manipuliert.

ten Strich getrennt sind (z. B. „3+4 | 8“ oder „6 | 2+3“). Die Probanden sind instruiert, den Zahlenterm zu berechnen und dann mit der gegenüber dargebotenen Zahl zu vergleichen, wobei die Reaktion in Richtung der größeren Zahl zu erfolgen hat. In der Abbildung 4.2 auf der nächsten Seite ist das Bearbeitungsprinzip grafisch dargestellt. Als Leistungskennwerte können die mittlere Reaktionszeit und der prozentuale Fehleranteil berechnet werden. Die hier verwendeten computerisierten Versionen in den Studien 1 und 2 enthielten bestimmte experimentelle Manipulationen, welche im Methodenteil des jeweiligen Studienabschnitts detailliert beschrieben sind. In der Studie 1 wurden Itemschwierigkeit und schwierigkeitsbezogene Itemhomogenität manipuliert, während in der Studie 2 die Testlänge (kurze versus lange Version) manipuliert wurde.

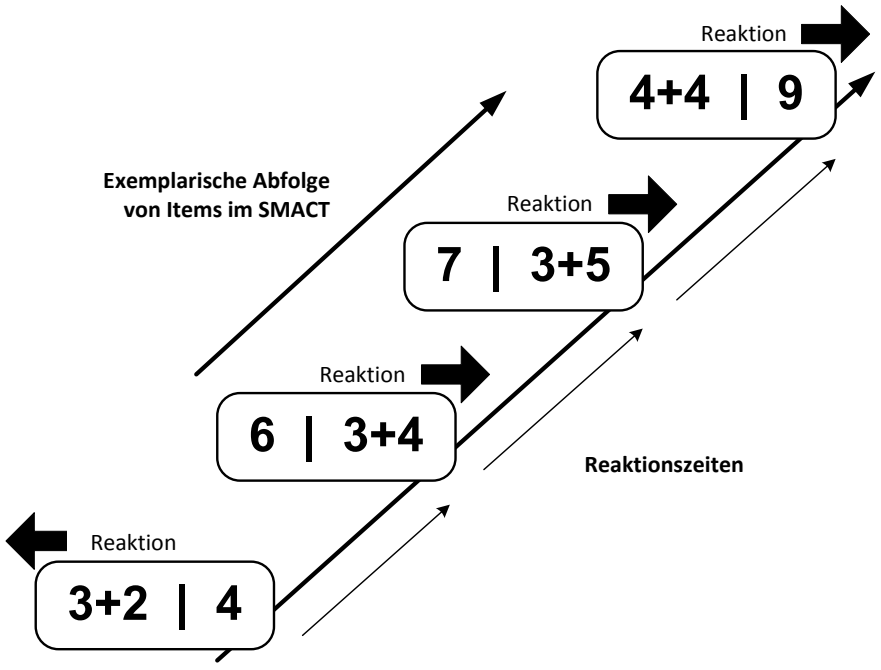

Abbildung 4.2 – Die Aufgabe besteht darin, den Additionsterm zu berechnen und die errechnete Zahlgröße anschließend mit der gegenüber dargebotenen Zahl zu vergleichen. Die Reaktion erfolgt in Richtung des größeren Zahlwerts. Die Itemschwierigkeit kann sowohl über die Größe des Additionsterms als auch über die Distanz zwischen errechnetem und dargebotenem Zahlenwert manipuliert werden.

Test d2 Aufmerksamkeits-Belastungs-Test

Der Test ist ein klassischer Durchstreichtest, der nach dem Bourdon-Prinzip konstruiert wurde (2002). Der Test besteht aus 14 Zeilen, welche aus den Kombinationen von Buchstaben (d und p) und Strichen bestehen. Die Buchstaben sind jeweils unterhalb und/oder oberhalb mit einem, zwei, drei oder vier Strichen versehen. Als Zielreize sind alle d mit genau zwei Strichen definiert, alle anderen Zeichen sind Distraktoren. Die Probanden werden instruiert, alle Zielreize zu markieren und alle anderen Reize auszulassen (also nicht zu markieren). Für jede der 14 Zeilen sind genau 20 Sekunden Bearbeitungszeit vorgesehen, welche der Versuchsleiter mitzuteilen hat. Der Versuchsleiter hat bei jeder Zeile nach genau 20 Sekunden zu sagen: „Stopp – nächste Zeile“.

Die Testdauer beträgt insgesamt 4 Minuten und 40 Sekunden. Mit Instruktion beträgt die gesamte Testzeit ungefähr 8 Minuten.

Test d2-C Aufmerksamkeits-Belastungs-Test

Die computerisierte Version (Brickenkamp, Merten, & Hänsgen, 1996) ist eine PC-Adaptation der klassischen d2-Papier-Bleistift-Version. Auf einem Bildschirm wird jeweils eine Zeile mit einer zufälligen Abfolge von neun Zeichen dargeboten. Die Zeichen bestehen entweder aus einem kleinen d oder einem kleinen p. Die Buchstaben sind zufällig ober- oder unterhalb mit einem, zwei, drei oder vier Punkt(en) versehen. Über bzw. unter einem Buchstaben befinden sich maximal zwei Punkte. Unter jedem Zeichen steht in aufsteigender Reihenfolge eine der Ziffern Eins bis Neun; diese Ziffern sind jedem zu bearbeitendem Zeichen fest zugeordnet. Der Proband soll zeilenweise, von links nach rechts, alle d mit zwei Punkten erkennen und anschließend die unter dem Zielzeichen stehende Ziffer markieren (Tippen der zugeordneten Ziffer auf der Nummernzeile der Tastatur, Antippen der entsprechenden Ziffer bei Touchscreen oder Anklicken der Ziffer per linker Maustaste). Durch wiederholtes Drücken derselben Zifferntaste kann der Proband seine Eingabe korrigieren. Allerdings erhält der Proband keine Rückmeldung, ob seine aktuelle Reaktion richtig oder falsch war. Die Darbietung der Zeilen erfolgt zusammengefasst in insgesamt 12 Serien; eine Serie dauert 30 Sekunden (innerhalb einer Serie sollen mehrere Zeilen bearbeitet werden). Ein Zeile bleibt somit mindestens 30 Sekunden bestehen oder so lange, bis der Proband selbstständig die nächste Zeile aufruft (bei Tastaturgebrauch der Leertaste, bei Touchscreen-Anwendung auf „weiter“ tippen, bei Mausanwendung mit der linken Taste auf „weiter“ klicken). Eine Serie endet mit dem ersten Tastendruck nach Ablauf der Bearbeitungszeit. Einschließlich der Instruktion und einer Übungsphase beträgt die Durchführungszeit des Tests ca. 10 Minuten.

Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II)

Es handelt sich um die computerisierte Version eines kognitiven Leistungstests (Moosbrugger & Goldhammer, 2005), der die allgemeine Konzentrationsfähigkeit von Probanden erfassen soll. Der Test kann sowohl in normaler bzw. in adaptiver Variante administriert werden, als auch in drei unterschiedlichen Schwierigkeitslevels, die drei verschiedene situative Belastungsgrade induzieren (FAKT-E, -S, -SR). In der vorliegenden Arbeit kam ausschließlich die Version „FAKT-E“ mit adaptiver Darbietungszeit zum Einsatz. Die Aufgabenform entspricht einer klassischen serialen Wahlreaktionsaufgabe, bei der die Probanden entscheiden sollen, ob es sich bei dem dargebotenen Zeichen nach einer vorher instruierten Regel um einen Zielreiz oder einen Distraktor handelt. Es gibt zwei Zielreize: (1) alle Kreise, die im Inneren einen weiteren Kreis haben, in dem sich drei Punkte befinden und (2) alle Kreise, in denen sich ein Quadrat befindet, in dem zwei Punkte enthalten sind. Alle anderen Reizkonfigurationen sind Distraktoren (z. B. Kreise, in denen Kreise sind, die mehr oder weniger als drei Punkte enthalten und alle Kreise, in den Quadrate sind, die mehr oder weniger als zwei Punkte enthalten).

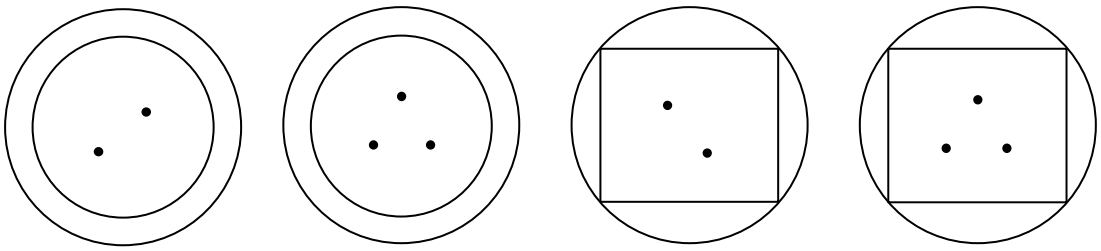

Abbildung 4.3 – Beispielitems im Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II).

Auf Zielreize ist die linke Taste (Taste 1 auf der Ziffernreihe) zu drücken, auf Distraktoren ist die rechte Taste (Taste 0 der Ziffernreihe) zu drücken. Es gibt keine Fehlerrückmeldung und auch keine Möglichkeit, gemachte Fehler zu korrigieren. In

der Abbildung 4.3 auf der vorherigen Seite sind einige Items aus dem FAKT-II nebeneinander dargestellt. Im Test erfolgt die Darstellung der Items als Einzeldarbietung. Dabei wird das jeweilige Item im Zentrum des Bildschirms präsentiert, während gleichzeitig ein Pfeil darauf verweist. Über die gesamte Testzeit ist die Regel (Kreis mit 3 Punkten, Quadrat mit 2 Punkten: 1 drücken) in der linken oberen Bildschirmhälfte eingeblendet. Der FAKT-II erhebt die Konzentrationsfähigkeit mit drei unterschiedlichen Leistungskennwerten. Dabei handelt es sich um die Mengenleistung (KL), die Genauigkeit der Bearbeitung (KG) und die Homogenität (KH). Die reine Testzeit in der hier verwendeten Testform „FAKT-E“ beträgt maximal 6 Minuten.

Raven Advanced Progressive Matrices (APM)

Das Verfahren erfasst die allgemeine Intelligenz im Sinne induktiv-schlussfolgernden Denkens (Raven, 1998). Ein Item setzt sich aus neun Feldern zusammen. Diese sind in einer 3 x 3 Matrix angeordnet. In acht Feldern sind figurale Symbole enthalten. Logisch ergänzt werden soll das neunte Feld. Acht Antwortalternativen werden dem Probanden dafür vorgegeben. Nur eine der dargebotenen Alternativen stellt die richtige Lösung dar. Es ist ein Set mit Übungsaufgaben vorhanden und ein Set mit Testaufgaben. In beiden Sets steigt der Schwierigkeitsgrad der Aufgaben an. Die 36 Aufgaben des Sets mit Testaufgaben sollen innerhalb von 40 Minuten bearbeitet werden. Der Kennwert für die allgemeine Intelligenz ist die Anzahl der richtig gelösten Aufgaben.

***Wiener Matrizen-Test* (WMT)**

Bei dem Verfahren handelt es sich um einen Intelligenztest zur Erfassung der allgemeinen kognitiven Leistungsfähigkeit im Sinne induktiv-schlussfolgernden Denkens (Formann & Piswanger, 1979). Ein Item besteht aus neun Feldern, die in einer 3 x 3 Matrix angeordnet sind. Acht der Felder enthalten figurale Symbole. Das neunte enthält ein Fragezeichen. Neben dieser Matrix werden acht mögliche Ergänzungen dargeboten. Aufgabe des Probanden ist es, die das Feld mit dem Fragezeichen logisch

ergänzende Variante anzukreuzen. Von den angebotenen Alternativen stellt nur eine die richtige Lösung dar. Der Schwierigkeitsgrad der Aufgaben steigt im Laufe der Bearbeitung an. Insgesamt sind 24 solcher Aufgaben im WMT enthalten, für deren Bearbeitung der Proband 25 Minuten Zeit erhält. Aus der Summe richtig gelöster Aufgaben ergibt sich der Kennwert für die allgemeine Intelligenz.

Wiener Matrizen-Test als computerisierte Version (WMT)

Bei der PC-Adaptation der klassischen Papier-Bleistift-Version werden die Items auf dem Monitor einzeln abgebildet, sodass rechts neben der 3 x 3 Matrix die insgesamt acht zur Auswahl stehenden figuralen Symbole präsentiert sind. Die Auswahl der dargebotenen Antwortalternativen erfolgt per Maus, indem die Alternative mit einem Mausklick zunächst markiert und mit einem weiteren Mausklick bestätigt wird. Direkt nach dem zweiten Mausklick wird das nächste Item auf dem Bildschirm präsentiert. Innerhalb der 24 Items kann der Proband selbstständig mit den Tasten „zurück“ und „weiter“ zu noch unbearbeiteten oder bereits bearbeiteten Items wechseln und ggf. auch Korrekturen vornehmen. Einschließlich der Instruktion und einer Übungsphase beträgt die Durchführungszeit des Tests ca. 27 Minuten.

Intelligenz-Struktur-Test (I-S-T 2000 R) als computerisierte Version

Hierbei handelt es sich um eine klassische Testbatterie, welche zum Zwecke einer umfangreichen psychometrischen Messung von Intelligenz konstruiert wurde (Amthauer, Brocke, Liepmann, & Beauducel, 2001). Im Handbuch steht, dass der I-S-T 2000 R ein vielseitig einsetzbarer und ökonomischer Intelligenztest sei, welcher hinsichtlich seiner Vorgängerversion durch aktualisierte Normdaten und Untersuchungen zu Gütekriterien erweitert worden sei. Tatsächlich ist der Test nicht unbedingt ökonomisch, da die Testdauer für das Grundmodul ca. 77 Minuten und zusammen mit dem Erweiterungsmodul Wissenstest ca. 130 Minuten beträgt. Jedoch können auch nur Teilaspekte oder gar einzelne Module zur diagnostischen Verwendung eingesetzt

werden, da die Normdaten sich auch auf Teilaspekte beziehen. Das macht den Test zumindest vielseitig. Das Grundmodul enthält folgende Untertests für den Parameter „verbale Intelligenz“: Satzergänzungen (SE), Analogien (AN), Gemeinsamkeiten (GE), für den Parameter „numerische Intelligenz“: Rechenaufgaben (RE), Zahlenreihen (ZR), Rechenzeichen (RZ) und für den Parameter „figurale Intelligenz“: Figurenauswahl (FA), Würfelaufgaben (WÜ), Matrizen (MA). Für den Parameter „Merkfähigkeit“ enthält das Grundmodul zwei spezielle Tests zur Merkfähigkeit: Im Test „verbale Merkfähigkeit“ müssen vorgegebene Wörter und deren Zuordnung zu Oberbegriffen eingeprägt werden. Im Test „figurale“ Merkfähigkeit werden Figurenpaare vorgegeben, die eingeprägt werden müssen. Die Abfrage erfolgt, indem nur ein Element des Paares vorgegeben wird, während das andere aus einem Angebot von fünf Alternativen heraus ergänzt werden muss. Zusätzlich gibt auch ein Erweiterungsmodul, den Wissenstest, in dem Fragen zu verschiedenen Wissensgebieten gestellt werden, deren Antwort aus einem Angebot von jeweils fünf Alternativen ausgewählt werden muss. Zur Anwendung für die Untersuchung kam die Testform „Standard A“ ohne Verwendung des Erweiterungsmoduls Wissenstest. Die in dieser Kombination angegebene Bearbeitungszeit beträgt 116 Minuten.

Wilde-Intelligenztest 2 (WIT-2) als computerisierte Version

Es handelt sich um eine Testbatterie zur Messung der allgemeinen Leistungsfähigkeit, welche das Ziel hat, klar unterscheidbare und theoretisch fundierte kognitive Teilfähigkeiten zu messen (Kersting, K., & Jäger, 2008). Die Testautoren geben als Besonderheit an, dass der WIT-2 sich sowohl für die Intelligenzdiagnostik im Allgemeinen als auch für die praktische Eignungsdiagnostik im Speziellen eignet. Tatsächlich ist der WIT-2 speziell auf angewandte diagnostische Fragestellungen ausgerichtet, wie sie sich in der beruflichen Personalauswahl und Personalentwicklung, der Berufs- und Studienberatung und im schulpsychologischen Dienst finden. Eine ganz besondere

Eigenschaft des WIT-2 ist es den Testautoren zufolge, dass ein ganz erheblicher Anteil an Validierungsdaten im Kontext von realen Bewerbersituationen erhoben wurde und nicht allein anhand studentischer Stichprobenpopulationen. Der WIT-2 enthält dabei die folgenden Dimensionen: Schlussfolgerndes Denken, Sprachliches Denken, Rechnerisches Denken, Räumliches Denken, Merkfähigkeit, Arbeitseffizienz, Wissen Informationstechnologie, Wissen Wirtschaft. Diese Dimensionen ergeben sich aus Kombinationsmöglichkeiten der Untertests. Der WIT-2 bietet folgende Untertests an: Analogien (AL), Zahlenreihen (ZN), Abwicklungen (AW), Gleiche Wortbedeutung (GW), Grundrechnen (GR), Eingekleidete Rechenaufgaben (ER), Merkfähigkeit (MF), Spiegelbilder (SP), E-Mails Bearbeiten (EM), Wissenstest Wirtschaft (WW) und Wissenstest Informationstechnologie (WI). Eine ganz wichtige Eigenschaft des WIT-2 ist außerdem noch erwähnenswert, nämlich dass eine gezielte Übungsphase für alle Testteilnehmer vorgesehen ist, um Chancengleichheit zu gewährleisten. Die Empfehlung der Testautoren ist, dem Testteilnehmer die Informationsbroschüre zum WIT-2 vor der eigentlichen Testung zukommen zu lassen, um dem Testteilnehmer Gelegenheit zur Übung zu geben. Damit soll der unlautere Vorteil von Lesern von Testknacker-Büchern nivelliert werden, indem Chancengleichheit für alle Testteilnehmer gewährleistet wird. Zur Anwendung für die Untersuchung kam die Testform „A“ ohne die Dimensionen Arbeitseffizienz, Wissen Wirtschaft und Wissen Informationstechnologie zu erheben. Damit wurden keine Aufgaben aus den Untertests E-Mails Bearbeiten (EM), Wissenstest Wirtschaft (WW) und Wissenstest Informationstechnologie (WI) zu Bearbeitung präsentiert Für diese Auswahl ist eine Bearbeitungszeit von 80 Minuten angegeben.

Cognitive Failures Questionnaire (CFQ)

Entwickelt wurde dieser Fragebogen von Broadbent et al. (1982). Hier wurde er in der deutschen Fassung nach Klumb (1995) verwandt. Der CFQ besteht aus 25 Items und erfragt kleinere Handlungs- und Aufmerksamkeitsfehler in alltäglichen Situationen

bezogen auf das letzte Jahr. Gemäß Klumb (1995) weist der CFQ eine brauchbare interne Konsistenz auf (Cronbachs Alpha = .78). Konstruiert wurde der CFQ ursprünglich als Fragebogen, der die Selbsteinschätzung von Aufmerksamkeit und Gedächtnisleistung im Alltag mit einem einzelnen Gesamtwert misst. Es gibt zahlreiche Studien, die versucht haben, eine faktorenanalytische Struktur zu finden, jedoch sind die Befunde entweder theoretisch nicht interpretierbar oder nicht stichprobenrobust replizierbar. Fest steht, dass sich keine voneinander unabhängigen (orthogonalen) Faktoren finden lassen, da alle Studien immer nur korrelierte Faktoren finden konnten. Eine Gegenüberstellung der deutschen und englischen Items findet sich bei Flehmig, Steinborn et al. (2007b). Der für diese Arbeit verwendete CFQ kann im Anhang auf Seite XIV eingesehen werden.

Short Inventory of Minor Lapses (SIML)

Der Fragebogen stellt eine brauchbare wenn auch weniger bekannte Alternative zum CFQ dar und wurde von Reason und Kollegen (Reason, 1993; Reason & Lucas, 1984) entwickelt. Der SIML ist ein 25-Item Fragebogen, der mit dem Ziel konstruiert wurde, Fehler und Aufmerksamkeitslapsus im alltäglichen Leben zu messen. Während der CFQ ein unidimensionales Konstrukt zu messen intendiert, so ist der SIML auf Basis theoretischer Überlegungen konstruiert worden: Gemäß dem Modell der Handlungsregulation (Hacker & Sachse, 1998; Reason, 1993) können Fehlleistungen auf drei hierarchischen Ebenen verortet werden, einer wissensbasierten Ebene, einer regelbasierten Ebene und einer fähigkeitsbasierten Ebene. Die hier verwendete deutsche Version des Fragebogens stammt von Köhne (2001), der die ursprüngliche Version um weitere Items erweiterte. Der so erweiterte SIML hat die Messintention, die individuelle Fehlerneigung auf der regelbasierten und auf der fähigkeitsbasierten Ebene zu messen. Im Anhang dieser Arbeit ist der eingesetzte SIML auf Seite XVI dargestellt.

Leistungsmotivationsinventar (LMI)

Das LMI erfasst Motivationstendenzen in 17 berufserfolgsrelevanten Dimensionen mit jeweils 10 Items, die weitgehend berufsbezogen formuliert sind. Dazu zählen Beharrlichkeit, Dominanz, Engagement, Erfolgszuversicht, Flexibilität, Flow, Furchtlosigkeit, Internalität, Kompensatorische Anstrengung, Leistungsstolz, Lernbereitschaft, Schwierigkeitspräferenz, Selbstständigkeit, Selbstkontrolle, Statusorientierung, Wettbewerbsorientierung und Zielsetzung. Die Auswertung ist dimensionsspezifisch und als Gesamtwert möglich. In der vorliegenden Untersuchung kam die Langversion zum Einsatz. Die von Schuler und Prochaska (2001) publizierten Beschreibungen jedes der 17 theoretischen Konstrukte stehen weitgehend im Einklang mit ihrer definitorischen Bedeutung, wie sie im Alltag auch von nichtpsychologischen Laien verwendet werden, daher werden diese hier nicht näher erläutert. Im Grunde handelt es sich bei dem Fragebogen nicht um ein auf theoretischen Überlegungen basierendes Instrument, sondern um eine Sammlung von aus motivationspsychologischer Sicht relevanten Eigenschaften. Die Skalenbezeichnungen sind aus einer taxonomischen Sicht absolut augenscheinvalide, aus messtheoretischer Sicht jedoch nicht zwangsläufig voneinander unabhängig. Die Bearbeitungszeit der 170 Items ist mit ca. 35 Minuten, einschließlich der Instruktion, ausgewiesen.

5 Empirischer Teil

5.1 Studie 1 - Itemschwierigkeit

5.1.1 Methode

Es bestand die Fragestellung, ob die Itemschwierigkeit in Konzentrationstests eine kritische experimentelle Variable darstellt, welche bei gleicher Aufgabenanforderung zu schwierigkeitsbasierten Änderungen der Validität führt. Die kritische experimentelle Manipulation bestand also in einer vergleichenden Analyse der Korrelationen von leichten versus schwierigen Testversionen zu Schulnoten (Realschule, Abitur, Vordiplom). Die prädiktive Validität wurde dabei retrograd erhoben. Es wurde auch verglichen, ob es einen Unterschied macht, wenn leichte und schwierige Items in homogenen versus heterogenen Testvarianten administriert werden. Daher kann auch ein direkter Vergleich von hinsichtlich der Itemschwierigkeit homogener versus heterogener Variante bezüglich der Korrelation zu Schulnoten vorgenommen werden.

Stichprobe. Die Untersuchung fand an der Professur Diagnostik und Intervention an der Technischen Universität Dresden statt und enthielt eine Reihe von Tests und Fragebögen. Separate Teile des Datensatzes wurden bereits von Flehmig (2010) im Rahmen einer persönlichkeitspsychologischen Fragestellung analysiert, jedoch betreffen die hier relevanten Analysen eine ganz andere Fragestellung. Insgesamt 106 Probanden (Alter: M = 24.0 Jahre; SD = 4.4 Jahre, 59 % weiblich) nahmen an zwei

experimentellen Sitzungen teil, die im Abstand von einer Woche erfolgten. Die meisten Probanden waren Rechtshänder (> 90 %) und hatten keine Sehschwierigkeiten bzw. hatten eine Brille.

Fragebogen. Die verwendeten Fragebögen sind im Kapitel 4 ausführlich beschrieben. Es kam der CFQ von Broadbent et al. (1982) zum Einsatz, konkret die deutsche Version in der Übersetzung von Klumb (1995). Obwohl in einigen Studien oft Zusammenhänge zwischen bestimmten Prädiktoren und verschiedenen Subskalen des CFQ berichtet werden (Flehmig, Steinborn et al., 2007a, 2007b; Goldhammer, Moosbrugger, & Krawietz, 2009), muss doch gesagt werden, dass es bislang keine eindeutige Interpretation des CFQ gibt. Es ist daher die Entscheidung gefallen, dass in den vorliegenden Studien lediglich der von Broadbent et al. (1982) intendierte Gesamtscore des CFQ verwendet wird und von einer Evaluation möglicher Unterskalen abgesehen wird. Außerdem wurde der SIML von Reason eingesetzt, welcher ein ähnliches Konstrukt misst (Reason, 1993). Die zwei Subskalen des SIML, regelbasierte und fähigkeitsbasierte Alltagsfehler, wurden separat als Skalenwerte verwendet, da es sich um theoretisch motivierte Skalen handelt, die nicht zwangsläufig statistisch unabhängig sein müssen.

Testverfahren. Zur Erfassung der Konzentrationsleistung wurde die Papier-und-Bleistift Version des Test d2 Aufmerksamkeits-Belastungs-Test (Brickenkamp, 2002) verwendet. Zur Erfassung der allgemeinen kognitiven Leistungsfähigkeit (Intelligenz) wurde der Advanced Progressive Matrices (APM) von Raven (1998) eingesetzt, dessen Validität als Goldstandard dienen sollte, an denen die Validität der anderen Tests gemessen werden sollte. Eine Variante des SMACT wurde zur Messung von Konzentration eingesetzt, wobei es ein Set mit 128 leichten und 128 schwierigen Items gab, die jeweils in geblockter und gemixter Bedingung dargeboten wurden. In geblockter Bedingung wurden leichte und schwierige Items separat dargeboten, in gemixter Bedingung wurden leichte und schwierige Items zufällig durchmischt dargeboten. Es

wurden insgesamt 1600 Items dargeboten. Die Testdauer betrug ca. 50 Minuten.

Ablauf der Untersuchung. Bis auf den APM wurden die kognitiven Leistungstests zweimal (Retest-Intervall = 1 Woche) bearbeitet. Fragebögen wurden in der ersten Sitzung administriert. Die Untersuchung fand in einem abgeschirmten Laborraum der Technischen Universität Dresden statt und der SMACT wurde auf einem Standardcomputer (Bildschirm 19 Zoll, 75 Hz Bildschirmfrequenz) dargeboten. Die Personen saßen ca. 60 cm vom Bildschirm entfernt und sollten bequem sitzen. Sowohl der Test d2 als auch der APM wurden in einer Papier-Bleistift-Version bearbeitet. Zwischen allen Tests gab es kurze Pausen mit einer Dauer von ca. einer bis zwei Minuten.

5.1.2 Ergebnisse

Deskriptive Analysen. Zunächst wurde die Stichprobe hinsichtlich ihrer deskriptiven Eigenschaften hin überprüft. Es fanden sich keine Besonderheiten, die nicht auch durch eine natürliche Variation des Leistungsverhaltens erklärt werden könnten. Auf eine ausführliche Erläuterung soll daher an dieser Stelle verzichtet werden, jedoch können die stichprobenbezogenen Angaben (Mittelwert, Variabilität, Verzerrungen) aller verwendeten Instrumente in der Tabelle A.9 im Anhang eingesehen werden.

Reliabilität. Vom APM gab es keine wiederholte Testung, daher kann die Retest-Reliabilität nicht beurteilt werden. Der Test d2 zeigte erwartungsgemäß eine hohe Retest-Reliabilität für die Leistungsmenge (r = .93**), weniger jedoch für die Leistungsgüte. Der SMACT wies hinsichtlich dieses basalen Gütekriteriums hervorragende Werte auf, sowohl für die leichte (r > .80**) als auch schwierige (r > .94**) Version. Dies gilt jedoch nicht für den prozentualen Fehleranteil bei leichter (r > .39**) und schwieriger (r > .57**) Version.

Validität. In Tabelle 5.1 auf der nächsten Seite sind die Korrelationen zwischen den klassischen Leistungstests und Schulnoten aufgeführt.

Tabelle 5.1 – Ergebnisse der Studie 1: Korrelationen zwischen standardisierten Leistungstests und Schulnoten

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Psychometrische Standardverfahren						
APM	-.06	**-.29***	-.06	-.16	**-.30***	**-.25***
d2 - GZ	-.20	-.19	-.14	-.17	**-.25***	**-.29***
d2 - GZ (Retest)	-.19	-.13	-.05	-.20	-.24	**-.25***
d2 - F%	**.28***	**.28***	.11	**.25***	.23	**.30***
d2 - F% (Retest)	.22	**.37****	.20	.22	.25	**.25***

Anmerkungen. Korrelation zwischen APM und Realschulnoten-Gesamt n = 72; Realschule-Mathe n = 74; Realschule-Physik n = 74; Abitur-Gesamt n = 65; Abitur-Mathe n = 68; Abitur-Physik n = 38; Korrelationen zwischen d2-GZ/d2-GZ (Retest) und Realschulnoten-Gesamt n = 72; Realschule-Mathe n = 74; Realschule-Physik n = 74; Abitur-Gesamt n = 65; Abitur-Mathe n = 68; Abitur-Physik n = 38; Korrelation zwischen d2-F%/d2-F% (Retest) und Realschulnoten-Gesamt n = 72; entsprechend Realschule-Mathe n = 74; Realschule-Physik n = 74; Abitur-Gesamt n = 65; Abitur-Mathe n = 68; Abitur-Physik n = 38; nur signifikante Korrelationen über der Relevanz-Schwelle (r = .25) sind fett markiert; ** für p < .01 (einseitige Testung), * für p < .05 (einseitige Testung).

Wenn im Folgenden von prädiktiver Validität gesprochen wird, so bezieht sich dies auf die statistische Vorhersage, nicht auf die zeitliche Reihung von Prädiktor und Kriterium. Für zeitbezogene Aspekte der Validität werden die Termini retrograde (rückwärtsbezogene) und anterograde (vorwärtsbezogene) Validierung verwendet. Ebenfalls gilt für alle nachfolgend berichteten Schulnoten, dass diese für die Berechnungen in dem üblichen Schulnotensystem und nicht in Punktangaben vorlagen. Entsprechend der Erwartungshaltung war die allgemeine Intelligenz im Sinne induktiv-schlussfolgernden Denkens ein guter Prädiktor für Schulnoten, insbesondere der Mathematiknote. Die Validität zeigte sich im Test für die Mathematiknote in der Realschule (r = .29*) und im Abitur (r = .30*). Die Gesamtmenge im Test d2 zeigte sich weniger valide, wobei die Korrelationen zu Abiturnoten, für Test und Retest ungefähr im Erwartungsbereich lagen, insbesondere für Mathematik (r = .25*, r = .24) und Physik (r = .29*, r = .25*). Die Reaktionszeit im SMACT wies über alle Bedingungen hinweg Korrelationen in vergleichbarer Höhe zum APM auf, was den SMACT als

hervorragendes Instrument auswies ($r > .30$**). Wie aus Tabelle 5.2 hervorgeht, waren sowohl leichte als auch schwierige Items ungefähr gleich gut in ihrer Vorhersage, die höchsten Vorhersagen fanden sich jedoch für die experimentellen Bedingungen, in denen leichte und schwierige Aufgaben in gemixter Form dargeboten wurden, was konzeptuell einer heterogenen Itemwahl entspricht. Angemerkt werden muss noch,

Tabelle 5.2 – Ergebnisse der Studie 1: Korrelationen zwischen Reaktionszeit im SMACT und Schulnoten als Funktion von niedriger versus hoher Itemschwierigkeit und Itemhomogenität

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Mittlere Reaktionszeit						
SMACT - Gesamt	**.33****	**.32****	**.33****	.27	**.36****	.33
SMACT - Gesamt (Retest)	**.33****	**.32****	**.32****	**.33****	**.34****	.33
SMACT - Block	**.31****	**.31****	**.31****	.21	**.32****	.30
SMACT - Block (Retest)	**.31****	**.31****	**.30****	.30	.31	.30
SMACT - MIX	**.35****	**.32****	**.34****	.32	**.39****	**.35***
SMACT - MIX (Retest)	**.31****	**.31****	**.30****	.30	.31	.30
SMACT - Block leicht	**.37****	**.32****	**.33****	.22	**.36****	.24
SMACT - Block leicht (Retest)	**.28***	.24	.24	.26	.26	.25
SMACT - Block schwer	.23	.25	.25	.17	.24	.26
SMACT - Block schwer (Retest)	.26	**.27***	**.27***	.25	**.27***	.26
SMACT - Mix leicht	**.33****	**.34****	**.31****	**.33****	**.45****	**.39***
SMACT - Mix leicht (Retest)	.23	.25	.23	.31*	**.32****	.31
SMACT - Mix schwer	**.31****	**.27***	**.31***	.27	**.30***	.26
SMACT - Mix schwer (Retest)	**.34****	**.29***	**.31****	**.32***	**.31***	.29

Anmerkungen. N = 106, Die Stichprobengröße ist leicht unterschiedlich für die 6 Kriterien, die Signifikanzniveaus sind entsprechend unterschiedlich; Nur signifikante Korrelationen über der Relevanz-Schwelle ($r = .25$) sind fett markiert; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

dass die SMACT-Fehlerprozente überwiegend nicht prädiktiv waren (siehe Anhang Tabelle A.4). Die Fragebögen CFQ und SIML leisteten keine erwähnenswerten Vorher-

sagen durch selbstberichtete Konzentration in Alltagssituationen. Die Korrelationen dazu sind im Anhang in Tabelle A.6 einzusehen.

Regressionsmodelle. Um abschätzen zu können, inwiefern die Leistung in Konzentrationstests einen inkrementellen Beitrag zu Intelligenz als Goldstandard leisten kann, wurden zwei zweistufige (blockweise) multiple Regressionsanalysen durchgeführt. In der ersten Regressionsanalyse wurde die APM-Leistung als erster Prädiktor definiert und die Konzentrationsleistung (Test d2-GZ und SMACT-RT-Gesamtleistung vom ersten Tag) in einem zweiten blockweisen Schritt hinzugefügt. Als Kriterium wurde die Mathematiknote aus der Realschule verwendet. Die Ergebnisse zeigen, dass die multiple Vorhersage (Intelligenz und Konzentration) keinen inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .29$ auf $R = .36$). Die Änderung in F ist nicht signifikant. In einer zweiten Regressionsanalyse wurde die APM-Leistung als erster Prädiktor definiert und die Konzentrationsleistung (Test d2-GZ und SMACT-RT-Gesamtleistung vom ersten Tag) in einem zweiten blockweisen Schritt hinzugefügt. Als Kriterium wurde diesmal die Mathematiknote aus dem Abitur verwendet. Diesmal zeigen die Ergebnisse, dass die multiple Vorhersage (Intelligenz und Konzentration) einen durchaus inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .30$ auf $R = .40$), obwohl die Änderung im Model nur marginal signifikant war. Die zweite Analyse belegt, dass es eine gute diagnostische Strategie sein kann, Personen anhand eines kurzen Matrizentests und eines Konzentrationstests zu prüfen, um Selektion und Klassifikation innerhalb schulischer oder akademischer Organisationen zu betreiben. Weitere Angaben zu den Regressionsanalysen finden sich im Anhang in Tabelle A.18.

5.1.3 Diskussion

Die Ergebnisse der Studie 1 replizieren zunächst die bereits bekannten Ergebnisse, dass der APM von Raven und der Test d2 von Brickenkamp sehr gut als Testinstrumente geeignet sind, um Schulnoten vorherzusagen. Die hier vorliegende Studie zeichnet sich dadurch aus, dass die Retest-Validität nachgewiesen werden konnte, was in nur wenigen Studien überhaupt berichtet wird, aber von großer Bedeutung ist. Denn gerade bei Korrelationen in moderater Höhe stellt sich oft das Problem der Replizierbarkeit ein, dass hier durch einen wiederholten Nachweis der Validität an zwei Messzeitpunkten nachgewiesen werden konnte. Hinsichtlich der experimentellen Fragestellung muss geschlussfolgert werden, dass anhand der hier vorliegenden Daten keine Entscheidung darüber getroffen werden kann, ob die Validität eines Rechenkonzentrationstests besser für leichte oder schwierige Items ist, da sich vergleichbare Korrelationen für alle experimentellen Bedingungen des SMACT fanden. Trotzdem kann dem SMACT als Testmedium eine beachtliche Eigenschaft zur Vorhersage bescheinigt werden.

5.2 Studie 2 - Testverlängerung und Testverlauf

5.2.1 Methode

Die Fragestellung in Studie 2 bestand darin festzustellen, ob die Testlänge in Konzentrationstests ebenfalls eine kritische Variable darstellt, deren potentieller Einfluss auf die prädiktive Validität herausgefunden werden sollte. Dazu wurden eine kurze und eine lange Version des SMACT programmiert, welche dann vergleichend auf ihre prädiktive Validität zur Vorhersage von Schulnoten eingesetzt wurden. Die kritische experimentelle Manipulation bestand demzufolge in einer vergleichenden Analyse der Korrelationen von kurzer versus langer Testvariante zu Schulnoten (Realschule, Abitur). Das Kriterium (Schulleistungen) wurde retrograd erhoben.

Stichprobe. Auch für Studie 2 wurden Daten, die an der Professur Diagnostik und Intervention an der Technischen Universität Dresden erhoben worden sind unter den für diese Arbeit relevanten Fragestellungen untersucht. Flehmig (2010) hatte im Rahmen einer persönlichkeitspsychologischen Fragestellung bereits Teile des Datensatzes analysiert. An der Untersuchung nahmen ursprünglich 103 Personen (Alter M = 26.6 Jahre; SD = 7.3 Jahre; 50 % Frauen) teil. Drei Personen erschienen jedoch nicht zur zweiten Sitzung (Retest) und mussten als Drop-out verzeichnet werden. Die zweite Sitzung fand im Abstand von einer Woche statt. Die meisten Teilnehmer gaben an, Rechtshänder (> 89 %) zu sein und keine Sehschwierigkeiten zu haben bzw. eine adäquate Sehhilfe zu nutzen. Alle Teilnehmer äußerten, sich gesund zu fühlen. Mittels Aushängen an der Technischen Universität Dresden wurde die Stichprobe geworben. Es wurde auf ein ausgewogenes Geschlechterverhältnis geachtet. Am Versuch teilnehmende Personen erhielten ein Feedback zu ihren Ergebnissen in den Fragebögen und den Tests sowie auf Wunsch wurden Versuchspersonenstunden bescheinigt.

Fragebogen. Es kam der CFQ (Broadbent et al., 1982) zum Einsatz, konkret die deutsche Version (Flehmig et al., 2007b; Klumb, 1995). Außerdem wurde der SIML von Reason herangezogen, um wieder die zwei Subskalen des SIML, regelbasierte und fähigkeitsbasierte Alltagsfehler, als Skalenwerte verwenden zu können.

Testverfahren. Zur Erfassung der Konzentrationsleistung wurde wieder die gleiche Papier-und-Bleistift-Version des Test d2 Aufmerksamkeits-Belastungs-Test (Brickenkamp, 2002) verwendet, wie sie bereits in Studie 1 verwendet wurde. Zur Erfassung der Intelligenz wurde diesmal der Wiener Matrizen-Test (Formann & Piswanger, 1979) als Papier-und Bleistift-Variante eingesetzt, welcher das gleiche Konstrukt wie der Test von Raven misst. Die nachgewiesenen psychometrischen Gütekriterien sind dem Raven-Test (Raven, 1998) gleichwertig, wobei der WMT allerdings eine höhere Testökonomie besitzt. Um die Effekte der Testlänge auf die Validität untersuchen zu können,

wurde eine kurze und eine lange Variante des SMACT eingesetzt. Um Testreihenfolgeeffekte kontrollieren zu können, wurde die Kurzversion vor und nach der Langversion administriert. Es wurden 48 verschiedene Items verwendet, welche in einer Sitzung 34 Mal dargeboten wurden, sodass insgesamt 1632 Items in zufälliger Reihenfolge pro Testsitzung präsentiert wurden. Um zusätzlich auch Effekte des Verlaufs auf die Validität messen zu können, wurden Leistungskennwerte (mittlere Reaktionszeit und prozentualer Fehleranteil) für jeden von vier gleichlangen Testabschnitten der Langversion ermittelt.

Ablauf der Untersuchung. Alle Tests wurden zweimal im Abstand von einer Woche bearbeitet, wobei Fragebögen nur einmal am Ende der ersten Sitzung ausgefüllt wurden. Die Untersuchung fand in einem von Außengeräuschen abgeschirmten Laborraum statt. Die Testreihenfolge war an beiden Tagen gleich. Es wurde zuerst der Test d2 (Papier-Bleistift-Version) bearbeitet, dann der WMT (Papier-Bleistift-Version) und danach der SMACT mit seinem Vor-, Haupt- und Nachtest (Kurzversion, Langversion, Kurzversion). Der SMACT wurde auf einem Standardcomputer (Monitor 19 Zoll, 75 Hz Bildschirmfrequenz) dargeboten. Die Personen saßen ca. 60 cm vom Bildschirm entfernt. Der Wechsel von einem zum nächsten Test war für die Versuchsteilnehmer mit einer kurzen Pause mit einer Dauer von ca. einer bis zwei Minuten verbunden.

5.2.2 Ergebnisse

Deskriptive Analysen. Wiederum wurde die Stichprobe hinsichtlich ihrer deskriptiven Eigenschaften für alle Instrumente überprüft. Es fanden sich auch hier keine nennenswerten Besonderheiten, die nicht auch durch eine natürliche Variation des Leistungsverhaltens erklärt werden könnten. Ausführliche Erläuterungen (stichprobenbezogener Mittelwert, Variabilität und Verzerrungen) zu den verwendeten Instrumenten finden sich in der Tabelle A.10 im Anhang.

Reliabilität. Die Kennwerte der Retest-Reliabilität waren sehr gut für den WMT (r = .82**). Der Test d2 zeigte erwartungsgemäß eine hohe Retest-Reliabilität für die Leistungsmenge (r = .93**), etwas geringer jedoch für die Leistungsgüte (r > .79**). Sowohl Kurz- als auch Langversion des SMACT wiesen sehr gute Reliabilitätskoeffizienten auf. Dies gilt sowohl für die mittlere Reaktionszeit von kurzer (r > .82**) und langer (r > .90**) Version. Bezogen auf den prozentualen Fehleranteil zeigte sich bei kurzer Version (r > .68**) ein mäßiger und bei langer Version (r > .83**) ein sehr guter Reliabilitätskoeffizient, wobei die Reaktionszeit der primäre Leistungskennwert ist.

Validität. In Tabelle 5.3 sind die Korrelationen zwischen den klassischen Leistungstests und Schulnoten aufgeführt. Entsprechend der Erwartungshaltung war

Tabelle 5.3 – Ergebnisse der Studie 2: Korrelationen zwischen klassischen Leistungstests und Schulnoten

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Psychometrische Standardverfahren						
WMT	**-.18***	**-.30****	-.14	**-.22***	**-.25****	-.14
WMT (Retest)	**-.25****	**-.30****	**-.25***	**-.24***	**-.29****	**-.24***
d2 - GZ	**-.31****	**.21***	-.18	**-.31****	**-.21***	**-.24***
d2 - GZ (Retest)	**-.30****	**-.18***	**-.19***	**-.31****	**-.23***	**-.24***
d2 - F%	.08	.12	-.02	.04	.06	-.09
d2 - F% (Retest)	.07	.08	-.03	.02	.00	-.17

Anmerkungen. Korrelationen zwischen klassischen Leistungstests und Realschulnoten-Gesamt n = 87; Realschule-Mathe n = 88; Realschule-Physik n = 87; Abitur-Gesamt n = 88; Abitur-Mathe n = 87; Abitur-Physik n = 53; ** für p < .01 (einseitige Testung), * für p < .05 (einseitige Testung).

die allgemeine Intelligenz im Sinne induktiv-schlussfolgernden Denkens, gemessen diesmal mit dem WMT, ein guter Prädiktor für Schulnoten, insbesondere der Mathematiknote. Die Validität zeigte sich im Test für die Mathematiknote in der Realschule (r = .30**) und im Retest (r = .30**) sowie im Abitur (r = .25**) und im Retest (r = .29**). Die Gesamtmenge im Test d2 zeigte sich weniger valide, wobei die Korrelationen zu

Abiturnoten, für Test und Retest, ungefähr im Erwartungsbereich lagen, insbesondere für Mathematik (r = .21*, r = .23*) und Physik (r = .24*, r = .24*). Die Reaktionszeit im SMACT wies über alle Bedingungen hinweg Korrelationen in vergleichbarer Höhe wie zum APM auf, was den SMACT als hervorragendes Instrument auswies (r > .30**).

Wie aus Tabelle 5.4 hervorgeht, war die SMACT-Kurzversion nicht schlechter in ihrer Vorhersage als die Langversion, wobei es auf die Testreihenfolge ankommt. Die

Tabelle 5.4 – Ergebnisse der Studie 2: Korrelationen zwischen Reaktionszeit im SMACT und Schulnoten als Funktion von Testlänge (kurz versus lang) und Testverlauf (Abschnitt 1-4)

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Reaktionszeiten						
SMACT - Kurztest (Prä)	**.34****	**.42****	.18	**.32****	.25	.14
SMACT - Kurztest (Retest)	**.28****	.23	.21	.25	.13	.06
SMACT - Langtest	.25	.26	.20	.15	.08	-.03
SMACT - Langtest (Retest)	.19	.12	.16	.17	.08	.01
SMACT - T1	**.28****	**.31****	.21	.19	.10	-.03
SMACT - T1 (Retest)	.23	.15	.22	.21	.10	.04
SMACT - T2	.25	.27	.20	.13	.06	-.03
SMACT - T2 (Retest)	.19	.14	.14	.14	.06	.00
SMACT - T3	.22	.24	.18	.12	.07	-.02
SMACT - T3 (Retest)	.16	.11	.12	.15	.09	-.01
SMACT - T4	.21	.21	.19	.12	.07	-.04
SMACT - T4 (Retest)	.18	.09	.15	.16	.08	.01
SMACT - Kurztest (Post)	.24	.18	.18	.14	.08	-.05
SMACT - Kurztest (Post; Retest)	.12	.09	.04	.09	-.03	.02

Anmerkungen. Korrelation zwischen SMACT und Realschulnoten-Gesamt n = 87; entsprechend Realschule-Mathe n = 88; Realschule-Physik n = 87; Abitur-Gesamt n = 88; Abitur-Mathe n = 87; Abitur-Physik n = 53; ** für p < .01 (einseitige Testung), * für p < .05 (einseitige Testung).

Kurzversion zu Beginn der Testung wies beispielsweise den höchsten Zusammenhang

zur Realschule-Gesamtnote auf (r = .34** und r = .28** im Retest), während die Kurzversion zum Ende der Testung einen geringeren bzw. gar keinen Zusammenhang zu Schulnoten mehr aufwies (r = .24 und r = .12 im Retest). Die Vorhersage der Langversion war dagegen viel schlechter als diejenige der Kurzversion zu Beginn der Testung. Interessanterweise zeigte die Analyse des Testverlaufs in der Langversion, dass die Vorhersage im ersten Testabschnitt noch relativ gut war (r = .28** und r = .23 im Retest), während sich die Vorhersage für den weiteren Verlauf immer weiter abschwächte, sodass eigentlich keine signifikanten Korrelationen im weiteren Verlauf der Langversion mehr auftraten. Diese Ergebnisse sind klar zu interpretieren. Die Kurzversion des SMACT ist genügend lang, um einerseits genau zu messen (Reliabilität) und andererseits valide vorherzusagen (Validität). Die Langversion ist hinsichtlich dieser Kriterien nur im ersten Testabschnitt genügend. Die Kurzversion an letzter Testposition ist zwar noch reliabel, aber nicht mehr valide. Es gibt keine andere Interpretation als die, dass die der Testleistung zugrundeliegende Merkmalsdimension (Konzentrationsfähigkeit) ab einem gewissen Grade an Testermüdung bzw. Motivationsverlust nicht mehr hinreichend zwischen den Probanden differenziert. Innerhalb dieser Studie waren die Fehlerprozente überhaupt nicht prädiktiv, wie der Tabelle A.5 im Anhang zu entnehmen ist. Wie bereits in Studie 1 haben CFQ und SIML keine erwähnenswerten Vorhersagen durch selbstberichtete Konzentration in Alltagssituationen leisten können. Die Korrelationen zu den Schulnoten sind im Anhang in Tabelle A.7 einzusehen.

Regressionsmodelle. Um abschätzen zu können, inwiefern die Leistung in Konzentrationstests einen inkrementellen Beitrag zu Intelligenz als Goldstandard leisten kann, wurden zwei zweistufige (blockweise) multiple Regressionsanalysen durchgeführt. In der ersten Regressionsanalyse wurde die WMT-Leistung als erster Prädiktor definiert und die Konzentrationsleistung (Test d2-GZ und SMACT-Kurzversion vom ersten Tag) in einem zweiten blockweisen Schritt hinzugefügt. Als Kriterium wurde

die Mathematiknote aus der Realschule verwendet. Die Ergebnisse zeigen, dass die multiple Vorhersage (Intelligenz und Konzentration) einen inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .30$ auf $R = .45$). Die Änderung in F ist signifikant. In einer zweiten Regressionsanalyse wurde die WMT-Leistung als erster Prädiktor definiert und die Konzentrationsleistung (Test d2-GZ und SMACT-Kurzversion vom ersten Tag) in einem zweiten blockweisen Schritt hinzugefügt. Als Kriterium wurde diesmal die Mathematiknote aus dem Abitur verwendet. Diesmal zeigen die Ergebnisse, dass die multiple Vorhersage (Intelligenz und Konzentration) keinen inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .25$ auf $R = .31$). Die Änderung in F ist nicht signifikant. Zumindest zeigt die erste Analyse ein signifikantes Ergebnis und lässt vermuten, dass die Kombination von Matrizentest und Konzentrationstest eine effektive, weil zeitökonomische diagnostische Strategie innerhalb schulpsychologischer oder betriebspsychologischer Kontexte darstellen kann. Im Anhang finden sich in Tabelle A.19 weitere Angaben zu den Regressionsanalysen.

5.2.3 Diskussion

Die Ergebnisse der Studie 2 replizieren das Ergebnis der Studie 1, dass eine Kombination von Intelligenz- und Konzentrationsmessung einander ergänzende Komponenten enthält, welche die effektive Vorhersage in praktischen beruflichen Kontexten verbessern können. Auch die Tatsache, dass Validität zweimal (im Test und im Retest) nachgewiesen werden konnte, ist ein nennenswertes Kennzeichen von Studie 2. Zur Frage der Testlänge, Konzentrationstests betreffend, muss gesagt werden, dass eine sehr langandauernde Testdauer nicht von Vorteil ist. Auch wenn aus den Annahmen der klassischen Testtheorie direkt hervorgeht, dass sich die Gütekriterien (Reliabilität und indirekt auch Validität) durch Testverlängerung verbessern lassen, so findet sich eine empirische Entsprechung dieser Vorhersage nicht. Sehr wahrscheinlich werden

die positiven Effekte auf die Messgenauigkeit durch die simultan eintretenden Veränderungen der Merkmalsmessung aufgehoben. Es gibt keine andere Interpretation als jene, dass die der Testleistung zugrundeliegende Merkmalsdimension (Konzentrationsfähigkeit) ab einem gewissen Grade an Testermüdung bzw. durch Motivationsverlust nicht mehr hinreichend zwischen den Probanden differenziert. Die sich daraus ergebende Empfehlung lautet also: Eine Testlänge von 10 bis 15 Minuten ist ausreichend und die Testposition innerhalb einer diagnostischen Gesamtstrategie ist entscheidend. Wichtige Tests sollten also nicht an hinterer Testposition (insbesondere bei langer Testung) administriert werden. Im Allgemeinen kann dem SMACT (wie bereits in Studie 1) als Testmedium eine sehr gute Eignung zur Vorhersage bescheinigt werden.

5.3 Studie 3 - Testdiversifikation

5.3.1 Methode

Sollte man lieber eine ganze Testbatterie oder einen Matrizentest verwenden, um allgemeine kognitive Leistungsfähigkeit (Intelligenz) zu messen? Die Anwendung einer Testbatterie bringt den Vorteil mit sich, dass sich durch die vielen Untertests ein ganzes Spektrum an Fähigkeiten messen lässt, hat aber den Nachteil, dass die Anwendung nicht zeitökonomisch ist. Die Anwendung eines klassischen Matrizentests ist zeitökonomischer, aber es wird nur ein einzelnes Fähigkeitsmerkmal gemessen (schlussfolgerndes Denken), dass allerdings den Ruf genießt, allgemeine kognitive Fähigkeit in Reinform zu messen. In der Studie 3 bestand die Fragestellung darin festzustellen, ob Testdiversifikation oder Testspezifität die bessere diagnostische Strategie darstellt, wenn es um die prädiktive Validität zur Vorhersage von Schulleistungen geht. Dazu wurden sowohl eine klassische Testbatterie zur Messung von Intelligenz (I-S-T 2000 R) als auch ein klassischer Matrizentest (WMT) verwendet,

welche dann vergleichend auf ihre prädiktive Validität zur Vorhersage von Schulnoten eingesetzt wurden. Die kritische experimentelle Manipulation bestand demzufolge in einer vergleichenden Analyse der Korrelationen beider diagnostischen Strategien zu Schulleistungen (retrograde Validierung: Realschule, Abitur) und akademischen Leistungen (konkurrente Validierung: Vordiplom).

Stichprobe. An der Untersuchung nahmen insgesamt 388 Personen (Alter M = 24.9 Jahre; SD = 3.0 Jahre; 50 % Frauen) teil. Die meisten Teilnehmer waren Rechtshänder (> 87 %) und hatten keine Sehschwierigkeiten bzw. eine entsprechende Sehhilfe dabei und alle Teilnehmer gaben ebenfalls an, sich soweit gesund zu fühlen. Die Mehrzahl der Teilnehmer verfügte über die Fachhochschulreife (Fachabitur) und diese waren Studierende der Hochschule Mittweida. Mehr als die Hälfte (53 %) der Teilnehmer gaben an, bereits über einen Berufsabschluss zu verfügen. Die Teilnehmer für die Stichprobe wurde durch Aushänge auf dem Campus der Hochschule Mittweida sowie im Rahmen von Lehrveranstaltungen geworben. Die Teilnehmer erhielten auf Wunsch auch ein Feedback über ihre Ergebnisse.

Fragebogen. Es kam erneut der CFQ in der deutschen Version zum Einsatz. Außerdem wurde der SIML von Reason eingesetzt, wobei wieder die zwei Subskalen des SIML, regelbasierte und fähigkeitsbasierte Alltagsfehler, als Skalenwerte verwendet wurden. Außerdem wurde das Leistungsmotivationsinventar (LMI) von Schuler und Prochaska (2001) eingesetzt, um die Dimensionen der Leistungsbereitschaft erfassen zu können. Es wurden noch eine Reihe anderer Fragebögen im Persönlichkeitsbereich erhoben, welche aber nicht Gegenstand dieser Untersuchung sind.

Testverfahren. Zur Erfassung der Konzentrationsleistung wurde diesmal die computerisierte Version des d2 Aufmerksamkeits-Belastungs-Tests verwendet, welche eine Adaptation der klassischen Testvariante an die heutigen technischen Möglichkeiten darstellt. Weiterhin wurde auch der Frankfurter Adaptiver Konzentrationsleistungs-

Test (FAKT-II) in seiner computerisierten Variante eingesetzt (Moosbrugger & Goldhammer, 2005). Zur Erfassung der Intelligenz mittels Matrizentest wurde wieder der Wiener Matrizen-Test (WMT), diesmal als computerisierte Variante eingesetzt. Zur Erfassung von Intelligenz mittels Testbatterie wurde der Intelligenz-Struktur-Test (I-S-T 2000 R) von Amthauer et al. (2001) eingesetzt. Der Test hat zum Ziel, allgemeine Intelligenz zu messen, kann aber auch flexibel zur Erfassung bestimmter Teilfähigkeiten eingesetzt werden.

Ablauf der Untersuchung. Die Untersuchung fand an zwei Tagen im Abstand von je einer Woche statt. Am ersten Untersuchungstag wurden die Versuchsteilnehmer in Einzeltestungen in einem Labor untersucht. Die Untersuchung fand in einem von Außengeräuschen abgeschirmten Laborraum statt. Die Tests wurden auf einem handelsüblichen Computer mit Bildschirm (Monitor 19 Zoll, 75 Hz Bildschirmfrequenz) administriert. Die Personen saßen ca. 60 cm vom Bildschirm entfernt. Der Wechsel von einem zum nächsten Test war für die Versuchsteilnehmer mit einer kurzen Pause verbunden. Die Tests und der Fragebogen zur Leistungsmotivation (LMI) wurden vollständig am PC durchgeführt. Damit sich die Versuchsteilnehmer mit der Tastatur, der Maus und dem Monitor vertraut machen konnten, hatten sie zunächst die Aufgabe, ihre soziodemografischen Daten im Hogrefe Testsystem (HTS) einzutragen. Der Versuchsleiter kontrollierte und unterstützte die Versuchsteilnehmer dabei. Danach begann die vom Versuchsleiter vorbereitete Testabfolge. Zwischen den Teilen gab es einen Pausenbildschirm, der es den Teilnehmern während der Testung erlaubte, sich zwischen den einzelnen Testungen kurzzeitig zu erholen. Dabei betrug die Pausenlänge ca. eine bis zwei Minuten. Am ersten Tag wurden die Testverfahren in folgender Reihenfolge administriert: d2-C, I-S-T 2000 R und WMT als computerisierte Version. Am zweiten Tag wurden die Fragebögen und der Konzentrationstest in folgender Reihenfolge appliziert: CFQ, FAKT-II, LMI, SIML.

5.3.2 Ergebnisse

Deskriptive Analysen. Es fanden sich keine nennenswerten Besonderheiten. Ausführliche Erläuterungen (stichprobenbezogener Mittelwert, Variabilität, Verzerrungen) zu verwendeten diagnostischen Instrumenten finden sich in Tabelle A.11 im Anhang.

Reliabilität. Es gab keine Reliabilitätsmessungen in der Studie 3, jedoch wurden bis auf den I-S T 2000 R nur solche diagnostische Instrumente verwendet, für die eine hohe Reliabilität im Retest auch nachgewiesen wurde.

Validität. In Tabelle 5.5 sind die Korrelationen zwischen Schulnoten und dem WMT und I-S-T 2000 R aufgeführt. Angaben zu den Korrelationen der Subtests des I-S-T 2000 R können der Tabelle A.14 im Anhang entnommen werden. Entgegen den Erwartungen waren die Korrelationen zwischen WMT und Schulnoten geringer als in der Studie 2. Aufgrund der hohen Stichprobengröße waren die Korrelationen aber bereits bei r = .14** (Mathematiknote in Realschule und Abitur) sehr signifikant. Verglichen mit WMT waren die Zusammenhänge zwischen dem Gesamtwert des

Tabelle 5.5 – Ergebnisse der Studie 3: Korrelationen zwischen klassischen Leistungstests und Schulnoten

Skalen	Realschule			Abitur			Vordiplom
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren							
WMT	-.06	**-.14****	**-.10***	-.08	**-.14****	**-.13***	**-.27****
I-S-T - Schlussfolg. Denken	-.03	**-.17****	**-.10***	-.03	**-.13****	-.07	**-.22****
I-S-T- Verbal	.02	.00	.03	.00	.07	.03	-.10
I-S-T- Numerisch	-.04	**-.18****	**-.15****	-.04	**-.18****	-.08	**-.25****
I-S-T- Figural	-.03	**-.16****	-.07	-.03	**-.13****	-.09	-.11
I-S-T- Merken gesamt	-.05	**-.10***	-.08	-.02	-.05	-.05	**-.28****

Anmerkungen. Korrelation zwischen WMT und Realschulnoten-Gesamt n = 351; Realschule-Mathe n = 374; Realschule-Physik n = 373; Abitur-Gesamt n = 366; Abitur-Mathe n = 378; Abitur-Physik n = 258; Vordiplom n = 239; Korrelationen zwischen I-S-T 2000 R und Realschulnoten-Gesamt n = 352; Realschule-Mathe n = 375; Realschule-Physik n = 374; Abitur-Gesamt n = 367; Abitur-Mathe n = 379; Abitur-Physik n = 259; Vordiplom n = 239; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

I-S-T 2000 R und Schulnoten in ungefähr gleicher Größenordnung gelagert (r = .17** und r = .13** für Mathematiknote in Realschule und Abitur; r = .22** für Vordiplom). Bei den Untertests des I-S-T 2000 R (siehe dazu die Tabelle A.14 im Anhang) zeichnete sich lediglich der Test Rechenaufgaben aus, der Vorhersagen zu Schulnoten im Bereich von r = .20** bis r = .30** zu machen imstande war. Wie in Tabelle 5.6 ersichtlich ist, war die Vorhersage des Test d2-C nicht gut, zumindest wenn man sie mit der erreichten Vorhersage in den Studien 1 und 2 vergleicht. Dies gilt auch für den FAKT-II, denn auch dieser Konzentrationstest zeigte sich nur wenig vorhersagekräftig. Auch in Studie 3 zeigten die beiden Fragebögen zur selbstberichteten Konzentration,

Tabelle 5.6 – Ergebnisse der Studie 3: Korrelationen zwischen klassischen Leistungstests und Schulnoten

Skalen	Realschule			Abitur			Vordiplom
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren							
d2 - GZ	-.05	-.07	-.07	-.04	-.08	-.10	-.08
d2 - F%	.08	**.10***	.06	.05	**.09***	**.13***	**.13***
FAKT - KL	-.07	-.09	**-.15****	-.01	-.04	**-.12***	-.08
FAKT - KG	**-.14****	**-.10***	**-.14****	**-.10****	-.01	**-.15****	-.09
FAKT - KH	**-.13***	**-.11***	**-.15****	-.06	.00	**-.17****	-.11

Anmerkungen. Korrelation zwischen d2 und Realschulnoten-Gesamt n = 351; entsprechend Realschule-Mathe n = 374; Realschule-Physik n = 373; Abitur-Gesamt n = 366; Abitur-Mathe n = 378; Abitur-Physik n = 258; Vordiplom n = 239; Korrelation zwischen FAKT II und Realschulnoten-Gesamt n = 318; Realschule-Mathe n = 338; Realschule-Physik n = 337; Abitur-Gesamt n = 330; Abitur-Mathe n = 341; Abitur-Physik n = 236; Vordiplom n = 208; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

CFQ und SIML, keine erwähnenswerten Vorhersagen. Die Tabelle A.8 im Anhang weist die entsprechenden Korrelationen dazu aus. Da einige rezente Studien, (z. B. Spinath et al., 2006, Schwinger et al., 2009, Spinath et al., 2010, Steinmayr et al., 2011) Zusammenhänge mit motivationalen Variablen und Schulleistungen gezeigt haben, wurde in Studie 3 unter anderem auch das Leistungsmotivationsinventar eingesetzt. Es bestand die explorative Frage, ob und welche Facetten von Leistungsmotivation ein

Prädiktor für Schul- und Studienleistung sein können. Die Tabelle 5.7 zeigt die Korrelationen des LMI zu den Noten. Interessanterweise fanden sich keinerlei nennenswerte Zusammenhänge mit Variablen wie Dominanz, Leistungsstolz, Wettbewerbsorientierung, oder Leistungsstreben. Dagegen waren drei Variablen prädiktiv: Beharrlichkeit, Engagement und Selbstkontrolle (mit dem Vordiplom, r = .19**, r = .16**, r = .16**).

Tabelle 5.7 – Ergebnisse der Studie 3: Korrelationen zwischen Leistungsmotivation und Schulnoten

Skalen	Realschule			Abitur			Vordiplom
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Psychometrisches Standardverfahren							
LMI - Gesamtwert	-.05	**-.13***	-.05	**-.13***	-.05	-.11	**-.20****
LMI - Beharrlichkeit	-.10	**-.18****	-.07	**-.16****	**-.11***	-.13	**-.19****
LMI - Dominanz	-.04	.02	.04	-.06	.08	.01	-.08
LMI - Engagement	**-.15****	**-.16****	-.08	**-.11***	-.05	**-.13***	**-.16***
LMI - Erfolgszuversicht	.01	**-.12***	-.05	-.05	-.03	-.03	**-.22****
LMI - Flexibilität	.06	-.03	.05	.04	.04	.05	-.07
LMI - Flow	-.04	**-.12***	-.08	**-.14***	.00	**-.17****	-.07
LMI - Furchtlosigkeit	-.03	-.08	-.07	-.07	.00	-.03	**-.15***
LMI - Internalität	.00	.00	-.03	-.04	.01	.02	**.15***
LMI - Kompens. Anstrengung	-.10	-.04	-.01	**-.15****	**-.15****	**-.16***	-.05
LMI - Leistungsstolz	-.07	-.09	-.05	**-.19****	-.07	**-.16***	-.12
LMI - Lernbereitschaft	.00	-.07	.00	-.05	.03	-.10	-.08
LMI - Schwierigkeitspräferenz	-.02	**-.18****	-.07	-.10	-.10	-.12	**-.20****
LMI - Selbstständigkeit	.09	.02	.03	.03	.10	.05	-.05
LMI - Selbstkontrolle	**-.20****	**-.15****	**-.16****	**-.22****	**-.11***	**-.17****	**-.16***
LMI - Statusorientierung	.08	.02	.09	.03	-.04	-.03	-.06
LMI - Wettbewerbsorientierung	-.04	-.05	-.05	-.05	-.05	-.05	-.12
LMI - Zielsetzung	-.02	-.09	-.01	-.07	-.07	-.11	**-.19****

Anmerkungen. Korrelation zwischen LMI und Realschulnoten-Gesamt n = 319; Realschule-Mathe n = 339; Realschule-Physik n = 338; Abitur-Gesamt n = 331; Abitur-Mathe n = 342; Abitur-Physik n = 236; Vordiplom n = 209; ** für $p < .01$ (zweiseitige Testung), * für $p < .05$ (zweiseitige Testung).

Regressionsmodelle. In den Studien 1 und 2 zeigten sich Hinweise darauf, dass Matrizentests und Rechenaufgaben (im SMACT) eine gute Kombination zur Vorhersage von Schulleistungen darstellen, weshalb es eine ausreichende Motivation für die erneute Prüfung gab, ob sich diese Befunde in Studie 3 replizieren lassen. Immerhin waren auch hier wieder Matrizentest und Rechenaufgaben-Test die besten Prädiktoren zur Vorhersage des relevanten Kriteriums. Um also zu untersuchen, inwiefern die Leistung im I-S-T 2000 R-Untertest Rechenaufgaben einen inkrementellen Beitrag zur Intelligenz (wie gemessen mit dem WMT) leisten kann, wurden zwei mehrstufige (blockweise) multiple Regressionsanalysen durchgeführt, in der die WMT Leistung als erster Prädiktor definiert wurde und die Leistung im Rechentest in einem zweiten Schritt hinzugefügt wurde. Als Kriterium diente in der ersten Analyse die Mathematiknote der Realschule, in der zweiten Analyse die Mathematiknote im Abitur. Die Ergebnisse in der ersten Analyse zeigen, dass die multiple Vorhersage (Intelligenz und Rechenaufgaben) einen inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .14$ auf $R = .25$). Die Änderung in F ist dabei signifikant. In der zweiten Regressionsanalyse mit der Mathematiknote aus dem Abitur als Kriterium zeigte die multiple Vorhersage (Intelligenz und Rechenaufgaben) ebenso einen inkrementellen Beitrag gegenüber einer Einfachvorhersage (von $R = .14$ auf $R = .24$). Auch hier war die Änderung in F signifikant. Details zur Regressionsanalyse finden sich in Tabelle A.20 im Anhang. Damit zeigen beide Analysen ein signifikantes Ergebnis und replizieren in Teilen die Ergebnisse aus den Studien 1 und 2 und zeigen in konzeptuell ähnlicher Weise, dass die Kombination von induktiv-schlussfolgerndem Denken und Kopfrechnen, als klassische Kulturtechnik, eine effektive Kombination als Prädiktor zur Erfolgsvorhersage darstellt.

Um zu prüfen, ob die motivationalen Variablen einen zusätzlichen Beitrag zu allgemeiner Intelligenz leisten, wurden drei Regressionsanalysen berechnet, indem

jeweils die Leistung im WMT als Hauptprädiktor eingesetzt wurde und die drei motivationalen Variablen (Beharrlichkeit, Engagement und Selbstkontrolle) dagegen als inkrementelle Prädiktoren blockweise verwendet wurden. Die Mathematiknote der Realschule, die Mathematiknote aus dem Abitur und die Note vom Vordiplom wurden jeweils separat als Kriterium in einer der drei Analysen herangezogen. Die erste Analyse mit der Realschul-Mathematiknote als Kriterium ergab, dass Motivation einen inkrementellen Beitrag dazu leisten kann, Studienleistungen vorherzusagen. Das multiple Modell (R = .26) war dem Einfachmodell (R = .14) überlegen. Die Änderung in F war dabei signifikant. Die zweite Analyse mit dem Kriterium Abitur-Mathematiknote wiederum zeigte, dass Motivation keinen inkrementellen Beitrag gegenüber der Einfachvorhersage (von R = .14 auf R = .20) zeigte. Die Änderung in F erwies sich als nicht signifikant. Die dritte Analyse mit der Vordiplomnote als Kriterium konnte wiederum zeigen, dass die Motivation hinsichtlich der Studienleistungen einen inkrementellen Beitrag leisten kann. Das multiple Modell (R = .38) war dem Einfachmodell (R = .29) überlegen. Es zeigte sich eine signifikante Änderung in F. Die Ergebnisse der drei Regressionsanalysen weisen somit auch darauf hin, dass die motivationalen Variablen Beharrlichkeit, Engagement und Selbstkontrolle gemeinsam mit induktiv-schlussfolgerndes Denken eine wirksamere Vorhersage erlauben. Details zu den Regressionsanalysen mit den motivationalen Variablen finden sich in Tabelle A.21 im Anhang.

5.3.3 Diskussion

Zusammenfassend kann geschlussfolgert werden, dass die Ergebnisse der Studie 3 indizieren, dass ein kleiner, aber effektiver Matrizentest wie der WMT einer großen Testbatterie hinsichtlich der statistischen Vorhersagekraft absolut ebenbürtig sein kann. Der WMT ist also ein Instrument, das jedem praktisch arbeitenden Diagnostiker wert-

volle Dienste leisten kann. Die hohe Validität bei gleichzeitig hoher Testökonomie macht dieses Instrument zu einem beinahe unverzichtbaren Handwerkszeug eines jeden Schulpsychologen und Betriebspsychologen mit personaldiagnostischem Aufgabenbereich. Außerdem konnten die Ergebnisse der Studien 1 und 2 repliziert werden, da sich auch in Studie 3 zeigte, dass die effektive Kombination von Matrizenaufgaben und Kopfrechnen zu einem inkrementellen Vorhersagevorteil führt, der die effektive Vorhersage in praktischen beruflichen Kontexten verbessern kann. Die Tatsache, dass weder der Test d2-C noch der FAKT-II in der Lage waren, vernünftige Vorhersagen zu machen, ist schwierig zu erklären. Die Stichprobenvarianz kann nur bedingt als post-hoc-Erklärung herhalten, denn dann hätten sich ja auch keine signifikanten Korrelationen zwischen I-S-T 2000 R und Matrizentest ergeben dürfen. Dies ist ein Vorteil der hier vorliegenden Studie, dass die Evaluation der Validität von Testinstrumenten vergleichend anhand derselben Stichprobe vorgenommen wurde.

5.4 Studie 4 - Validität in realer Auswahlsituation (I-S-T 2000 R)

5.4.1 Methode

Die Testautoren des Wilde-Intelligenz-Tests (WIT-2), Kersting et al. (2008, S. 12) führen in ihrem Handbuch folgendes Statement an: „Während zahlreiche Testentwickler für die Entwicklung und Normierung auf die Daten von so genannten „anfallenden“ Stichproben zurückgreifen, wurde der überwiegende Teil der Daten für den WIT-2 im Kontext des Ernstfalls von beruflichen Bewerbungssituationen erhoben (und nicht anhand anonymer Gruppen von Schülern oder Studierenden). Die Testgütekriterien sind somit für diejenigen Subpopulationen bestimmt worden, für die der Test laut diagnostischer Zielsetzung eingesetzt werden soll“. Gemäß dieser Sichtweise sollten

Testverfahren, welche für die berufliche Praxis entwickelt worden sind, auch an der definierten Zielgruppe validiert werden. Da die meisten Testverfahren nicht in einem angewandten Kontext validiert wurden, steht die Frage unbeantwortet im Raum, ob die Validität eines Verfahrens dadurch unterschätzt oder überschätzt wird. Diese Frage sollte beantwortet werden, indem der I-S-T 2000 R in einer echten Bewerbersituation eingesetzt wurde. Die Probanden wurden im Kontext des Bewerbungsverfahrens für den Masterstudiengang an der Hochschule Mittweida getestet. Kritisch ist dabei der Vergleich der Validität des I-S-T 2000 R in einer Messung im psychologischen Experimentallabor. Da es sich in den Studien 3 und 4 um die gleichen Stichprobenpopulationen handelt, wurden die Korrelationen zwischen Testleistung und Schulnoten in den Studien 3 und 4 vergleichend evaluiert. Die Validierung erfolgte in Studie 3 wieder retrograd für schulische Leistung (Abitur) und konkurrent für akademische Leistung (Bachelor). Zusätzlich wurde auch die Konzentrationsleistung gemessen, diesmal jedoch wurde der Komplexe Konzentrationstest (KKT) von Westhoff und Scholz (2007) eingesetzt, welcher die Leistung der Probanden bei numerischen, figuralen und verbalen Aufgabenmaterial erfasst.

Stichprobe. An der Untersuchung nahmen insgesamt 105 Personen (52 Frauen, 53 Männer) im Alter von 21 bis 37 Jahren (M = 25.3; SD = 3.4) teil, die sich für den Masterstudiengang Betriebswirtschaftslehre bzw. Industrial Management an der Hochschule Mittweida beworben hatten. Zum Zeitpunkt der Teilnahme verfügten 96 % der Personen bereits über einen Hochschulabschluss (Bachelor oder Diplom). Alle Teilnehmer hatten normales Sehvermögen bzw. hatten ihre Sehhilfe dabei und gaben an, sich gesund zu fühlen. Im Rahmen des Auswahlprozesses für eine limitierte Anzahl von (Master-) Studienplätzen wurde die Stichprobe an der Hochschule Mittweida gewonnen und unter Berücksichtigung des Datenschutzes gespeichert. Die Teilnehmer erhielten auf Wunsch ein Feedback zu ihren Ergebnissen. Nach Abschluss des

Bewerbungsverfahrens wurden die Daten anonymisiert, sodass kein Rückschluss auf persönliche Daten möglich ist. Damit verbunden war jedoch auch keine nachträgliche anterograde Validierung des I-S-T 2000 R mehr möglich, in der die vorhergesagte Leistung im Masterstudiengang durch die Testleistung von vornherein hätte vorhergesagt werden können.

Fragebogen. Da sich in den Studien 1 bis 3 keine erwähnenswerten Vorhersagen durch selbstberichtete Konzentration in Alltagssituationen fanden, wurden diese Fragebögen, CFQ und SIML, nicht mehr eingesetzt.

Testverfahren. Zur Erfassung von allgemeiner kognitiver Leistungsfähigkeit (Intelligenz) wurde der I-S-T 2000 R eingesetzt. Die Testadministration fand in der gleichen Weise statt, wie es in der Studie 3 der Fall war, außer dass es sich bei der Untersuchung in Studie 4 um eine echte Bewerbersituation handelte. Den Probanden in der hier berichteten Studie 4 war also bewusst, dass sich die erbrachte Leistung auf die zukünftige Lebenssituation auswirken kann, während die Probanden in der Vergleichsstichprobe der Studie 3 keine derartigen Auswirkungen zu fürchten hatten. Der Komplexe Konzentrationstest (KKT) wurde eingesetzt, um die Konzentrationsfähigkeit der Probanden zu erfassen.

Ablauf der Untersuchung. Die Untersuchung fand verteilt an mehreren Tagen an der Hochschule Mittweida statt. Die Tests wurden auf einem handelsüblichen Computer mit Bildschirm (Monitor 19 Zoll, 75 Hz Bildschirmfrequenz) administriert und die Personen saßen ca. 60 cm vom Bildschirm entfernt. Der Wechsel von einem bestimmten Test zum nächsten Test war für die Versuchsteilnehmer immer mit einer kurzen Pause (ca. eine bis zwei Minuten) verbunden. Die Tests wurden vollständig in computerisierter Form durchgeführt, wobei der I-S-T 2000 R mittels des Hogrefe Testsystems (HTS) und der KKT über das Wiener Testsystem (WTS) präsentiert worden ist. Die Probanden hatten Gelegenheit, sich mit den Gegebenheiten vertraut zu

machen und waren zunächst aufgefordert, ihre Angaben zu Person, Alter und anderen soziodemographischen Variablen im System einzutragen. Danach begann die vom Versuchsleiter vorbereitete Testabfolge, die in der Administration des KKT und darauf folgend des I-S-T 2000 R bestand. Zwischen den Untertests gab es eine kurze Pause (ca. eine bis zwei Minuten), zu der die Probanden durch eine am Bildschirm dargebotene Instruktion aufgefordert wurden.

5.4.2 Ergebnisse

Deskriptive Analysen. Es fanden sich keine nennenswerten Besonderheiten. Ausführliche Erläuterungen (stichprobenbezogener Mittelwert, Variabilität, Verzerrungen) zu verwendeten diagnostischen Instrumenten finden sich im Anhang in Tabelle A.12.

Reliabilität. Da es aufgrund der Tatsache, dass die Testung in einer realen Bewerbersituation erfolgte, nur einen Testzeitpunkt gab, konnte die Retest-Reliabilität der eingesetzten Instrumente nicht ermittelt werden. Für den KKT zeigte Flehmig et al. (2007), dass dieser in besonderem Maße retest-reliabel ist. Für den I-S-T 2000 R liegen diesbezüglich keine Befunde vor.

Validität. Wieder sind in Tabelle 5.8 auf der nächsten Seite die Korrelationen zwischen Leistungstests (I-S-T 2000 R und KKT) und den relevanten Erfolgskriterien (schulische und akademische Leistungen) aufgeführt. Die Angaben zu den Korrelationen der Subtests des I-S-T 2000 R können im Anhang der Tabelle A.15 entnommen werden. Diesmal zeigte sich die Gesamtleistung des I-S-T 2000 R als valide hinsichtlich der Vorhersage von Abiturnoten, sowohl für die Gesamtnote (r = .30**) als auch die Mathematiknote (r = .25*) und die Physiknote (r = .35**). Jedoch konnte die I-S-T-Gesamtleistung keine Studienleistung (Bachelor) vorhersagen. Von den Untertests des I-S-T 2000 R taten sich die Tests Analogien und Rechenaufgaben hervor, ansonsten war keiner der Untertests für sich allein stehend prädiktiv. Was die Leistung im

Tabelle 5.8 – Ergebnisse der Studie 4: Korrelationen zwischen klassischen Leistungstests und Schulnoten im Kontext einer realen Bewerberauslesesituation an der Hochschule Mittweida

Skalen	Abitur			Bachelor
	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren				
I-S-T - Schlussfolgerndes Denken	**-.30****	**-.25****	**-.35****	.04
I-S-T- Verbale Intelligenz	-.33**	-.23*	-.33**	-.01
I-S-T- Numerische Intelligenz	-.19*	-.23*	-.19	.03
I-S-T- Figurale Intelligenz	-.19*	-.13	-.34**	.07
I-S-T- Merken gesamt	-.10	-.08	-.18	-.23*
KKT - Gesamtleistung	-.21*	-.10	-.14	-.03
KKT - Figurales Material	-.02	.01	-.06	.01
KKT - Numerisches Material	-.30**	-.15	-.20	-.02
KKT - Verbales Material	-.21*	-.11	-.11	-.05

Anmerkungen. Korrelation zwischen I-S-T 2000 R und Abitur-Gesamt n = 101; Abitur-Mathe n = 97; Abitur-Physik n = 58; Bachelor n = 93; Korrelationen zwischen KKT und Abitur-Gesamt n = 101; Abitur-Mathe n = 97; Abitur-Physik n = 58; Bachelor n = 93; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

KKT angeht, so zeigte sich hier, dass der Untertest numerisches Material (Berechnen von Kettenaufgaben) für sich allein stehend eine bessere Vorhersage leisten konnte (r = .30** zur Abiturnote) als der KKT-Gesamtwert und alle anderen KKT-Untertests (figural und verbal). Dies zeigt erneut, dass Rechenaufgaben als Aufgabenform in Intelligenz- und Konzentrationstests sehr gut geeignet sind, schulische Leistungen vorherzusagen. Verglichen zu einer Experimentalsituation (Studie 3) ist die Validität des I-S-T-Gesamtwerts in einer realen Bewerbersituation deutlich valider (r = .03 versus r = .30**), wobei sich für die Untertests ein differenzierteres Muster ergibt. Die nachfolgenden Abbildungen 5.1 auf der nächsten Seite und 5.2 auf Seite 120 zeigen die Korrelationen vergleichend visualisiert, jeweils für Studie 3 (Experimentallabor) und Studie 4 (reale Bewerbersituation).

Die Prüfung der Korrelationskoeffizienten auf Unterschiedlichkeit ergab unter

Berücksichtigung der unterschiedlichen Stichprobengrößen, hochsignifikante Unterschiede für Schlussfolgerndes Denken und Verbale Intelligenz. Auf der Ebene der I-S-T-Subtests sind die Unterschiede in den Korrelationen für Analogien und Rechenzeichen hochsignifikant, für Satzergänzung und Gemeinsamkeiten signifikant. Im Anhang sind der Tabelle A.17 die Korrelationen, die Prüfgrößen z sowie die p-Werte zu entnehmen.

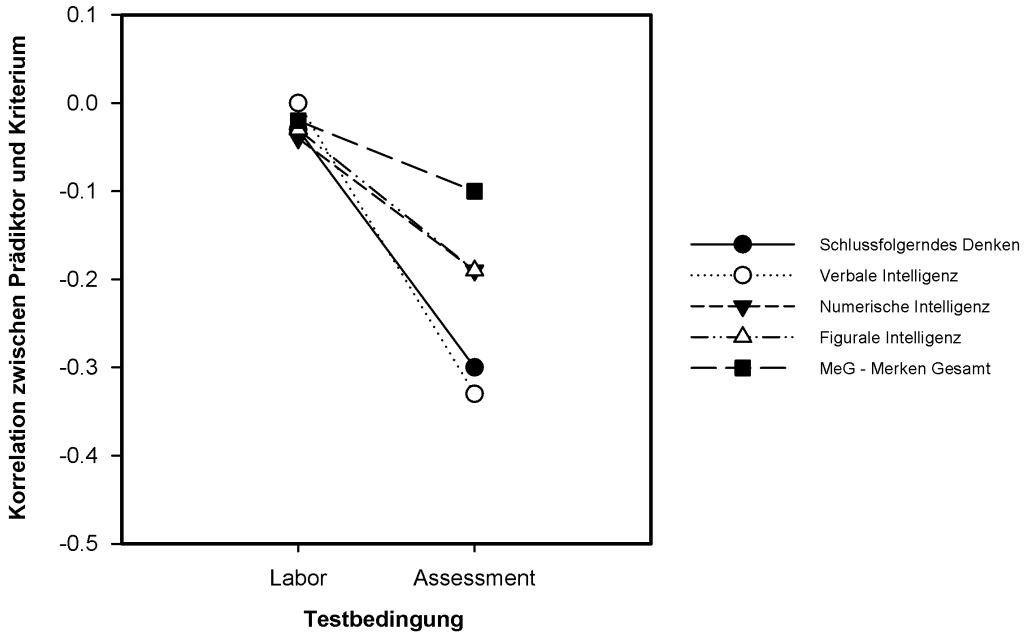

Abbildung 5.1 – Für den Vergleich wurden die Korrelationen der aggregierten Parameter Schlussfolgerndes Denken, verbale, numerische und figurale Intelligenz sowie die Gesamtleistung Merken des I-S-T 2000 R als Prädiktoren und die Abitur-Gesamtnote als Kriterium definiert.

Regressionsmodell. In den Studien 1, 2 und 3 zeigte sich, dass ein Intelligentest in Kombination mit Rechenaufgaben eine gute Kombination zur Vorhersage von Schulleistungen darstellt, weshalb auch hier geprüft werden sollte, ob sich die Kombination Intelligenz und Rechenaufgaben erneut bewähren würde. Da weder der WMT noch der APM-Test von Raven in dieser Studie eingesetzt wurden, wird die Analyse der inkrementellen Validität natürlich nicht aus der Perspektive der Testökonomie durch-

geführt. Dies soll lediglich der Frage dienen, ob die Validität von Intelligenz durch

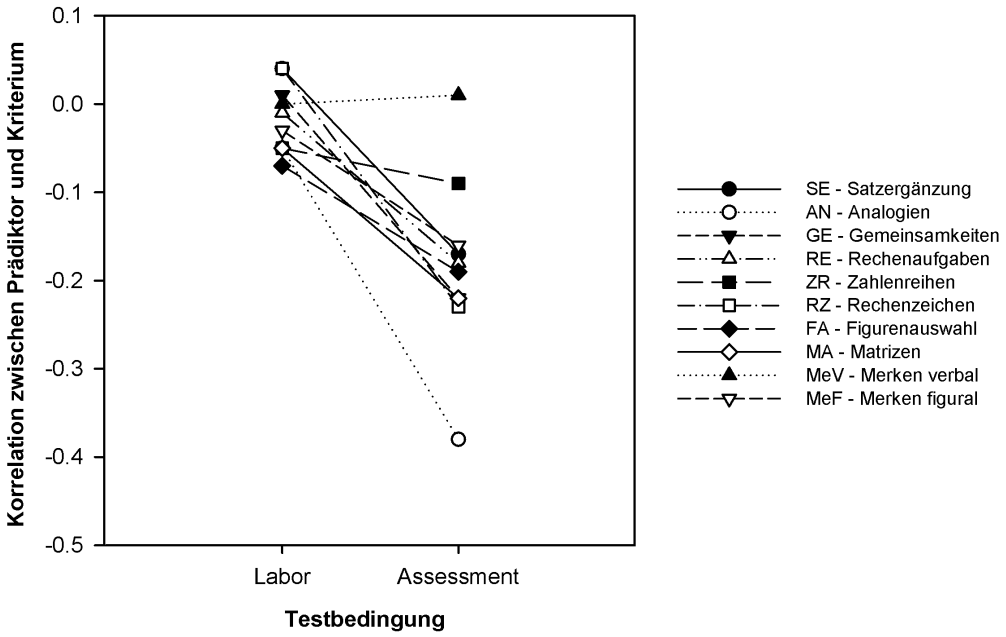

Abbildung 5.2 – Für den Vergleich wurden die Korrelationen der 8 Subtests sowie der 2 Subtest des Moduls Merkaufgaben als Prädiktoren zur Abitur-Gesamtnote als Kriterium betrachtet.

das Hinzufügen von Rechenaufgaben inkrementell verbessert werden kann. Um dies zu untersuchen, wurde eine mehrstufige multiple Regressionsanalyse durchgeführt, in der die I-S-T-Gesamtleistung (Intelligenz) als erster Prädiktor einging und die Leistung im KKT-Rechentest (numerisches Material) in einem zweiten Schritt hinzugefügt wurde. Als vorherzusagendes Kriterium wurde die Abitur-Gesamtnote definiert. Die Ergebnisse zeigen in Studie 4, dass die multiple Vorhersage (Intelligenz und Rechenaufgaben) keinen inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .30$ auf $R = .34$). Die Änderung in F ist nicht signifikant. Damit zeigt das Ergebnis, dass aus praktischer Sicht die I-S-T-Gesamtleistung ausreichend für die Vorhersage des Kriteriums ist und nicht durch die Hinzunahme des KKT-Rechentests ein zusätzlicher Beitrag generiert wird. Aus theoretischer Sicht kann angenommen werden, dass sehr wohl ein eigenständiger inkrementeller Beitrag des KKT-Rechentests vorliegt, jedoch

sehr ähnliche Aspekte des Verhaltens prädiktiv zum Kriterium sind und somit eine redundante Vorhersage erfolgt.

5.4.3 Diskussion

Zusammenfassend kann geschlussfolgert werden, dass die prädiktive Validität des I-S-T 2000 R in einer realen Bewerbersituation höher ist, verglichen zu einer Testsituation im psychologischen Experimentallabor. Schließlich lag die Korrelation zwischen I-S-T-Gesamtleistung in Studie 3 bei r = .03, während sie in Studie 4 bei r = .30** lag, wobei beide Stichproben als vergleichbar angesehen werden können. Man kann also davon ausgehen, dass die mit Stress verbundene reale Bewerbersituation nicht zu einer Nivellierung individueller Differenzen führt. Die im psychologischen Labor ermittelten Kennwerte der Validität stellen also eher eine Unterschätzung der in der Realität zu erwartenden Validität dar, was die hier vorliegenden Ergebnisse empirisch belegen. Da die Ergebnisse an einer gleichartigen Stichprobe ermittelt wurden, gibt es auch methodisch nicht viel zu kritisieren. Aber es sollte an dieser Stelle klar werden, dass die von Kersting angeführte Möglichkeit der reduzierten Stichprobenvarianz in praktischen Kontexten der Eignungsfeststellung dennoch auftreten kann. Dieser Problematik kann sich die vorliegende Arbeit nicht widmen. Die vorliegende Arbeit hat sich auf den Wirkmechanismus fokussiert, der in realen Bewerberkontexten auftreten kann, nämlich auf die Effekte von Leistungsdruck und Stress und die damit verbundene Frage, ob erhöhter Leistungsdruck in einer gesamten Stichprobenpopulation zu einer Erhöhung oder Nivellierung individueller Differenzen führt und damit zu einer Verbesserung versus einer Verschlechterung der testpsychologischen Gütekriterien Reliabilität und Validität.

5.5 Studie 5 - Validität in realer Auswahlsituation (WIT-2)

5.5.1 Methode

In der finalen Studie 5 wurde erneut untersucht, ob sich eine valide Diagnostik auch in einer realen Bewerbersituation durchführen lässt. Während in der Studie 4 der I-S-T 2000 R verwendet wurde, dessen Validität zur Vorhersage von Schulnoten dann mit der Validität des I-S-T 2000 R in einer „artifiziellen“ Situation im Experimentallabor verglichen wurde, so kommt in Studie 5 der WIT-2 zum Einsatz. Wie bereits erörtert, ist der WIT-2 explizit dafür konstruiert worden, den Anforderungen diagnostischer Strategien innerhalb von echten Personalauswahlverfahren in praktischen Berufsfeldern zu genügen. Daher bietet es sich an, den WIT-2 auf seine prädiktive Validität zur Vorhersage von Schulnoten hin zu überprüfen. Da es keine Vergleichsstichprobe gibt, kann lediglich beurteilt werden, ob der WIT-2 als Instrument in einem Kontext valide ist, in dem Probanden einer echten Auswahlsituation mit entsprechenden Konsequenzen ausgesetzt sind. Da anzunehmen ist, dass die Motivation der Probanden hier höher ist als in einer äquivalenten Laborsituation, stellt sich die Frage, ob eine solche Situation mit einer zufriedenstellenden Validität einhergeht. Wie in Studie 4 wurde zusätzlich noch der KKT eingesetzt, um die Validität eines Konzentrationstests zu evaluieren.

Stichprobe. An der Untersuchung nahmen insgesamt 97 Personen (59 Frauen, 38 Männer) im Alter von 21 bis 35 Jahren (M = 25.2; SD = 3.1) teil, die sich für den Masterstudiengang Betriebswirtschaftslehre bzw. Industrial Management an der Hochschule Mittweida beworben hatten. Zum Zeitpunkt der Teilnahme verfügten 78 % der Personen bereits über einen Hochschulabschluss (Bachelor oder Diplom). Alle Teilnehmer hatten normales Sehvermögen bzw. hatten ihre Sehhilfe dabei und gaben

an, sich gesund zu fühlen. Im Rahmen des Auswahlprozesses für eine limitierte Anzahl von (Master-) Studienplätzen wurde die Stichprobe an der Hochschule Mittweida gewonnen und die generierten Daten unter Berücksichtigung des Datenschutzes gespeichert. Die Teilnehmer erhielten auf Wunsch ein Feedback zu ihren Ergebnissen. Nach Abschluss des Bewerbungsverfahrens wurden die Daten anonymisiert, sodass kein Rückschluss auf persönliche Daten möglich ist. Damit verbunden war jedoch auch keine nachträgliche anterograde Validierung des WIT-2 mehr möglich, in der die vorhergesagte Leistung im Masterstudiengang durch die Testleistung von vornherein hätte vorhergesagt werden können.

Fragebogen. Da sich in den Studien 1 bis 3 keine erwähnenswerten Vorhersagen durch selbstberichtete Konzentration in Alltagssituationen fanden, wurden diese Fragebögen, CFQ und SIML, nicht mehr eingesetzt.

Testverfahren. Zur Erfassung von allgemeiner kognitiver Leistungsfähigkeit (Intelligenz) wurde der WIT-2 eingesetzt. Die Testadministration fand in der gleichen Weise statt, wie es in der Studie 4 der Fall war. Ebenso wie in Studie 4 handelte es sich bei der hier berichteten Untersuchung um eine echte Bewerbersituation. Den Probanden in der hier berichteten Studie 5 war also bewusst, dass sich die erbrachte Leistung auf die zukünftige Lebenssituation auswirken kann, während die Probanden in der Vergleichsstichprobe der Studie 3 keine derartigen Auswirkungen zu fürchten hatten. Der Komplexe Konzentrationstest (KKT) wurde eingesetzt, um die Konzentrationsfähigkeit der Probanden zu erfassen.

Ablauf der Untersuchung. Die Untersuchung fand verteilt auf mehrere Tage an der Hochschule Mittweida statt. Die Tests wurden auf einem handelsüblichen Computer mit Bildschirm (Monitor 19 Zoll, 75 Hz Bildschirmfrequenz) administriert und die Personen saßen ca. 60 cm vom Bildschirm entfernt. Der Wechsel von einem bestimmten Test zum nächsten Test war für die Versuchsteilnehmer immer mit einer kurzen

Pause (ca. eine bis zwei Minuten) verbunden. Die Tests wurden vollständig in computerisierter Form durchgeführt, wobei der WIT-2 mittels des Hogrefe Testsystems (HTS) und der KKT über das Wiener Testsystem (WTS) präsentiert worden ist. Die Probanden hatten Gelegenheit, sich mit den Gegebenheiten vertraut zu machen und waren zunächst aufgefordert, ihre Angaben zu Person, Alter und anderen soziodemographischen Variablen im System einzutragen. Danach begann die vom Versuchsleiter vorbereitete Testabfolge, die in der Administration des KKT und darauf folgend des WIT-2 bestand. Zwischen den Untertests gab es eine kurze Pause (ca. eine bis zwei Minute), zu der die Probanden durch eine am Bildschirm dargebotene Instruktion aufgefordert wurden.

5.5.2 Ergebnisse

Deskriptive Analysen. Es fanden sich keine nennenswerten Besonderheiten. Ausführliche Erläuterungen (stichprobenbezogener Mittelwert, Variabilität, Verzerrungen) zu verwendeten diagnostischen Instrumenten finden sich im Anhang in Tabelle A.13.

Reliabilität. Aufgrund der mit der Bewerbersituation verbundenen einmaligen Testung wurde die Retest-Reliabilität nicht erfasst, jedoch sind beide Verfahren (WIT-2 und KKT) nachgewiesenermaßen retest-reliabel.

Validität. Wie aus Tabelle 5.9 hervorgeht, zeigte sich der WIT-2 entgegen den Erwartungen kaum valide hinsichtlich der Gesamtnote in Realschule und Abitur. Das ist erstaunlich, da der Test ja ausdrücklich für die Anwendung in praktischen Personalauswahlkontexten entwickelt wurde. Sehr interessant ist jedoch der Befund, dass zwar keiner der WIT-2-Untertests ein guter Prädiktor für die Gesamtnote war, jedoch für die Physiknote. Aus der Tabelle A.16 im Anhang geht hervor, dass insbesondere die Untertests „Zahlenreihen“, „Analogien“, und „Grundrechnen“ gute Prädiktoren für die Physiknote im Abitur sind (r = .32*, r = .31*, r = .28*).

Tabelle 5.9 – Ergebnisse der Studie 5: Korrelationen zwischen klassischen Leistungstests und Schulnoten im Kontext einer realen Personalauslesesituation

Skalen	Realschule			Abitur			Bachelor
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren							
WIT - Schlussfolg. Denken	-.06	**-.22***	-.02	-.11	.01	**.28***	-.03
WIT - Sprachliches Denken	.01	.01	.07	-.02	.13	**.27***	-.05
WIT - Rechnerisches Denken	.11	.05	.10	.04	.00	.22	.05
WIT - Räumliches Denken	-.08	-.15	-.04	-.10	.14	.05	.02
WIT - Merkfähigkeit	-.13	**-.20***	-.04	-.14	-.13	.09	.04
KKT - Gesamtleistung	.08	.02	.15	.05	**.21***	**.51****	.00
KKT - Figurales Material	.05	.08	.05	.02	**.29****	**.34***	.01
KKT - Numerisches Material	.13	-.08	.18	.05	.01	**.53****	.03
KKT - Verbales Material	.02	.05	.14	.06	**.25***	**.43****	-.04

Anmerkungen. Korrelation zwischen WIT-2 und Realschulnoten-Gesamt n = 66; Realschule-Mathe n = 72; Realschule-Physik n = 75; Abitur-Gesamt n = 91; Abitur-Mathe n = 85; Abitur-Physik n = 43; Bachelor n = 76; Korrelationen zwischen KKT und Realschulnoten-Gesamt n = 66; Realschule-Mathe n = 72; Realschule-Physik n = 75; Abitur-Gesamt n = 91; Abitur-Mathe n = 85; Abitur-Physik n = 43; Bachelor n = 76; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

Von den Dimensionen kognitiver Leistung (die sich aus Untertests zusammensetzen) waren folgende von guter Vorhersagekraft hinsichtlich der Abitur-Physiknote: Schlussfolgerndes Denken (r = .28*), Sprachliches Denken (r = .27*) und Rechnerisches Denken (r = .22). Die Dimensionen Merkfähigkeit und Räumliches Denken waren keine guten Prädiktoren für Schulleistungen. Interessanterweise zeigte sich der KKT ebenfalls prädiktiv hinsichtlich der Physiknote: Die KKT-Gesamtleistung korrelierte gut mit der Physiknote im Abitur (r = .51**), aber auch die Untertests (figural, numerisch, verbal) waren sehr prädiktiv (r = .34*, r = .53**, r = .43**), relativ zu den Erwartungen, die man aufgrund der eingeschränkten Stichprobenvarianz haben kann.

Regressionsmodell. Über alle Studien hinweg zeigten sich Rechenaufgaben als bedeutsame Aufgabenform zur statistischen Vorhersage von schulischen und akademischen Leistungen. Die Zusammenhänge zeigten sich unabhängig von der spe-

zifischen Aufgabenform und der damit verbundenen Variation in den kritischen Parametern Itemschwierigkeit, Testlänge und Aufgabenformat. Auch in Studie 5 waren Rechenaufgaben wieder besonders prädiktiv, was sowohl für den WIT-2-Untertest „Grundrechnen“ gilt als auch für den KKT-Untertest „numerisches Aufgabenmaterial“, bei dem es sich um Kettenaufgaben handelt, die progressiv in ihrer Itemschwierigkeit ansteigen. Auch in der hier berichteten Studie 5 soll erneut geprüft werden, ob sich die Kombination Intelligenz und Rechenaufgaben erneut bewähren kann. Da der WIT-2 keine theoretisch interpretierbare Option für den Gesamtwert hat, wurde die Leistungsdimension „Schlussfolgerndes Denken“ als Intelligenz definiert (sie entspricht dem Konstrukt, welches auch im WMT und im Raven-Test erfasst wird). Es wurde eine mehrstufige multiple Regressionsanalyse durchgeführt, in der WIT-2 Schlussfolgerndes Denken (Intelligenz) als erster Prädiktor einging, und die Leistung im KKT-Rechentest (numerisches Material) in einem zweiten Schritt hinzugefügt wurde. Als vorherzusagendes Kriterium wurde die Abitur-Physiknote definiert. Die Ergebnisse zeigen, dass die multiple Vorhersage (Intelligenz und Rechenaufgaben) einen inkrementellen Beitrag gegenüber einer Einfachvorhersage bringt (von $R = .28$ auf $R = .53$). Dabei war die Änderung in F signifikant. Details zur Regressionsanalyse finden sich in Tabelle A.23 im Anhang.

5.5.3 Diskussion

Als ein besonderer Vorzug des Wilde-Intelligenz-Test 2 (WIT-2) gilt im Allgemeinen, dass dieser eine große Bandbreite an unterschiedlichsten Teilfähigkeiten erfasst, welche in den meisten beruflichen Situationen relevant sind. Die hier vorliegenden Ergebnisse belegen, dass es vor allem die Dimensionen „Schlussfolgerndes Denken“, „Sprachliches Denken“ und „Rechnerisches Denken“ waren, welche sich durch eine hohe prädiktive Validität auszeichnen. Die Dimensionen „Räumliches Denken“

und „Merkfähigkeit“ waren nicht prädiktiv. Bemerkenswert ist, dass sich in Studie 5 lediglich die Physiknote im Abitur gut vorhersagen ließ. Dies trifft auch für die Validität des Komplexen Konzentrationstests (KKT) zu, denn der KKT-Gesamttest als auch alle drei Untertest korrelierten hoch mit der Abitur-Physiknote (r = .34* bis .53**). Es ist schwer zu erklären, warum sich weder die Gesamtnote in Realschule und Abitur vorhersagen ließ, dafür aber die Abitur-Physiknote. Da es sich bei der Stichprobe um die Teilnehmer in einer realen Bewerbersituation handelte, welche an der Hochschule Mittweida stattfand, sollte klar sein, dass die Stichprobenvarianz eingeschränkt ist. Dieses Problem einer eingeschränkten Stichprobenvarianz lässt sich kaum vermeiden, wenn man das Ziel verfolgt, die Validität von Testleistungen in realen Situationen eines Auswahlverfahrens zu erfassen, welche auch die realen Konsequenzen für die Bewerber mit sich bringt. Gemäß Kersting, Althoff und Jäger (2008, S. 12) ist ein genereller Nachteil dieser Methode, dass die Daten mit einer Streubreiteneinschränkung behaftet sind. Wir müssen daher auch hier davon ausgehen, dass die tatsächliche Vorhersagekraft des WIT-2 unterschätzt wird. Daher ist es umso erstaunlicher, dass der KKT in dieser Stichprobe so hohe Korrelationen zur Abitur-Physiknote aufzeigte, welche auch die gefundenen Korrelationen in der Studie 4 übertrafen. In jedem Fall zeigte sich wiederholt, dass die Kombination von Intelligenz und Konzentration die prädiktive Validität einer diagnostischen Strategie verbessern kann, weshalb dem Diagnostiker im schulpsychologischen oder personalpsychologischen Bereich empfohlen wird, schlussfolgerndes Denken und Rechenkonzentration in testökonomischer Weise zu erfassen.

6 Diskussion

Was ist letztlich eine gute Teststrategie bei der diagnostischen Erfassung von Konzentrationsleistungen? Das war die grundlegende Frage, welche in diesem Sinne auch die Motivationsgrundlage der vorliegenden Arbeit darstellt (Westhoff, 1995). An allerhöchster Stelle stand dabei natürlich die Frage nach dem praktischen Beitrag einer optimierten Testdiagnostik in (ganz) natürlichen Situationen des Berufsalltags, in denen Personen aus Gründen der Bewerberüberzahl oder bestimmter benötigter Fähigkeiten einer Auslese- oder Entscheidungssituation unterzogen werden. Hier kann es für den praktizierenden Personalpsychologen durchaus hilfreich sein, wenn hinsichtlich Auswahl und Anwendung von Testverfahren sowohl Empfehlungen als auch obligatorische Verfahrensregeln gegeben werden können, um eine optimale Diagnostik zu gewährleisten. Um aber über bekanntes testtheoretisches Allgemeinwissen hinaus gehen zu können, sind experimentelle Analysen erforderlich, in denen die Effekte verschiedener kritischer Variablen wie Testlänge, Itemschwierigkeit und Itemhomogenität systematisch auf ihre prädiktive Validität hin untersucht werden. Dies war das operative Ziel der vorliegenden Dissertationsschrift. In insgesamt fünf experimentellen Studien konnte gezeigt werden, dass die Anwendung von Rechen- bzw. Durchstreichkonzentrationstests als Bestandteil einer diagnostischen Teststrategie einen inkrementellen Beitrag zur Vorhersage von relevanten Außenkriterien (Schulerfolg) leistet, welche über die allein durch einen Intelligenztest erreichte Vorhersage hinausgeht. Hinsichtlich der kritischen experimentellen Variablen zeigte sich

dabei, dass die Variablen Testverlauf, Testlänge und Itemschwierigkeit eine wichtige Rolle spielen.

Machen Testlänge und Testverlauf einen Konzentrationstest reliabel und valide? Aus der klassischen Testtheorie lässt sich ableiten, dass sich die Reliabilität (und damit verbunden auch die Validität) eines Konzentrationstests erhöhen lässt, wenn der entsprechende Test verlängert wird (Lienert & Raatz, 1998). Dies sollte deshalb geschehen, weil mit Erhöhung der Itemmenge auch die Messgenauigkeit erhöht wird. Ist ein Test genügend lang (und damit genügend reliabel), so sollten Reliabilität und Validität auch im Testverlauf ansteigen, weil das zu messende Merkmal (ausdauernde Konzentration) mit zunehmender Testermüdung stärker zwischen den Probanden differenziert werden kann (Hagemeister, 1994). Die empirischen Ergebnisse der Studie 2 zeigen jedoch, dass dies nicht der Fall ist: Die Kurzversion eines Rechenkonzentrationstests war sowohl reliabler als auch valider (hinsichtlich der Vorhersage von Schulleistungen) als derselbe Test in einer Langversion. Außerdem zeigte sich, dass die prädiktive Validität in der Langversion am besten im ersten Testabschnitt ist, dann aber über die weiteren Testabschnitte schlechter wird. Dies zeigt, dass bereits ein relativ kurzer Test geeignet ist, um Konzentration ökonomisch zu erfassen und dass Ermüdungseffekte sich eher nachteilig auf die Validität auswirken. Führt eine Erhöhung der Itemschwierigkeit zu einer besseren prädiktiven Validität? Dies würde sich aus Modellen vorhersagen lassen, welche annehmen, dass sich Probanden mit zunehmender Schwierigkeit der Aufgabe besser differenzieren lassen. Auch dies war nicht der Fall. Die Ergebnisse der Studie 1 zeigen, dass ein Rechenkonzentrationstest mit leichten Items etwas valider ist als eine äquivalente Version mit schwierigeren Items. Diese Ergebnisse zeigen, dass die Befundlage komplexer ist, als man bislang glaubte und dass empirische Grundlagenforschung auch im Bereich der Validierungsforschung notwendig ist.

In der Arbeit wurde das Prinzip der inkrementellen Validitätsprüfung durchweg verfolgt. Das bedeutet, dass Intelligenz als Gold-Standard verwendet wurde, anhand derer die inkrementelle Validität von Konzentrationsleistungen beurteilt wurde. Die allgemeine Strategie bestand darin, dass Intelligenz als Basisprädiktor verwendet wurde und die Konzentrationstests als zusätzliche Prädiktoren in einem inkrementellen Regressionsmodell hinzugefügt wurden. Es ging im Prinzip um die folgende Frage: Sollte man für eine praktische diagnostische Strategie lieber eine ganze Testbatterie verwenden oder einen einzelnen kritischen Test? Eine weitere wichtige Frage der Testdiagnostik ist dementsprechend die Frage nach der optimalen Diversifikation von Untertests innerhalb von Testbatterien (Schmidt-Atzert et al., 2006; Westhoff & Hagemeister, 2005). Grundsätzlich gilt, dass eine Testbatterie (mit vielen verschiedenen Untertests) immer besser sein sollte als ein einzelner Test, da die Wahrscheinlichkeit hoch ist, dass genau das gesuchte kritische Merkmal (z. B. der ultimative Test, welcher Schulerfolg bestens vorherzusagen imstande ist) durch einen der Untertest erfasst wird. Wenn man jedoch weiß, dass es ein ganz kritisches kognitives Fähigkeitskonstrukt gibt, das optimal trennscharf ist hinsichtlich der Vorhersage eines bestimmten Kriteriums, dann wäre es wohl die bessere (weil ökonomischere) Strategie, wenn man gezielt nur diesen kritischen Test für eine diagnostische Strategie verwendet. In Studie 3 wurde der Wiener Matrizen-Test (WMT, induktives Denken) als auch eine ganze Testbatterie von Intelligenzaufgaben (I-S-T 2000 R) auf ihre prädiktive Validität hin untersucht. Die Ergebnisse zeigen, dass der Gesamtscore der Testbatterie (I-S-T 2000 R) nur geringfügig valider ist als der WMT (als direktes Maß für „g“), dafür aber ungleich zeitaufwendiger als dieser. Außerdem zeigte sich auch innerhalb der Untertests, dass es gerade das Merkmal induktives Denken ist (Matrizen, Analogien), welches kritisch für die prädiktive Validität zu sein scheint. Tatsächlich scheint es so zu sein, wie Spearman (1927) es bereits behauptet hat, dass gerade jene Aufgaben

besonders valide sind, welche induktives und deduktives Schlussfolgern erfordern, da diese Prozesse an fast allen intellektuell anspruchsvollen Aufgaben des Lebens beteiligt sind.

Obwohl die Validität von Konzentrationstests im Vordergrund stand, war es wichtig, das auch die Eigenschaften von Intelligenztests als Gold-Standard überprüft wurden. Für die praktische Diagnostik kann also geschlussfolgert werden, dass man lieber einen trennscharfen Test „induktiven Denkens“ verwenden sollte, als mit einer ganzen Testbatterie ökonomische Ressourcen auf den Dimensionen Zeit und Geld zu vergeuden. Es ist aber nicht nur eine Frage ökonomischen Testens, sondern eine Frage der optimalen Wahl von Prädiktoren für eine Teststrategie. Die Ergebnisse bestätigen aber eigentlich nur, dass die Matrizenaufgabe, wie sie im Raven-Test (Raven, 1998) und im WMT (Formann & Piswanger, 1979) realisiert ist, einer der besten Prädiktoren überhaupt ist, weshalb sie dem praktizierenden Schul- bzw. Personalpsychologen uneingeschränkt empfohlen werden kann. Eine für die Praxis relevante Problemstellung ist jedoch darüber hinaus auch die Frage, ob die in der grundlagenorientierten Laborforschung verorteten Befunde überhaupt auf die Praxis generalisiert werden können. Tatsächlich gibt es nach ausgiebiger Literaturrecherche keine einzige publizierte Studie, welche die Validität von Testverfahren in einem echten realen Kontext untersucht hat. Natürlich lassen sich zahlreiche Studien finden, welche die Testverfahren an einer entsprechenden Zielgruppe (z. B. Auszubildenden, Bankangestellte, etc.) untersucht haben und dabei (gleichzeitig oder später) auch die Vorhersagekraft hinsichtlich eines berufsrelevanten Kriteriums untersucht haben. Da die Probanden nicht in einer echten Auswahlsituation standen, in der sie mit einer Konsequenz ihrer Testleistung rechnen müssen, kann die Validität in einer echten Bewerbersituation gar nicht beurteilt werden. Es sollte daher ein besonderes Ziel dieser Dissertationsschrift sein, sich dieser Frage zu widmen. In den Studien 4 und 5 wurde die prädiktive

Validität von Testleistungen in realen Situationen untersucht, in denen die Probanden innerhalb einer realen Eignungsfeststellung (Auswahl für einen Studienplatz in einem Masterstudiengang) getestet wurden.

Die praktische Validität eines jeden Testverfahrens würde es erfordern, dass sich die Gütekriterien im psychologischen Labor und in der Lebenspraxis nicht unterscheiden, sonst könnte man die Ergebnisse aller Validierungsstudien von Testverfahren überhaupt nicht mehr interpretieren. Idealerweise sollte die Validität eines Leistungstests im Experimentallabor als auch im Feld (d. h. in realen Kontexten) nicht nennenswert unterschiedlich sein. Eine solche Fragestellung erfordert einen direkten Vergleich der prädiktiven Validität einer diagnostischen Strategie in jeweils realer (lebensechtes Personalauswahlverfahren) versus artifizieller (Laborstudie) Situation. Zur Beantwortung dieser Frage wurden daher die Studien 4 und 5 durchgeführt, deren Korrelationen zu Schulnoten dann verglichen wurden. Während die Probanden der Studie 3 im Experimentallabor getestet wurden, so wurden die Probanden der Studie 4 innerhalb eines echten Auswahlverfahrens getestet. Das Auswahlverfahren fand an der Hochschule Mittweida statt. Die Studie 5 diente noch einmal der Generalisierung von Studie 4 mittels eines anderen psychometrischen Leistungstestverfahrens, dem Wilde-Intelligenz-Test (WIT-2). Zusätzlich zu Testleistungen wurden auch Selbstauskünfte über Aufmerksamkeits- und Konzentrationsfähigkeit der Probanden als weiteres Kriterium erhoben. Wie in der Studie von Forster und Lavie (2007) kam auch hier der CFQ (Broadbent et al., 1982) zum Einsatz, welcher als Instrument zur Messung von Aufmerksamkeitslapsus und Fehlertendenzen in alltäglichen Situationen allgemein anerkannt ist. Auch der SIML (Reason, 1993) kam zum Einsatz, welcher ein ähnliches Konstrukt misst. Da sich allerdings überhaupt keine beachtenswerten Korrelationen zwischen Selbstbericht und Schulnoten fanden, werden die Ergebnisse nicht weiter erläutert. Sie können vom interessierten Leser im Anhang in den Tabellen

A.6, A.7 und A.8 jedoch eingesehen und entsprechend bewertet werden.

Die vorliegende Arbeit enthält auch empirische Daten, die hinsichtlich der Frage nach der Wiederholbarkeit publizierter Befunde zur Validität von Leistungstests relevant sind. In der Forschergemeinde finden Bemühungen im Allgemeinen nur wenig Anerkennung, wenn sie sich lediglich auf die Replikation bereits vorhandenen Wissens beziehen. Dennoch haben die jüngsten Debatten über die hohe Anzahl fehlgeschlagener Replikationsversuche zu einem allgemeinen Umdenken geführt. Erwähnenswert sind zunächst die Korrelationen zwischen den handelsüblichen Aufmerksamkeits- und Konzentrationstests und Schulleistungen. So waren die Korrelationen zwischen schulischer Gesamtnote und dem Test d2 (Leistungsmenge) sowohl für Studie 1 im Test als auch im Retest valide. Das konnte in Studie 2 allerdings nicht in gleicher Höhe bestätigt werden. Dies zeigt, dass selbst klassische Testverfahren die relevanten Kriterien nur in moderater Höhe vorhersagen können. Auch in Studie 3 zeigten sich unbefriedigende Zusammenhänge zwischen sämtlichen handelsüblichen Testverfahren und Schulnoten. Weder waren der FAKT-II noch der Aufmerksamkeitstest d2 gute Prädiktoren von Schulleistungen. Verglichen zu den anderen handelsüblichen Testverfahren zeigte sich die Leistung im WMT als über alle Studien hinweg brauchbarer Prädiktor von Schulleistungen. Dies ist ein durchaus interessanter Befund, da belegt werden kann, dass Matrizentests im Allgemeinen und der WMT im Speziellen eine gewisse Robustheit gegenüber Stichprobenunterschieden besitzen.

Da motivationale Variablen viel stärker Einfluss auf Konzentrationsleistungen ausüben als auf Intelligenz, wurde diese Selbstberichtvariable in Studie 3 explizit mit aufgenommen. Konkret sind es die Konstrukte Persistenz, Selbstkontrolle und Selbstregulation (Mojzisch & Schulz-Hardt, 2007; Humphreys & Revelle, 1984). Erwähnenswert hinsichtlich bekannter Testverfahren sind schließlich noch die Ergebnisse der Validität von Dimensionen der Leistungsmotivation (Schuler & Prochaska, 2001),

welche in der Studie 3 untersucht wurden. Obwohl sie nicht zentraler Gegenstand der hier vorliegenden Arbeit sind, sollen die Ergebnisse dem interessierten Leser nicht vorenthalten werden. Es wurde diesbezüglich von zahlreichen Autoren vermutet, dass sich die Determinanten von schulischem bzw. akademischem Erfolg von der Kindheit bis zum Erwachsenenalter grundlegend ändern: Bei Kindern ist allgemeine Intelligenz die wichtigste Determinante schulischen Erfolgs, bei Jugendlichen ist ebenfalls Intelligenz von großer Bedeutung, aber selbstregulatorische Variablen wie Interesse und Motivation spielen eine viel größere Rolle als in der Kindheit. Bei jungen Erwachsenen hingegen hat sich die Erfahrungswelt so weit entfaltet, dass sich neben kognitiven Fähigkeiten eben auch kompensatorische Strategien zur Zielerreichung als weitere Determinante herausgebildet haben. Gemäß einer solchen Sichtweise spielt die allgemeine kognitive Leistungsfähigkeit also immer (über alle Lebensalter hinweg) eine gleich wichtige Rolle, während der relative Einfluss von motivationalen Variablen kontinuierlich ansteigt. Die Ergebnisse der Studie 3 unterstützen genau diese Hypothese. Über alle Skalen hinweg zeigte sich ein Muster, welches einen sich vergrößernden Einfluss von Dimensionen der Leistungsmotivation indiziert. Moderate Korrelationen von r =.15** bis r = .20** mit Schulnoten fanden sich für die Dimensionen Beharrlichkeit, Engagement, Schwierigkeitspräferenz und Selbstkontrolle. Interessanterweise fanden sich keine Zusammenhänge mit status-relevanten Dimensionen wie Dominanz, Furchtlosigkeit, Statusorientierung usw. Man kann also sagen, dass Persistenz und Selbstkontrolle viel wichtigere Prädiktoren sind als Statusorientierung, dass aber allgemeine Intelligenz trotzdem immer ein wichtiger Prädiktor bleibt. Dies ist konsistent mit Studien, welche die Frage nach der Rolle von entweder Motivation oder kognitiver Leistungsfähigkeit als Determinanten schulischen Erfolgs thematisieren. In diesem Sinne konnte mit der hier vorliegenden Arbeit auch ein Beitrag im Bereich der motivationalen Forschung bzw. Motivationsdiagnostik erbracht werden.

6.1 Sind schwierige Tests besser als leichte Tests?

Die Idee, dass eine Anhebung der Itemschwierigkeit zu einer besseren Differenzierung individueller Differenzen führt, ist bereits von zahlreichen Autoren vertreten worden. Empirisch gibt es jedoch kaum Belege für oder gegen diese Vermutung. Westhoff und Kluck (1984) haben zum Beispiel argumentiert, dass die Konzentrationsleistung einer Person eine Funktion dreier Parameter sei, (1) der Schwierigkeit der Reizgrundlage, (2) der Schwierigkeit der Reaktionsanforderungen und (3) dem individuellen Geübtheitsgrad des Probanden. Entsprechend sollte eine Anhebung der Reiz- oder Reaktionsanforderungen sowie eine Verringerung (oder Verhinderung) des Geübtheitsgrades zu einer besseren Differenzierung der Probanden und über diesen Weg zu einer besseren prädiktiven Validität führen. Faktisch bedeutet das für die vorliegende Arbeit, dass die Testgütekriterien ansteigen sollten, wenn ein Test schwieriger wird. Interessanterweise wird diese Hypothese in Studie 1 nicht bestätigt, sondern genau das Gegenteil war der Fall. In Studie 1 wurde eine Stichprobe von 106 Probanden mit verschiedenen Testverfahren getestet, deren Ergebnisse dann zur Vorhersage von Schulnoten (Realschule, Abitur) als relevantes Kriterium verwendet wurden. Die kritische experimentelle Manipulation bestand in einem Vergleich von leichten und schwierigen Items hinsichtlich ihrer prädiktiven Validität in einem Rechenkonzentrationstest, dem SMACT. Entgegen der Hypothese zeigten sich jedoch höhere Zusammenhänge für die leichte Testversion verglichen zur schwierigen Testversion (s. Ergebnisteil zu Studie 1). Es kann also empirisch nicht bestätigt werden, dass schwierigere Items in Konzentrationstests aus psychometrischer Sicht besser geeignet seien als leichte Items.

Beachtenswert sind aber insbesondere die Ergebnisse der Analyse sogenannter Mixkosten der Itemschwierigkeit. Es wurde nämlich nicht nur die Itemschwierigkeit verglichen, sondern auch der Effekt der Separierung versus Mischung leichter und

schwieriger Items (homogene versus heterogene Items). Tatsächlich wurden in Studie 1 also leichte und schwierigere Items in geblockter und gemixter Bedingung verglichen. Dabei wurde offenkundig, dass die prädiktive Validität besser in der (bezüglich der Schwierigkeit) heterogenen Bedingung ist, verglichen zur homogenen Bedingung. Dies belegt, dass es neben der Itemschwierigkeit auch noch darauf ankommt, ob leichte und schwierigere Items in homogenen oder heterogenen Testbedingungen administriert werden. Auf jeden Fall erlauben diese Ergebnisse die Konklusion, dass schwierigere Items nicht unbedingt valider sind als leichte Items, aber dass vor allem bezüglich der Itemschwierigkeit heterogene (versus homogene) Testbedingungen zu hoher prädiktiver Validität führen. Weiterführende Forschung ist hierzu allerdings nötig, um diese Befunde zu bestätigen und zu generalisieren. Wie kann man diesen sogenannten Mixkosteneffekt der Itemschwierigkeit auf die prädiktive Validität erklären? Eine mögliche Erklärung besteht darin, dass die Variabilität von Itemschwierigkeit (versus Homogenität) kognitive Ressourcen beansprucht, welche zu messbaren individuellen Differenzen führt. Diesbezüglich ist auch gezeigt worden, dass es zu Kosten kommt, wenn die zeitliche Rhythmisierung durch zeitlich variable preparatorische Intervalle erschwert wird (Steinborn & Langner, 2011, 2012). Ebenfalls kommt es zu Leistungseinbußen, wenn in einem Reaktionszeittest die Signaleigenschaften der Reize (z. B. Modalität, Intensität) variieren (Steinborn et al. 2008, 2009, 2010).

Um jedoch den genauen Mechanismus zu verstehen, muss man sich die grundlagenorientierte Forschung anschauen. Diese Modelle werden nachfolgend kurz erläutert. Unabhängig von der jeweils spezifischen Erklärung des kognitiven Mechanismus ist es aus diagnostischer Sicht heraus eine empirische Frage, ob bestimmte experimentelle Manipulationen zu einer Erhöhung oder Verringerung der prädiktiven Validität führen. Im vorliegenden Fall zeigte sich, dass es besser ist, leichte und schwierigere Items zu mischen, als diese in separierten Blöcken zu administrieren, mit entspre-

chenden Implikationen für die Praxis. Es wäre eine besondere Herausforderung für zukünftige Studien, die Effekte von Itemschwierigkeit und Testlänge in Interaktion mit energetischen Faktoren (Ermüdung, Schlafentzug, zirkadianer Rhythmus) auf ihre psychometrischen Eigenschaften hin zu untersuchen (Bratzke et al. 2009, 2012a, 2012b).

Aus der Sicht eines klassischen Modells serialer Reaktionsanforderungen, dem Tracking-Modell von Rabbitt (1969), überwachen Personen in selbstbestimmten Arbeitsanforderungen ihre eigene Arbeitsgeschwindigkeit und Genauigkeit kontinuierlich. Ist die Taktung jedoch durch die experimentelle Situation vorgegeben, dann verringert sich entsprechend der notwendige Aufwand für die Selbstüberwachung. In solchen taktgebundenen Situationen hängt es dann davon ab, ob Probanden durch die vorgegebene Geschwindigkeit überfordert oder unterfordert sind. Das supervisorische Aufmerksamkeitssystem (SAS) ist gemäß dieses Modells für die Überwachung verantwortlich und alarmiert den Probanden bei Abweichungen von intern gesetzten Standards. Wird eine Abweichung festgestellt, dann kommt es zur Korrekturmaßnahme, die z. B. in einem langsameren Arbeitsverhalten resultieren könnte. Aus der Sicht dieses Modells wäre eine Mischung von leichten und schwierigen Aufgaben immer mit einer größeren Anforderung an das supervisorische System verbunden, wobei das Modell keine Aussage darüber macht, ob auch individuelle Differenzen dadurch vergrößert würden. Aus den Annahmen des kognitiv-energetischen Modells (Sanders, 1998, Kap. 9) geht hervor, dass eine Erhöhung kognitiver Anforderungen vor allem das Effort-System beansprucht. Streng genommen ist es die subjektive Wahrnehmung von erhöhten Anforderungen. Ob es zu einer Differenzierung von Probanden oder zu einer Nivellierung kommt, geht jedoch auch aus diesem Modell nicht hervor.

Tatsächlich gibt es derzeit keine Modelle, welche klare Aussagen darüber machen, ob und welche Arten von kognitiven Anforderungen zu einer Vergrößerung

individueller Differenzen oder zu einer Verkleinerung derselben führen. Es ist letztlich eine empirische Frage, welches Schwierigkeitsniveau angemessen bzw. optimal ist, um höchstmögliche prädiktive Validität zu gewährleisten. Empirisch gesehen können die Ergebnisse der Studie 1 der vorliegenden Dissertation nur so interpretiert werden, dass die Korrelation zum relevanten Außenkriterium „Schulerfolg“ in der Bedingung mit schwierigen Items nicht besser war als in der Bedingung mit leichten Items eines Rechenkonzentrationstests. Dies spricht zunächst dafür, dass die gewählte Itemschwierigkeit in dem Bereich der hier verwendeten Items schon optimal war. Über die Effekte noch leichterer oder noch viel schwierigerer Items kann nicht viel ausgesagt werden. Spricht dieses Ergebnis dafür, dass leichte Rechenaufgaben eher zur Konstruktion eines Konzentrationstests geeignet wären als schwierigere Rechenaufgaben? Oder lassen die Ergebnisse den Schluss zu, dass die Itemschwierigkeit überhaupt nicht kritisch für die prädiktive Validität ist? Die Annahmen allgemeinpsychologischer Modelle der Konzentrationsleistung, welche in der vorliegenden Arbeit referiert wurden, geben nur wenige Antworten auf diese Frage. Vielmehr sollte zukünftige Forschung einen noch intensiveren Ansatz verfolgen, indem die Itemschwierigkeit nicht nur in zwei Stufen, sondern in mehreren graduell abgestuften Schwierigkeiten administriert wird. Um Generalisierung zu erreichen, wäre es auch angeraten, dass eine solche experimentelle Anordnung auch mehrere Aufgabenformen simultan bei mehreren Schwierigkeitsstufen untersucht.

Westhoff und Graubner (2003) haben einen Ansatz vorgelegt, in dem sie sowohl drei verschiedene Aufgabenformen (verbales, figurales, numerisches Material) als auch fünf Stufen der Itemschwierigkeit im Komplexen Konzentrationstest (KKT) verwendeten. Die Itemschwierigkeit wurde dabei jedoch auf unterschiedliche Weise variiert. Beim Pfeiltest (figural) wurde die Ähnlichkeit von Zielreiz und Distraktor erhöht, so dass es für die Probanden schwieriger wurde, den Zielreiz rein wahrneh-

mungsmäßig zu erkennen. Beim Rechentest (numerisch) wurde die Kettenlänge der Additionsaufgaben erhöht. Hier ergab sich die Schwierigkeit also aus einem ganz anderen Faktor, da nicht so sehr das Erkennen, sondern die Anforderungen an das Arbeitsgedächtnis immer größer wurden. Beim Wortumkehrtest (verbal) wiederum war es die Wortlänge der Wörter, die dem Grundwortschatz entnommen wurden, welche die Schwierigkeit des Tests ausmachte. Dies zeigt, dass es zahlreiche Möglichkeiten gibt, die Schwierigkeit von Items zu manipulieren und die Frage stellt sich entsprechend, ob diese Möglichkeiten einander äquivalent sind. Ein Problem dieser Studie besteht darin, dass die Autoren nur die prädiktive Validität der Untertests angeben, nicht jedoch die Effekte der Itemschwierigkeit in den verschiedenen Untertests. In der vorliegenden Studie jedoch wurde an dieser Stelle angesetzt und die prädiktive Validität separat für leichte und schwierigere Items in einem Rechenkonzentrationstests untersucht. Damit stellt die vorliegende Arbeit einen Erkenntnisfortschritt dar, welcher in weiterführenden Arbeiten ausgebaut werden kann.

Die Tatsache, dass nur eine Aufgabenform (Rechenkonzentrationstest) verwendet wurde, könnte als Limitierung der Studie 1 angesehen werden. Die Wahl kann aber mit theoretischen Argumenten verteidigt werden. Erstens ist es bei vielen Aufgabenvarianten sehr schwierig, bestimmte experimentelle Manipulationen der Itemschwierigkeit vorzunehmen, bei Rechenkonzentrationstests hingegen ist es denkbar einfach. Zweitens ist gezeigt worden, dass eine Manipulation der visuellen Erkennbarkeit von Zielreizen unter Distraktoren in Durchstreichtests nicht zu einer Verbesserung, sondern zu einer Verschlechterung der Validität führt. Eine bekannte Studie ist jene von Förster und Lavie (2007), die gezeigt haben, dass eine Erhöhung der perzeptuellen Belastung zu einer Nivellierung individueller Differenzen führt, welche über diesen Weg dann auch die Validität (Konzentrationsfehler im Alltag) verschlechtert. Ähnliche Befunde werden von Brickenkamp (2002) berichtet, welcher in experimentellen

Studien herausfand, dass die Erhöhung der Zielreiz-Distraktor Ähnlichkeit zu einer Verschlechterung der Gütekriterien führte. Diese frühen Studien haben letztlich zur Entwicklung des Test d2 geführt, welcher heute noch als eines der Standardinstrumente der diagnostischen Aufmerksamkeitsforschung angesehen wird. Wohlgemerkt gibt es bezüglich der Validität des d2 Aufmerksamkeits- Belastungstests auch andere Meinungen, jedoch ist nicht von der Hand zu weisen, dass der Test trotz seiner Kürze (d. h. seiner Testökonomie) beachtliche psychometrische Eigenschaften, bezogen auf das Tempomaß, besitzt. So gibt es kaum einen vergleichbar kurzen Test, welcher Werte der Retest-Reliabilität von $r > .90$ aufweisen kann. Die handelsüblichen Rechenkonzentrationstests können bezüglich dieser Werte kaum mithalten, weder der Revisionstest von Marschner (1971) noch der Konzentrations-Leistungstest von Düker und Lienert (1965) kommen mit ihren Gütekriterien nur annähernd in diese Region. Lediglich das Testprinzip des Serialen Mentalen Additions- und Komparationstests (SMACT) lässt da hoffen, weshalb diese Aufgabe hier gewählt wurde (Steinborn et al., 2008, 2010).

Tatsächlich hat der SMACT einige besondere Merkmale, die gegenüber bisherigen Rechentests sehr vorteilhaft hinsichtlich kritischen experimentellen Manipulationen sind. Auch basiert der SMACT auf einem theoretisch stark motivierten Fundament. Wie bereits in vorherigen Abschnitten erörtert, hat Düker (1930, 1949) ein theoretisches Rahmenwerk zur Konstruktion von Rechenkonzentrationstests vorgelegt. Das Modell thematisiert die serielle Koordination kognitiver Operationen und ist daher nicht auf numerisches Aufgabenmaterial beschränkt, sondern kann theoretisch auch auf verbales oder figurales Material ausgeweitet werden, sofern das jeweilige Aufgabenprinzip entsprechende kognitive Operationen enthält. Das diagnostische Konzept „Konzentration“ wird dabei als die Fähigkeit zur Handlungsregulation zum Zwecke der Leistungsoptimierung aufgefasst, in Dükers (1949) Definition als die „Fähigkeit zur Anspannung zum Zwecke der Koordination von Teilhandlungen“. Die

konstituierenden Elemente sind dabei (1) „Anspannung“, (2) „Koordination“ und (3) „Rückmeldung“, deren effektives Zusammenwirken die bei einer Aufgabe erforderlichen Einzeltätigkeiten (z. B. Auffassen, Behalten, Reproduzieren, Vergleichen, Abstrahieren, Antworten) in optimaler Weise ablaufen lässt. Damit unterscheidet sich Dükers Koordinationskonzept sehr stark vom Kraepelinschen Willenskonzept (Brandstätter, 1995; Kraepelin, 1902). Während bei Kraepelin die Konzentrationsleistung noch durch reine volitionale Fähigkeiten bestimmt wird, so sind es bei Düker selbstregulatorische Fähigkeiten der proaktiven und retroaktiven Handlungskontrolle. In Alltagsworten ausgedrückt ist es bei Kraepelin die Fähigkeit, sich „durchbeißen“ zu können, während man bei Düker „clever“ sein muss, um gut zu sein.

Da der SMACT in prototypischer Weise die Konzeption von Düker repräsentiert und sich darüber hinaus zahlreiche experimentelle Manipulationen anwenden lassen, wurde er in der vorliegenden Arbeit als Testmedium verwendet. Wie bereits ausgeführt, verlangt auch der SMACT die Koordination von Einzeltätigkeiten, da zunächst ein Additionsterm berechnet werden muss, welcher daraufhin zu einer dargebotenen Zahl verglichen werden muss. Genauer gesagt, muss zunächst wahrgenommen und festgestellt werden, auf welcher Seite des Bildschirms sich der Additionsterm befindet (auf der Gegenseite befindet sich der zu vergleichende Zahlenterm). Dann muss der Term berechnet werden. Der Proband muss die berechnete Zahl im Gedächtnis behalten, um sie dann sogleich mit der gegenüberliegenden Zahl vergleichen zu können. Dann muss die Reaktion programmiert und ausgelöst werden. In diesem Sinne ist der SMACT in idealer Weise geeignet, um Koordination gemäß Düker zu messen. Da Düker und Lienert nie eine Variation der Items des KLT vorgenommen hatten, um die Effekte von Testlänge und Itemschwierigkeit auf die Gütekriterien Reliabilität und Validität zu überprüfen, stellt die hier vorgelegte Studie einen gewaltigen Erkenntnisfortschritt dar. Gemäß der theoretischen Annahmen von Düker und Lienert hätte

eine Anhebung der Itemschwierigkeit zu einer Steigerung der prädiktiven Validität führen müssen, was jedoch nicht der Fall war. Dies ist schwer zu erklären, außer dass man annimmt, dass die leichte Testversion des SMACT bereits schwer genug war, um optimale Reliabilität und Validität zu gewährleisten. Der Fakt, dass die Korrelation zum relevanten Kriterium „Schulnoten“ (insbesondere Mathematik) in gleicher Höhe lag wie die Korrelation des WMT (als Indikator für „g“) mit Schulnoten, spricht für diese Erklärung. Die Tatsache jedoch, dass eine gemischte Darbietung leichter und schwieriger Items (gemixte Bedingung) höhere Korrelationen aufwies als eine separate Darbietung (geblockte Bedingung), zeigt weiterhin, dass auch die Fähigkeit der Probanden, die eigene Arbeitsgeschwindigkeit supervisorisch zu überwachen und möglichst optimal zu rhythmisieren, eine koordinative Grundfähigkeit ist, welche hohe prädiktive Validität besitzt.

6.2 Sind lange Tests besser als kurze Tests?

Es ist eine testheoretische Grundannahme, dass die Messgenauigkeit mit zunehmender Anzahl an Messpunkten zunehmen sollte. Konkret bedeutet dies, dass die psychometrischen Gütekriterien Reliabilität und Validität ansteigen sollten, wenn mehr Verhaltenseinheiten in einem Test aggregiert werden. Aus dieser Logik heraus erscheint es trivial, diese Hypothese auch noch empirisch prüfen zu wollen. Es ist daher umso erstaunlicher, dass die Ergebnisse von Studie 2 diesen „Fakt“ nicht empirisch bestätigen. In der Studie wurde eine Stichprobe von 103 Probanden mit einer Reihe von Testverfahren getestet, deren Ergebnisse dann retrograd zur Vorhersage der Schulnoten (Realschule, Abitur) verwendet wurden. Als experimentelle Manipulation wurden jeweils eine Kurz- und eine Langversion des Rechenkonzentrationstests „SMACT“ verwendet, deren prädiktive Validität zum Kriterium „Schulerfolg“ dann verglichen wurde. Im Ergebnis zeigten sich höhere Zusammenhänge für die Kurzver-

sion (Test: r = .34**; Retest: r = .28**) verglichen mit der Langversion (Test: r = .25; Retest: r = .19). Erwähnenswert sind diesbezüglich auch die Ergebnisse der Analyse des Testverlaufs. Dazu wurde die Langversion des SMACT in vier Abschnitte unterteilt, welche dann vergleichend auf den Zusammenhang mit Schulnoten hin analysiert wurden. Die Korrelation zum Kriterium war im ersten Abschnitt am größten, fiel jedoch kontinuierlich über den Testverlauf ab. Dies zeigt, dass auch eine Verlängerung des Testverlaufs eher nachteilige Auswirkungen auf die Validität hat. Diese Ergebnisse erlauben nur jene Schlussfolgerung, dass kurze Tests eine bessere prädiktive Validität besitzen als lange Tests und dass auch eine Analyse des Testverlaufs bei langen Tests keine Verbesserung der Validität bringt.

Es stellt sich die Frage, wie solche Ergebnisse zu interpretieren sind und mit welchem Mechanismus diese erklärt werden können. Aus den Annahmen der klassischen Testtheorie geht bekanntlich hervor, dass sich die Reliabilität als Funktion der Testlänge (d. h. Anzahl der Items im Test) erhöhen sollte. Dies resultiert daher, dass die Schätzung des Mittelwerts der Kennwerte einer Person mit zunehmender Anzahl an Items genauer wird. Da mit einer besseren Messgenauigkeit auch die Grundlagen für eine hohe Validität (d. h. die Korrelation mit Außenkriterien) hergestellt werden, sollte eine Testverlängerung also indirekt auch zu einer Verbesserung der Validität führen. Zumindest sollte sich keine Verschlechterung der Validität aufgrund einer genaueren Messung (Reliabilität) ergeben. Die Tatsache, dass die prädiktive Validität (d. h. die Korrelation der SMACT-Testleistung zu Schulnoten) durch Testverlängerung nicht verbessert, sondern verschlechtert wurde, ist nicht im Einklang mit den theoretischen Vorhersagen der klassischen Testtheorie und muss erklärt werden. Gemäß Lienert und Raatz (1998) wäre es denkbar, dass sich das zugrundeliegende (d. h. durch die Messung zu erfassende) Merkmal durch Testverlängerung selbst ändert. Eine schleichende Änderung der Leistungsfähigkeit oder Leistungsbereitschaft könnte also dann

zu Einbußen in der Validität führen, wenn individuelle Differenzen dadurch nivelliert werden. Würde die gleiche Änderung zu einer Differenzierung von individuellen Differenzen führen, dann sollte genau der gegenteilige Effekt eintreten, eine zusätzliche Verbesserung der Gütekriterien, welche nicht durch testtheoretische Effekte heraus erklärt werden kann.

Es muss an dieser Stelle gesagt werden, dass die Retest-Reliabilität der Kurz- und Langversion des SMACT ungefähr gleich hoch war, während die Korrelation zum Außenkriterium (Validität) in der Kurzversion besser war als in der Langversion. Dies spricht dafür, dass eintretende Ermüdung und Motivationszerfall nicht zwangsläufig zu einer schlechteren Differenzierung der Probanden führt, sondern dass die im Laufe verlängerter Testung hinzu kommenden Leistungsmerkmale nicht kritisch für die Vorhersage von Schulerfolg sind oder gar das kritische Leistungsmerkmal zunehmend maskieren. Die Annahmen der klassischen Testtheorie, insbesondere jene bezüglich des Effektes der Testverlängerung auf die Messgenauigkeit gelten nur dann, wenn sichergestellt ist bzw. plausibel angenommen werden darf, dass sich das zu messende Merkmal nicht über den Testverlauf verändert. Sofern aber eine Veränderung explizit (d. h. in theoretischen Vorhersagen) angenommen wird, ist die Voraussetzung der Merkmalsstabilität nicht mehr gegeben. Dies ist ein Dilemma, das schon von zahlreichen Autoren debattiert wurde, ohne jedoch dass es bislang eine Lösung dafür gibt. Merkmalsstabilität ist konkret dann nicht mehr gegeben, wenn die energetischen Voraussetzungen durch den Akt der Leistungsmessung selbst verändert werden, so dass testbedingte Leistungseinbußen (Ermüdung, Motivationszerfall etc.) oder auch Leistungsverbesserungen (Übung, Teststrategien etc.) eintreten. In den Testvarianten der frühen Konzentrationsforschung wurde Veränderung tatsächlich auch intendiert, um die Willenseigenschaften einer Person zu messen. Das bekannteste Beispiel ist der Pauli-Test (1938), ein 90 Minuten dauernder Rechenkon-

zentrationstest, der Angaben darüber enthält, wie Leistungsänderungen registriert und ausgewertet werden können.

Um Verlaufsmerkmale in der Langversion des SMACT zu erfassen, wurde der Test in Studie 2 einfach in vier Abschnitte geteilt, welche jeweils die gleiche Anzahl an Items enthalten. Das ist eine pragmatische Vorgehensweise, die von vielen Autoren schon allein deshalb empfohlen wird, weil sie die Probleme vermeidet, welche mit spezielleren Verfahren erkauft werden. So ist vorgeschlagen worden, dass man nicht die Anzahl bearbeiteter Items, sondern die Bearbeitungszeit als Grundlage für die Einteilung zugrunde legt. Kraepelin schlug vor, die Leistungsverläufe einer jeden Person individuell zu betrachten und zu interpretieren. Er nahm dabei an, dass verschiedene Phasen des Verlaufs als Indikator für verschiedene Prozesse angesehen werden können, die er als Erwärmungsphase, stabile Leistungsphase und Ermüdungsphase bezeichnete. Gemäß Kraepelin kommt es zunächst zu einer Leistungssteigerung bis eine optimale Arbeitsgeschwindigkeit erreicht ist, welche dann über eine gewisse Zeit aufrechterhalten wird. Gegen Ende der Testung kommt es zu Ermüdung, die durch eine langsame und fehlerhafte Arbeitsweise gekennzeichnet ist. Die Erfassung und Interpretation dieser Leistungsindikatoren kann nur durch einen „erfahrenen" Beobachter erfolgen, welcher imstande ist, Leistungsschwankungen zu interpretieren (Brickenkamp, 1966). Derartige Analysen im Sinne einer klinischen Urteilsbildung sind in der vorliegenden Arbeit nicht intendiert, da zu viele interpretatorische Freiheitsgrade bestehen. Es wäre aber durchaus interessant, innerhalb einer separaten Forschungsfrage Urteilerübereinstimmung und prädiktive Validität von subjektiven Expertenurteilen hinsichtlich Schulleistungen zu untersuchen bzw. die erforderliche Methodik zu erarbeiten, die für eine sachgemäße und einwandfreie Vorgehensweise nötig ist (Flehmig et al., 2007; Westhoff & Kluck, 1984).

Abschließend stellt sich die Frage nach der Generalisierbarkeit der hier vorge-

legten Ergebnisse, indem die prädiktive Validität einer Kurzversion und einer Langversion eines Rechenkonzentrationstests (SMACT) vergleichend untersucht wurde. Es zeigte sich, dass die Kurzversion besser als die Langversion abschnitt, was gegen die Notwendigkeit einer Testverlängerung zum Zwecke der Verbesserung psychometrischer Gütekriterien spricht. Die Frage ist, ob sich diese Ergebnisse auf andere Testvarianten oder andere Aufgabenformen verallgemeinern lassen. Generalisierung von Ergebnissen auf andere Aufgabenformen, Stichproben und Situationen ist ein Grundprinzip wissenschaftlichen Erkenntnisstrebens. Der KKT (Westhoff & Graubner, 2003) ist sogar explizit dafür konstruiert worden, die „generalisierte“ Konzentrationsfähigkeit einer Person unabhängig vom speziellen Aufgabenformat (figural, numerisch, verbal) zu messen. Die Antwort ist nicht einfach. Einerseits ist es eine Limitierung der aktuellen Studie, dass der Vergleich von Kurzversion und Langversion nur bei einer einzigen Variante eines Rechenkonzentrationstest (SMACT) vorgenommen wurde und nicht bei drei oder vier verschiedenen Testvarianten (z. B. Durchstreich- und Sortierverfahren, etc.). Andererseits soll zur Verteidigung der Vorgehensweise gesagt sein, dass ein oder zwei Generalisierungsstudien bei einer so zeitaufwendigen experimentellen Anordnung zwangsläufig auf Kosten der anderen Fragestellungen hätte gehen müssen. Mit anderen Worten, um die Frage der Testverlängerung ausführlicher bearbeiten zu können, hätten die anderen Teilfragestellungen dieser Arbeit ausgelassen werden müssen. Es ist aber sehr wahrscheinlich, dass sich die prädiktive Validität bei anderen Aufgabenformen in gleicher Weise verhalten würde, wenn man Kurz- und Langversion vergleichen würde, sofern es sich ebenfalls um eine Analyse eines sehr langen Tests handelt, welcher die Konzentration über längere Zeit extrem beansprucht. Weitere Studien sind nötig, um den Befund zu erhärten, dass sich die Validität von Konzentrationstests hinsichtlich Schulleistungen bei Testverlängerung verschlechtert.

Wie lassen sich diese Ergebnisse schlussendlich in die vorhandene Theorienland-

schaft einordnen, auf deren Basis die diagnostische Forschung zu Aufmerksamkeit und Konzentration erfolgt? Hier muss man zunächst einmal zwischen drei Arten von Theorien unterscheiden, (1) der grundlagenorientierten, (2) der anwendungsorientierten und (3) der testtheoretischen, die jeweils separat zu debattieren sind. Dies macht wirklich gute diagnostische Forschung am Ende sehr schwierig, da ebendiese vielen Gesichtspunkte alle simultan zu berücksichtigen sind. Kognitionspsychologische Ansätze zielen vornehmlich darauf ab, die zugrundeliegenden Mechanismen zu identifizieren und via geeigneter experimenteller Manipulation zu dissoziieren, ohne dabei Rücksicht auf diagnostische Aspekte (d. h. die Effekte experimenteller Manipulationen auf psychometrische Gütekriterien) zu nehmen (Van Breukelen et al., 1996). Diese Klasse von Theorien ist leicht zu prüfen, da kausale Mechanismen postuliert sind. Anwendungsorientierte Ansätze wiederum sind häufig taxonomischer Art oder bedienen sich der faktorenanalytischen Methode, um die korrelative Struktur von Tests als Repräsentanten verschiedener Konstrukte (Aufmerksamkeit, Konzentration, Arbeitsgedächtnis etc.) zu beschreiben (Schmidt-Atzert et al., 2008). Diese Klasse von Theorien ist sehr schwierig auf ihre Richtigkeit zu prüfen, da die in faktorenanalytischen Verfahren verwendeten Tests oft handelsübliche Verfahren sind, welche zwar für bestimmte Konstrukte (z. B. Aufmerksamkeit etc.) stehen, aber häufig ganz unterschiedlich in Länge und Art der Leistungsregistrierung sind. Miller und Ulrich (2013) führen an, dass man Tests unterschiedlicher Länge und Anzahl an Items gar nicht in einer gemeinsamen Faktorenanalyse analysieren darf, da die Reliabilität dieser unterschiedlich langen Verfahren einen Einfluss auf die Faktorenstruktur hat. Die dritte Klasse (testtheoretischen Modelle) macht Annahmen über Gütekriterien, auf die es in der diagnostischen Forschung ankommt.

Das Problem ist, dass sich inhaltliche Vorhersagen (z. B. dass die Fähigkeit „Aufmerksamkeit“ ein bestimmtes relevantes Kriterium vorhersagt) eigentlich nur aus den

kognitionspsychologischen Modellen ableiten lassen, die reale Prädiktor–Kriterium Relation jedoch oftmals eine testtheoretische Frage ist, bei der es darauf ankommt, ob die kognitionspsychologisch motivierten Prozesse in richtiger (bzw. optimaler) Weise in einem Test implementiert wurden. Die Arbeitsgruppe um Westhoff und Kollegen hat genau diese Schnittstelle bedient, indem der komplizierte Spagat zwischen Grundlagenforschung und angewandt-diagnostischer Forschung versucht wurde. Dies hat zur Entwicklung des Akkumodells der Konzentration geführt (Westhoff, 1995, S. 389), welches sowohl kognitiv–energetische Mechanismen aus Grundlagentheorien als auch angewandt-diagnostische und psychometrische (testtheoretische) Aspekte berücksichtigt. Indem es außerdem Leitfäden und Handanweisungen zur objektiven und subjektiven Erfassung von Konzentrationsfähigkeit in der Schule ableitet, schafft es die Grundlage für das Verständnis von Konzentrationsschwierigkeiten in Feldern der diagnostischen Berufspraxis. Das Akku-Modell berücksichtigt weiterhin auch experimentelle Effekte auf die Leistung (z. B. Itemschwierigkeit, Verlauf, Pausenmanagement, Testlänge etc.), von denen Testlänge und Testverlauf die wohl relevantesten Variablen sind. Aus dem Akku-Modell heraus lässt sich direkt ableiten, dass sich während lang andauernder Testung ein energetisches Defizit aufbaut, das größer ist bei Probanden mit einem schlechteren Akku. Warum fand sich dann keine Verbesserung der Validität über den Verlauf (Langversion des SMACT)? Gemäß Westhoff kann der hier vorliegende Befund so erklärt werden: Die Aufrechterhaltung einer optimalen Fokussierung während der Bearbeitung eines Konzentrationstests wird vor allem durch die Verwendung homogener und sehr leichter Aufgabenformen erschwert, da diese schnell recht langweilig werden und keinen intrinsischen Anreizwert besitzen.

Wenn das Aufgabenmaterial keinen intrinsischen Anreizwert besitzt, müssen sich Probanden die ganze Zeit über endogen energetisieren, was schnell zur Überlastung des volitionalen Systems führt (Akku-Crash). Die in Studie 2 verwendeten Items

waren womöglich nicht so leicht, dass ihre Übungskurven schon an der Asymptote waren, sodass sich Ermüdung und item-spezifische Übungseffekte gegeneinander aufgewogen haben. Das ist ein grundsätzliches Problem bei der Verwendung komplexer Aufgabenformen in Reaktionszeitparadigmen, wie auch Konzentrationstests es sind. Mit zunehmender Übung sinken auch die Anforderungen an die Koordinierungsfähigkeit der Probanden. Gemäß Westhoff (1995) kann man sich den von Düker postulierten und von Krumm et al. (2011) analysierten Koordinationsmechanismus wie einen Akku in einer Kamera vorstellen. Dieser versorgt die programmierten Aktionsmuster der Kamera mit Energie. Das ist der energetische Aspekt, auf den auch schon Bäumler (1967) hingewiesen hat. Diese Versorgung mit Energie ermöglicht es, dass die Aktionsmuster der Kamera effizient ablaufen können. Wie bereits erörtert, beschreibt Westhoff (1995, S. 389) diesen Mechanismus der Koordination als absichtsvolle und bewusste Handlung, welche (1) Aktionsmuster mit Energie versorgt, (2) ihren seriellen Ablauf proaktiv plant und programmiert sowie (3) die ordentliche Ausführung dabei kontinuierlich überwacht. Gemäß Westhoff wird die Leistung der Probanden von sowohl inneren (Befinden, Motivation etc.) und äußeren Gegebenheiten (Umgebungsgeräusche, Ablenkungen etc.) moduliert. Testlänge und Testverlauf sind demgemäß zwar wichtige situative Determinanten der Konzentrationsleistung, tatsächlich steigt die Messgenauigkeit eines Tests jedoch nur dann zwingend mit der Testlänge an, wenn man annimmt, dass ein zu messendes Merkmal über die Zeit konstant bleibt.

Womöglich führt jede Veränderung durch Ermüdung eher zur Nivellierung als zu einer Differenzierung individueller Differenzen, womit sich der Befund in Studie 2 erklären ließe, dass sich die Validität mit zunehmender Ermüdung nicht verbesserte, sondern verschlechterte. Welche theoretischen Implikationen ergeben sich daraus? Zunächst einmal zeigt es, dass die Analyse von Testlänge und Testverlauf ein grundsätzliches Dilemma darstellt, welches sich nicht auf die aktuelle Studie 2

beschränkt. Da eine Analyse des Testverlaufs explizit auf Veränderung abzielt und danach fragt, ob die experimentell evozierte Veränderung zu einer Verbesserung der Validität führt, kann einerseits folgerichtig geschlussfolgert werden, dass dies in Studie 2 nicht der Fall war. Eintretende Testermüdung führte zu einer Verschlechterung der Validität. Andererseits gibt es noch weitere Prozesse, welche sich ebenfalls verändert haben könnten und welche die Ergebnisse mehr oder weniger stark beeinflusst haben könnten. Einigen Autoren zufolge ist es immer ungewiss und schwer experimentell zu kontrollieren, ob und inwiefern sich Übungseffekte und Ermüdungseffekte gegenseitig beeinflusst haben könnten. Aus diagnostischer Sicht ist es jedoch eine empirische Frage, ob sich die psychometrischen Gütekriterien eines Tests verbessern oder verschlechtern, wenn ein Test mit der Intention verlängert wird, Ermüdungseffekte zu untersuchen. Dies ist in Studie 2 geschehen. Um Übungseffekte zu minimieren, wurde dabei ein Set an sehr vielen Items verwendet. Die praktischen Implikationen liegen jedoch klar auf der Hand. Ein kurzer Test von ca. 10 Minuten Dauer ist, wie die Daten implizieren, einem längeren Test eindeutig vorzuziehen, was auch aus ökonomischen Gründen angeraten ist.

6.3 Sind Testbatterien besser als Einzeltests?

Wenn man bereits weiß, welche kognitive Fähigkeit ein bestimmtes Kriterium (z. B. Schulleistungen) bestmöglich vorhersagen kann, ist es immer angeraten, genau dieses Fähigkeitskonstrukt auch in einem Test zu erfassen. Die diagnostische Strategie in einer realen Auswahlsituation wäre denkbar einfach. Man verwendet einfach genau jenes Testverfahren, welches diese kognitive Fähigkeit erfasst. Wenn man aber nicht genau weiß, welches der vielen spezifischen kognitiven Fähigkeitskonstrukte der trennschärfste Prädiktor ist, dann hat man gar keine andere Wahl, als eine ganze Testbatterie zu verwenden, welche die nötige Breite an erfassbaren Konstrukten bietet.

Da die diagnostische Forschung schon eine fast 100-jährige Tradition hat, hat sich bis zum heutigen Tage eine große Fülle von Studienergebnissen angesammelt, in denen die verschiedensten Varianten kognitiver Konstrukte und ihrer spezifischen Operationalisierungen in Relation zu einer Vielzahl von relevanten Außenkriterien gesetzt wurden. Wir wissen heute sehr gut, dass die allgemeine kognitive Leistungsfähigkeit, oft einfach als Intelligenz bezeichnet, der beste Prädiktor für Leistungen, die unter anderem Intelligenz verlangen, ist. Zahlreiche Autoren haben diesbezüglich geäußert, dass man in einer praktischen Auswahlsituation eigentlich gar nicht lange über diagnostische Strategien nachdenken muss, da es ausreichend sei, einfach allgemeine kognitive Leistungsfähigkeit zu erfassen. Andere Instrumente zur Erfassung spezifischer Fähigkeiten (z. B. Aufmerksamkeits- und Konzentrationstests) oder gar Leistungsmotivation seien nicht nötig, da nur selten ein inkrementeller Beitrag zu erwarten sei. Diese Auffassung hat übrigens auch schon Spearman (1927) vertreten. Sofern eine diagnostische Auswahl von Personen für ganz normale Berufsfelder vorgenommen werden soll, in denen der „gesunde Menschenverstand“ gefordert ist, ist es geboten, die Fähigkeit zur Deduktion und zur Induktion zu erfassen, also schlussfolgerndes Denken.

Gemäß Spearman (1927) ist schlussfolgerndes Denken diejenige Fähigkeit, welche überall dort beteiligt ist, wo „gesunder Menschenverstand“ gefordert ist. Der von seinem Schüler John Raven entwickelte Progressive Matrizentest (Raven, 1998) gilt dafür als prototypische Aufgabensituation. Trotz zahlreicher Nachweise hoher Validität für eine ganz breite Palette von Kriterien gibt es auch Autoren, welche den Standpunkt vertreten, dass man nicht nur einen Indikator für Intelligenz messen sollte, sondern besser eine ganze Reihe an spezielleren Fähigkeiten, welche in ihrer Summe auch eine gute psychometrische Schätzung für Intelligenz ergeben würden. Damit ergibt sich die bis heute ungeklärte Forschungsfrage, welche in Studie 3 überprüft wurde: Sollte

eine diagnostische Strategie lieber eine gute Mischung relevanter Aufgabenformen enthalten oder sollte man eine kritische Aufgabenform verwenden, um Personen in bestimmten Situationen optimal auswählen zu können. Diese Frage ist nicht einfach zu beantworten und auch nicht einfach empirisch zu prüfen, da es natürlich immer darauf ankommt, welche Kriterien vorhergesagt werden sollen. In der vorliegenden Arbeit wurde durchgängig schulische Leistung als Kriterium verwendet, um Vergleichbarkeit über die Studien 1 bis 5 hinweg zu gewährleisten. Konkret wurde in Studie 3 verglichen, ob die prädiktive Validität bezüglich Schulnoten besser für eine Testbatterie (welche eine ganze Reihe von Tests enthält) oder für einen einzelnen Matrizentest (welcher schlussfolgerndes Denken erfasst) ist. Als Testbatterie diente der I-S-T 2000 R und als progressiver Matrizentest wurde der Wiener Matrizen-Test verwendet.

Es ist schon allein aus testökonomischen Gesichtspunkten wichtig zu wissen, ob es nötig ist, Probanden einer 90-minütigen Testung zu unterziehen oder ob man die gleiche Vorhersage auch bereits mit einem 20-minütigen Test erreichen kann. Aus inhaltlicher Sicht ist es die diagnostische Grundsatzfrage, ob man lieber diversifiziert oder optimiert testet. Westhoff und Kollegen gehen noch weiter, indem sie argumentieren, dass man auch ein einziges „optimales“ Konstrukt besser mittels einer multimethodalen diagnostischen Strategie erfasst, als nur eine einzige Operationalisierung eines Konstrukts zu verwenden. Gemäß dem Standpunkt einer diversifizierten diagnostischen Strategie ist also eine multimethodale Vorgehensweise angesagt, um den Einfluss von materialspezifischen Erfahrungsunterschieden zu minimieren. Gemäß dem Standpunkt eines optimalen Tests sollte sich Forschung und Praxis darauf konzentrieren, das für ein bestimmtes Kriterium optimale Konstrukt zu finden und dies in seiner optimalen operationalen Umsetzung als Test zu verwenden. Die Ergebnisse weisen beide Testverfahren, den I-S-T 2000 R und den WMT, als gleichermaßen valide aus. Während bei der Vorhersage von schulischen Leistungen der I-S-T 2000 R

ein klein wenig besser ist, so ist der WMT etwas besser bei der Vorhersage akademischer Leistungen, konkret der Leistungen im Vordiplom. Die Korrelationen waren allerdings im Allgemeinen geringer als in den Studien 1 und 2. Die Mathematiknote in der 10. Klasse (r = .17** für I-S-T 2000 R und r = .14** für WMT) und im Abitur (r = .13** für I-S-T 2000 R und r = .14** für WMT) wurde am besten vorhergesagt, relativ zu den anderen Fachnoten. Die akademischen Leistungen (Vordiplom-Note) ließen sich allerdings besser vorhersagen (r = .22** für I-S-T 2000 R und r = .27** für WMT). Es muss jedoch diskutiert werden, warum die Korrelationen im Allgemeinen so niedrig ausfielen und warum sie für die akademischen Leistungen höher waren als für die schulischen Leistungen.

Warum fielen die Korrelationen im Allgemeinen so niedrig aus? Ein sehr wahrscheinlicher Grund für die niedrigen Korrelationen (verglichen zu den Studien 1 und 2) liegt in der geringeren Stichprobenvarianz. Wie im Methodenteil genauer beschrieben, wurde die Stichprobe in den Studien 1 und 2 an der Technischen Universität Dresden erhoben. Die Stichprobe der Studie 3 dagegen wurde an der Hochschule Mittweida erhoben. Obwohl man intuitiv annehmen würde, dass an Fachhochschulen erhobene Daten größere Stichprobenvarianz haben müssten, ist es in der Realität genau umgekehrt. Das hat verschiedene Gründe unter anderem, dass viele Personen erst zur Fachhochschule wechseln, wenn sie merken, dass die Anforderungen der Universität zu groß werden. Warum sind dann aber die Korrelationen von Tests und akademischer Leistung viel höher (und in ihrer Höhe vergleichbar zu den Studien 1 und 2)? Die Literatur zeigt eindeutig, dass Schulnoten trotz der retrograden Validierung die besseren Kriterien sind als akademische Leistungen in zeitlich–konkurrenter Validierung. Dies wird damit erklärt, dass in späteren (post-schulischen) Ausbildungen immer eine fähigkeits- und interessenspezifische Selbstselektion stattfindet, da Personen ihre Ausbildungswege (zu einem gewissen Grade) selbst wählen können. Gemäß einer solchen

Passungshypothese von Person und Situation (d. h. dem gewähltem Ausbildungsweg) hätte das Korrelationsmuster aber umgekehrt ausfallen müssen. Die Korrelationen zu schulischen Leistungen hätten höher sein müssen. Eine plausible Erklärung ist jene, dass die Notengebung an der Fachhochschule eine genügende Differenzierung der Studenten anstrebt, die unabhängig von der Stichprobenvarianz ist. Mit anderen Worten muss ein Dozent seine Studierenden über das Notenspektrum verteilen und tut dies auf die gleiche Weise wie ein Testkonstrukteur, indem er Items selektiert und seine Klausuren trennschärfer gestaltet.

Was sind also die praktischen Implikationen der Studie 3? Die Ergebnisse zeigen ganz klar, dass es ökonomischer ist, einfach einen Matrizentest zu verwenden als eine ganze Testbatterie. Die prädiktive Validität wäre ungefähr vergleichbar. Der Wiener Matrizen-Test benötigt nur 20 Minuten Bearbeitungszeit, ist aber genauso valide wie der Raven Matrizentest, wie in vielen Studien gezeigt wurde. Was den I-S-T 2000 R betrifft, so könnte ein praktizierender Diagnostiker sich die Frage stellen, ob nicht doch auch einige der Untertests separat verwendet werden könnten. Das ist technisch gesehen möglich, da alle Untertests des I-S-T 2000 R separat validiert wurden und die Korrelation zu relevanten Außenkriterien im Handbuch angegeben ist. In Studie 3 fand sich unter anderem eine Korrelation des I-S-T-Untertests Rechenaufgaben zu Mathematik (r = .23** für Realschule und r = .22** für Abitur, r = .30** für Vordiplom). Dies zeigt, dass Rechenaufgaben sehr gut geeignet sind, um die allgemeine kognitive Fähigkeit zu messen, wie es schon Düker und Lienert (1949) postuliert haben. Allerdings wäre es wohl eine bessere diagnostische Strategie, gleich den KLT zu verwenden, anstatt den Untertest einer Testbatterie zu nehmen. Im Großen und Ganzen fielen die Korrelationen zwischen Leistung in den I-S-T-Untertests und Schulnoten eher dürftig aus. Dies indiziert, dass die prädiktive Validität des Gesamtwerts des I-S-T 2000 R allein aufgrund der extremen Aggregation von Testinformationen entsteht und nicht

durch inhaltliche Trennschärfe. Damit ist gemeint, dass die Untertests als Bausteine nicht unbedingt dazu beitragen, jenen Aspekt an kognitiver Leistungsfähigkeit zu erfassen, welcher optimal hinsichtlich der Vorhersage von Schulleistungen ist, sondern eher ein breites Spektrum an Fähigkeiten, die sich nur lose zu einer Gesamtfähigkeit addieren. Möglicherweise war das von den Testautoren auch so intendiert, um eine möglichst hohe Breite an Konstrukten erfassen zu können. Sofern der ökonomische Aspekt jedoch von Bedeutung ist – und das ist er in der beruflichen Praxis - sollte sich weitere Forschung mehr darauf konzentrieren, optimale Prädiktoren für spezifische Kriterien zu entwickeln.

6.4 Sind Tests auch unter „realen“ Bedingungen valide?

Die meisten wissenschaftlichen Untersuchungen zur prädiktiven Validität von psychometrischen Verfahren finden im psychologischen Labor einer Universität statt. Die Stichprobe besteht dabei aus jungen, mehrheitlich weiblichen Menschen mit verglichen zum Bevölkerungsdurchschnitt hohem Bildungsstand. Daraus ergeben sich einige Probleme für die Validität, welche häufig Gegenstand von Debatten sind (Kersting et al., 2008, S. 12). In den Handbüchern der meisten handelsüblichen Testverfahren sind die Stichproben auf den Bevölkerungsdurchschnitt genormt und sind repräsentativer, jedoch handelt es sich bei der Validierung eines Tests durch seine eigenen Testautoren immer um eine Art Produktentwicklungsforschung. Obwohl den meisten Autoren ein ehrliches Motiv nachgesagt werden kann, entspricht diese Vorgehensweise dennoch nicht der Norm, welche zum Beispiel in anderen Produktentwicklungsbereichen angesagt ist. Streng genommen dürfte ein Test nur als valide gelten, wenn seine Gütekriterien durch unabhängige Forschungsinstitute geprüft worden sind. Die schwerwiegendste, wenngleich am wenigsten debattierte Frage ist jedoch jene nach der Validität unter realen Bedingungen. Die meisten Validierungsstudien basieren auf

Probanden, welche freiwillig (für Versuchspersonenstunden oder Geld) an der Studie teilnehmen, ohne dass sie für schlechte Leistungen mit ernsthaften Konsequenzen zu rechnen hätten. Aus diesem Grund ist es nötig, vergleichend zu untersuchen, ob die prädiktive Validität eines bestimmten psychometrischen Verfahrens besser in einer Laborsituation oder in einer realen Auswahlsituation ist (Kersting et al., 2008, S. 12). Intuitiv würde man womöglich erwarten, dass die Teilnehmer einer lebensechten Bewerberauslese immer ihr Bestes geben, während man sich bei den Probanden einer universitären Laborstudie dessen nicht sicher sein kann. Womöglich geben dort einige auch ihr Bestes, aber nicht alle.

In der diagnostischen Forschung spricht man diesbezüglich vom Problem der Mindestschätzung der Leistung eines Probanden in einem psychometrischen Test (Moosbrugger & Goldhammer, 2005, 2006). Konkret besagt dies, dass die kognitive Leistungsmessung in einer Stichprobe nur dann zu einer genügend hohen Messgenauigkeit führt, wenn alle Mitglieder der Stichprobe auch tatsächlich ihre maximal mögliche Leistung gezeigt haben. Damit ergibt sich auch das empirische Problem, was die Motivation der folgenden Studie darstellt: Führt eine Maßnahme unter der alle Probanden sich maximal anstrengen (verglichen zu einer Situation in der diese Maßnahme nicht appliziert wird) zu besserer Leistung oder nicht? Führt dies auch zu einer besseren Differenzierung der Probanden oder nicht? Die Frage ist nicht so trivial, wie sie auf den ersten Blick scheint. Eine Bewerberauswahlsituation hat sehr wahrscheinlich stimulierende Effekte auf die aktuelle Leistung eines Probanden, weshalb eine Verbesserung der Leistung zu erwarten wäre. Andererseits könnte der mit der Situation einhergehende Stress aber auch zu einer Verschlechterung der Leistung führen. Nimmt man an, dass die Bewerbungssituation die Testleistung der Probanden steigert und nicht mindert, wäre die sich daran anschließende Frage jene, ob individuelle Differenzen durch diese Steigerung vergrößert oder verringert werden. Lassen

sich Probanden eines Tests durch die reale Testung (verglichen zur Labortestung) also besser oder schlechter differenzieren? Zur Beantwortung dieser Frage wurde der I-S-T 2000 R in einer realen Bewerberauslesesituation eingesetzt (Amthauer et al., 2001), welche an der Hochschule Mittweida stattfand. Die Testung war dabei ein Bestandteil einer umfangreichen diagnostischen Strategie, in der viele Informationen (Zeugnisse, Lebenslauf, Interview, Leistungstests) ausgewertet und zu einem Gesamturteil integriert wurden. Diese Daten wurden mit den Daten der Studie 3 verglichen, indem der I-S-T 2000 R innerhalb einer ganz normalen Laboruntersuchung eingesetzt wurde.

6.5 Validität unter realen Bedingungen - Generalisierung

Die Ergebnisse der Studie 4, in welcher der I-S-T 2000 R in einer realen Auswahlsituation eingesetzt wurde, sollten nochmals mit einem anderen Verfahren generalisiert werden. So fand eine weitere Untersuchung statt, in welcher eine große Stichprobe in einer realen Testsituation analysiert wurde. Die Fragen in dieser Studie waren den Fragen der Studie 4 sehr ähnlich. Lassen sich Probanden eines Tests durch die reale Testung in einer tatsächlichen Bewerbersituation genügend differenzieren? Zur Beantwortung dieser Frage wurde diesmal der Wilde-Intelligenztest (WIT-2) in einer realen Bewerberauslesesituation eingesetzt, welche erneut an der Hochschule Mittweida stattfand. Wie in der Studie 4 war die Testung ebenfalls ein Bestandteil einer umfangreichen diagnostischen Strategie, in der multiple Informationen (Zeugnisse, Lebenslauf, Interview, Leistungstests) ausgewertet und zu einem Gesamturteil integriert wurden. Während die Ergebnisse der Studie 4 (I-S-T 2000 R, reale Situation) mit den Ergebnissen der Studie 3 (I-S-T 2000 R, Laborsituation) verglichen werden konnten, so gab es keine derartige Vergleichsstichprobe, mit der die Ergebnisse der Studie 5

hätten verglichen werden können. Das schränkt die vergleichende Interpretation zwar ein, jedoch können die Korrelationen zu den relevanten Kriterien unabhängig davon interpretiert werden. Die Fragestellung war also entsprechend so formuliert, ob in der vorliegenden Studie 5 hinreichend hohe Korrelationen beobachtet werden können, die eine Aussage über die prädiktive Validität des WIT-2 in realen Bewerbersituationen zulassen. Darüber hinaus wurde in Studie 5 auch die Konzentrationsfähigkeit mit dem Komplexen Konzentrationstest (KKT) erfasst, der von Westhoff und Graubner (2003) publiziert wurde.

Ein Vergleich einer großen Intelligenzbatterie mit einem relativ ökonomischen Verfahren zu Messung von Konzentrationsfähigkeit war somit möglich. Im Ergebnis zeigten sich überhaupt keine signifikanten Korrelationen zwischen dem WIT-2 (Kersting et al., 2008) und Schulleistungen. Dieses Ergebnis ist erstaunlich, da der WIT-2 ja explizit dafür konstruiert wurde, die allgemeine kognitive Leistung in einem praktischen Kontext zu erfassen. Da keiner der Untertest irgendeine prädiktive Validität aufwies, erübrigt sich weitere Diskussion bezüglich der Untertests. Die Ergebnisse stehen überdies im Kontrast zu den im Testhandbuch beschriebenen Werten zur prädiktiven Validität. Tatsächlich wurden Korrelationen zu zahlreichen Kriterien untersucht und es wurden teilweise hohe Korrelationen zu Schulleistungen, sowohl retrograd als auch zeitgleich, berichtet. Die übliche Erklärung für ein solches Ergebnis besteht in der Kritik an der Stichprobe, die in den meisten psychologischen Untersuchungen zu homogen bezüglich kognitiver Leistungsfähigkeit und auch der relevanten Kriterien ist. In der hier vorliegenden Studie wurde eine Stichprobe von Studierenden der Hochschule Mittweida verwendet, welche verglichen mit der Gesamtbevölkerung sicherlich auch zu homogen ist, jedoch womöglich heterogener als eine universitäre studentische Stichprobe, bestehend aus Studierenden des Fachs Psychologie. Unter der Annahme, dass dies an der Stichprobe liegt, sollten sich ebenfalls keine signifikanten

Korrelationen zwischen der Leistung im KKT und diesen Kriterien ergeben. Jedoch zeigten sich dort signifikante Korrelationen in bedeutsamer Höhe, welche den KKT als Instrument höchster prädiktiver Validität ausweisen. Die Gesamtleistung des KKT korrelierte zum Beispiel sehr bedeutsam mit der Physiknote im Abitur (r = .51**), während die KKT-Untertests ebenfalls korrelierten, wenngleich in etwas weniger bedeutsamer Höhe (r = .34* für figural; r = .53** für numerisch; r = .43** für verbal). Der KKT hat also, verglichen zum WIT-2, eine ungleich bessere Vorhersagekraft hinsichtlich schulischer Leistungen. Dies gilt sowohl für den Gesamttest als auch für die drei Untertests, wobei der Rechentest sich als der beste herausstellte.

Aus dem oben genannten Grund kann die fehlende prädiktive Validität des WIT-2 nicht auf Merkmale der Stichprobe zurückgeführt werden, sonst hätte die Leistung im KKT auch nicht mit den relevanten Kriterien korrelieren dürfen. Dies muss betont werden, da es tatsächlich zeigt, dass ein einfacher und ökonomischer Konzentrationstest viel bessere psychometrische Eigenschaften hat als eine zeitaufwendige und unökonomische Testbatterie. Da die Stichprobe in Studie 5 hinsichtlich demographischer Merkmale und dem Grad der Homogenität vergleichbar zu Studie 4 ist, stellt sich die Frage, warum es in Studie 5 keinerlei korrelative Beziehungen zu Noten als relevantem Kriterium gab. Dies impliziert einen direkten Vergleich der prädiktiven Validität von I-S-T 2000 R und WIT-2 in ein und derselben Stichprobe. Angesichts der Testlänge beider Verfahren wäre es ein anspruchsvolles Unterfangen, das über die hier angestrebten Ziele hinausgeht. Möglicherweise sind Aufmerksamkeits- und Konzentrationstests bestens dafür geeignet, um in realen Auswahlsituationen eingesetzt zu werden, in denen Probanden unter Stress bzw. Leistungsdruck stehen. Einige Autoren wie beispielsweise Schmidt-Atzert et al. (2008) kritisieren die Verwendung komplexer Aufgabenformen in Konzentrationstests, weil dadurch eine Vermengung der Konstrukte Aufmerksamkeit/Konzentration und Arbeitsgedächtnis zugelassen wird.

7 Theoretische Implikationen

Am Ende einer jeden empirischen Analyse ist die Frage zu stellen: Welchen Wissenszuwachs haben die durchgeführten Studien erbracht? Was wissen wir nun mehr, was wir vorher nicht wussten? Die vorliegenden fünf Studien zur Validität kognitiver Leistungstests hatten das Ziel, nicht nur den Wissensstand einer kleinen Gemeinde von Spezialisten im Bereich der Testkonstruktion zu erweitern, sondern darüber hinaus auch praktische Empfehlungen an praktizierende Testpsychologen geben zu können. Vordergründig ging es um die Validität von Aufmerksamkeits- und Konzentrationstests als inkrementellem Bestandteil einer multimethodalen Forschungsstrategie. Da Lienert und Raatz (1998) darauf hingewiesen haben, dass es nicht die eine Validität gibt, sondern dass ein Testverfahren immer nur hinsichtlich bestimmter relevanter Außenkriterien valide sein kann, wurde der Geltungsbereich der vorliegenden Studien auf Schulleistungen eingegrenzt. In der Tat sind Schulnoten das am häufigsten verwendete Kriterium bei der Bewertung der prädiktiven Validität von Tests. Obwohl ein umfangreicher Fundus an empirischen Studien bereits nahe legt, dass die handelsüblichen Testverfahren relevante Kriterien (Schulnoten etc.) vorhersagen können, gibt es keine eindeutigen Erkenntnisse darüber, wie man Leistungstests im Allgemeinen konstruieren sollte, damit sie den Anforderungen an Reliabilität, Validität und Testökonomie entsprechen. Diese Erkenntnis bleibt auch so lange verwehrt, so lange Testentwicklung eine Produktforschung ist, welche zum Ziel hat, das eigene Produkt zu bewerben und mittels empirischer Ergebnisse gegenüber konkurrierenden

Verfahren vorteilhaft erscheinen zu lassen. Die Fragen der hier vorliegenden Arbeit setzten jedoch an folgenden übergreifenden Problempunkten an: Wie lang sollte ein Konzentrationstest idealerweise sein? Wie schwierig sollten die Items in Konzentrationstests sein? Wie sollten allgemeine und spezielle diagnostische Konstrukte (z. B. Intelligenz und Konzentration) miteinander kombiniert werden, um optimale Validität zu gewährleisten?

Die übliche Forschungsstrategie zur Erforschung der Validität von Aufmerksamkeitstests besteht normalerweise darin, verschiedene handelsübliche Tests auszuwählen und diese an einer großen Stichprobe von Probanden zu administrieren. Deren Leistungskennwerte werden dann mit den relevanten Kriterien korreliert, wobei schulische Leistung das am häufigsten verwendete Kriterium ist. Der hier gewählte experimentell-korrelative Ansatz wird mit dem Argument verteidigt, dass Korrelationen zwischen handelsüblichen Tests und relevanten Kriterien keine Schlussfolgerungen über die zugrundeliegenden kognitiven Mechanismen zulassen, da oft sowohl gleiche Konstrukte mit unterschiedlichen Aufgabenformen als auch verschiedene Konstrukte mit gleichen Aufgabenformen erfasst werden. Diese Situation ist als unzureichend zu beurteilen. Der methodische Fokus lag also nicht auf der „Anwendung“ von Tests, sondern auf der Analyse kritischer experimenteller Variablen mit Einfluss auf die Validität. Während testtheoretische Modelle eine Verbesserung psychometrischer Gütekriterien vorhersagen, fand sich in der Studie 2 keine entsprechende Verbesserung der Validität im SMACT. Als Erklärung liegt nahe, dass die Messgenauigkeit bereits nach 100 bis 200 Durchgängen gegen die Asymptote (das maximale Niveau der Reliabilität) akkumuliert und daher keine weiteren Verbesserungen zu erwarten sind. Eine Verbesserung der Validität aufgrund von verbesserter Reliabilität ist demzufolge nicht zu erwarten. Bei einer sehr geeigneten Aufgabenform wie der mentalen Addition ist, so zeigen die Ergebnisse der Studie 2, bereits nach 10 Minuten Testzeit ein relativ hohes

Niveau der Messgenauigkeit erreicht, weshalb die Langversion keine Verbesserung gegenüber der Kurzversion erbracht hatte.

Auf jeden Fall sprechen die Ergebnisse der Studie 2 gegen die oft formulierte These, dass sich Reliabilität und Validität erhöhen, wenn Probanden in kritischen Situationen (z. B. unter Stress, Ermüdung etc.) getestet werden. Die dahinter stehende Idee leitet sich aus der alten theoretischen Sichtweise ab, nach welcher die Leistung beim fortdauernden Addieren durch die Willenseigenschaften einer Person determiniert ist, welche bei ständiger Belastung geschwächt werden. Individuelle Differenzen resultieren gemäß einer solchen Sichtweise aus der Fähigkeit, die Willenseigenschaften möglichst lange auf hohem Niveau aufrechterhalten zu können. Eine moderne Version des Willensansatzes findet sich bei Muraven und Baumeister (2000), welche annehmen, dass Wille als eine Art unspezifische Ressource aufgefasst werden kann, welche sich durch fortdauernde Beanspruchung verringert und auch für andere (Transfer-) Aufgaben nicht mehr zur Verfügung steht. Auch das Modell der psychischen Sättigung von Schulz-Hardt (2001) basiert auf einer ähnlichen Idee. Während diese Ansätze hohe theoretische Plausibilität besitzen, haben sie empirisch gesehen eher ein Problem. Sie berücksichtigen keine proaktive Selbstregulation kognitiver Ressourcen, sondern gehen allein von einem reaktiven Modell aus, wonach aufgabenbezogene Anforderung zu Ressourcenverwendung führt, welche bei Überbeanspruchung zu Funktionseinbußen führt. Allerdings hat Westhoff (1995) die Möglichkeit einer proaktiven Regulation in seiner Konzeption des Akku-Modells bereits berücksichtigt, da er davon ausgeht, dass individuelle Differenzen in der Akkuleistung durch vorwegnehmende Strategien kompensiert werden können. Solche proaktiven Regulationsformen können aus dem periodischen Einlegen von Mikropausen bestehen, dem Einlegen einer Arbeitsgeschwindigkeit unterhalb des maximal möglichen Werts oder auch aus einer Minimierung von Schwankungen der Arbeitsgeschwindigkeit während

der Bearbeitung von Konzentrationstests (Westhoff & Hagemeister, 2005; Westhoff & Kluck, 1984).

Auch die These, dass Reliabilität und Validität mit zunehmender Itemschwierigkeit ansteigen sollten, konnte in Studie 1 nicht bestätigt werden. Dafür gibt es allerdings keine einfache Erklärung. Es bleibt ein Rätsel, warum die schwierigere Version des SMACT keine bessere Validität gegenüber der leichteren Version zeigte. Da sowohl die leichte als auch die schwierige Version relativ gute Korrelationen zu Schulnoten aufwiesen, kann man auch hier nur so argumentieren, dass sich beide Itemschwierigkeiten im Bereich der optimalen Validität befunden haben. Zum einen ist das eine Schwäche in der Konzeption von Studie 1, schließlich hätte die Itemschwierigkeit gemäß des Prinzips der Maximierung der Primärvarianz besser kontrastiert werden können. Zum anderen muss zur Verteidigung angeführt werden, dass es keine Anhaltspunkte dafür gab, dass bereits die leichte Version für eine optimale Validität ausreichend schwierig ist. Immerhin unterscheiden sich beide Testvarianten des SMACT hinsichtlich der Reaktionszeit um mehr als 1000 Millisekunden, weshalb davon auszugehen ist, dass es einen relativ breiten Bereich der Itemschwierigkeit gibt, der optimale Korrelationen zu Schulnoten zu erbringen imstande ist. Was die Ergebnisse der Studie 1 auf jeden Fall indizieren, ist, dass die Dinge aus theoretischer Sicht oft recht einfach und klar sind, deren empirische Überprüfung jedoch oft das Problem mit sich bringt, dass zahlreiche modulierende Variablen am Werk sind, die nicht berücksichtigt und oft sogar nicht bekannt sind. Miller und Ulrich (2013) weisen diesbezüglich mit Nachdruck darauf hin, dass selbst die Reliabilität eines kognitiven Tests von einer Vielzahl von Einflussfaktoren abhängt und daher praktisch immer einer empirischen Überprüfung im individuellen Fall bedarf. Die theoretische Implikation ist daher die, dass man sich niemals auf die formelle Vorhersage einer Verbesserung der Messgenauigkeit (und indirekt damit auch der Validität) verlassen sollte, sondern

dass dies im individuellen Fall einer bestimmten Aufgabenform immer empirisch zu überprüfen ist.

Das relevante Kriterium bestand in der vorliegenden Arbeit aus den Schulleistungen der Realschule und des Abiturs. In Studie 3 wurde außerdem, wie in der Studie von Krumm et al. (2012), die Vordiplom-Note erhoben. Dass die Schulleistungen dabei retrograd erhoben wurden, stellt kein Problem dar, da es bei der Testevaluation lediglich auf die vergleichende prädiktive Validität zwischen Tests bzw. kritischen experimentellen Bedingungen (z. B. Testlänge etc.) ankommt. Es sollte klar sein, dass der Terminus „prädiktive Validität“ sich auf die statistische Vorhersage bezieht, welche wiederum eine hypothetische Simulation der Vorhersage von Erfolg durch Testleistungen in realen Kontexten darstellt. Die Methode der retrograden Validierung wird von manchen Autoren kritisiert, jedoch ist dies aus folgenden Gründen unbegründet: In vielen Testhandbüchern (u. a. des I-S-T 2000 R und des WIT-2) wird oft angeführt, dass die Korrelationen bei vergleichend retrograder versus konkurrenter versus anterograder Vorhersage von Schulnoten ungefähr gleich hoch ist. Es ist also mit sehr hoher Wahrscheinlichkeit davon auszugehen, dass die zeitlichen Gegebenheiten keinen großen Unterschied machen. Streng genommen würde eine retrospektive Erhebung des Kriteriums immer eine Unterschätzung (aber nie eine Überschätzung) des wahren Zusammenhangs bedeuten, da sich das kognitive Leistungsmerkmal über die Zeit geändert haben könnte. Potentielle Zusammenhänge würden also als eine Mindestschätzung interpretiert werden müssen. Insgesamt gesehen stehen die vorliegenden Ergebnisse gut im Einklang mit den Annahmen der klassischen Testtheorie, wenn man berücksichtigt, dass die kritischen Variablen Testlänge und Itemschwierigkeit einen bestimmten Optimalbereich haben, welcher jedoch durch das spezifische Aufgabenmaterial bestimmt ist und deren Generalisierung natürlich wiederum einer erneuten empirischen Prüfung im konkreten Fall bedarf.

8 Praktische Konsequenzen

Der Einsatz von geeigneten psychometrischen Testverfahren im Rahmen einer diagnostischen Strategie zur Klassifikation oder Auswahl von Personen in Organisationen erfordert Kenntnisse der verantwortlichen Psychologen und anderer Angehöriger relevanter Berufsgruppen. Aus diesem Grunde wird im Allgemeinen empfohlen, dass nur ausgebildete Personen mit Zusatzausbildung in der DIN 33430 derartige Verfahren zusammenstellen und anwenden sollten. Der operative Gegenstand dieser Norm sind Qualitätskriterien und methodische Standards bei der Anwendung diagnostischer Verfahren zum Zwecke der Beratung, Klassifikation oder Selektion von Personen in beruflichen Domänen. Die in der DIN 33430 festgelegten Empfehlungen sind sehr allgemein gehalten und hinsichtlich zahlreicher Aspekte spezifikationsfähig bzw. –bedürftig. Dies betrifft sowohl die konkrete Zusammenstellung von Testverfahren sowie deren Reihenfolge und Zeitbedarf, als auch deren Gewichtung als Prädiktoren für die Vorhersage des relevanten Kriteriums (z. B. in Bewerbersituationen der Ausbildungs- oder Berufserfolg). Da es zur Messung von kognitiven Fähigkeiten zahlreiche geeignete Verfahren gibt, ist auch die Auswahl des besten Verfahrens eine empirische Frage, auf die in der DIN 33430 keine Antwort gegeben wird. In der vorliegenden Arbeit wurde die Nützlichkeit von Aufmerksamkeits- und Konzentrationstests unter anderem dahingehend untersucht, ob sich einfache oder eher komplexe Items besser eignen, ob kurze oder lange Testformen verwendet werden sollten, ob man lieber eine homogene Testform oder eine ganze Testbatterie heterogener Untertests verwenden sollte und ob

die Validität psychometrischer Verfahren in echten Auswahlsituationen besser ist im Vergleich zur üblichen Messung im psychologischem Experimentallabor. Auf diese Fragen können anhand empirischer Daten klare Antworten formuliert werden, wobei klar sein sollte, dass ein ungleich aufwendigeres Forschungsprogramm notwendig ist, um diese Fragen endgültig beantworten zu können (Flehmig et al., 2007).

Die Ergebnisse der insgesamt fünf Studien legen sehr nahe, dass die Komplexität von Konzentrationstests im mittleren Bereich liegen sollte. Als einfach gelten dabei Items unter einer durchschnittlichen Reaktionszeit von 500 Millisekunden, als mittelschwer gelten Items mit einer Reaktionszeit von 500 bis 2000 Millisekunden, und als schwer gelten Items mit einer Reaktionszeit von 2000 bis 6000 Millisekunden, wobei Kettenaufgaben mit durchschnittlichen Reaktionszeiten über 6000 Millisekunden als superschwer einzuordnen sind (Miller & Ulrich, 2013). Im Bereich der Grundlagenforschung (mentale Chronometrie) bzw. in den häufig verwendeten Reaktionszeitparadigmen in diesem Bereich liegen die mittleren Reaktionszeiten oft in einem Bereich zwischen ungefähr 300 und 900 Millisekunden, je nachdem ob nur Einfachreaktionszeiten oder (mehr oder weniger komplexe) Wahlreaktionszeiten abgefragt werden. In der Forschung zum mentalen Addieren, Subtrahieren und Multiplizieren liegen die Reaktionszeiten deutlich höher. Hinsichtlich der Testlänge indizieren die vorliegenden Ergebnisse, dass eine Testzeit von 15 Minuten ausreichend ist und eine längere Testung keine nennenswerten Verbesserungen in den psychometrischen Gütekriterien Reliabilität und Validität bringt. Im Gegenteil, eine Testverlängerung um die dreifache Testzeit führt sogar zu einer Verschlechterung der prädiktiven Validität, was womöglich durch den Einfluss von Ermüdung hervorgerufen wird. Die vorliegenden Ergebnisse stützen die These also nicht, wonach Ermüdbarkeit ein relevantes Merkmal mit diagnostischer Relevanz ist, zumindest nicht für die Vorhersage von Schulleistungen. Hinsichtlich der Frage nach der Testdiversifikation versus der Verwendung einer homogenen

Aufgabenform zeigen die vorliegenden Ergebnisse, dass psychometrische Intelligenz als Prädiktor von höherer Vorhersagekraft hinsichtlich Schulleistungen ist, wenn ein einschlägiger Matrizentest (Raven Test, Wiener Matrizen-Test) verwendet wird, statt einer ganzen Testbatterie (I-S-T 2000 R, WIT-2).

Die praktische Empfehlung für den tätigen Schulpsychologen oder Bildungsberater lautet demnach ganz eindeutig: Die Verwendung eines Matrizentests zur Erfassung der allgemeinen kognitiven Fähigkeit ist zu empfehlen und ein kurzer Konzentrationstest kann die diagnostische Strategie sinnvoll erweitern, ohne dabei auf Kosten der Testökonomie zu gehen. Der klassische Test d2 Aufmerksamkeits- und Belastungstest bietet sich aufgrund seiner psychometrischen Eigenschaften an. Schorr (1995) hatte in einer Befragung von praktizierenden Diagnostikern herausgefunden, dass der Test d2 zu den am häufigsten eingesetzten testdiagnostischen Verfahren zählt. Obwohl der Test auch vielfach kritisiert wurde, weisen die vorliegenden Ergebnisse das Verfahren als valide im Vergleich zu den anderen Testverfahren aus. Wenn hier von Validität gesprochen wird, ist dies also immer relativ zu der Validität anderer Testverfahren in derselben oder einer vergleichbaren Stichprobe zu sehen. Tatsächlich fanden sich über alle Studien hinweg, wenn überhaupt, nur moderate Korrelationen zwischen Leistungswerten im Test und Schulnoten. Wie bereits erwähnt, liegt das an der eingeschränkten Stichprobe, welche hier verwendet wurde und welche in den meisten psychologischen Studien üblicherweise verwendet wird. Vergleichend ist der Test d2 aber ein gutes Instrument, denn weder der FAKT-II noch spezielle Untertests des I-S-T 2000 R und des WIT-2 (welche Konzentration erfassen) konnten mit dem Test d2 mithalten. Die höchste prädiktive Validität hatte jedoch der SMACT, ein auf Rechnen basierender Konzentrationstest, welcher in mehreren experimentellen Varianten auf seine psychometrischen Eigenschaften hin untersucht wurde. Da der SMACT noch nicht an einer ausgewiesenen Normstichprobe getestet worden ist und auch

handelsüblich noch nicht erhältlich ist, kann das Verfahren bislang noch nicht für den praktisch arbeitenden Diagnostiker empfohlen werden, da der Einsatz dieses Verfahrens Kenntnisse erfordert, welche bei praktizierenden Psychologen zwar erwünscht sind, jedoch nicht zwingend vorausgesetzt werden können.

Ein Verfahren, welches die Testökonomie und die Idee einer multimethodalen Konzentrationsdiagnostik simultan berücksichtigt, ist der Komplexe Konzentrationstest (KKT), welcher von Westhoff und Graubner (2003) entwickelt wurde. Der Test enthält Aufgaben in unterschiedlichem Format (numerisch, figural, verbal), welche in ihrer Itemschwierigkeit in insgesamt fünf Stufen progressiv gestaffelt sind. Westhoff und Graubner (2003) intendierten mit dieser Verfahrensweise eine praktische Anwendungsbreite. Schmidt-Atzert et al. (2006) und Schmidt-Atzert et al. (2008) haben jedoch darauf hingewiesen, dass die Verwendung von schwierigen Aufgaben das Konstrukt Konzentration verändern, indem mit zunehmender Schwierigkeit Anteile vom Konstrukt Arbeitsgedächtnis erfasst werden. Cronbach (1975) hat darauf hingewiesen, dass ein jeder Test nicht um seiner selbst willen konstruiert wird, sondern aufgrund praktischer Erfordernisse in beruflichen Feldern der psychologischen Diagnostik. Damit ist ein psychometrischer Test immer nur so viel wert, wie er imstande ist, gute Vorhersagen zu machen. Die Optimierung der prädiktiven Validität bezüglich relevanter Kriterien ist also das primäre Ziel jeder Testkonstruktion und die Optimierung der Konstruktvalidität ist eine diesem Ziel untergeordnete Zielsetzung. Ein in der beruflichen Praxis sinnvoll einzusetzendes diagnostisches Konzept besteht folglich auch in der Kombination eines testökonomischen Matrizentests (z. B. WMT) mit dem KKT. Den gegebenen Kenntnisstand nutzend, sollte es das Ziel weiterführender Forschung sein, Testkombinationen herauszufinden, welche zu einer Optimierung der Vorhersage führen und dennoch ökonomisch sind.

Die rasante Entwicklung technischer Möglichkeiten erlaubt außerdem eine direk-

te Verbindung zwischen praktischer Anwendung und technologieorientierter Testentwicklungsforschung. So gibt es neuere Entwicklungen, die z. B. eine psychometrische Erfassung von momentaner Aufmerksamkeit ermöglichen, während Nutzer auf einer Internetseite surfen (z. B. Berendt & Brenstein, 2001). Die von Westhoff (1995) formulierte Vision rückt immer näher, dass praktizierende Diagnostiker und universitäre Forscher zusammenarbeiten und ihre empirisch (im Experimentallabor und im Feld) gewonnenen Daten einander zur Verfügung stellen. Der damit einhergehende Wissenszuwachs würde riesig sein, wenn man testübergreifende Aspekte wie Testschwierigkeit, Testlänge und Testdiversifikation sowohl in laborexperimentellen Situationen als auch in realen Auswahlverfahren vergleichend analysieren könnte. Die vorliegende Arbeit hat einen Schritt in diese Richtung getan, doch ist eine viel umfangreichere Forschungsarbeit notwendig, um die hier berichteten Ergebnisse generalisieren zu können. Um in der Zukunft eine den wissenschaftlichen Standards genügende Forschung leisten zu können, darf Testentwicklung keine Produktforschung mehr sein, muss die Validität von neuen Testverfahren vergleichend zu anderen Testverfahren festgestellt werden, wobei es wichtig ist, dass die vergleichende Analyse der Validität konkurrierender Verfahren von unabhängigen Arbeitsgruppen nachgewiesen wird. Das Ziel einer zukünftigen Testentwicklungsforschung muss überdies in der Aufklärung der allgemeinen Prinzipien bestehen, welche die Gütekriterien Reliabilität und Validität beeinflussen. Es sollte also nicht eine Frage sein, ob der Test d2 oder der FAKT-II ein valides Instrument ist, sondern ob bestimmte Konstruktionsprinzipien und deren Parametereinstellungen (Itemschwierigkeit, Testlänge, Testdiversifikation etc.) zu hohen Gütekriterien führen.

9 Literaturverzeichnis

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

Ashcraft, M. H. (1992). Cognitive arithmetic: a review of data and theory. *Cognition*, *44*, 75-106. doi: 10.1016/0010-0277(92)90051-I

Bäumler, G. (1967). *Statistische, experimentelle and theoretische Beiträge zur Frage der Blockierung bei fortlaufenden Reaktionstätigkeiten*. Würzburg: Julius-Maximilians-Universität Würzburg.

Beauducel, A., & Brocke, B. (1993). Intelligence and speed of information processing: Further results and questions on Hick's paradigm and beyond. *Personality and Individual Differences*, *15*, 627-636. doi: 10.1016/0191-8869(93)90004-M

Berendt, B., & Brenstein, E. (2001). Visualizing individual differences in Web navigation: STRATDYN, a tool for analyzing navigation patterns. *Behavior Research Methods Instruments & Computers*, *33*(2), 243-257. doi: 10.3758/BF03195371

Bertelson, P., & Jofefe, R. (1963). Blockings in prolonged serial responding. *Ergonomics*, *6*, 109-116. doi: 10.1080/00140136308930682

Bills, A. G. (1931). Blocking: A new principle of mental fatigue. *American Journal of Psychology*, *43*, 230-245. doi: 10.2307/1414771

Bills, A. G. (1937). Facilitation and inhibition in mental work. *Psychological Bulletin*, *34*, 286-309. doi: 10.1037/h0055788

Bourdon, B. (1895). Observations comparatives sur la reconnaisance, la discrimination

et l'assoication. *Revue Philosophique*, *40*, 153-185.

Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5., vollst. überarb. u. aktualis. Aufl.). Berlin: Springer.

Brandstätter, H. (1995). Die Arbeitskurve nach Kraepelin-Pauli - doch ein Willenstest? *Zeitschrift für Arbeits- und Organisationspsychologie*, *39*, 54-66.

Bratzke, D., Rolke, B., Steinborn, M. B., & Ulrich, R. (2009). The effect of 40 hours of constant wakefulness on task switching efficiency. *Journal of Sleep Research (18)*, 167-172. doi: 10.1111/j.1365-2869.2008.00729.x

Bratzke, D., Steinborn, M. B., Rolke, B., & Ulrich, R. (2012a). Effects of sleep loss and circadian rhythm on executive inhibitory control in the Stroop and Simon tasks. *Chronobiology International*, *29*(1), 55-61. doi: 10.3109/07420528.2011.635235

Bratzke, D., Steinborn, M. B., Rolke, B., & Ulrich, R. (2012b). Reply to: Testing of executive function. *Chronobiology International*, *29*(1), 1285-1285. doi: 10.3109/07420528.2012.728412

Bresin, K., Robinson, M. D., Ode, S., & Leth-Steensen, C. (2011). Driven, distracted, or both? A performance-based ex-Gaussian analysis of individual differences in anxiety. *Journal of Personality*, *79*, 875-904. doi: 10.1111/j.1467-6494.2010.00709.x

Brickenkamp, R. (1966). Anmerkungen zur Interpretierbarkeit von Leistungsschwankungen im Test d2. *Diagnostica*, *19*, 125-131.

Brickenkamp, R. (2002). *Test d2. Aufmerksamkeits-Belastungs-Test (9. Auflage)*. Göttingen: Hogrefe.

Brickenkamp, R., Merten, T., & Hänsgen, K. D. (1996). *d2-C: Aufmerksamkeits-Belastungs-Test (32BIT)*. Göttingen: Hogrefe.

Broadbent, D. E., Cooper, P. F., Fitzgerald, P., & Parkes, K. R. (1982). The Cognitive Failures Questionnaire (CFQ) and its correlates. *British Journal of Clinical Psychology*, *21*, 1-16. doi: 10.1111/j.2044-8260.1982.tb01421.x

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105. doi: 10.1037/h0046016

Colom, R., Escorial, S., Shih, P. C., & Privado, J. (2007). Fluid intelligence, memory span, and temperament difficulties predict academic performance of young adolescents. *Personality and Individual Differences*, *42*, 1503-1514. doi: 10.1016/j.paid.2006.10.023

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Memory & Cognition*, *19*, 151-158. doi: 10.3758/BF03197111

Cronbach, L. J. (1975). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302. doi: 10.1037/h0040957

de Zeeuw, P., Aarnoudse-Moens, C., Bijlhout, J., Konig, C., Uiterweer, A. P., Papanikolau, A., et al. (2008). Inhibitory performance, response speed, intraindividual variability, and response accuracy in ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*, 808-816. doi: 10.1097/CHI.0b013e318172eee9

Dodge, R. (1917). The laws of relative fatigue. *Psychological Review*, *24*, 89-113. doi: 10.1037/h0075549

Düker, H. (1930). Experimentelle Untersuchung über den Einfluss des Taktes auf fortlaufende Arbeit bei kurzer Arbeitszeit. *Psychotechnische Zeitschrift*, *5*, 82-86.

Düker, H. (1949). Über ein Verfahren zur Untersuchung der psychischen Leistungsfähigkeit. *Psychologische Forschung*, *23*, 10-24. doi: 10.1007/BF00406639

Düker, H., & Lienert, G. A. (1965). *Konzentrations-Leistungs-Test (KLT)*. Göttingen: Hogrefe.

Düker, H., Lienert, G. A., Lukesch, H., & Mayrhofer, S. (2002). *Konzentrations-Leistungs-Test (revidierte Fassung)*. Göttingen: Hogrefe.

Eid, M., Gollwitzer, M. & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim: Beltz.

Falkenstein, M., Hoormann, J., Hohnsbein, J., & Kleinsorge, T. (2003). Short-term mobilization of processing resources is revealed in the event-related potential. *Psychophysiology*, *40*, 914-923. doi: 10.1111/1469-8986.00109

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*, 340-347. doi: 10.1162/089892902317361886

Finomore, V., Matthews, G., Shaw, T., & Warm, J. (2009). Predicting vigilance: A fresh look at an old problem. *Ergonomics*, *52*, 791-808. doi: 10.1080/00140130802641627

Flehmig, H. C. (2010). *Reaktionszeitvariabilität als Indikator von Konzentrationsleistungen*. Dresden: Technische Universität Dresden. Abgerufen von http://nbn-resolving.de/urn:nbn:de:bsz:14-qucosa-39404

Flehmig, H. C., Steinborn, M. B., Langner, R., Scholz, A., & Westhoff, K. (2007). Assessing intraindividual variability in sustained attention: Reliability, relation to speed and accuracy, and practice effects. *Psychology Science*, *49*, 132-149.

Flehmig, H. C., Steinborn, M. B., Langner, R., & Westhoff, K. (2007a). Lapses of attention in reaction time tasks predict people's beliefs about their everyday life attention and memory. In S. Bülthoff, A. Chatziastros, H. Mallot & R. Ulrich (Eds.), *Proceedings of the 10th Tübingen Perception Conference*. Kirchentellingsfurt: Knirsch.

Flehmig, H. C., Steinborn, M. B., Langner, R., & Westhoff, K. (2007b). Neuroticism and the mental noise hypothesis: Relation to lapses of attention and slips of action in everyday life. *Psychology Science*, *49*, 343-360.

Flehmig, H. C., Steinborn, M. B., Westhoff, K., & Langner, R. (2010). Neuroticism and speed-accuracy tradeoff in serial mental addition and comparison performance.

Journal of Individual Differences, *31*, 130-137. doi: 10.1027/1614-0001/a000021

Formann, A. K., & Piswanger, K. (1979). *Wiener Matrizen-Test. Ein Rasch-skalierter sprachfreier Intelligenztest*. Weinheim: Beltz.

Forster, S., & Lavie, N. (2007). High perceptual load makes everybody equal - Eliminating individual differences in distractibility with load. *Psychological Science*, *18*, 377-381. doi: 10.1111/j.1467-9280.2007.01908.x

Goldhammer, F., Moosbrugger, H., & Krawietz, S. A. (2009). FACT-2-The Frankfurt Adaptive Concentration Test: Convergent validity with self-reported cognitive failures. *European Journal of Psychological Assessment*, *25*, 73-82. doi: 10.1027/1015-5759.25.2.73

Green, C. E., Chen, C. E., Helms, J. E., & Henze, K. T. (2011). Recent reliability reporting practices in Psychological Assessment: Recognizing the people behind the data. *Psychological Assessment*, *23*, 656-669. doi: 10.1037/a0023089

Gulian, E., Matthews, G., Glendon, A. I., Davies, D. R., & Debney, L. M. (1989). Dimension of driver stress. *Ergonomics*, *32*, 585-602. doi: 10.1080/00140138908966134

Hacker, W., & Sachse, P. (1998). *Allgemeine Arbeitspsychologie - Psychische Regulation von Arbeitstätigkeiten*. Göttingen: Hogrefe.

Hagemeister, C. (1994). *Individuelle Fehlerneigung beim konzentrierten Arbeiten*. Berlin: Köster.

Hagemeister, C. (2003). Übung in Konzentrationstests erkennen: Nutzung von Merkmalskombinationen. *Wirtschaftspsychologie*, *1*, 173-176.

Hagemeister, C. (2006). *Erkennen von Übung in Konzentrationstests*. Dresden: TUPress.

Hagemeister, C. (2007). How useful is the power law of practice for recognizing practice in concentration tests? *European Journal of Psychological Assessment*, *23*, 157-165. doi: 10.1027/1015-5759.23.3.157

Hagemeister, C., Scholz, A., & Westhoff, K. (2002). Wie kann man Geübtheit in Kon-

zentrationstests erkennbar machen? *Zeitschrift für Personalpsychologie*, *2*, 94-102. doi: 10.1026//1617-6391.1.2.94

Healy, A. F., Kole, J. A., Buck-Gengler, C. J., & Bourne, L. E. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied*, *10*, 188-199. doi: 10.1037/1076-898X.10.3.188

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11-26. doi: 10.1080/17470215208416600

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biological Psychology*, *45*, 73-93. doi: 10.1016/S0301-0511(96)05223-4

Hockey, G. R. J., & Earle, F. (2006). Control over the scheduling of simulated office work reduces the impact of workload on mental fatigue and task performance. *Journal of Experimental Psychology: Applied*, *12*, 50-65. doi: 10.1037/1076-898X.12.1.50

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance - A theory of the relationship between individual-differences and information-processing. *Psychological Review, 91*(2), 153-184. doi: 10.1037/0033-295x.91.2.153

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. New York: Appleton-Century-Crofts.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*, 188-196. doi: 10.1037/h0056940

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam: Elsevier.

Jentzsch, I., & Dudschig, C. (2009). Why do we slow down after an error? Mechanisms underlying the effects of posterror slowing. *Quarterly Journal of Experimental Psychology*, *62*, 209-218. doi: 10.1080/17470210802240655

Jentzsch, I., & Leuthold, H. (2006). Control over speeded actions: A common processing

locus for micro- and macro-trade-offs? *Quarterly Journal of Experimental Psychology*, *59*, 1329-1337. doi: 10.1080/17470210600674394

Karsten, A. (1928). Psychische Sättigung. *Psychologische Forschung*, *10*, 142-254. doi: 10.1007/BF00492011

Kersting, M., K., A., & Jäger, A. O. (2008). *Wilde-Intelligenz-Test 2*. Göttingen: Hogrefe.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., et al. (2010). Control and interference in task switching - A review. *Psychological Bulletin*, *136*, 849-874. doi: 10.1037/a0019842

Klumb, P. L. (1995). Cognitive failures and performance differences - Validation studies of a German version of the cognitive failures questionnaire. *Ergonomics*, *38*, 1456-1467. doi: 10.1080/00140139508925202

Köhne, R . (2001). *Konzentration und Interferenzneigung* (Unveröffentlichte Diplomarbeit). Technische Universität Dresden.

Kraepelin, E. (1902). Die Arbeitscurve. *Philosophische Studien*, *19*, 459-507.

Kraepelin, E. (1903). Über Ermüdungsmessungen. *Archiv für die gesamte Psychologie*, *1*, 9-30.

Krumm, S. (2008). *Determinanten von Konzentrationstest-Leistungen: Koordination, Arbeitstempo oder mentale Geschwindigkeit?* Marburg: Phillipps-Universität Marburg.

Krumm, S., Lipnevich, A. A., Schmidt-Atzert, L., & Bühner, M. (2012). Relational integration as a predictor of academic achievement. *Learning and Individual Differences*, *22*, 759-769. doi 10.1016/j.lindif.2012.05.002

Krumm, S., Schmidt-Atzert, L., Bracht, M., & Ochs, L. (2011). Coordination as a crucial component of performance on a sustained attention test. *Journal of Individual Differences*, *32*, 117-128. doi: 10.1027/1614-0001/a000044

Krumm, S., Schmidt-Atzert, L., & Eschert, S. (2008). Investigating the structure of attention: How do test characteristics of paper-pencil sustained attention tests

influence their relationship with other attention tests? *European Journal of Psychological Assessment*, *24*, 108-116. doi: 10.1027/1015-5759.24.2.108

Krumm, S., Schmidt-Atzert, L., Michalczyk, K., & Danthiir, V. (2008). Speeded paper-pencil sustained attention and mental speed tests: Can performance be discriminated? *Journal of Individual Differences*, *29*, 205-216. doi: 10.1027/1614-0001.29.4.205

Krumm, S., Schmidt-Atzert, L., Schmidt, S., Zenses, E. M., & Stenzel, N. (2012). Attention tests in different stimulus presentation modes. *Journal of Individual Differences*, *33*, 146-159. doi: 10.1027/1614-0001/a000085

Krumm, S., Ziegler, M., & Bühner, M. (2008). Reasoning and working memory as predictors of school grades. *Learning and Individual Differences*, *18*, 248-257. doi: 10.1016/j.lindif.2007.08.002

Langner, R., & Eickhoff, S. B. (2013). Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin*, *139*, 870-900. doi: 10.1037/a0030694

Langner, R., Eickhoff, S. B., & Steinborn, M. B. (2011). Mental fatigue modulates dynamic adaptation to perceptual demand in speeded detection. *PLoS One*, *6*, e28399. doi: 10.1371/journal.pone.0028399

Langner, R., Steinborn, M. B., Chatterjee, A., Sturm, W., & Willmes, K. (2010). Mental fatigue and temporal preparation in simple reaction-time performance. *Acta Psychologica*, *133*, 64-72. doi: 10.1016/j.actpsy.2009.10.001

Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: A response time distributional approach. *Acta Psychologica*, *104*, 167-190. doi: 10.1016/S0001-6918(00)00019-6

Lewin, K. (1928). Die Bedeutung der „Psychischen Sättigung“ für einige Probleme der Psychotechnik. *Psychotechnische Zeitschrift*, *3*, 182-188.

Lienert, G. A., & Raatz, U. (1998). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883-914. doi: 10.1037/0278-7393.18.5.883

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*, 376-400. doi: 10.1037/0033-295X.109.2.376

Los, S. A. (1996). On the origin of mixing costs: Exploring information processing in pure and mixed blocks of trials. *Acta Psychologica*, *94*, 145-188. doi: 10.1016/0001-6918(95)00050-X

Los, S. A. (1999a). Identifying stimuli of different perceptual categories in mixed blocks of trials: evidence for cost in switching between computational processes. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 3-23. doi: 10.1037/0096-1523.25.1.3

Los, S. A. (1999b). Identifying stimuli of different perceptual categories in pure and mixed blocks of trials: evidence for stimulus-driven switch costs. *Acta Psychologica*, *103*, 173-205. doi: 10.1016/S0001-6918(99)00031-1

Los, S. A., Knol, D. L., & Boers, R. M. (2001). The foreperiod effect revisited: Conditioning as a basis for nonspecific preparation. *Acta Psychologica*, *106*, 121-145. doi: 10.1016/S0001-6918(00)00029-9

Luo, D. S., Thompson, L. A., & Detterman, D. K. (2003). The causal factor underlying the correlation between psychometric g and scholastic performance. *Intelligence*, *31*, 67-83. doi: 10.1016/S0160-2896(02)00113-7

Marschner, G. (1971). Der Revisionstest als allgemeiner Leistungstest. *Diagnostica*, *17*, 14-26.

Matthews, G., Campbell, S. E., Falconer, S., Joyner, L. A., Huggins, J., Gilliland, K., et al. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, *2*, 315-340. doi: 10.1037/1528-

3542.2.4.315

Mierke, K. (1955). Wille und Leistung. Göttingen: Hogrefe. Miller, J. (2006). A likelihood ratio test for mixture effects. *Behavior Research Methods*, *38*, 92-106.

Miller, J., & Ulrich, R. (2003). Simple reaction time and statistical facilitation: a parallel grains model. *Cognitive Psychology*, *46*, 101-151. doi: 10.1016/S0010-0285(02)00517-0

Miller, J., & Ulrich, R. (2013). Mental chronometry and individual differences: modeling reliabilities and correlations of reaction time means and effect sizes. *Psychonomic Bulletin & Review*, *20*, 819-858. doi: 10.3758/s13423-013-0404-5

Mojzisch, A., & Schulz-Hardt, S. (2007). Being fed up: a social cognitive neuroscience approach to mental satiation. *Annals of the New York Academy of Sciences*, *1118*, 186-205. doi: 10.1196/annals.1412.006

Moosbrugger, H., & Goldhammer, F. (2005). *FAKT-II. Frankfurter Adaptiver Konzentrationsleistungs-Test (2. Auflage)*. Bern: Huber.

Moosbrugger, H., & Goldhammer, F. (2006). Aufmerksamkeits- und Konzentrationsdiagnostik. In K. Schweizer (Ed.), *Leistung und Leistungsdiagnostik* (pp. 83-102). Heidelberg: Springer. doi: 10.1007/3-540-33020-8_6

Moosbrugger, H., Goldhammer, F., & Schweizer, K. (2006). Latent factors underlying individual differences in attention measures - Perceptual and executive attention. *European Journal of Psychological Assessment*, *22*, 177-188. doi: 10.1027/1015-5759.22.3.177

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: does self-control resemble a muscle? *Psychological Bulletin*, *126*, 247-259. doi: 10.1037/0033-2909.126.2.247

Notebaert, W., Houtman, F., Opstal, F. V., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: an orienting account. *Cognition*, *111*, 275-279. doi: 10.1016/

j.cognition.2009.02.002

Oehrn, A. (1895). Experimentelle Studien zur Individual-Psychologie. *Psychologische Arbeiten*, *1*, 92-151.

Oswald, W. D., & Roth, E. (1987). *Der Zahlen-Verbindungs-Test (ZVT): Ein sprachfreier Test zur Messung der kognitiven Leistungsgeschwindigkeit*. Göttingen: Hogrefe.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological Bulletin*, *116*, 220-244. doi: 10.1037/0033-2909.116.2.220

Pashler, H., & Baylis, G. C. (1991). Procedural learning: Locus of practice effects in speeded choice tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 20-32. doi: 10.1037/0278-7393.17.1.20

Pauli, R. (1938). Die Arbeitskurve als ganzheitlicher Prüfungsversuch (als Universaltest). *Archiv für die gesamte Psychologie*, *108*, 412-424.

Pieters, J. P. M. (1983). Sternberg additive factor method and underlying psychological processes: Some theoretical considerations. *Psychological Bulletin*, *93*, 411-426. doi: 10.1037/0033-2909.93.3.411

Pieters, J. P. M. (1985). Reaction time analysis of simple mental tasks: A general approach. *Acta Psychologica*, *59*, 227-269. doi: 10.1016/0001-6918(85)90046-0

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25-42. doi: 10.1146/annurev.ne.13.030190.000325

Rabbitt, P. M. A. (1969). Psychological refractory delay and response-stimulus interval duration in serial, choice-response tasks. *Acta Psychologica*, *30*, 195-. doi: 10.1016/0001-6918(69)90051-1

Rabbitt, P. M. A., & Rogers, M. (1965). Age and choice between responses in a self-paced repetitive task. *Ergonomics*, *8*, 435-444. doi: 10.1080/00140136508930824

Rammsayer, T., & Stahl, J. (2004). Extraversion-related differences in response organization: evidence from lateralized readiness potentials. *Biological Psychology*, *66*,

35-49. doi: 10.1016/j.biopsycho.2003.08.003

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

Raven, J. C. (1998). *Advanced Progressive Matrices*. Oxford: Oxford University Press.

Reason, J. T. (1993). Self-report questionaires in cognitive psychology: Have they delivered the goods? . In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention, selection, awareness and control* (pp. 406-423). Oxford: Clarendon.

Reason, J. T., & Lucas, D. (1984). Absent-mindedness in shops: its incidence, correlates and consequences. *British Journal of Clinical Psychology*, *23*, 121-131. doi: 10.1111/j.2044-8260.1984.tb00635.x

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology*, *83*, 274-278. doi: 10.1037/h0028573

Ridderinkhof, K. R., Scheres, A., Oosterlaan, J., & Sergeant, J. A. (2005). Delta plots in the study of individual differences: New tools reveal response inhibition deficits in AD/HD that are eliminated by methylphenidate treatment. *Journal of Abnormal Psychology*, *114*, 197-215. doi: 10.1037/0021-843X.114.2.197

Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence*, *32*, 573-589. doi: 10.1016/j.intell.2004.06.005

Sanders, A. F. (1983). Towards a model of stress and human performance. *Acta Psychologica*, *53*, 61-97. doi.org/10.1016/0001-6918(83)90016-1

Sanders, A. F. (1998). *Elements of human performance*. Mahwah, NJ: Lawrence Erlbaum.

Sanders, A. F., & Hoogenboom, W. (1970). On the effects of continuous active work on performance. *Acta Psychologica*, *33*, 414-431. doi: 10.1016/0001-6918(70)90151-4

Schmidt-Atzert, L., & Bühner, M. (1997). Komponenten von Konzentrationsleistungen: Eine Untersuchung mit systematisch variierten Durchstreichtest-Varianten.

Diagnostica, *43*, 314-326.

Schmidt-Atzert, L., Bühner, M., & Enders, P. (2006). Messen Konzentrationstests Konzentration? Eine Analyse der Komponenten von Konzentrationsleistungen. *Diagnostica*, *52*, 33-44. doi: 10.1026/0012-1924.52.1.33

Schmidt-Atzert, L., Büttner, G., & Bühner, M. (2004). Theoretische Aspekte von Aufmerksamkeits-/Konzentrationsdiagnostik. In G. Büttner & L. Schmidt-Atzert (Eds.), *Diagnostik von Aufmerksamkeit und Konzentration* (pp. 3-22): Hogrefe.

Schmidt-Atzert, L., & Deter, B. (1993). Die Vorhersage des Ausbildungserfolgs bei verschiedenen Berufsgruppen durch Leistungstests. *Zeitschrift für Arbeits- und Organisationspsychologie*, *37*, 191-196.

Schmidt-Atzert, L., Krumm, S., & Bühner, M. (2008). Ableitung eines Strukturmodells und systematische Einordnung von Tests. *Zeitschrift für Neuropsychologie*, *19*, 59-82. doi: 10.1024/1016-264X.19.2.59

Schmiedek, F., Oberauer, K., Wilhelm, O., Suss, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414-429. doi: 10.1037/0096-3445.136.3.414

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

Schuler, H., & Prochaska, M. (2001). *Leistungsmotivationsinventar (LMI). Dimensionen berufsbezogener Leistungsorientierung.* Göttingen: Hogrefe.

Schulz-Hardt, S., Meinken, I., Rott, A., & Frey, D. (2001). Psychische Sättigung: Eine neue experimentelle Untersuchung zu einem alten Konstrukt. *Zeitschrift für Experimentelle Psychologie*, *18*, 188-200. doi: 10.1026//0949-3946.48.3.188

Schulz-Hardt, S., Rott, A., Meinken, I., & Frey, D. (2001). Ein weiterentwickeltes Modell

psychischer Sättigung. *Psychologische Rundschau*, *52*, 141-149. doi: 10.1026//0033-3042.52.3.141

Schweizer, K. (2005). An overview of research into the cognitive basis of intelligence. *Journal of Individual Differences*, *26*, 43-51. doi: 10.1027/1614-0001.26.1.43

Schweizer, K., & Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence*, *32*, 329-347. doi: 10.1016/j.intell.2004.06.006

Schweizer, K., Moosbrugger, H., & Goldhammer, F. (2005). The structure of the relationship between attention and intelligence. *Intelligence*, *33*, 589-611. doi: 10.1016/j.intell.2005.07.001

Schwinger, M., Steinmayr, R., & Spinath, B. (2009). How do motivational regulation strategies affect achievement: Mediated by effort management and moderated by intelligence. *Learning and Individual Differences*, *19*(4), 621-627. doi: 10.1016/j.lindif.2009.08.006

Sergeant, J. (2000). The cognitive-energetic model: An empirical approach to Attention-Deficit Hyperactivity Disorder. *Neuroscience and Biobehavioral Reviews*, *24*, 7-12. doi: 10.1016/S0149-7634(99)00060-3

Sergeant, J. (2005). Modeling attention-deficit/hyperactivity disorder: A critical appraisal of the cognitive-energetic model. *Biological Psychiatry*, *57*, 1248-1255. doi: 10.1016/j.biopsych.2004.09.010

Sergeant, J., & van der Meere, J. (1990). Additive factor method applied to psychopathology with special reference to childhood hyperactivity. *Acta Psychologica*, *74*, 277-295. doi: 10.1016/0001-6918(90)90009-5

Smit, J. C., & Van der Ven, A. G. H. S. (1995). Inhibition and speed in concentration tests: The Poisson Inhibition Model. *Journal of Mathematical Psychology*, *39*, 265-274. doi: 10.1006/jmps.1995.1027

Soetens, E., Boer, L. C., & Hueting, J. E. (1985). Expectancy or automatic facilitation?

Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *10*, 581-598. doi: 10.1037/0096-1523.10.4.581

Spearman, C. (1927). *The abilities of man*. New York: MacMillan.

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*, 363-374. doi: 10.1016/j.intell.2005.11.004

Spinath, B., Freudenthaler, H. H., & Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences*, *48*(4), 481-486. doi: 10.1016/j.paid.2009.11.028

Stahl, J., & Rammsayer, T. (2004). Differences in the transmission of sensory input into motor output between introverts and extraverts: Behavioral and psychophysiological analyses. *Brain and Cognition*, *56*, 293-303. doi: 10.1016/j.bandc.2004.07.004

Stahl, J., & Rammsayer, T. H. (2008). Extroversion-related differences in speed of premotor and motor processing as revealed by lateralized readiness Potentials. *Journal of Motor Behavior*, *40*, 143-154. doi: 10.3200/JMBR.40.2.143-154

Steinborn, M. B. (2004). *Kennzeichen der Übung in einem Verfahren zur Messung von Konzentrationsfähigkeit* (Unveröffentlichte Diplomarbeit). Technischen Universität Dresden.

Steinborn, M. B., Flehmig, H. C., Bratzke, D., & Schroter, H. (2012). Error reactivity in self-paced performance: Highly-accurate individuals exhibit largest post-error slowing. *Quarterly Journal of Experimental Psychology*, *65*, 624-631. doi: 10.1080/17470218.2012.660962

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2008). Predicting school achievement from self-paced continuous performance: Examining the contri-

butions of response speed, accuracy, and response speed variability. *Psychology Science Quarterly*, *50*, 613-634.

Steinborn, M. B., Flehmig, H. C., Westhoff, K., & Langner, R. (2009). Differential effects of prolonged work on performance measures in self-paced speed tests. *Advances in Cognitive Psychology*, 105-113. doi: 10.2478/v10053-008-0070-8

Steinborn, M. B., & Langner, R. (2011). Distraction by irrelevant sound during foreperiods selectively impairs temporal preparation. *Acta Psychologica*, *136*, 405-418. doi: 10.1016/j.actpsy.2011.01.008

Steinborn, M. B., & Langner, R. (2012). Arousal modulates temporal preparation under increased time uncertainty: Evidence from higher-order sequential foreperiod effects. *Acta Psychologica*, *139*, 65-76. doi: 10.1016/j.actpsy.2011.10.010

Steinborn, M. B., Rolke, B., Bratzke, D., & Ulrich, R. (2010). The effect of a cross-trial shift of auditory warning signals on the sequential foreperiod effect. *Acta Psychologica*, *134*(1), 94-104. doi: 10.1016/j.actpsy.2009.12.011

Steinmayr, R., Bipp, T., & Spinath, B. (2011). Goal orientations predict academic performance beyond intelligence and personality. *Learning and Individual Differences*, *21*(2), 196-200. doi: 10.1016/j.lindif.2010.11.026

Sternberg, S. (2001). Separate modifiability, mental modules, and the use of pure and composite measures to reveal them. *Acta Psychologica*, *106*, 147-246. doi: 10.1016/S0001-6918(00)00045-7

Sturm, W., & Willmes, K. (2001). On the functional neuroanatomy of intrinsic and phasic alertness. *Neuroimage*, *14*, 76-84. doi: 10.1006/nimg.2001.0839

Sturm, W., Willmes, K., Orgass, B., & Hartje, W. (1997). Do specific attention deficits need specific training? *Neuropsychological Rehabilitation*, *7*, 81-103. doi: 10.1080/713755526

Sturm, W., & Zimmerman, P. (2000). Aufmerksamkeitsstörungen. In W. Sturm, M.

Herrmann & C. W. Wallesch (Eds.), *Lehrbuch der klinischen Neuropsychologie* (pp. 345-365). Lisse: Swets & Zeitlinger.

Stuss, D. T., Meiran, N., Guzman, A., Lafleche, G., & Willmer, J. (1996). Do long tests yield a more accurate diagnosis of dementia than short tests? A comparison of 5 neuropsychological tests. *Archives of Neurology*, *53*, 1033-1039. doi: 10.1001/archneur.1996.00550100119021

Ulrich, R., Mattes, S., & Miller, J. (1999). Donders's assumption of pure insertion: an evaluation on the basis of response dynamics. *Acta Psychologica*, *102*, 43-75. doi: 10.1016/S0001-6918(99)00019-0

Ulrich, R., & Miller, J. (2008). Response grouping in the psychological refractory period (PRP) paradigm: models and contamination effects. *Cognitive Psychology*, *57*, 75-121. doi: 10.1016/j.cogpsych.2007.06.004

Ulrich, R., & Miller, J. O. (1994). Effects of truncation on reaction time analysis. *Journal of Experimental Psychology: General*, *123*, 34-80. doi: 10.1037/0096-3445.123.1.34

Van Breukelen, G. J. P., Roskam, E. E. C. I., Eling, P. A. T. M., Jansen, R. W. T. L., Souren, D. A. P. B., & Ickenroth, J. G. M. (1996). A model and diagnostic measures for response time series on tests of concentration: Historical background, conceptual framework, and some applications. *Brain and Cognition*, *27*, 147-179. doi: 10.1006/brcg.1995.1015

Van der Ven, A. G. H. S., Smit, J. C., & Jansen, R. W. T. L. (1989). Inhibition in prolonged work tasks. *Applied Psychological Measurement*, *13*, 177-191. doi: 10.1177/014662168901300209

van Drunen, P. (1993). Von der Psychotechnik zur Psychodiagnostik. In H. Lück & R. Miller (Eds.), *Illustrierte Geschichte der Psychologie* (pp. 254-256). München: Beltz.

von Voss, G. (1899). Über Schwankungen der geistigen Arbeitsleistung. *Psychologische Arbeiten*, *2*, 399-449.

Weaver, H. B. (1942). Study of discriminative serial action: Manual response to color. *Journal of Experimental Psychology*, *31*, 177-201. doi: 10.1037/h0054894

Westhoff, K. (1985). Eine erste Prüfung einer Konzentrationstheorie. *Diagnostica*, *31*, 265-278.

Westhoff, K. (1989). Übungsabhängigkeit von Leistungen in Konzentrationstests. *Diagnostica*, *35*, 122-130.

Westhoff, K. (1995). Aufmerksamkeit und Konzentration. In M. Amelang (Ed.), *Verhaltens- und Leistungsunterschiede. Band 2 des Bereichs Differentielle Psychologie und Persönlichkeitsforschung der Enzyklopädie der Psychologie* (pp. 375-402). Göttingen: Hogrefe.

Westhoff, K., & Dewald, D. (1990). Effekte der Übung in der Bearbeitung von Konzentrationstests. *Diagnostica*, *36*, 1-15.

Westhoff, K., & Hagemeister, C. (2002). Konzentrationstraining. In K. J. Klauer (Ed.), *Kognitives Training* (pp. 515-538). Göttingen: Hogrefe.

Westhoff, K., & Graubner, J. (2003). Konstruktion eines komplexen Konzentrationstests. *Diagnostica*, *49*, 110-119. doi: 10.1026//0012-1924.49.3.110

Westhoff, K., & Hagemeister, C. (1992). Reliabilität von Fehlern in Konzentrationstests. *Diagnostica*, *38*, 116-129.

Westhoff, K., & Hagemeister, C. (2005). *Konzentrationsdiagnostik*. Lengerich: Pabst Publishers.

Westhoff, K., & Kluck, M. L. (1984). Ansätze einer Theorie konzentrativer Leistungen. *Diagnostica*, *30*, 167-183.

Westhoff, K., & Scholz, A. (2007). *Komplexer Konzentrationstest - KKT*. Wien: Schuhfried.

Anhang

Anhang A: Verwendete Fragebogen

Anhang B: Instruktion für den SMACT

Anhang C: Präsentierte Items im SMACT (Studie 1 und 2)

Anhang D: Korrelationen zwischen Fehleranteil im SMACT und Schulnoten

Anhang E: Korrelationen zwischen selbstberichteter Aufmerksamkeit und Schulnoten

Anhang F: Stichprobenkennwerte der Studien 1 bis 5

Anhang G: Korrelationen zwischen den Subtests des I-S-T 2000 R sowie des WIT-2 und Schulnoten

Anhang H: Vergleich der Korrelationskoeffizienten aus verschiedenen Stichproben zum I-S-T 2000 R

Anhang I: Inkrementelle Regressionsanalysen der Studien 1 bis 5

Anhang J: Erklärung

Anhang A: Verwendete Fragebogen

Cognitive Failures Questionnaire (CFQ)

Die folgenden Fragen beziehen sich auf kleinere Fehler, die jedem Menschen von Zeit zu Zeit passieren. Einige der Fehler kommen häufiger vor als andere.
Wir möchten wissen, wie oft so etwas bei Ihnen im letzten Jahr vorgekommen ist.
Bitte kreuzen Sie die passende Zahl an. Die Zahlen geben nicht an, dass Ihnen etwas 1- 2- 3- 4- oder keinmal passiert ist, sondern kennzeichnen, wie oft etwas passiert; je höher die Zahl, desto häufiger. Die Zahl 3 kreuzen Sie z. B. an, wenn Ihnen etwas ziemlich oft passiert.

	sehr oft	ziemlich oft	gelegentlich	ziemlich selten	fast nie
1. Lesen Sie etwas und stellen fest, dass Sie nicht darüber nachgedacht haben und es noch einmal lesen müssen?	4	3	2	1	0
2. Stellen Sie fest, dass Sie vergessen haben, warum Sie von einem Teil Ihrer Wohnung in einen anderen gegangen sind?	4	3	2	1	0
3. Übersehen Sie Wegweiser an der Straße?	4	3	2	1	0
4. Verwechseln Sie rechts und links, wenn Sie die Richtung angeben?	4	3	2	1	0
5. Stoßen Sie unabsichtlich gegen andere Personen?	4	3	2	1	0
6. Stellen Sie fest, dass Sie vergessen, ob Sie eine Lampe oder den Herd ausgemacht haben oder die Türe abgeschlossen haben?	4	3	2	1	0
7. Überhören Sie die Namen von Leuten, wenn Sie diese treffen?	4	3	2	1	0
8. Sagen Sie etwas und erkennen anschließend, dass dies als beleidigend aufgefasst werden könnte?	4	3	2	1	0
9. Überhören Sie Leute, die zu Ihnen sprechen, wenn Sie gerade dabei sind, etwas anderes zu tun?	4	3	2	1	0
10. Geraten Sie in Wut und bereuen es?	4	3	2	1	0
11. Lassen Sie versehentlich Briefe tagelang liegen, die umgehend beantwortet werden müssten?	4	3	2	1	0

Fragebogen wird fortgesetzt ...

	sehr oft	ziemlich oft	gelegentlich	ziemlich selten	fast nie
12. Vergessen Sie, welche Richtung Sie einschlagen müssen auf einem Weg, den Sie gut kennen, aber selten benutzen?	4	3	2	1	0
13. Übersehen Sie das, was Sie in einem Supermarkt wollen, obwohl es da ist?	4	3	2	1	0
14. Fragen Sie sich plötzlich, ob Sie ein Wort richtig gebraucht haben?	4	3	2	1	0
15. Fällt es Ihnen schwer, Entscheidungen zu treffen?	4	3	2	1	0
16. Vergessen Sie Verabredungen?	4	3	2	1	0
17. Vergessen Sie, wo Sie etwas hingelegt haben wie zum Beispiel eine Zeitung oder ein Buch?	4	3	2	1	0
18. Werfen Sie unbeabsichtigt etwas weg, das Sie behalten wollten, und behalten das, was Sie eigentlich wegwerfen wollten?	4	3	2	1	0
19. Träumen Sie vor sich hin, wenn Sie eigentlich etwas anderem zuhören sollten?	4	3	2	1	0
20. Vergessen Sie die Namen von Leuten?	4	3	2	1	0
21. Beginnen Sie mit der Erledigung einer Sache, und Sie machen auf einmal ohne Absicht etwas ganz anderes?	4	3	2	1	0
22. Fällt Ihnen etwas nicht ein, obwohl es Ihnen auf der Zunge liegt?	4	3	2	1	0
23. Vergessen Sie, was Sie in Geschäften einkaufen wollten?	4	3	2	1	0
24. Lassen Sie Gegenstände fallen?	4	3	2	1	0
25. Beobachten Sie, dass Sie etwas sagen wollen, aber nicht mehr wissen, was es war?	4	3	2	1	0

Short Inventory of Minor Lapses (SIML)

Die folgenden Fragen beziehen sich auf kleinere Fehler, die jedem Menschen von Zeit zu Zeit passieren. Einige der Fehler kommen häufiger vor als andere.
Wir möchten wissen, wie oft so etwas bei Ihnen im letzten Jahr vorgekommen ist.
Bitte kreuzen Sie die passende Zahl an. Die Zahlen geben nicht an, dass Ihnen etwas 1- 2- 3- 4- oder keinmal passiert ist, sondern kennzeichnen, wie oft etwas passiert; je höher die Zahl, desto häufiger. Die Zahl 3 kreuzen Sie z. B. an, wenn Ihnen etwas ziemlich oft passiert.

	sehr oft	ziemlich oft	gelegentlich	ziemlich selten	fast nie
1. Wie oft vergessen Sie etwas zu sagen, was Sie erwähnen wollten?	4	3	2	1	0
2. Wie oft machen Sie etwas falsch, weil Sie Ihr Verhalten nicht geändert bzw. angepasst haben?	4	3	2	1	0
3. Wie oft haben Sie das Gefühl, dass Sie etwas tun wollen, jetzt oder später, Sie sich aber nicht mehr daran erinnern können, was es ist?	4	3	2	1	0
4. Wie oft bemerken Sie, dass Sie einen Fehler gemacht haben?	4	3	2	1	0
5. Wie oft bemerken Sie, dass Sie an etwas denken, woran Sie lieber nicht denken würden?	4	3	2	1	0
6. Wie oft setzen Sie eine bekannte Verhaltensweise ein, wo ein neues Verhalten angezeigt wäre?	4	3	2	1	0
7. Wie oft kommt es vor, dass Sie sich nicht erinnern können, was Sie gerade getan haben oder wo Sie gerade gewesen sind?	4	3	2	1	0
8. Wie oft möchten Sie Ihre Aufgaben so gut wie möglich erledigen?	4	3	2	1	0
9. Wie oft lassen Sie einen notwendigen Schritt beim Ausführen einer Aufgabe aus (z. B. vergessen, Tee in die Teekanne zu tun)?	4	3	2	1	0
10. Wie oft setzen Sie eine Verhaltensweise ein, die sich nicht bewährt hat?	4	3	2	1	0

Fragebogen wird fortgesetzt ...

	sehr oft	ziemlich oft	gelegentlich	ziemlich selten	fast nie
11. Wie oft kommt es vor, dass Sie sich nicht auf Anhieb an Namen von Ihnen vertrauten Personen erinnern können, deren Namen Sie eigentlich kennen?	4	3	2	1	0
12. Wie oft geben Sie sich selbst die Schuld in Situationen, wo Fehler passieren, obwohl unklar ist, wer die Fehler verursacht hat?	4	3	2	1	0
13. Wie oft denken Sie, dass Sie Ihre Aufmerksamkeit gerade auf etwas richten, es in Wirklichkeit aber nicht tun?	4	3	2	1	0
14. Wie oft machen Sie einen Fehler, den Sie durch Ihr Wissen hätten vermeiden können?	4	3	2	1	0
15. Wie oft haben Sie das Gefühl „Wieso bin ich hier?“, wenn Sie bemerken, dass Sie vergessen haben, wofür Sie herkamen?	4	3	2	1	0
16. Wie oft kontrollieren Sie Ihre eigene Arbeit?	4	3	2	1	0
17. Wie oft kommt es vor, dass Sie bemerken, dass Sie irgend etwas noch einmal tun, obwohl Sie es schon getan haben?	4	3	2	1	0
18. Wie oft haben Sie Probleme mit neuen Situationen?	4	3	2	1	0
19. Wie oft kommt es vor, dass Sie etwas tun, was nicht notwendig ist (z. B. das Licht anknipsen, obwohl es heller Tag ist)?	4	3	2	1	0
20. Wie oft passieren Ihnen Fehler, weil Sie sich nicht genügend angestrengt haben?	4	3	2	1	0
21. Wie oft stellen Sie fest, dass Sie vergessen haben, etwas zu tun, was Sie vorhatten?	4	3	2	1	0
22. Wie oft verhalten Sie sich in neuen Situationen unpassend?	4	3	2	1	0
23. Sie tun etwas und werden abgelenkt. Wie oft passiert es Ihnen, dass Sie nach der Ablenkung etwas anderes tun?	4	3	2	1	0

Fragebogen wird fortgesetzt …

	sehr oft	ziemlich oft	gelegentlich	ziemlich selten	fast nie
24. Wie oft sagen andere, dass Sie anderen die Schuld geben, obwohl Sie selbst einen Fehler gemacht haben?	4	3	2	1	0
25. Wie oft bemerken Sie, dass Sie nach etwas suchen, das Sie gerade weggelegt haben?	4	3	2	1	0
26. Wie oft machen Sie Fehler, weil Sie nicht alle verfügbaren Informationen berücksichtigt haben?	4	3	2	1	0
27. Sie tun etwas und werden unterbrochen. Wie oft machen Sie nach der Unterbrechung unabsichtlich mit etwas anderem weiter, obwohl sie eigentlich die erste Handlung zu Ende führen wollten?	4	3	2	1	0
28. Wie oft werden Sie von anderen Personen auf Fehler hingewiesen, die Sie nicht bemerkt haben?	4	3	2	1	0
29. Wie oft schweifen Sie mit den Gedanken ab, wenn Sie sich auf etwas konzentrieren müssen?	4	3	2	1	0
30. Wie oft machen Sie Fehler, weil Sie so handeln, wie Sie immer gehandelt haben?	4	3	2	1	0
31. Wie oft tun Sie das Richtige, aber an den falschen Gegenständen (oder Personen) (z. B. Süßigkeiten auspacken, die Süßigkeiten wegwerfen und das Papier in den Mund stecken)?	4	3	2	1	0
32. Wie oft nehmen Sie es mit der Ausführung einer Aufgabe nicht so genau?	4	3	2	1	0
33. Wie oft bemerken Sie, dass Sie nach etwas suchen, das Sie mit sich herumtragen?	4	3	2	1	0

Anhang B: Instruktion für den SMACT

Allgemeine Instruktion für die Bearbeitung des SMACT

Es folgt zunächst eine Übung.

Es erscheinen Rechenaufgaben. Auf einer Seite ist stets eine Zahl, auf der anderen dagegen befindet sich ein Additionsterm.

Drücke die linke und die rechte Taste jeweils in der Richtung des größeren Zahlwerts.

Ein Beispiel: Bei folgender Aufgabe

2+5 | 9

drücke die rechte Taste, denn rechts ist der größere Zahlwert.

Drücke die Leertaste, um zu beginnen.

Nach der Erwärmungsphase folgte noch eine Instruktion, nach dieser wurde der Test dargeboten.

Es erscheinen jetzt die eigentlichen Rechenaufgaben.

Arbeite so schnell wie möglich und dabei auch so genau wie möglich.

Drücke die Leertaste, um zu beginnen.

Anhang C: Präsentierte Items im SMACT (Studie 1 und 2)

Verwendete Items aus Studie 1

Tabelle A.1 – Liste der leichten Items im Serialen Mentalen Additions- und Komparationstests (SMACT) in Studie 1

Nr.	L	Item	R	Nr.	L	Item	R	Nr.	L	Item	R
1		5+6 | 19	x	44		9+5 | 22	x	87		4 | 3+9	x
2		4+7 | 19	x	45		8+7 | 23	x	88		5 | 6+7	x
3		3+8 | 19	x	46		9+6 | 23	x	89		5 | 5+8	x
4		2+9 | 19	x	47		9+7 | 24	x	90		5 | 4+9	x
5		5+7 | 20	x	48		9+8 | 25	x	91		6 | 6+8	x
6		4+8 | 20	x	49	x	6+5 | 3		92		6 | 5+9	x
7		3+9 | 20	x	50	x	7+4 | 3		93		7 | 7+8	x
8		6+7 | 21	x	51	x	8+3 | 3		94		7 | 6+9	x
9		5+8 | 21	x	52	x	9+2 | 3		95		8 | 7+9	x
10		4+9 | 21	x	53	x	7+5 | 4		96		9 | 8+9	x
11		6+8 | 22	x	54	x	8+4 | 4		97	x	19 | 6+5	
12		5+9 | 22	x	55	x	9+3 | 4		98	x	19 | 7+4	
13		7+8 | 23	x	56	x	7+6 | 5		99	x	19 | 8+3	
14		6+9 | 23	x	57	x	8+5 | 5		100	x	19 | 9+2	
15		7+9 | 24	x	58	x	9+4 | 5		101	x	20 | 7+5	
16		8+9 | 25	x	59	x	8+6 | 6		102	x	20 | 8+4	
17	x	5+6 | 3		60	x	9+5 | 6		103	x	20 | 9+3	
18	x	4+7 | 3		61	x	8+7 | 7		104	x	21 | 7+6	
19	x	3+8 | 3		62	x	9+6 | 7		105	x	21 | 8+5	
20	x	2+9 | 3		63	x	9+7 | 8		106	x	21 | 9+4	
21	x	5+7 | 4		64	x	9+8 | 9		107	x	22 | 8+6	
22	x	4+8 | 4		65	x	19 | 5+6		108	x	22 | 9+5	
23	x	3+9 | 4		66	x	19 | 4+7		109	x	23 | 8+7	

Tabelle wird fortgesetzt …

Tabelle A.1 - Liste der leichten Items im Serialen Mentalen Additions- und Komparationstests (SMACT) in Studie 1 (Fortsetzung)

Nr.	L	Item	R	Nr.	L	Item	R	Nr.	L	Item	R
24	x	6+7 | 5		67	x	19 | 3+8		110	x	23 | 9+6	
25	x	5+8 | 5		68	x	19 | 2+9		111	x	24 | 9+7	
26	x	4+9 | 5		69	x	20 | 5+7		112	x	25 | 9+8	
27	x	6+8 | 6		70	x	20 | 4+8		113		3 | 6+5	x
28	x	5+9 | 6		71	x	20 | 3+9		114		3 | 7+4	x
29	x	7+8 | 7		72	x	21 | 6+7		115		3 | 8+3	x
30	x	6+9 | 7		73	x	21 | 5+8		116		3 | 9+2	x
31	x	7+9 | 8		74	x	21 | 4+9		117		4 | 7+5	x
32	x	8+9 | 9		75	x	22 | 6+8		118		4 | 8+4	x
33		6+5 | 19	x	76	x	22 | 5+9		119		4 | 9+3	x
34		7+4 | 19	x	77	x	23 | 7+8		120		5 | 7+6	x
35		8+3 | 19	x	78	x	23 | 6+9		121		5 | 8+5	x
36		9+2 | 19	x	79	x	24 | 7+9		122		5 | 9+4	x
37		7+5 | 20	x	80	x	25 | 8+9		123		6 | 8+6	x
38		8+4 | 20	x	81		3 | 5+6	x	124		6 | 9+5	x
39		9+3 | 20	x	82		3 | 4+7	x	125		7 | 8+7	x
40		7+6 | 21	x	83		3 | 3+8	x	126		7 | 9+6	x
41		8+5 | 21	x	84		3 | 2+9	x	127		8 | 9+7	x
42		9+4 | 21	x	85		4 | 5+7	x	128		9 | 9+8	x
43		8+6 | 22	x	86		4 | 4+8	x				

Anmerkungen. Für die Bearbeitung war eine Reaktion L (= Links) oder R (= Rechts) gefordert. Die jeweils korrekte Antwort war im Test nicht sichtbar, ist hier in der Tabelle mit x markiert. Die Differenz zwischen Ergebnis und Term beträgt immer 7.

Tabelle A.2 – Liste der schwierigen Items im Serialen Mentalen Additions- und Komparationstests (SMACT) in Studie 1

Nr.	L	Item	R	Nr.	L	Item	R	Nr.	L	Item	R
1		5+6 | 12	x	44		9+5 | 15	x	87		11 | 3+9	x
2		4+7 | 12	x	45		8+7 | 16	x	88		12 | 6+7	x
3		3+8 | 12	x	46		9+6 | 16	x	89		12 | 5+8	x
4		2+9 | 12	x	47		9+7 | 17	x	90		12 | 4+9	x
5		5+7 | 13	x	48		9+8 | 18	x	91		13 | 6+8	x
6		4+8 | 13	x	49	x	6+5 | 10		92		13 | 5+9	x
7		3+9 | 13	x	50	x	7+4 | 10		93		14 | 7+8	x
8		6+7 | 14	x	51	x	8+3 | 10		94		14 | 6+9	x
9		5+8 | 14	x	52	x	9+2 | 10		95		15 | 7+9	x
10		4+9 | 14	x	53	x	7+5 | 11		96		16 | 8+9	x
11		6+8 | 15	x	54	x	8+4 | 11		97	x	12 | 6+5	
12		5+9 | 15	x	55	x	9+3 | 11		98	x	12 | 7+4	
13		7+8 | 16	x	56	x	7+6 | 12		99	x	12 | 8+3	
14		6+9 | 16	x	57	x	8+5 | 12		100	x	12 | 9+2	
15		7+9 | 17	x	58	x	9+4 | 12		101	x	13 | 7+5	
16		8+9 | 18	x	59	x	8+6 | 13		102	x	13 | 8+4	
17	x	5+6 | 10		60	x	9+5 | 13		103	x	13 | 9+3	
18	x	4+7 | 10		61	x	8+7 | 14		104	x	14 | 7+6	
19	x	3+8 | 10		62	x	9+6 | 14		105	x	14 | 8+5	
20	x	2+9 | 10		63	x	9+7 | 15		106	x	14 | 9+4	
21	x	5+7 | 11		64	x	9+8 | 16		107	x	15 | 8+6	
22	x	4+8 | 11		65	x	12 | 5+6		108	x	15 | 9+5	
23	x	3+9 | 11		66	x	12 | 4+7		109	x	16 | 8+7	
24	x	6+7 | 12		67	x	12 | 3+8		110	x	16 | 9+6	
25	x	5+8 | 12		68	x	12 | 2+9		111	x	17 | 9+7	

Tabelle wird fortgesetzt ...

Tabelle A.2 - Liste der schwierigen Items im Serialen Mentalen Additions- und Komparationstests (SMACT) in Studie 1 (Fortsetzung)

Nr.	L	Item	R	Nr.	L	Item	R	Nr.	L	Item	R
26	x	4+9 | 12		69	x	13 | 5+7		112	x	18 | 9+8	
27	x	6+8 | 13		70	x	13 | 4+8		113		10 | 6+5	x
28	x	5+9 | 13		71	x	13 | 3+9		114		10 | 7+4	x
29	x	7+8 | 14		72	x	14 | 6+7		115		10 | 8+3	x
30	x	6+9 | 14		73	x	14 | 5+8		116		10 | 9+2	x
31	x	7+9 | 15		74	x	14 | 4+9		117		11 | 7+5	x
32	x	8+9 | 16		75	x	15 | 6+8		118		11 | 8+4	x
33		6+5 | 12	x	76	x	15 | 5+9		119		11 | 9+3	x
34		7+4 | 12	x	77	x	16 | 7+8		120		12 | 7+6	x
35		8+3 | 12	x	78	x	16 | 6+9		121		12 | 8+5	x
36		9+2 | 12	x	79	x	17 | 7+9		122		12 | 9+4	x
37		7+5 | 13	x	80	x	18 | 8+9		123		13 | 8+6	x
38		8+4 | 13	x	81		10 | 5+6	x	124		13 | 9+5	x
39		9+3 | 13	x	82		10 | 4+7	x	125		14 | 8+7	x
40		7+6 | 14	x	83		10 | 3+8	x	126		14 | 9+6	x
41		8+5 | 14	x	84		10 | 2+9	x	127		15 | 9+7	x
42		9+4 | 14	x	85		11 | 5+7	x	128		16 | 9+8	x
43		8+6 | 15	x	86		11 | 4+8	x				

Anmerkung. Für die Bearbeitung war eine Reaktion L (= Links) oder R (= Rechts) gefordert. Die jeweils korrekte Antwort war im Test nicht sichtbar, ist hier in der Tabelle mit x markiert. Die Differenz zwischen Ergebnis und Term beträgt immer 1.

Verwendete Items aus Studie 2

Tabelle A.3 – Liste der Items im SMACT in Studie 2

Nr.	L	Item	R	Nr.	L	Item	R	Nr.	L	Item	R
1	x	1+1 | 1		17	x	2+2 | 3		33	x	3+3 | 5	
2		1 | 1+1	x	18		3 | 2+2	x	34		5 | 3+3	x
3		1+1 | 3	x	19		2+2 | 5	x	35		3+3 | 7	x
4	x	3 | 1+1		20	x	5 | 2+2		36	x	7 | 3+3	
5		1+2 | 4	x	21	x	2+3 | 4		37	x	3+4 | 6	
6	x	4 | 1+2		22		4 | 2+3	x	38		6 | 3+4	x
7	x	1+2 | 2		23		2+3 | 6	x	39		3+4 | 8	x
8		2 | 1+2	x	24	x	6 | 2+3		40	x	8 | 3+4	
9	x	1+3 | 3		25	x	2+4 | 5		41	x	3+5 | 7	
10		3 | 1+3	x	26		5 | 2+4	x	42		7 | 3+5	x
11		1+3 | 5	x	27		2+4 | 7	x	43		3+5 | 9	x
12	x	5 | 1+3		28	x	7 | 2+4		44	x	9 | 3+5	
13	x	2+1 | 2		29	x	3+1 | 3		45	x	4+2 | 5	
14		2 | 2+1	x	30		3 | 3+1	x	46		5 | 4+2	x
15		2+1 | 4	x	31		3+2 | 6	x	47		4+2 | 7	x
16	x	4 | 2+1		32	x	6 | 3+2		48	x	7 | 4+2	

Anmerkungen. Für die Bearbeitung war eine Reaktion L (= Links) oder R (= Rechts) gefordert. Die jeweils korrekte Antwort war im Test nicht sichtbar, ist hier in der Tabelle mit x markiert.

Anhang D: Korrelationen zwischen Fehleranteil im SMACT und Schulnoten

Tabelle A.4 – Ergebnisse der Studie 1: Korrelationen zwischen Fehleranteil im SMACT und Schulnoten als Funktion von niedriger versus hoher Itemschwierigkeit und Itemhomogenität

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Fehlerprozentwerte						
SMACT - Gesamt	-.21	.10	-.18	-.17	-.08	-.06
SMACT - Gesamt (Retest)	**-.32****	-.12	**-.35****	-.24	-.02	-.14
SMACT - Block	-.15	.12	-.09	-.13	-.06	-.10
SMACT - Block (Retest)	**-.36****	-.12	**-.33****	**-.35****	-.22	-.15
SMACT - MIX	-.24	-.07	-.23	-.18	-.06	-.05
SMACT - MIX (Retest)	-.28	-.11	**-.34****	-.14	-.08	-.08
SMACT - Block leicht	.02	.18	.03	.06	-.06	-.05
SMACT - Block leicht (Retest)	-.28	-.24	**-.31****	-.26	-.28	-.29
SMACT - Block schwer	-.20	.04	-.14	-.20	-.11	-.10
SMACT - Block schwer (Retest)	**-.33****	-.04	-.28	-.32	-.15	-.07
SMACT - Mix leicht	-.21	-.08	-.25	-.23	.02	-.28
SMACT - Mix leicht (Retest)	-.24	-.10	**-.31****	-.10	-.08	-.01
SMACT - Mix schwer	-.22	.06	-.20	-.15	-.07	.07
SMACT - Mix schwer (Retest)	-.26	-.11	**-.33****	-.14	-.07	-.09

Anmerkungen. Korrelationen zwischen prozentualem Fehleranteil im SMACT und Realschulnoten-Gesamt n = 72; Realschule-Mathe n = 74; Realschule-Physik n = 74; Abitur-Gesamt n = 65; Abitur-Mathe n = 68; Abitur-Physik n = 38; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

Tabelle A.5 – Ergebnisse der Studie 2: Korrelationen zwischen Fehleranteil im SMACT und Schulnoten als Funktion von Testlänge (kurz versus lang) und Testverlauf (Abschnitt 1-4)

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Fehlerprozentwerte						
SMACT - Kurztest (Prä)	-.08	.00	-.11	-.14	-.06	-.08
SMACT - Kurztest (Retest)	-.07	-.07	-.18	-.15	-.01	-.15
SMACT - Langtest	-.15	-.04	-.17	-.16	-.04	-.12
SMACT - Langtest (Retest)	-.10	.01	-.14	-.12	-.08	-.04
SMACT - T1	-.08	.06	-.11	-.13	-.11	-.11
SMACT - T1 (Retest)	-.05	-.02	-.16	-.15	-.16	-.02
SMACT - T2	-.13	-.05	-.16	-.16	-.04	-.14
SMACT - T2 (Retest)	-.05	.04	-.10	-.12	-.07	-.09
SMACT - T3	-.18	-.13	-.17	-.12	-.03	-.06
SMACT - T3 (Retest)	-.12	.03	-.12	-.10	-.08	-.03
SMACT - T4	-.14	-.03	-.15	-.17	.04	-.14
SMACT - T4 (Retest)	-.12	-.01	-.12	-.06	.00	-.01
SMACT - Kurztest (Post)	-.14	-.06	-.16	-.24	-.08	-.10
SMACT - Kurztest (Post; Retest)	-.01	.09	-.07	-.08	-.01	-.03

Anmerkungen. Korrelation zwischen SMACT-Retest und Realschulnoten-Gesamt n = 87; entsprechend Realschule-Mathe n = 88; Realschule-Physik n = 87; Abitur-Gesamt n = 88; Abitur-Mathe n = 87; Abitur-Physik n = 53; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

Anhang E: Korrelationen zwischen selbstberichteter Aufmerksamkeit und Schulnoten

Tabelle A.6 – Ergebnisse der Studie 1: Korrelationen zwischen selbstberichteter Aufmerksamkeit und Schulnoten

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Selbstberichtete Aufmerksamkeit						
CFQ	-.06	-.09	-.05	-.16	-.06	-.20
SIML - rule	.04	.02	-.02	-.13	.02	-.06
SIML - skill	.16	.20*	.17	.08	.25*	.18

Anmerkungen. Korrelation zwischen CFQ und Realschulnoten-Gesamt n = 70; Realschule-Mathe n = 72; Realschule-Physik n = 72; Abitur-Gesamt n = 65; Abitur-Mathe n = 67; Abitur-Physik n = 38; Korrelationen zwischen SIML-rule und Realschulnoten-Gesamt n = 72; Realschule-Mathe n = 74; Realschule-Physik n = 74; Abitur-Gesamt n = 65; Abitur-Mathe n = 68; Abitur-Physik n = 38; Korrelation zwischen SIML-skill und Realschulnoten-Gesamt n = 72; Realschule-Mathe n = 74; Realschule-Physik n = 74; Abitur-Gesamt n = 65; Abitur-Mathe n = 68; Abitur-Physik n = 38; * für $p < .05$ (einseitige Testung).

Tabelle A.7 – Ergebnisse der Studie 2: Korrelationen zwischen selbstberichteter Aufmerksamkeit und Schulnoten

Skalen	Realschule			Abitur		
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik
Selbstberichtete Aufmerksamkeit						
CFQ	.05	.06	.13	.09	.02	.01
SIML - rule	.03	.06	.17	.09	.11	.11
SIML - skill	.11	.10	.15	.02	-.06	-.05

Anmerkungen. Korrelation zwischen CFQ und Realschulnoten-Gesamt n = 87; Realschule-Mathe n = 88; Realschule-Physik n = 87; Abitur-Gesamt n = 88; Abitur-Mathe n = 87; Abitur-Physik n = 53; Korrelationen zwischen SIML-rule und Realschulnoten-Gesamt n = 87; Realschule-Mathe n = 88; Realschule-Physik n = 87; Abitur-Gesamt n = 88; Abitur-Mathe n = 87; Abitur-Physik n = 53; Korrelation zwischen SIML-skill und Realschulnoten-Gesamt n = 87; Realschule-Mathe n = 88; Realschule-Physik n = 87; Abitur-Gesamt n = 88; Abitur-Mathe n = 87; Abitur-Physik n = 53; * für $p < .05$ (einseitige Testung).

Tabelle A.8 – Ergebnisse der Studie 3: Korrelationen zwischen selbstberichteter Aufmerksamkeit und Schulnoten

Skalen	Realschule			Abitur			Vordiplom
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Selbstberichtete Aufmerksamkeit							
CFQ	.00	.03	.06	.06	-.01	.06	.10
SIML - rule	-.05	-.01	.00	-.02	-.02	-.03	.09
SIML - skill	-.02	.04	.03	.09*	.03	.07	.09

Anmerkungen. Korrelation zwischen CFQ und Realschulnoten-Gesamt n = 334; Realschule-Mathe n = 356; Realschule-Physik n = 355; Abitur-Gesamt n = 349; Abitur-Mathe n = 360; Abitur-Physik n = 248; Vordiplom n = 228; Korrelationen zwischen SIML-rule und Realschulnoten-Gesamt n = 340; Realschule-Mathe n = 362; Realschule-Physik n = 361; Abitur-Gesamt n = 356; Abitur-Mathe n = 367; Abitur-Physik n = 249; Vordiplom n = 234; Korrelation zwischen SIML-skill und Realschulnoten-Gesamt n = 335; Realschule-Mathe n = 357; Realschule-Physik n = 356; Abitur-Gesamt n = 351; Abitur-Mathe n = 362; Abitur-Physik n = 249; Vordiplom n = 233; * für $p < .05$ (einseitige Testung).

Anhang F: Stichprobenkennwerte der Studien 1 bis 5

Tabelle A.9 – Stichprobenkennwerte der Studie 1: Itemschwierigkeit

		Sitzung 1				Sitzung 2			
		M	*SD*	Schiefe	Kurtosis	*M*	*SD*	Schiefe	Kurtosis
APM		27.4	4.9	-0.9	1.3	-	-	-	-
d2	GZ	497.0	84.2	-0.2	-0.4	544.5	75.5	-0.3	-0.8
	F%	3.2	2.5	1.5	3.1	2.0	2.0	2.2	7.5
SMACT - Block (leichte Items)	RT	975.4	176.6	1.1	2.9	801.6	168.4	2.2	10.9
	F%	1.2	1.5	5.4	41.3	0.9	0.9	1.2	1.2
SMACT - Block (schwierige Items)	RT	2116	478.1	1.0	1.3	1774	387.2	1.0	0.7
	F%	3.3	2.6	1.9	4.0	2.4	1.7	1.5	3.2
SMACT - MIX (leichte Items)	RT	1184	203.6	0.5	-0.2	907.0	162.0	1.1	2.9
	F%	0.9	0.9	2.3	8.2	0.5	0.6	2.0	7.1
SMACT - MIX (schwierige Items)	RT	2003	445.9	1.1	1.7	1732	381.0	1.1	1.5
	F%	4.0	3.1	1.7	3.4	3.4	2.6	1.3	2.3

Tabelle A.10 – Stichprobenkennwerte der Studie 2: Testverlängerung und Testverlauf

		Sitzung 1				Sitzung 2			
		M	*SD*	Schiefe	Kurtosis	*M*	*SD*	Schiefe	Kurtosis
WMT		18.3	4.3	-0.8	0.3	20.7	3.7	-1.7	2.9
d2	GZ	504.9	75.6	-0.1	-0.7	552.6	69.4	-0.4	-0.5
	F%	3.1	3.0	2.3	7.1	1.7	1.9	2.7	10.3
SMACT (Kurzversion)	RT	1540	371	1.2	1.5	1139	191	0.6	-0.4
	F%	2.6	2.0	2.1	7.1	1.6	1.4	1.7	4.9
SMACT (Langversion)	RT	1286	237	0.7	0.4	1041	185	1.0	2.5
	F%	2.4	1.6	1.7	4.3	1.7	1.3	1.8	3.3
SMACT (Kurzversion)	RT	1165	217.6	0.8	0.6	1019	194	1.7	5.2
	F%	1.9	1.5	1.2	1.7	1.5	1.2	1.6	4.2

Tabelle A.11 – Stichprobenkennwerte der Studie 3: Testdiversifikation

	M	*SD*	Schiefe	Kurtosis
Wiener Matrizen-Test (WMT)				
Schlussfolgerndes Denken	15.1	3.5	-0.2	-0.2
Intelligenz-Struktur-Test (I-S-T 2000 R)				
Schlussfolgerndes Denken	111.6	17.5	0.0	-0.4
Verbale Intelligenz	36.3	6.3	-0.6	1.0
Numerische Intelligenz	40.3	9.8	-0.1	-0.6
Figurale Intelligenz	35.0	7.8	0.0	-0.5
Merken gesamt	16.3	4.0	-0.6	0.1
Merken verbal	8.0	2.3	-1.2	0.9
Merken figural	8.3	2.6	-0.3	-0.4
Satzergänzung	12.9	2.8	-0.4	0.4
Analogien	11.5	2.6	-0.2	0.6
Gemeinsamkeiten	12.0	3.0	-0.6	0.5
Rechenaufgaben	12.8	3.8	0.0	-0.6
Zahlenreihen	13.2	4.6	-0.3	-0.8
Rechenzeichen	14.3	3.6	-0.4	0.0
Figurenauswahl	12.3	3.3	-0.1	-0.4
Würfelaufgaben	11.3	4.3	-0.2	-0.8
Matrizenaufgaben	11.4	2.7	0.0	-0.2
d2-Aufmerksamkeits-Belastungs-Test (d2)				
GZ	497.2	77.3	0.0	0.2
F%	1.3	1.1	1.5	3.3
Frankfurter Adaptiver Konzentrationsleistungs-Test (FAKT-II)				
KL	120.9	26.2	-0.4	0.2
KG	87.8	5.1	-0.2	-0.4
KH	63.1	32.6	1.1	1.3

Tabelle A.12 – Stichprobenkennwerte der Studie 4: Validität in realer Auswahlsituation (I-S-T 2000 R)

	M	*SD*	Schiefe	Kurtosis
Intelligenz-Struktur-Test (I-S-T 2000 R)				
Schlussfolgerndes Denken	112.3	19.2	-0.4	-0.3
Verbale Intelligenz	37.0	6.3	0.2	3.2
Numerische Intelligenz	40.4	10.2	-0.2	-0.6
Figurale Intelligenz	34.9	8.5	-0.4	-0.6
Merken gesamt	16.7	4.3	-0.5	-0.2
Merken verbal	8.1	2.4	-1.2	0.6
Merken figural	8.6	2.8	-0.3	-0.6
Satzergänzung	12.9	3.0	-0.7	1.3
Analogien	12.1	2.5	-0.1	0.4
Gemeinsamkeiten	12.0	2.9	-0.8	0.8
Rechenaufgaben	12.6	3.6	0.0	-0.6
Zahlenreihen	13.6	4.8	-0.6	-0.6
Rechenzeichen	14.3	3.7	-0.2	-0.7
Figurenauswahl	11.8	3.8	-0.6	0.1
Würfelaufgaben	11.4	4.3	-0.4	-0.5
Matrizenaufgaben	11.7	2.7	-0.2	-0.3
Komplexer Konzentrationstest (KKT)				
Gesamt	0.3	0.9	0.3	0.1
Figurales Material	1188.9	292.4	0.2	-0.4
Numerisches Material	188.1	52.3	0.6	0.4
Verbales Material	225.9	62.0	0.5	0.0

Tabelle A.13 – Stichprobenkennwerte der Studie 5: Validität in realer Auswahlsituation (WIT-2)

	M	*SD*	Schiefe	Kurtosis
Wilde-Intelligenztest (WIT-2)				
Schlussfolgerndes Denken	105.3	7.4	-0.1	0.1
Sprachliches Denken	109.0	7.4	-0.7	0.4
Rechnerisches Denken	103.9	9.4	0.3	-0.4
Räumliches Denken	105.3	10.1	0.3	-0.3
Merkfähigkeit	14.1	3.0	-0.4	-0.2
Zahlenreihen	9.8	4.4	0.1	-0.3
Analogien	11.1	4.0	-0.2	-0.7
Abwicklungen	11.2	3.8	0.2	-0.3
Wortbedeutung	13.6	3.5	-0.7	0.5
Grundrechnen	8.9	4.1	0.3	-0.4
Rechenaufgaben	9.7	4.4	0.5	-0.3
Spiegelbilder	12.4	5.1	-0.4	-0.5
Komplexer Konzentrationstest (KKT)				
Gesamt	0.3	0.9	0.6	0.0
Figurales Material	1176.4	322.1	0.8	0.3
Numerisches Material	189.6	50.4	0.2	0.2
Verbales Material	229.0	55.5	0.3	-0.5

Anhang G: Korrelationen zwischen den Subtests des I-S-T 2000 R sowie des WIT-2 und Schulnoten

Tabelle A.14 – Ergebnisse der Studie 3: Korrelationen zwischen den Subtests des I-S-T 2000 R und Schulnoten

Skalen	Realschule			Abitur			Vordiplom
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren							
I-S-T- Merken verbal	-.03	-.06	-.02	.00	-.02	-.01	**-.23****
I-S-T- Merken figural	-.05	**-.10***	**-.11***	-.03	-.07	-.08	**-.22****
I-S-T- Satzergänzung	.06	.03	.06	.04	.05	.04	-.01
I-S-T- Analogien	-.05	-.02	.00	-.05	.01	-.05	-**.12***
I-S-T- Gemeinsamkeiten	.04	-.01	.01	.01	**.11***	.07	-.10
I-S-T- Rechenaufgaben	**-.11***	**-.23****	**-.21****	-.07	**-.22****	-.07	**-.30****
I-S-T- Zahlenreihen	-.01	**-.13****	-.08	-.05	**-.13****	**-.12***	**-.22****
I-S-T- Rechenzeichen	.01	-.08	-.07	.04	**-.09***	.00	-.10
I-S-T- Figurenauswahl	**-.09***	**-.22****	-.08	-.07	**-.13****	-.08	**-.17****
I-S-T- Würfelaufgaben	.03	-.07	-.03	.04	**-.09***	.01	-.01
I-S-T- Matrizenaufgaben	-.02	-.09	-.06	-.05	-.06	**-.16****	-.08

Anmerkungen. Korrelation zwischen I-S-T 2000 R und Realschulnoten-Gesamt n = 352; Realschule-Mathe n = 375; Realschule-Physik n = 374; Abitur-Gesamt n = 367; Abitur-Mathe n = 379; Abitur-Physik n = 259; Vordiplom n = 239; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

Tabelle A.15 – Ergebnisse der Studie 4: Korrelationen zwischen den Subtests des I-S-T 2000 R und Schulnoten im Kontext einer realen Bewerberauslesesituation an der Hochschule Mittweida

Skalen	Abitur			Bachelor
	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren				
I-S-T- Merken verbal	.01	-.04	-.11	-.28**
I-S-T- Merken figural	-.16	-.09	-.17	-.11
I-S-T- Satzergänzung	-.17*	-.12	-.23*	.00
I-S-T- Analogien	-.38**	-.24**	-.26*	-.09
I-S-T- Gemeinsamkeiten	-.22*	-.18*	-.20	.05
I-S-T- Rechenaufgaben	-.18*	-.25**	-.23*	.04
I-S-T- Zahlenreihen	-.09	-.18*	-.07	.04
I-S-T- Rechenzeichen	-.23*	-.15*	-.25*	-.01
I-S-T- Figurenauswahl	-.19*	-.18*	-.33**	.04
I-S-T- Würfelaufgaben	-.07	-.02	-.21	.07
I-S-T- Matrizenaufgaben	-.22*	-.19*	-.30*	.04

Anmerkungen. Korrelation zwischen I-S-T 2000 R und Abitur-Gesamt n = 101; Abitur-Mathe n = 97; Abitur-Physik n = 58; Bachelor n = 93; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

Tabelle A.16 – Ergebnisse der Studie 5: Korrelationen zwischen den Subtests des WIT-2 und Schulnoten im Kontext einer realen Bewerberauslesesituation an der Hochschule Mittweida

Skalen	Realschule			Abitur			Bachelor
	Gesamt	Mathe	Physik	Gesamt	Mathe	Physik	Gesamt
Psychometrische Standardverfahren							
WIT - Zahlenreihen	.01	**-.22***	-.03	-.09	-.14	**.32***	.02
WIT - Analogien	-.06	-.12	.07	-.08	.12	**.31***	-.10
WIT - Abwicklungen	-.12	-.15	-.10	-.08	.11	.07	-.02
WIT - Wortbedeutung	.07	.12	.06	.05	.08	.16	.02
WIT - Grundrechnen	.07	.10	.14	.03	-.06	**.28***	.02
WIT - Rechenaufgaben	.12	-.01	.02	.05	.05	.14	.11
WIT - Spiegelbilder	-.04	-.11	.00	-.09	-.18	.08	.05

Anmerkungen. Korrelation zwischen WIT-2 und Realschulnoten-Gesamt n = 66; Realschule-Mathe n = 72; Realschule-Physik n = 75; Abitur-Gesamt n = 91; Abitur-Mathe n = 85; Abitur-Physik n = 43; Bachelor n = 76; ** für $p < .01$ (einseitige Testung), * für $p < .05$ (einseitige Testung).

Anhang H: Vergleich der Korrelationskoeffizienten aus verschiedenen Stichproben zum I-S-T 2000 R

Tabelle A.17 – Vergleich der Korrelationskoeffizienten aus verschiedenen Stichproben zum I-S-T 2000 R

	r aus Studie 3 (N = 388)	*r* aus Studie 4 (N = 105)	Prüfgröße *z*	*p*-Wert
Intelligenz-Struktur-Test (I-S-T 2000 R)				
Schlussfolgerndes Denken	-.03	-.30	-2.51	0.01
Verbale Intelligenz	.00	-.33	-3.079	0.00
Numerische Intelligenz	-.04	-.19	-1.368	0.09
Figurale Intelligenz	-.03	-.19	-1.458	0.07
Merken gesamt	-.02	-.10	-.721	0.24
Merken verbal	.00	.01	0.09	0.46
Merken figural	-.03	-.16	-1.18	0.12
Satzergänzung	.04	-.17	-1.901	0.03
Analogien	-.05	-.38	-3.143	0.00
Gemeinsamkeiten	.01	-.22	-2.098	0.02
Rechenaufgaben	-.07	-.18	-1.005	0.16
Zahlenreihen	-.05	-.09	-.361	0.36
Rechenzeichen	.04	-.23	-2.462	0.01
Figurenauswahl	-.07	-.19	-1.098	0.14
Würfelaufgaben	.04	-.07	-.989	0.16
Matrizen	-.05	-.22	-1.559	0.06

Anmerkungen. Der Test auf Unterschiedlichkeit der Korrelationskoeffizienten (r) aus unterschiedlichen Stichproben basiert auf dem Verfahren von Eid, Gollwitzer & Schmitt, 2011, S. 547

Anhang I: Inkrementelle Regressionsanalysen der Studien 1 bis 5

Tabelle A.18 – Inkrementelle Regressionsanalysen der Studie 1

Kriterium	Prädiktoren	R	Änderung in R^2	p-Werte
Mathematiknote Realschule				
	APM	.29	.086	.012
	APM; Test d2-GZ; SMACT-RT-Gesamtleistung	.36	.045	.172
Mathematiknote Abitur				
	APM	.30	.092	.013
	APM; Test d2-GZ; SMACT-RT-Gesamtleistung	.40	.065	.095

Anmerkungen. Die Berechnung erfolgte auf Basis einer zweistufigen, blockweisen, multiplen Regressionsanalyse. Im ersten Schritt erfolgte die Einfachvorhersage zwischen Prädiktor- und Kriteriumsvariable. Im zweiten Schritt erfolgte durch Hinzugabe weiterer Prädiktorvariablen die Prüfung auf einen inkrementellen Beitrag bezüglich des zu messenden Kriteriums.

Tabelle A.19 – Inkrementelle Regressionsanalysen der Studie 2

Kriterium	Prädiktoren	R	Änderung in R^2	p-Werte
Mathematiknote Realschule				
	WMT	.30	.089	.005
	WMT; Test d2-GZ; SMACT-RT-Kurzversion	.45	.109	.005
Mathematiknote Abitur				
	WMT	.25	.063	.019
	WMT; Test d2-GZ; SMACT-RT-Kurzversion	.31	.029	.226

Anmerkungen. Die Berechnung erfolgte auf Basis einer zweistufigen, blockweisen, multiplen Regressionsanalyse. Im ersten Schritt erfolgte die Einfachvorhersage zwischen Prädiktor- und Kriteriumsvariable. Im zweiten Schritt erfolgte durch Hinzugabe weiterer Prädiktorvariablen die Prüfung auf einen inkrementellen Beitrag bezüglich des zu messenden Kriteriums.

Tabelle A.20 – Inkrementelle Regressionsanalysen der Studie 3

Kriterium	Prädiktoren	R	Änderung in R^2	p-Werte
Mathematiknote Realschule				
	WMT	.14	.019	.008
	WMT; I-S-T- Rechenaufgaben	.25	.044	.000
Mathematiknote Abitur				
	WMT	.14	.020	.006
	WMT; I-S-T- Rechenaufgaben	.24	.036	.000

Anmerkungen. Die Berechnung erfolgte auf Basis einer zweistufigen, blockweisen, multiplen Regressionsanalyse. Im ersten Schritt erfolgte die Einfachvorhersage zwischen Prädiktor- und Kriteriumsvariable. Im zweiten Schritt erfolgte durch Hinzugabe einer weiteren Prädiktorvariable die Prüfung auf einen inkrementellen Beitrag bezüglich des zu messenden Kriteriums.

Tabelle A.21 – Inkrementelle Regressionsanalysen der Studie 3 (LMI)

Kriterium	Prädiktoren	R	Änderung in R^2	p-Werte
Mathematiknote Realschule				
	WMT	.14	.020	.010
	WMT; Beharrlichkeit; Engagement; Selbstkontrolle	.26	.048	.001
Mathematiknote Abitur				
	WMT	.14	.020	.009
	WMT; Beharrlichkeit; Engagement; Selbstkontrolle	.20	.020	.071
Vordiplom-Note				
	WMT	.29	.082	.000
	WMT; Beharrlichkeit; Engagement; Selbstkontrolle	.38	.060	.003

Anmerkungen. Die Berechnung erfolgte auf Basis einer zweistufigen, blockweisen, multiplen Regressionsanalyse. Im ersten Schritt erfolgte die Einfachvorhersage zwischen Prädiktor- und Kriteriumsvariable. Im zweiten Schritt erfolgte durch Hinzugabe weiterer Prädiktorvariablen die Prüfung auf einen inkrementellen Beitrag bezüglich des zu messenden Kriteriums.

Tabelle A.22 – Inkrementelle Regressionsanalysen der Studie 4

Kriterium	Prädiktoren	R	Änderung in R^2	p-Werte
Gesamtnote Abitur				
	I-S-T - Schlussfolgerndes Denken	.30	.088	.003
	I-S-T - Schlussfolgerndes Denken; KKT - Numerisches Material	.34	.025	.101

Anmerkungen. Die Berechnung erfolgte auf Basis einer zweistufigen, blockweisen, multiplen Regressionsanalyse. Im ersten Schritt erfolgte die Einfachvorhersage zwischen Prädiktor- und Kriteriumsvariable. Im zweiten Schritt erfolgte durch Hinzugabe einer weiteren Prädiktorvariable die Prüfung auf einen inkrementellen Beitrag bezüglich des zu messenden Kriteriums.

Tabelle A.23 – Inkrementelle Regressionsanalysen der Studie 5

Kriterium	Prädiktoren	R	Änderung in R^2	p-Werte
Physiknote Abitur				
	WIT - Schlussfolgerndes Denken	.28	.078	.071
	WIT - Schlussfolgerndes Denken; KKT - Numerisches Material	.53	.203	0.02

Anmerkungen. Die Berechnung erfolgte auf Basis einer zweistufigen, blockweisen, multiplen Regressionsanalyse. Im ersten Schritt erfolgte die Einfachvorhersage zwischen Prädiktor- und Kriteriumsvariable. Im zweiten Schritt erfolgte durch Hinzugabe einer weiteren Prädiktorvariable die Prüfung auf einen inkrementellen Beitrag bezüglich des zu messenden Kriteriums.

Anhang J: Erklärung

Erklärung gemäß § 6 (2) Punkt 5 der Promotionsordnung der Fakultät für Human- und Sozialwissenschaften der Technischen Universität Chemnitz:

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die von mir angegebenen Hilfsmittel benutzt habe.

Frank Schumann

Chemnitz, den 02.11.2015