Entwicklung und Erprobung eines Kurztests zum Konditionalen Schlussfolgern

Dissertation
zur Erlangung des akademischen Grades
doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Fakultät für Sozial- und Verhaltenswissenschaften der
Friedrich-Schiller-Universität Jena
von Dipl.-Psych. Hendryk Frank Böhme
geboren am 19. September 1977 in Altenburg

Gutachter

1. Prof. Dr. Rolf Steyer

2. Prof. Dr. Edgar Erdfelder

Tag der mündlichen Prüfung: 16. August 2010

Vorwort

Von April 2006 bis März 2008 betreute ich ein Forschungsprojekt des Lehrstuhls für Methodenlehre und Evaluationsforschung der Friedrich-Schiller-Universität Jena. Ziel dieses Projektes war die Entwicklung eines mehrdimensionalen Fähigkeitstests für Chipdesigner, der zur Personalentwicklung eingesetzt werden sollte. Im Rahmen der Arbeiten an diesem Projekt kristallisierte sich eine Lücke in der Diagnostik Schlussfolgernden Denkens als geeignetes Thema für eine Dissertation heraus und damit begann meine Arbeit am KKS, dem Kurztest zum Konditionalen Schlussfolgern.

Diese Arbeit wäre jedoch nicht möglich gewesen ohne die Unterstützung vieler Anderer, bei denen ich mich an dieser Stelle bedanken möchte. Besonderer Dank gilt meinem Betreuer Rolf Steyer, der mir im Rahmen des o.g. Forschungsprojektes die Freiräume für die Erforschung dieses Themas ließ, der mich bei Fragen und Problemen nach besten Kräften unterstützte und der mir in der schwierigen Anfangsphase der Arbeit mit seinem „Blick von außen“ deren konzeptuelle Stärken vor Augen führte und mich so davor bewahrte, dieses vielversprechende Thema aufzugeben. Des Weiteren danke ich Edgar Erdfelder, dem zweiten Gutachter meiner Arbeit. Sein Interesse am Thema und seine wertvollen Hinweise zur Begrenzung des Arbeitsumfangs hatten sowohl auf meine Motivation als auch auf mein Zeitmanagement eine positive Wirkung. In gleicher Weise gebührt ein Dank unserem Projektpartner und Auftraggeber, der Cadence Design Systems GmbH, und dabei insbesondere dem Projektmanager Eyck Jentzsch, der uns große Freiheiten bei den Forschungsarbeiten einräumte. Ich bedanke mich ebenso bei den studentischen Hilfskräften Anna Grohmann, Tina Urbach, Juliane Graf, Anna Wollny und Erik Sengewald, deren individuelle Stärken und unermüdliche Motivation ein entscheidender Erfolgsfaktor für das Gelingen unseres gemeinsamen wissenschaftlichen Forschens waren. Ich danke Christiane Fiege, Steffi Pohl, Norman Rose und Tim Loßnitzer für ihre konstruktiven Rückmeldungen zu einzelnen Teilen der Arbeit sowie dem Doktorandenkolloquium „Disskont“ und der Arbeitsgruppe „Denken“. Neben dem fachlichen Austausch habe ich in den vergangenen Jahren auch den großen Wert emotionaler Unterstützung schätzen gelernt. Am stärksten erfuhr ich das von meinen Eltern und natürlich von Stephanie, die mir oft den Rücken für die vielen notwendigen Wochenendsitzungen freigehalten hat; ich hoffe, dass ich mich bald revanchieren kann.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

Development and Evaluation of a Conditional-Reasoning (Short) Test

The processes underpinning logical reasoning with conditionals (if-then propositions) are frequently subject to general psychological examination. In the field of differential psychology, however, the ability to make logically correct inferences from conditionals has rarely been addressed. As a result, there are also no psychodiagnostic instruments for the assessment of such a construct. The aim of the present paper was therefore to develop and evaluate a psychodiagnostic test of conditional reasoning. To this end, a level model which is based on cognitive-psychological theories and which allows differential-psychological conclusions to be drawn regarding conditional reasoning is presented. This level model is extended to include conditionals with additional negations and is thus applicable to the so-called negations paradigm. In the resulting *extended level-model*, four levels of *sophistication of conditional reasoning* (*SCR*) are assumed. Based on theoretical derivations and practical demands, 16 conditional-reasoning items were constructed and presented to a sample of 905 participants. Since particular response patterns (more specifically, response-pattern probabilities) are assumed for the four levels, the extended level-model can be specified as a latent class model. Data were thus examined using latent class analysis. In line with hypotheses, four classes of the latent variable SCR were identified. Following a theory-based ordering of these classes, the expected positive relationship with general reasoning was revealed. Based on these promising findings, implications for further testing of this 16-item set were derived with a particular focus on the determination and evaluation of test quality criteria. In addition to the major properties of objectivity, reliability, and validity, eight minor quality criteria (e.g., availability of norm data, usefulness, and economy) were examined. This examination was based on theoretical derivations which were empirically tested at crucial points in three further studies. Results demonstrate that examined quality criteria were met (almost) without exception. The 16-item set may thus justifiably be referred to as a ‘psychological test’. In summary, it can be concluded that it is possible, based on the extended level-model, to introduce a construct sophistication of conditional reasoning, which can be objectively, reliably, validly, and highly economically measured using the developed *Conditional-Reasoning (Short) Test*.

Zusammenfassung

Die Prozesse, die dem logischen Schlussfolgern bei Konditionalaussagen (Wenn-dann-Aussagen) zugrunde liegen, sind häufig Gegenstand allgemeinpsychologischer Untersuchungen. In der Differenziellen Psychologie wird die Fähigkeit, logisch korrekte Inferenzen bei Konditionalaussagen zu ziehen, hingegen kaum thematisiert. Folglich existieren auch keine psychodiagnostischen Testverfahren zur Messung eines solchen Konstruktes. Ziel dieser Arbeit ist daher die Entwicklung und Erprobung eines psychodiagnostischen Tests zum Konditionalen Schlussfolgern. Dazu wird ein auf kognitionspsychologischen Theorien aufbauendes Stufen-Modell vorgestellt, das differenzialpsychologische Aussagen zum Konditionalen Schlussfolgern erlaubt. Dieses Stufen-Modell wird auf Konditionalaussagen mit zusätzlichen Negationen erweitert und ist damit auf das sog. Negationsparadigma anwendbar. In dem resultierenden *erweiterten Stufen-Modell* werden vier Stufen der *Sophistiziertheit Konditionalen Schlussfolgerns* (*SKS*) postuliert. Abgeleitet aus den theoretischen Überlegungen und aufgrund pragmatischer Vorgaben wurden 16 Aufgaben zum Konditionalen Schlussfolgern konstruiert, die einer Stichprobe von 905 Personen vorgegeben wurden. Da für die vier Stufen spezifische Antwortmuster(-wahrscheinlichkeiten) postuliert werden, lässt sich das erweiterte Stufen-Modell als Latente-Klassen-Modell präzisieren. Folglich werden die Daten mittels Analyse latenter Klassen ausgewertet. Es resultieren vier hypothesenkonforme Klassen der latenten Variable SKS. Nach dem theoriegeleiteten Ordnen der Klassen zeigt sich ein erwartungsgemäß positiver Zusammenhang mit Reasoning. Aus diesen vielversprechenden Ergebnissen werden Implikationen für die weitere Erprobung dieser 16 Items und insbesondere für die Bestimmung von Testgütekriterien abgeleitet. Neben den klassischen Hauptgütekriterien Objektivität, Reliabilität und Validität werden zudem acht Nebengütekriterien (darunter Normierung, Nützlichkeit und Testökonomie) betrachtet und für die vorliegenden 16 Items überprüft. Dies erfolgt durch theoretische Herleitungen, die an entscheidenden Stellen in drei weiteren Studien empirisch überprüft werden. Die betrachteten Gütekriterien können (fast) ausnahmslos als erfüllt angesehen werden, sodass es gerechtfertigt scheint, die 16 Items als „Psychologischen Test“ zu bezeichnen. Insgesamt kann festgehalten werden, dass es möglich ist, auf Basis des erweiterten Stufen-Modells ein Konstrukt Sophistiziertheit Konditionalen Schlussfolgerns einzuführen, das mit dem entwickelten *Kurztest zum Konditionalen Schlussfolgern* (*KKS*) objektiv, reliabel, valide und äußerst ökonomisch erhoben werden kann.

1 Einleitung

„Logik ist die Anatomie des Denkens.“ (John Locke, 1632-1704)

Eines der Hauptergebnisse von Metaanalysen zur Vorhersage von Berufserfolg (z.B. Salgado et al., 2003; Schmidt & Hunter, 1998) ist die zentrale Bedeutung von Intelligenz. Andere Prädiktoren haben kaum zusätzliche Vorhersagekraft. Betrachtet man aktuelle Studien zu den geforderten Fähigkeiten in der heutigen Arbeitswelt (z.B. Beck, 2007), dann wird nicht der Begriff „Intelligenz“ verwendet, sondern meist „Analytische oder Logische Denkfähigkeit“. Damit wird differenzierter beschrieben, welche Intelligenz damit letztlich gemeint ist. Auch im Rahmen eines Forschungsprojektes[1] des Lehrstuhls für Methodenlehre und Evaluationsforschung der Friedrich-Schiller-Universität Jena wurde diese Form der Intelligenz als relevant für die Tätigkeit des Chipdesigners identifiziert (vgl. Böhme & Steyer, 2008). Im Rahmen der Logischen Denkfähigkeit scheint für diese Tätigkeit von besonderer Bedeutung, logisch korrekte Schlussfolgerungen bei Konditionalaussagen (Wenn-dann-Aussagen) zu ziehen. Man stelle sich dazu vor, bestimmte Regeln oder Zusammenhänge seien in Form solcher Konditionalaussagen gegeben, bspw. „Wenn beim Programmieren der Fehler *A* begangen wurde, dann erscheint am Ende einer Simulation die Fehlermeldung *B*“. Des Weiteren wird ein bestimmtes Faktum festgestellt. So erscheint bspw. am Ende einer Simulation die Fehlermeldung *B*. Nun stellt sich die Frage, welche Schlussfolgerungen logisch korrekt sind. Ist in diesem Falle bspw. „Fehler *A* wurde begangen“ ein korrekter logischer Schluss?

Ziel des erwähnten Forschungsprojektes war es, die im Rahmen einer Anforderungsanalyse identifizierten, für Chipdesigner tätigkeitsrelevanten Fähigkeiten und Kompetenzen messbar zu machen (vgl. Böhme & Steyer, 2008). Unter der Annahme, dass das Ziehen logisch korrekter Schlüsse bei Konditionalaussagen eine Fähigkeit oder eine Kompetenz darstellt, wurde eine Recherche nach entsprechenden Testverfahren durchgeführt. Das überraschende Ergebnis war, dass – zumindest im Erwachsenenbe-

[1] Projekt „Konstruktion Psychometrischer Fähigkeitstests für Chipdesigner“ (04/2006 – 03/2008, Kurzbeschreibung unter http://www.metheval.uni-jena.de/projekte_pp2.php)

reich – keinerlei deutschsprachige Tests zur Erfassung eines solchen Konstruktes existieren. Daraus resultierte die Notwendigkeit, einen entsprechenden Test selbst zu konstruieren. Dies ist das Ziel der vorliegenden Arbeit, in der theoretische Fundierung, konzeptionelle Entwicklung, praktische Umsetzung und empirische Erprobung dieses Tests vorgestellt werden. Die Einbettung in das erwähnte Forschungsprojekt ist insofern relevant, als generell bei einer Testkonstruktion häufig Entscheidungen zwischen verschiedenen, mitunter ähnlich attraktiven Alternativen getroffen werden müssen. Bei der in der Folge beschriebenen Testkonstruktion erfolgen einzelne Entscheidungen recht pragmatisch, das heißt vor dem Hintergrund des Forschungsprojektes und entsprechend den Wünschen des Auftraggebers. Dies wird an den jeweiligen Stellen der Arbeit vermerkt. Generell werden an die vorliegende Arbeit zwei Forderungen gestellt, die zwei wesentliche Ansprüche an das wissenschaftliche Vorgehen bei der Testkonstruktion repräsentieren. Die erste Forderung ist, dass die Testkonstruktion kognitionspsychologisch fundiert sein soll, die zweite, dass zur empirischen Hypothesenprüfung Latente-Variablen-Modelle verwendet werden sollen, um dem Messfehlerproblem von vornherein Rechnung zu tragen. Beide Forderungen werden an späterer Stelle ausführlich beschrieben und begründet. Es folgt ein Überblick über den Aufbau der Arbeit.

Nach dieser kurzen Einführung behandelt der zweite Teil der Arbeit die Frage nach der Konstruierbarkeit einer Fähigkeit oder Kompetenz, logisch korrekte Schlüsse bei Konditionalaussagen zu ziehen. Dazu werden aus kognitionspsychologischen Befunden differenzialpsychologische Überlegungen abgeleitet und zu einem (Kompetenz-)Stufen-Modell zusammengefasst. Auf diesem basiert die Entwicklung von 16 Items, die den *Kurztest zum Konditionalen Schlussfolgern* (*KKS*) bilden. Zum Abschluss dieses zweiten Kapitels wird eine empirische Studie vorgestellt, in der die zentralen Annahmen überprüft werden, die diesem Stufen-Modell und damit der Konstruktion des KKS zugrunde liegen. Daraus resultieren Implikationen für die weitere empirische Erprobung des KKS. Diese ist Gegenstand des dritten Teils der Arbeit, in dem insbesondere die Bestimmung seiner Testgütekriterien thematisiert wird. Schließlich werden im vierten Teil der Arbeit die empirischen Befunde zur Entwicklung und Erprobung des KKS kritisch diskutiert, ebenso wie die verwendeten Methoden. Daraus werden Implikationen für die Forschung und für die psychologische Praxis abgeleitet. Im fünften Teil ist die verwendete Literatur aufgelistet.

2 Entwicklung des Tests

In der psychometrischen Tradition der Testkonstruktion ist es üblich, dass Itemparameter wie bspw. Itemschwierigkeiten mithilfe der Lösungshäufigkeiten in einer repräsentativen Stichprobe geschätzt werden. Eine Aufgabe wird dann als schwierig bezeichnet, wenn wenige Personen sie lösen. Wiederum lösen sie aber nur wenige Personen, weil sie schwierig ist. Bereits Berg und Schaarschmidt (1984) kritisieren, dass man empirisch findet, was aus der Konstruktionsstrategie notwendigerweise folgt. Daher fordern bspw. Beckmann und Guthke (1999) oder auch Wilhelm (2000) eine stärker theoretisch-inhaltlich begründete Itempool-Konstruktion und damit Itemparameterbestimmung. Dies stellt die erste Forderung an die vorliegende Arbeit dar (vgl. auch Kapitel 1). Die in dieser Arbeit beschriebene Testentwicklung soll kognitionspsychologisch fundiert sein, resultierend in klaren Vorhersagen, welche Aufgabenmerkmale ein Item für welche Personen schwierig machen.

Doch im Rahmen dieser Testkonstruktion ist mehr als nur die kognitionspsychologische Fundierung von Bedeutung. Zunächst ist das interessierende Konstrukt zu bestimmen. Dieses wird bereits in der Einleitung als *Ziehen logisch korrekter Schlüsse bei Konditionalaussagen* vorgestellt und in der Folge als *Konditionales Schlussfolgern* bezeichnet. Nach einigen Ausführungen zu logischen Schlüssen bei Konditionalaussagen (sog. *Konditionalen Schlüssen*) aus Perspektive der Aussagenlogik werden Theorien zu den kognitiven Prozessen vorgestellt, die beim Konditionalen Schlussfolgern ablaufen. Schließlich werden differenzialpsychologische Befunde zum Konditionalen Schlussfolgern thematisiert, um der Frage nachzugehen, ob so etwas wie eine Fähigkeit zum Konditionalen Schlussfolgern konstruiert werden kann. Dazu erfolgen zudem ein kurzer Abriss aktueller Intelligenzstrukturtheorien sowie eine Betrachtung potenziell relevanter diagnostischer Testverfahren. Diese Begriffsbestimmung Konditionalen Schlussfolgerns aus verschiedenen Perspektiven ist Gegenstand des ersten Abschnitts dieses Kapitels. Aufbauend auf den bis dahin angestellten Überlegungen wird anschließend ein (Kompetenz-)Stufen-Modell zum Konditionalen Schlussfolgern vorgestellt und weiterentwickelt, welches die Basis für die Konstruktion von 16 Items bildet, die

die Identifikation der beschriebenen Stufen zum Ziel haben. Zum Abschluss dieses Kapitels wird eine erste empirische Erprobung des so konstruierten Tests präsentiert, in der die zentralen Annahmen des Stufen-Modells überprüft und die Ergebnisse vor dem theoretischen Hintergrund diskutiert werden. Daraus resultieren Implikationen für die Weiterentwicklung dieses Tests sowie für weitere empirische Erprobungen, insbesondere die Bestimmung von Testgütekriterien.

2.1 Konditionales Schlussfolgern

Ziel der vorliegenden Arbeit ist die Entwicklung eines Tests, mit dem eine Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern erfasst werden kann. Trotz intensiver Literaturrecherche lässt sich jedoch keine Konzeption einer solchen Fähigkeit oder Kompetenz finden (siehe auch Abschnitt 2.1.4), gleichwohl das Ziehen korrekter Schlussfolgerungen und insbesondere das „Nicht-Ziehen“ unlogischer Schlüsse weit über die Arbeitswelt hinaus von zentraler Bedeutung ist. So führt bspw. Dörner (1989) aus, dass das Unglück von Tschernobyl die Folge unlogischer (zum Teil auch Konditionaler) Schlussfolgerungen in der Leitzentrale darstellt. Die praktische Relevanz einer solchen Fähigkeit oder Kompetenz steht also außer Frage, sodass eine Konstruktion, auch über die Bedeutung für die Tätigkeit des Chipdesigners (vgl. Kapitel 1) hinaus, sinnvoll scheint. Um − mit dem Ziel einer Testentwicklung − eine Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern konstruieren zu können, soll das interessierende Konditionale Schlussfolgern zunächst aus verschiedenen Perspektiven vorgestellt werden. Dies ist Gegenstand des folgenden Abschnitts.

Ein erster wichtiger Schritt ist, zu entscheiden, welche Ergebnisse beim Konditionalen Schlussfolgern als korrekt gewertet werden. Dazu wird Konditionales Schlussfolgern zunächst aus Perspektive der Aussagenlogik betrachtet. Um der ersten Forderung an die vorliegende Arbeit (vgl. Abschnitt 2) Rechnung zu tragen, werden anschließend die kognitiven Prozesse thematisiert, die beim Konditionalen Schlussfolgern ablaufen. Hierzu werden die beiden Theorien vorgestellt und verglichen, die den aktuellen Stand der kognitionspsychologischen Forschung zum Konditionalen Schlussfolgern dominieren. Es folgen Ausführungen zu interindividuellen Unterschieden im Konditionalen Schlussfolgern und damit ein Perspektivwechsel hin zur Differenziellen Psychologie. Darauf baut dann die (theoretische) Einordnung einer Fähigkeit zum Kon-

ditionalen Schlussfolgern in verschiedene Intelligenzstrukturtheorien auf. Schließlich werden noch publizierte diagnostische Testverfahren zur Erfassung Konditionalen Schlussfolgerns vorgestellt. Letztlich sollen über die verschiedenen Perspektiven hinweg Implikationen für die Testkonstruktion abgeleitet werden, vorausgesetzt, Konditionales Schlussfolgern stellt überhaupt ein „testgeeignetes" Konstrukt dar.

2.1.1 Logisch korrekte Lösungen beim Konditionalen Schlussfolgern

Als ein erster wichtiger Punkt im Rahmen einer Testkonstruktion zum Konditionalen Schlussfolgern ist zu entscheiden, welche Ergebnisse des Schlussfolgerungsprozesses als korrekt bewertet werden. Dies müssen keineswegs *logisch* korrekte Schlussfolgerungen sein. So lehnen bspw. Gigerenzer und Gaissmaier (2006) die formale Logik als normatives Modell ab und gehen in ihrem Ansatz davon aus, dass es notwendig ist, Denkprozesse an Umweltbedingungen und soziale Strukturen anzupassen. Dazu bedarf es adaptiver Heuristiken, die mit logischer Korrektheit nichts zu tun haben müssen, aber dennoch als rational betrachtet werden. Auf Grundlage derartiger Annahmen einen Test zu konstruieren, scheint jedoch äußerst schwierig, da dann bspw. die Korrektheit einer Lösung situationsabhängig sein kann. Ein einheitliches System, unter welchen Bedingungen eine Lösung korrekt ist, würde unweigerlich sehr komplex. Eine weitere Möglichkeit, die Korrektheit von Lösungen beim Konditionalen Schlussfolgern zu bestimmen, stellt die Aussagenlogik dar. Menschliches Verhalten, das mit den Normen der Logik übereinstimmt, bezeichnet Anderson (1990, 1991) als *normativ rationales Verhalten*. Stanovich (1999) spricht in diesem Zusammenhang von der *Apologistischen Position*, bei der lediglich normativ rationales Verhalten als „gutes" Schlussfolgern bezeichnet wird. Der Vorteil der Aussagenlogik als Bezugssystem liegt darin, dass die Korrektheit von Lösungen eindeutig festgelegt ist. Bevor jedoch logisch korrekte Lösungen beim Konditionalen Schlussfolgern thematisiert werden, sollen wichtige aussagenlogische Prinzipien und Begriffe (siehe bspw. auch Kutschera & Breitkopf, 2007) eingeführt werden, die im Rahmen des sog. *Logischen Denkens* von Bedeutung sind. Auf viele dieser Prinzipien und Begriffe wird im weiteren Verlauf der Arbeit wieder Bezug genommen.

Prinzipiell geht es beim Logischen Denken (oder Schlussfolgern) um die Anwendung eines geregelten Verfahrens (sog. *Kalküle*), um von bestimmten Aussagen, den sog. *Prämissen* eines Schlusses, zu einer weiteren Aussage, der sog. *Konklusion* des

Schlusses, zu gelangen. Bei einem korrekten Schluss gewährleistet die Wahrheit der Prämissen die Wahrheit der Konklusion. Unter *Korrektheit eines Schlusses* oder besser unter der *Korrektheit der Schlussform* ist dabei die Übereinstimmung mit den Gesetzen und Regeln der formalen Logik zu verstehen. „Technisch gesprochen" ist ein solcher Logischer Schluss immer auch ein *Deduktiver Schluss* (Knauff, 2006). Als Folge davon ist es durchaus angebracht, auch Logisches Denken und *Deduktives Denken* (bzw. *Deduktives Schlussfolgern*) synonym zu gebrauchen. Formal wird Deduktives Denken als notwendig wahres, gehaltkonservierendes Schlussfolgern definiert. Es ist von *Induktivem Denken* abzugrenzen, dem nicht notwendig wahren, dafür aber gehalterweiternden Schlussfolgern (siehe z.B. Skyrms, 1989; Stegmüller, 1996). Beim Induktiven Denken ist die Konklusion immer mit einer gewissen Unsicherheit verbunden. Es handelt sich eigentlich um eine „unzulässige Verallgemeinerung" (Beckmann & Guthke, 1999). Bei einem gültigen deduktiven Schluss ist die Konklusion hingegen keine Erweiterung der Prämisseninformation, sondern gilt tautologisch, das heißt, wahre Voraussetzungen führen immer zu einem wahren Schluss. Dadurch haben deduktive Schlüsse eine wichtige Eigenschaft, die sie von induktiven Schlüssen unterscheidet: Deduktive Schlüsse sind beweisbar.

Die drei wichtigsten Varianten des Deduktiven Denkens sind nach Knauff (2006) *Syllogistisches Schlussfolgern*, *Relationales Schlussfolgern* und Konditionales Schlussfolgern. Unter Syllogistischem Schlussfolgern wird das Schlussfolgern bei Aussagen mit sog. *Quantoren* (alle, einige, keine) verstanden (für einen Überblick siehe z.B. Knauff, 2006). Als Relationales Schlussfolgern wird das Schlussfolgern bei Aussagen bezeichnet, in denen Elemente durch bestimmte Relationen (meist räumliche oder zeitliche) miteinander verknüpft sind[2]. Konditionales Schlussfolgern wird dem *Propositionalen Schlussfolgern* zugeordnet. Darunter versteht man das Schlussfolgern mit sog. *Junktoren*. Junktoren sind Operatoren, die Aussagen (im einfachsten Fall zwei *atomare Aussagen*[3]) miteinander verknüpfen. Durch die Wahrheitswerte der einzelnen Aussagen wird der Wahrheitswert der verknüpften Aussage eindeutig determiniert. Beim in dieser Arbeit interessierenden Konditionalen Schlussfolgern ist diese Verknüpfung das sog. *Konditional* („wenn…, dann"). Außerdem gibt es noch vier weitere derartige Verknüpfungen in der Aussagenlogik: die *Negation* („nicht"), die *Äquivalenz* („wenn…, dann

[2] Aus den Aussagen „A ist links von B" und „B ist links von C" kann bspw. geschlussfolgert werden, dass A links von C ist. Dabei handelt es sich um einen (räumlich) relationalen logischen Schluss.
[3] Unter einer „atomaren Aussage" wird in der Aussagenlogik eine Aussage verstanden, die nicht aus anderen Aussagen zusammengesetzt ist.

und nur genau dann“), die *Konjunktion* („und“) sowie die *Disjunktion* („oder“). Letztere lässt sich noch in die inklusive Disjunktion („Aussage A oder Aussage B oder beide“) und die exklusive Disjunktion („Aussage A oder Aussage B, nicht aber beide“) unterteilen.

Nachdem nun wichtige Begriffe eingeführt sind, sollen die aussagenlogischen Grundlagen speziell des Konditionalen Schlussfolgerns vorgestellt werden. Ausgangspunkt ist der Begriff des *Syllogismus*, unter dem man die Ableitung einer Konklusion aus einer Prämisse unter Zuhilfenahme einer weiteren Prämisse versteht (Beckmann & Guthke, 1999). Handelt es sich bei einer dieser beiden Prämissen um eine Konditionalaussage, sprechen Beckmann und Guthke (1999) von einem *Konditionalen Syllogismus*. Die Basis bildet demnach stets eine Konditionalaussage. Ein dafür im deutschen Sprachraum häufig gebrauchtes Beispiel soll auch hier illustrierend verwendet werden. Dabei wird folgende Konditionalaussage betrachtet:

Wenn es regnet, dann ist die Straße nass.

Die auf „wenn“ folgende Aussage („es regnet“) wird auch *Vorderglied* oder *Antezedens* genannt, die auf „dann“ folgende Aussage („die Straße ist nass“) *Hinterglied* oder *Konsequenz*. Der Junktor „wenn…, dann“ wird in der Aussagenlogik als *Implikation* bezeichnet. Das Antezedens („es regnet“) *impliziert* also die Konsequenz („die Straße ist nass“). Die übliche Notation für Aussagen sind in der Aussagenlogik kleine Buchstaben (meist „p“ und „q“), für die Implikation ist es ein in Richtung Konsequenz gerichteter Pfeil (→). Die symbolische Darstellung einer einfachen Konditionalaussage sieht demnach wie folgt aus: $p \rightarrow q$. Für das Beispiel entspricht p…„es regnet“ und q…„die Straße ist nass“. Die beim Konditionalen Schlussfolgern stets zugrundeliegende Konditionalaussage ($p \rightarrow q$) wird als *Haupt-* oder *Majorprämisse* bezeichnet.

Damit ist nun zwar die Konditionalaussage genauer beschrieben, für einen Konditionalen Syllogismus bedarf es jedoch noch einer weiteren Prämisse, der sog. *Neben-* oder *Minorprämisse*. Dabei wird entweder das Antezedens bestätigt oder negiert, oder aber die Konsequenz wird bestätigt oder negiert. Für das Beispiel ergeben sich daher die folgenden vier möglichen Nebenprämissen (die propositionale Annotation „¬“ bezeichnet dabei die Negation der Aussage p bzw. q):

1. Es regnet. (*p*)

2. Es regnet nicht. (¬*p*)

3. Die Straße ist nass. (*q*)

4. Die Straße ist nicht nass. (¬*q*)

Durch Kombination der Hauptprämisse mit diesen vier möglichen Nebenprämissen ergeben sich die vier „klassischen" Konditionalen Syllogismen, die auch als die *vier* (*logischen*) *Schlussfiguren des Konditionalen Schlussfolgerns* bezeichnet werden. Sie werden − inklusive ihrer (logisch) zulässigen Konklusionen − in der Folge vorgestellt.

***Modus Ponens* (*MP*)**

Mit Modus Ponens wird der logische Schluss bezeichnet, bei dem als Nebenprämisse das Antezedens bestätigt wird (*p*). Für das Beispiel wäre das folgende Prämissenkombination:

Beispiel:	oder allgemein:
Hauptprämisse: Wenn es regnet, dann ist die Straße nass.	$p \rightarrow q$
Nebenprämisse: Es regnet.	p
---	---
Logischer Schluss: Die Straße ist nass.	q

Der korrekte logische Schluss beim Modus Ponens ist die Bestätigung der Konsequenz. Generell gilt der Modus Ponens als die Inferenz beim Konditionalen Schlussfolgern, die von Probanden am häufigsten korrekt gelöst wird (z.B. Evans, Newstead & Byrne, 1993).

***Negation des Antezedens* (*NA*)**

Wie der Name schon sagt, besteht die Nebenprämisse bei dieser logischen Schlussfigur in der Negation des Antezedens (¬*p*), was für das Beispiel bedeutet:

Beispiel:	oder allgemein:
Hauptprämisse: Wenn es regnet, dann ist die Straße nass.	$p \rightarrow q$
Nebenprämisse: Es regnet nicht.	$\neg p$

Für dieses Schlussschema ist weder *q* noch ¬*q* alleiniger gültiger Schluss. Korrekt ist vielmehr die exklusive Disjunktion[4] *q oder* ¬*q*. Angewendet auf das Beispiel bedeutet das: „Die Straße ist nass oder die Straße ist nicht nass." Inhaltlich begründet kann es durchaus andere Bedingungen geben, die dazu führen, dass die Straße nass ist, obwohl es nicht regnet.

***Bestätigung der Konsequenz* (*BK*)**

Auch für diese Schlussfigur lässt sich bereits aus der Bezeichnung ableiten, dass die Nebenprämisse in der Bestätigung der Konsequenz (*q*) besteht, was für das Beispiel heißt:

Beispiel:	oder allgemein:
Hauptprämisse: Wenn es regnet, dann ist die Straße nass.	$p \rightarrow q$
Nebenprämisse: Die Straße ist nass.	q

Weder *p* noch ¬*p* ist hier alleiniger gültiger Schluss. Stattdessen ist ebenso wie bei der Negation des Antezedens eine exklusive Disjunktion der korrekte logische Schluss, also *p oder* ¬*p*. Auf das inhaltliche Beispiel angewendet, kann es andere Bedingungen geben, die dazu geführt haben, dass die Straße nass ist. Daher kann nicht eindeutig geschlussfolgert werden, dass es regnet.

***Modus Tollens* (*MT*)**

Die vierte logische Schlussfigur stellt der sog. Modus Tollens dar. Die Nebenprämisse bildet dabei die Negation der Konsequenz (¬*q*), was für das Beispiel bedeutet:

Beispiel:	oder allgemein:
Hauptprämisse: Wenn es regnet, dann ist die Straße nass.	$p \rightarrow q$
Nebenprämisse: Die Straße ist nicht nass.	$\neg q$
---	---
Logischer Schluss: Es regnet nicht.	$\neg p$

Im Falle des Modus Tollens ist ein eindeutiger logischer Schluss möglich, nämlich die Negation des Antezedens (¬*p*). Die exklusive Disjunktion *p oder* ¬*p* ist in diesem Fall

[4] Die Disjunktion muss exklusiv sein, da die inklusive Disjunktion implizieren würde:
q oder ¬*q oder* (*q und* ¬*q*).
Die Konjunktion einer Aussage und ihrer Negation, wie in diesem Fall *q und* ¬*q*, ist jedoch ein logischer Widerspruch und letztlich das Ergebnis einer *reductio ad absurdum* (siehe z.B. Knauff, 2006), welche per definitionem (im aussagenlogischen Sinne) falsch ist.

falsch, denn *p* impliziert *q*. Die Konjunktion *q und* ¬*q* ist jedoch per definitionem falsch (reductio ad absurdum, siehe auch Fußnote 4). Folglich ist der Schluss *p* nicht möglich und ¬*p* alleiniger logischer Schluss. Angewendet auf das inhaltliche Beispiel lässt sich formulieren: „Würde es regnen, wäre die Straße nass. Die Straße ist aber nicht nass, folglich kann es nicht regnen." Es liegt nahe, dass diese formal-logisch eindeutige Schlussfigur seltener korrekt gelöst wird als der Modus Ponens. Eine ausführliche Betrachtung soll jedoch erst an späterer Stelle erfolgen (siehe dazu vor allem Abschnitt 2.1.2). Tabelle 1 fasst die vier Schlussfiguren des Konditionalen Schlussfolgerns, die den Ausgangspunkt für sämtliche weiteren Testkonstruktionsüberlegungen bilden, noch einmal zusammen.

Tabelle 1: Überblick über die vier Schlussfiguren Konditionalen Schlussfolgerns bei einer Konditionalaussage der Form $p \rightarrow q$

Bezeichnung der Schlussfigur	Nebenprämisse	Logischer Schluss
Modus Ponens	Bestätigung des Antezedens (p)	Bestätigung der Konsequenz (q)
Negation des Antezedens	Negation des Antezedens ($\neg p$)	kein eindeutiger logischer Schluss ableitbar *
Bestätigung der Konsequenz	Bestätigung der Konsequenz (q)	kein eindeutiger logischer Schluss ableitbar *
Modus Tollens	Negation der Konsequenz ($\neg q$)	Negation des Antezedens ($\neg p$)

Anmerkung. Die Formulierung der mit * gekennzeichneten Aussage meint keineswegs, dass *q oder* ¬*q* (bei Bestätigung der Konsequenz: *p oder* ¬*p*) kein logisch eindeutiger Schluss wäre. Sie ist vielmehr pragmatisch zu verstehen und meint, dass für diese Schlussfiguren einzelne atomare Aussagen, wie ¬*q* (bei Bestätigung der Konsequenz: *p*) keine erschöpfende valide Inferenz darstellen. Mit der Formulierung „kein eindeutiger logischer Schluss ableitbar" soll auf spätere Formulierungen von Antwortoptionen bei den Testitems vorbereitet werden.

Für diese Testkonstruktion wurde also die Aussagenlogik als Referenzsystem für die Beurteilung der Korrektheit von Lösungen Konditionaler Syllogismen festgelegt. In einem nächsten Schritt werden nun die kognitiven Prozesse betrachtet, die der Generierung solcher Lösungen zugrunde liegen.

2.1.2 Kognitionspsychologische Theorien zum Konditionalen Schlussfolgern

Wie bereits zu Beginn von Kapitel 2 erwähnt, soll die Testkonstruktion gemäß der Forderung von Beckmann und Guthke (1999) oder auch von Wilhelm (2000) kognitionspsychologisch fundiert sein. Das heißt, Aussagen über Itemeigenschaften sollen aus den kognitiven Prozessen ableitbar sein, die beim Bearbeiten eines Items ablaufen. Nimmt man die Menge an Publikationen als Indikator für die Bedeutsamkeit, dann stellt Konditionales Schlussfolgern (zumindest in der Kognitionspsychologie) die bedeutsamste Form deduktiver Inferenzen dar (Evans, 1993; Evans & Over, 2004; Johnson-Laird & Byrne, 2002). Der folgende Abschnitt widmet sich den kognitiven Prozessen beim Konditionalen Schlussfolgern. Dabei thematisieren kognitionspsychologische Theorien meist das Ergebnis eines Konditionalen Schlusses, ohne Bewertung, ob das Ergebnis (logisch) korrekt ist oder nicht. Nachdem in Abschnitt 2.1.1 für die Testkonstruktion festgelegt wurde, dass aussagenlogische Korrektheit des Schlusses den Bewertungsschlüssel für die Itemantworten darstellt, werden die kognitionspsychologischen Befunde in diesem Abschnitt zusätzlich in Bezug auf die aussagenlogische Korrektheit des Schlusses betrachtet. Wird in der Folge von einer „korrekten“ Lösung oder Reaktion gesprochen, ist damit aussagenlogische Korrektheit des Schlusses gemeint.

Kognitionspsychologische Theorien zum Schlussfolgern lassen sich grob in die *nicht-rationalen* und in die *rationalen Ansätze* einteilen. Die nicht-rationalen Ansätze, unter denen besonders die *Atmosphärentheorie* (z.B. Begg & Denny, 1969) populär ist, gehen davon aus, dass beim Schlussfolgern hauptsächlich logisch-irrelevante Aspekte (wie bspw. Oberflächenmerkmale) berücksichtigt werden (Beckmann & Guthke, 1999). Es ist jedoch unwahrscheinlich, dass Schlussfolgern hauptsächlich, geschweige denn ausschließlich, durch nicht-rationales Verhalten erklärt werden kann (vgl. vor allem Stanovich, 1999). Zudem scheint es wenig sinnvoll, mit der Zielstellung einer Testkonstruktion die nicht-rationalen Ansätze zu verfolgen. Schließlich würde dies bedeuten, eben nicht Deduktives (also Logisches) Denken zu erfassen, sondern vielmehr die (Un-)Anfälligkeit gegenüber bestimmten logisch irrelevanten Aspekten. Im Folgenden werden daher die rationalen Ansätze weiter verfolgt. Zunächst werden die beiden wichtigsten kognitionspsychologischen Theorien zum Konditionalen Schlussfolgern, die *Theorie mentaler Modelle* (Johnson-Laird & Byrne, 2002) und die *Theorie einer mentalen Logik* (Braine & O’Brien, 1991) vorgestellt. Abschließend werden beide Theorien

hinsichtlich ihrer Bedeutung für die beabsichtigte Testkonstruktion bewertet. Entscheidendes Bewertungskriterium ist dabei die Erklärungskraft für empirische Befunde. Daher wird im Folgenden zunächst ein kurzer Überblick über häufige empirische Befunde zu den vier Schlussfiguren des Konditionalen Schlussfolgerns gegeben.

Empirische Befunde

In einer Vielzahl von Experimenten zeigt sich, dass der Modus Ponens von fast allen Probanden (üblicherweise mehr als 95%) korrekt gelöst wird (siehe z.B. Evans et al., 1993 für einen Überblick). Die zweite valide Schlussfigur, der Modus Tollens, wird deutlich seltener korrekt gelöst, in den meisten Studien jedoch von mehr als der Hälfte der Probanden (siehe ebenfalls Evans et al., 1993 für einen Überblick). Bei der Negation des Antezedens wie auch bei der Bestätigung der Konsequenz werden häufig alleinige Schlüsse gezogen (für die Konditionalaussage $p \rightarrow q$ bei Negation des Antezedens: $\neg q$, bei Bestätigung der Konsequenz: p), obwohl sie logisch falsch sind. Stattdessen reagieren diejenigen Personen (logisch) korrekt, die nicht diese alleinigen Schlüsse ziehen, sondern bei Negation des Antezedens *q oder* $\neg q$ bzw. bei Bestätigung der Konsequenz *p oder* $\neg p$ schlussfolgern (siehe auch Abschnitt 2.1.1). Für die Fragestellung der vorliegenden Arbeit ist es problematisch, dass in vielen kognitionspsychologischen Studien bei den Schlussfiguren Negation des Antezedens und Bestätigung der Konsequenz lediglich die Häufigkeiten für die alleinigen Schlüsse und damit falschen Inferenzen berichtet werden. Zur Häufigkeit korrekter Reaktionen (wie bspw. *q oder* $\neg q$ oder auch „*keine Aussage möglich*“) werden hingegen keine Ergebnisse berichtet. Die Häufigkeit korrekter Reaktionen kann daher nicht genau angegeben werden, ist folglich aber stets kleiner-gleich der Differenz zwischen 100% und dem Prozentwert für die berichteten (logisch) falschen Inferenzen. Betrachtet man nun diese Ergebnisse, so zeigt sich, dass die Zahl korrekter Reaktionen auf die Schlussfiguren Negation des Antezedens und Bestätigung der Konsequenz in nahezu allen betrachteten Studien noch unter der für den Modus Tollens liegt. Tabelle 2 illustriert diesen Befund anhand verschiedener Studien, in denen alle vier Schlussfiguren des Konditionalen Schlussfolgerns durch die Probanden bearbeitet werden. Prinzipiell werden diese empirischen Befunde auch von einer Vielzahl Studien gestützt, deren Stichprobenumfänge jedoch deutlich geringer sind (für einen Überblick siehe Evans et al., 1993).

Tabelle 2: Empirische Befunde zum Prozentsatz (logisch) korrekter Lösungen bei den vier Schlussfiguren Konditionalen Schlussfolgerns

Schlussfigur	Prozentsatz (logisch) korrekter Lösungen	
	Kleinbeck, 2005; $N = 325$	Evans, Handley, Neilens & Over, 2007; $N = 120$
Modus Ponens	95%	100%
Negation des Antezedens	$\leq 40\%$ *	$\leq 45\%$ *
Bestätigung der Konsequenz	$\leq 32\%$ *	$\leq 26\%$ *
Modus Tollens	67%	61%

Anmerkung. Die bei Kleinbeck (2005) berichteten Daten wurden von Beller über eine Vielzahl empirischer Studien aggregiert (vgl. Kleinbeck, 2005). Für den Modus Ponens ist der Stichprobenumfang allerdings geringer ($N = 285$) als angegeben.
* Da in den Originalarbeiten nur die Häufigkeiten (und damit Prozentsätze) der (logisch falschen) alleinigen Inferenzen (bei Negation des Antezedens: $\neg q$, bei Bestätigung der Konsequenz: p) berichtet werden, wird der Prozentsatz der (logisch) korrekten Lösungen jeweils als kleiner-gleich der Differenz zwischen 100% und dem Prozentsatz dieser (logisch falschen) alleinigen Inferenz angegeben.

Es bleibt also festzuhalten, dass die beiden kognitionspsychologischen Theorien dahingehend zu bewerten sind, inwieweit sie zwei zentrale Befunde erklären können: Zum einen, dass Modus-Ponens-Probleme leichter korrekt zu lösen sind als Modus-Tollens-Probleme und zum anderen, dass Aufgaben zu den invaliden Schlussfiguren (Negation des Antezedens, Bestätigung der Konsequenz) offenbar schwerer (logisch) korrekt zu lösen sind als Aufgaben zu den beiden validen Schlussfiguren (Modus Ponens und Modus Tollens). In der Folge werden nun die beiden angeführten kognitionspsychologischen Theorien jeweils in Bezug auf Deduktives Denken, Konditionales Schlussfolgern und schließlich hinsichtlich ihrer Erklärungskraft für die beschriebenen empirischen Befunde vorgestellt.

2.1.2.1 Die Theorie mentaler Modelle

Die Theorie mentaler Modelle (Johnson-Laird & Byrne, 1991) liefert nicht nur für die kognitiven Prozesse beim Konditionalen Schlussfolgern einen theoretischen Rahmen, sondern für Deduktive Denkprozesse insgesamt (Johnson-Laird, 2001), ebenso wie für Prozesse des Induktiven Denkens (Johnson-Laird, 1994). Für Deduktives Denken bezeichnet sie Wilhelm (2000) gar als die am besten theoretisch wie empirisch fundierte Theorie. In Bezug auf Deduktives Denken geht die Theorie mentaler Modelle von

einem dreiphasigen Prozess des Schlussfolgerns aus, dessen Ausgangspunkt eine Behauptung (in Form einer oder mehrerer Prämissen) ist:

Phase 1 (Modellkonstruktion[5]): Personen formulieren ein *mentales Modell* der Behauptung. Dieses repräsentiert eine Möglichkeit (oder Aussage) innerhalb eines Kontextes, in dem diese Behauptung wahr ist. Allerdings erfolgt diese Repräsentation nur dann, wenn das mentale Modell in der repräsentierten Möglichkeit (oder Aussage) auch tatsächlich wahr ist. Dies wird in der Theorie mentaler Modelle als „Prinzip der Wahrheit" bezeichnet.

Phase 2 (Modellinspektion): Das mentale Modell wird hinsichtlich neuer Informationen untersucht, die nicht explizit in den Prämisseninformationen enthalten sind. Es wird eine (vorläufige) Konklusion generiert. Eine solche Konklusion wird (zunächst) als *mögliche* Konklusion bezeichnet, da sie mit mindestens einem mentalen Modell übereinstimmt.

Phase 3 (Modellvariation): Es werden weitere (mentale) Modelle der Behauptung konstruiert. Diese werden dahingehend überprüft, ob eines dieser Modelle der vorläufig für gültig gehaltenen Konklusion widerspricht. Ist dies nicht der Fall, spricht man von einer notwendigerweise wahren, also deduktiv validen Konklusion. Existiert ein mentales Modell, das der vorläufigen Konklusion widerspricht, beginnt der Prozess in Phase 2 von neuem.

Zur Veranschaulichung, insbesondere der ersten Phase, soll nun das Konzept der mentalen Modelle weiter vertieft werden. Das Prinzip der Wahrheit impliziert, dass in mentalen Modellen nur Möglichkeiten repräsentiert werden, die im Sinne einer Wahrheitstafel (also in deren Zeilen) wahr sind, nicht jedoch Möglichkeiten, die im Sinne einer Wahrheitstafel falsch sind. Das folgende Beispiel einer inklusiven Disjunktion (nach Johnson-Laird & Byrne, 2002) soll das verdeutlichen:

Die Batterie ist leer oder der Stromkreis ist nicht geschlossen oder beides.

In diesem Beispiel würden drei mentale Modelle repräsentiert:

[5] Bezeichnungen der Phasen in Anlehnung an Knauff, 2006

leer

¬ geschlossen

leer ¬ geschlossen

Jede Zeile kennzeichnet dabei ein mentales Modell, „leer“ repräsentiert ein Modell des Abschnitts „die Batterie ist leer“ der Behauptung, „¬ geschlossen“ repräsentiert ein Modell des (negativ formulierten) Abschnitts „der Stromkreis ist nicht geschlossen“ der Behauptung. Die zugehörige Wahrheitstafel (Tabelle 3) illustriert, dass lediglich wahre Elemente einer Reihe als mentale Modelle repräsentiert werden. Gleichzeitig wird deutlich, dass auch Negationen (wie hier „¬ geschlossen“) wahre Elemente darstellen, insofern sie in dieser Form Teil der Behauptung sind.

Tabelle 3: Wahrheitstafel für das Beispiel einer inklusiven Disjunktion („*Die Batterie ist leer oder der Stromkreis ist nicht geschlossen oder beides.*“)

„leer“	„¬ geschlossen“
wahr	falsch
falsch	wahr
wahr	wahr
falsch	falsch

Was in einem mentalen Modell nicht repräsentiert wird, ist also entsprechend falsch. Für das erste der drei mentalen Modelle (s.o.) kann man also daraus, dass „¬ geschlossen“ nicht repräsentiert ist, „geschlossen“ schlussfolgern. Das Beispiel verdeutlicht zudem, dass es für eine Behauptung nicht nur ein mentales Modell, sondern eine bestimmte Menge von mentalen Modellen geben kann (für eine inklusive Disjunktion bspw. drei), wobei jedes Modell eine Möglichkeit repräsentiert. Für die Elemente einer Reihe der Wahrheitstafel (für das Beispiel siehe Tabelle 3), die falsch sind, machen Personen sog. *mentale „Fußnoten“*. Diese mentalen Fußnoten können verwendet werden, um vollständig explizite Modelle über das, was wahr ist, zu konstruieren:

leer geschlossen

¬ leer ¬ geschlossen

leer ¬ geschlossen

oder auch zu schlussfolgern, was falsch ist, nämlich:

¬ leer geschlossen

Die Theorie mentaler Modelle geht davon aus, dass die Erzeugung und Manipulation mentaler Modelle im Arbeitsgedächtnis passiert (vgl. z.B. Knauff, 2006; Wilhelm, 2000). Folglich ist ein Schluss umso schwieriger, je mehr Modelle im Arbeitsgedächtnis konstruiert und inspiziert werden müssen. Schlussfolgerungen, die auf lediglich einem mentalen Modell basieren, sind also leichter als auf mehreren Modellen basierende (Johnson-Laird & Byrne, 2002). Übersteigt die Anzahl der zu berücksichtigenden Modelle die *Arbeitsgedächtniskapazität*, kann es passieren, dass ein „widersprechendes" Modell nicht berücksichtigt wird und eine lediglich mögliche Konklusion als deduktiv valide angenommen wird. Dies nimmt die Theorie mentaler Modelle als Hauptfehlerquelle beim Schlussfolgern an. Weitere Fehlerquellen sind die Neigung von Personen, mentale Fußnoten zu vergessen (Johnson-Laird & Byrne, 2002) sowie die Abweichung von normativen Standards bei der Interpretation der Prämissen.

Die Modelltheorie Konditionalen Schlussfolgerns (Johnson-Laird & Byrne, 2002)

Wenngleich Konditionales Schlussfolgern bereits integraler Bestandteil der Theorie mentaler Modelle (kurz: Modelltheorie) von Johnson-Laird und Byrne (1991) ist, werden durch die (erweiterte) Modelltheorie Konditionalen Schlussfolgerns (Johnson-Laird & Byrne, 2002)[6] bis dahin existierende Lücken (vgl. z.B. Fillenbaum, 1993) geschlossen. So erhebt die Modelltheorie den Anspruch, semantische Aspekte beim Konditionalen Schlussfolgern zu berücksichtigen. Es werden zusätzlich zum Prinzip der Wahrheit (s.o.) fünf weitere Grundprinzipien angenommen:

1. Das Prinzip der Kernbedeutungen

Dabei wird von zwei basalen Konditionalaussagen ausgegangen. Diese lauten (bei zunächst neutralen Inhalten A und B): „Wenn A, dann B" und „Wenn A, dann möglicherweise B". Das Antezedens jeder dieser beiden Konditionalaussagen beschreibt eine Möglichkeit, die Konsequenz kann gegeben dieser Möglichkeit auftreten. Jede dieser beiden Konditionalaussagen hat dann eine Kernbedeutung. Für „Wenn A, dann B" ist das eine konditionale Interpretation, das heißt, A ist hinreichende Bedingung für B. Die Kernbedeutung der Konditionalaussage „Wenn A, dann möglicherweise B" ist eine

[6] Alle weiteren Ausführungen zur Modelltheorie Konditionalen Schlussfolgerns in diesem Abschnitt beziehen sich (soweit nicht anders gekennzeichnet) auf Johnson-Laird und Byrne (2002).

tautologische Interpretation, da diese Aussage nicht falsch sein kann, unabhängig davon, welche Wahrheitswerte A und B annehmen.

2. Das Prinzip der konjunktivischen Bedeutungen

Dieses Prinzip gilt für konjunktivische Konditionalaussagen, also Konditionalaussagen in der Möglichkeitsform. Die zugehörigen mentalen Modelle sind dann entweder faktische Möglichkeiten oder auch kontrafaktische Möglichkeiten. Unter faktischen Möglichkeiten sind Möglichkeiten zu verstehen, die zum gegenwärtigen Zeitpunkt tatsächlich noch eintreten können. Kontrafaktisch sind solche Möglichkeiten, die zu einem bestimmten Zeitpunkt faktisch möglich waren, mittlerweile jedoch nicht eingetreten sind. Das Prinzip der konjunktivischen Bedeutungen besagt nun, dass bei konjunktivischen Konditionalaussagen die gleichen Interpretationen (konditional, tautologisch) gelten wie bei indikativen Konditionalaussagen.

3. Das Prinzip der impliziten Modelle

Bei basalen Konditionalaussagen werden lediglich die mentalen Modelle explizit konstruiert, in denen das Antezedens in seiner ursprünglichen Form enthalten ist. Modelle, in denen das Antezedens nicht derart repräsentiert ist (sondern bspw. nur dessen Negation), sind allenfalls implizite Modelle. Mentale Fußnoten an den impliziten Modellen können verwendet werden, um vollständig explizite Modelle zu bilden. Allerdings neigen Personen dazu, mentale Fußnoten (mitunter sogar die impliziten Modelle selber) zu vergessen.

4. Das Prinzip der semantischen Modulation

Die Bedeutungen von Antezedens und Konsequenz wie auch deren (semantisch) sinnvolle Verknüpfbarkeit können bei der Bildung vollständig expliziter Modelle helfen.

5. Das Prinzip der pragmatischen Modulation

Der Kontext einer Konditionalaussage ist abhängig vom allgemeinen wie spezifischen Vorwissen und kann ebenfalls bei der Bildung vollständig expliziter Modelle helfen.

Von den beiden Kernbedeutungen wird fortan lediglich die basale Konditionalaussage der Form „Wenn …, dann …“ betrachtet. Diese entspricht der Implikation, die den (aussagenlogischen) Ausgangspunkt für Konditionales Schlussfolgern im Rahmen dieser Arbeit darstellt (siehe Abschnitt 2.1.1). Aus dem Prinzip der impliziten Modelle

leitet sich ab, dass die mentalen Modelle einer Konditionalaussage „Wenn A, dann B“ wie folgt lauten:

a b
…

„a“ indiziert dabei eine (konkrete) Möglichkeit von „A“ und „b“ eine (konkrete) Möglichkeit von „B“. Die Aufzählungspunkte (…) kennzeichnen implizite Modelle. Diese sind mit einer mentalen Fußnote versehen, mit deren Hilfe vollständig explizite Modelle der Konditionalaussage gebildet werden können:

a b
¬ a b
¬ a ¬ b

Entsprechend den Prinzipien der semantischen und pragmatischen Modulation geht die Modelltheorie davon aus, dass Konditionalaussagen je nach Inhalt und Kontext nicht nur konditional oder tautologisch, sondern außerdem auf acht weitere Arten interpretiert werden können, auf deren Aufzählung jedoch verzichtet werden soll (siehe dazu Johnson-Laird & Byrne, 2002). Von besonderer Bedeutung ist darunter allerdings die bikonditionale Interpretation, da Konditionalaussagen selbst bei neutralem Inhalt häufig als bikonditional interpretiert werden (vgl. z.B. Evans, 1982). Bei der bikonditionalen Interpretation wird das Antezedenz als notwendig und hinreichend für die Konsequenz angenommen, woraus folgende vollständig expliziten Modelle einer Konditionalaussage „Wenn A, dann B“ resultieren:

a b
¬ a ¬ b

Die bikonditionale Interpretation ist auch entwicklungspsychologisch interessant, da die Modelltheorie Konditionalen Schlussfolgerns davon ausgeht, dass Konditionalaussagen auf verschiedenen Stufen der Entwicklung unterschiedlich interpretiert werden. Im frühen Kindesalter werden sie als Konjunktionen (ein vollständig explizites Modell), später als bikonditional (zwei vollständig explizite Modelle; s.o.) und in der Adoleszenz sowie im Erwachsenenalter als konditional (drei vollständig explizite Modelle; s.o.) interpretiert. Bei komplexen Konditionalaussagen ist es jedoch möglich, dass Erwach-

sene die impliziten Modelle vergessen und damit die Konjunktions-Interpretation von Kindern zeigen.

Es stellt sich nun die Frage, welche der eingangs vorgestellten empirischen Befunde (siehe Abschnitt 2.1.2) die Modelltheorie Konditionalen Schlussfolgerns erklären kann. Der erste zentrale Befund, dass Modus-Ponens-Probleme häufiger korrekt gelöst werden als Modus-Tollens-Probleme, wird durch die Theorie explizit erklärt. Ausgangspunkt ist die zentrale Annahme der (allgemeinen) Theorie mentaler Modelle, dass Inferenzen, die bereits aus den mentalen Modellen gezogen werden können, leichter sind als solche, die aus vollständig expliziten Modellen gezogen werden müssen (Johnson-Laird & Byrne, 2002). Der Modus-Ponens-Schluss ist bereits im ersten mentalen Modell der Konditionalaussage („Wenn A, dann B“) repräsentiert:

a b
…

Folglich sollte er leichter sein als der Modus-Tollens-Schluss, für den drei vollständig explizite Modelle gebildet werden müssen.

Die Modelltheorie Konditionalen Schlussfolgerns liefert jedoch keine explizite Erklärung für den zweiten zentralen Befund, dass Aufgaben zu den beiden invaliden Schlussfiguren (Negation des Antezedens, Bestätigung der Konsequenz) schwerer zu lösen sind als Aufgaben zu den beiden validen Schlussfiguren (Modus Ponens, Modus Tollens). Somit bleibt festzuhalten, dass mit der Modelltheorie Konditionalen Schlussfolgerns zumindest einer der beiden zentralen empirischen Befunde erklärt werden kann.

2.1.2.2 Die Theorie einer mentalen Logik

Die Theorie einer mentalen Logik (Braine & O’Brien, 1991) gehört zu den sog. *Regeltheorien*, welche ebenfalls keine spezifischen Theorien zu Prozessen des Konditionalen Schlussfolgern sind, sondern Theorien für das gesamte Spektrum Deduktiver Denkprozesse. Unter den Regeltheorien sind die Ansätze von Braine (1978; Braine, Reiser & Rumain, 1984; Braine & O’Brien, 1998) und Rips (1983, 1994) die bekanntesten. Zunächst werden die Grundannahmen beider Ansätze kurz vorgestellt.

Die Theorie einer mentalen Logik nach Braine

Nach Braine (1978; Braine et al., 1984; Braine & O'Brien, 1998) verfügen Menschen über eine Menge natürlicher Inferenzschemata, die im Langzeitgedächtnis gespeichert sind. Das Besondere dabei ist, dass unter diesen Inferenzschemata auch solche vorkommen, die von der formalen Logik abweichen und eher einer „natürlichen, mentalen Logik" zuzuordnen sind. Was in der formalen Logik wahr ist, muss nicht zwingend in der mentalen Logik wahr sein und umgekehrt. Hierin vermutet Braine eine der Ursachen für Fehler, die Menschen beim Logischen Denken machen. Eine Auflistung sämtlicher dieser Inferenzschemata wäre an dieser Stelle zu umfangreich, findet sich aber bspw. bei O'Brien (2004). Der Prozess des logischen Schlussfolgerns beginnt jedoch zunächst mit einer mentalen Repräsentation der Prämissen im Arbeitsgedächtnis. Bereits hier können erste Fehler, sog. *Verständnisfehler* auftreten, z.B. weil die Prämissen einfach falsch verstanden werden und folglich auch nicht korrekt repräsentiert werden können. Anschließend werden die verfügbaren Inferenzschemata auf diese mentale Repräsentation angewendet. Ziel ist dabei, weitere logisch gültige Aussagen abzuleiten, also eine Konklusion zu ziehen. Ist das mit den verfügbaren Inferenzschemata nicht möglich, kann der Schluss nur auf der Basis von Heuristiken gezogen werden, die dann potenziell zu „unlogischen" Schlüssen führen. Man spricht in diesem Fall von sog. *Strategiefehlern*. Da fast der gesamte Prozess im Arbeitsgedächtnis abläuft, ist er von dessen Kapazität abhängig und wie in der Theorie mentaler Modelle (siehe Abschnitt 2.1.2.1) können auch Überlastungen des Arbeitsgedächtnisses zu Fehlern, sog. *Prozessfehlern*, führen.

Die Theorie mentaler Beweise nach Rips

Ähnlich der Theorie von Braine bildet auch bei Rips (1983, 1994) ein System verfügbarer Regeln die Basis logischen Schlussfolgerns. Allerdings sind diese Regeln strenger als bei Braine an den Gesetzen der formalen Logik orientiert (Knauff, 2006). Rips (1994) nennt dieses System *PSYCOP* (*PSYC*hology *O*f *P*roof). Regeln können dabei sowohl vorwärts (bei der Generierung von Konklusionen) als auch rückwärts (bei der Beurteilung des Wahrheitswertes von Konklusionen) verwendet werden. Gegeben seien zunächst aber wiederum eine mentale Repräsentation von Prämissen sowie die mentale Repräsentation einer Konklusion. Zwischen beiden soll nun eine Verbindung geschaffen werden, und zwar durch Anwendung der verfügbaren Regeln. Die zugrundeliegende Idee ist einfach: Je mehr Regeln notwendig sind, um die Verbindung zwischen Prämis-

sen und Konklusion herzustellen, desto fehleranfälliger (und damit schwieriger) ist die Inferenz. Gerade für schwierige Inferenzen ist es daher in PSYCOP von entscheidender Bedeutung, Zwischenziele zu setzen. Das Zulassen von Redundanzen zwischen Regeln sowie das Zulassen zahlreicher Gewichtungen von Regeln (je nach Einsatz in der Vergangenheit) machen bei PSYCOP den Unterschied zu formalen Deduktionssystemen aus. Auf diese Weise versucht Rips (1994), den Besonderheiten menschlichen Denkens Rechnung zu tragen.

Konditionales Schlussfolgern wird im Rahmen beider Ansätze thematisiert. Allerdings formulieren Braine und O'Brien (1991) ihre Theorie einer mentalen Logik (im Folgenden Logiktheorie genannt) explizit für Konditionales Schlussfolgern. Daher soll diese Theorie nachfolgend vorgestellt werden, wenngleich an einzelnen Stellen auch auf Überlegungen der Theorie mentaler Beweise (Rips, 1994) zum Konditionalen Schlussfolgern eingegangen wird.

Die Logiktheorie Konditionalen Schlussfolgerns (Braine & O'Brien, 1991)

Nach Braine und O'Brien (1991) sollte eine Theorie zum Schlussfolgern bei einem logischen Junktor (wie bspw. „wenn…, dann") Aussagen über drei Komponenten treffen: den lexikalischen Eintrag, den Ablauf des Schlussfolgerns selbst, also das dabei ablaufende „Programm", und pragmatische Verstehensprozesse. Diese drei Komponenten werden nun für die Logiktheorie zum Konditionalen Schlussfolgern (Braine & O'Brien, 1991)[7] betrachtet.

Lexikalischer Eintrag

Für den lexikalischen Eintrag wird angenommen, dass die Menge natürlicher Inferenzschemata zwei Inferenzschemata speziell für Konditionalaussagen umfasst: Eines ist der Modus Ponens, das andere der sog. *Konditionale Beweis*. Gegeben *p* und *q* seien zwei beliebige Aussagen, dann besagt das Modus-Ponens-Inferenzschema:

Gegeben *Wenn p, dann q* und *p*, kann man *q* schlussfolgern.

Das Inferenzschema für den Konditionalen Beweis besagt:

[7] Alle weiteren Ausführungen zur Logiktheorie Konditionalen Schlussfolgerns in diesem Abschnitt beziehen sich (soweit nicht anders gekennzeichnet) auf Braine und O'Brien (1991).

Um *Wenn p, dann…* abzuleiten, nehme man zunächst *p* an; für jede Aussage *q*, die aus der Annahme *p* in Verbindung mit anderen vorausgesetzten Informationen folgt, kann man schlussfolgern *Wenn p, dann q*.

Beide Inferenzschemata folgen aus der Aussagenlogik. Für die mentale Logik unterliegt der Konditionale Beweis jedoch drei Einschränkungen:

1. Nichts folgt aus einem Widerspruch, außer dass eine Voraussetzung falsch ist.
2. Eine Annahme kann nur dann Antezedens einer Konditionalaussage sein, wenn sie konsistent zu den vorausgesetzten Informationen (z.B. Prämissen oder vorher getroffenen Annahmen) ist.
3. Eine Voraussetzung, die in der konditionalen Argumentation wiederholt wird, kann nicht im Widerspruch zu der Annahme stehen, die Antezedens der Konditionalaussage ist.

Viele Inferenzen, die formal-logisch völlig korrekt sind, werden aufgrund dieser Einschränkungen in der mentalen Logik nicht gezogen. Das folgende Beispiel soll dies verdeutlichen. Um eine Konditionalaussage *Wenn p, dann q* abzuleiten, muss entsprechend dem Konditionalen Beweis *p* angenommen werden. Aus Einschränkung 2 folgt nun, dass diese Annahme nur dann möglich ist, wenn ¬*p* nicht Teil der Prämissen bzw. vorher getroffener Annahmen ist. In der formalen Logik ist dies keine notwendige Voraussetzung.

Schlussfolgerungs-„Programm“

Das Schlussfolgerungs-Programm von Personen besteht aus einem Routine-Teil und einem strategischen Teil. Der Routine-Teil läuft bei allen (Erwachsenen) gleich ab. Es werden die verfügbaren Inferenzschemata mit sämtlichen Aussagen der Prämissen abgeglichen. Alle direkt anwendbaren Schemata, für die also keine weiteren Annahmen eingeführt werden müssen, werden angewendet. Jede dadurch gezogene Inferenz wird zur Prämissen-Menge hinzugefügt. Das Lösen von Inferenzaufgaben, die mit diesem Routine-Teil des Schlussfolgerungs-Programms gelöst werden können, erfolgt nahezu fehlerfrei. Ist eine Inferenzaufgabe jedoch nicht mit dem Routine-Teil lösbar, muss der strategische Teil des Schlussfolgerungs-Programms genutzt werden. Bezüglich dieses Teils existieren große interindividuelle Unterschiede, die sich in entsprechenden Unter-

schieden in der Fähigkeit zum Schlussfolgernden Denken widerspiegeln. Ergänzend sei festgehalten, dass die verfügbaren Inferenzschemata introspektiv nicht zugänglich sind.

Pragmatische Verstehensprozesse

Nach der Logiktheorie werden die verfügbaren Inferenzschemata auf semantische Informationen angewendet. Diese wiederum sind das Ergebnis von Verstehensprozessen. Alle Faktoren, die das Verstehen beeinflussen, beeinflussen demnach auch die Informationen, auf denen die Inferenzen basieren. Die Logiktheorie nimmt dafür drei generelle Prinzipien an:

1. Das erste Prinzip geht davon aus, dass der Inhalt einer Proposition die Art beeinflusst, in der sie konstruiert wird. Eine Interpretation, die gegeben das allgemeine (Welt-)Wissen wie auch gegeben das spezifische Wissen einer Person plausibel ist, wird wahrscheinlicher mental repräsentiert als eine unplausible.
2. Das zweite Prinzip, das sog. *Kooperative Prinzip* (Grice, 1975, 1978), geht davon aus, dass bei Konversationen der Sprecher versucht, so informativ, sachdienlich, verständlich und genau wie möglich zu sein und der Zuhörer wiederum die Aussagen des Sprechers unter der Annahme interpretiert, dieser sei so informativ, sachdienlich, verständlich und genau wie möglich.
3. Das dritte Prinzip bezieht sich auf die sog. „*einladenden Inferenzen*“ (Geis & Zwicky, 1971; siehe auch später in diesem Abschnitt sowie in Abschnitt 2.2.1). Im Falle von Konditionalaussagen wird davon ausgegangen, dass diese je nach Inhalt dazu „einladen“, zusätzliche Informationen hinzu zu interpretieren. Menschen neigen dazu, die aus diesen Informationen resultierenden „einladenden Inferenzen“ zu ziehen, es sei denn, sie haben Grund zu der Annahme, dass diese unangebracht oder falsch sind.

Wie erklärt nun die Logiktheorie die in Abschnitt 2.1.2 angeführten empirischen Befunde? Die Logiktheorie geht davon aus, dass bei den theoretisch einfachsten Inferenzaufgaben die Lösung direkt aus dem lexikalischen Eintrag folgt. In Bezug auf Konditionalaussagen zählen dazu Modus-Ponens-Inferenzen. Dabei werden nahezu keine Fehler gemacht – weder von Kindern (Braine & Rumain, 1983; O’Brien, 1987) noch von Erwachsenen (vgl. z.B. Evans et al., 1993). Für Modus-Tollens-Inferenzen existiert kein expliziter lexikalischer Eintrag. Sie müssen über eine Aneinanderreihung einzelner Schemata (Annahme: p, Modus Ponens: q, Kontradiktion/Inkompatibilität, reductio ad

absurdum: $\neg p$; siehe z.B. Knauff, 2006) gezogen werden und sind folglich fehleranfälliger. Auch in Rips' (1994) PSYCOP-System existiert eine explizite Regel für den Modus Ponens, jedoch keine für den Modus Tollens. Dieser muss über andere (vergleichsweise einfache) Regeln hergeleitet werden (siehe dazu Rips, 1994). Damit kann die Logiktheorie (wie auch Rips' Theorie mentaler Beweise) den ersten zentralen empirischen Befund erklären: Modus-Ponens-Probleme sind leichter korrekt zu lösen als Modus-Tollens-Probleme. Die Logiktheorie erklärt auch den zweiten zentralen empirischen Befund, dass Aufgaben zu den invaliden Schlussfiguren (Negation des Antezedens, Bestätigung der Konsequenz) schwerer korrekt zu lösen sind als Aufgaben zu den beiden validen Schlussfiguren (Modus Ponens und Modus Tollens), und zwar mit dem Prinzip der einladenden Inferenzen. Hierzu bedarf es einiger genauerer Ausführungen:

Das Konzept der einladenden Inferenzen besagt, dass eine Konditionalaussage, z.B. „*Wenn Du den Rasen mähst, gebe ich Dir 5 Dollar*" (Beispiel nach Geis & Zwicky, 1971) Personen „einlädt", die Inferenz, „*Wenn Du den Rasen nicht mähst, gebe ich Dir keine 5 Dollar*" quasi automatisch als ebenfalls gültig anzunehmen. Ursache sind Erfahrungen in der zwischenmenschlichen Konversation. Für das Beispiel ist es durchaus nachvollziehbar, dass die erste Aussage die zweite inhaltlich impliziert. Die Interpretation der konditionalen Aussage wird bikonditional, da die einladende Inferenz akzeptiert wird. Gegeben dieser bikonditionalen Interpretation werden nun für alle vier Schlussfiguren einer Konditionalaussage $p \rightarrow q$ Inferenzen gezogen (q beim Modus Ponens, $\neg p$ beim Modus Tollens, $\neg q$ bei Negation des Antezedens, p bei Bestätigung der Konsequenz). Alle Personen, die die einladende Inferenz akzeptieren, ziehen also eine (formal-logisch falsche) Inferenz bei den Schlussfiguren Negation des Antezedens und Bestätigung der Konsequenz. Beim Modus Ponens und beim Modus Tollens verhalten sie sich hingegen ebenso wie Personen, die die einladende Inferenz nicht akzeptieren. Das erklärt, warum Aufgaben zu den Schlussfiguren Negation des Antezedens und Bestätigung der Konsequenz seltener korrekt gelöst werden als Aufgaben zu Modus Ponens und Modus Tollens. Damit bleibt festzuhalten, dass die Logiktheorie beide zentralen empirischen Befunde zum Konditionalen Schlussfolgern erklären kann.

2.1.2.3 Vergleich beider kognitionspsychologischer Theorien

Nach diesem kurzen Abriss der beiden umfassendsten kognitionspsychologischen Theorien zu Deduktiven Denkprozessen im Allgemeinen wie zu Prozessen des Konditionalen Schlussfolgerns im Speziellen soll ein Vergleich beider Theorien diese Ausführun-

gen abschließen. Dieser zunächst allgemeine Vergleich mündet in den Vergleich von Modelltheorie und Logiktheorie (zum Konditionalen Schlussfolgern). Ziel ist die Auswahl einer der beiden Theorien als kognitionspsychologische Grundlage für die beabsichtigte Testkonstruktion.

Für die üblichen logischen Kalküle gelangen beide Theorien meist zu identischen Entscheidungen bei der Beurteilung der Korrektheit von Argumenten (Rips, 1994). Sie liefern letztlich also oft nur unterschiedliche Erklärungen für das gleiche empirische Phänomen (Wilhelm, 2000). Entsprechend scheinen beide Theorien auch nicht völlig unvereinbar, wie Arbeiten von Falmagne (1993; Falmagne & Gonsalves, 1995) und Roberts (1993) zeigen. Dennoch existiert ein langjähriger Disput zwischen beiden Theorien (z.B. Bonatti, 1994; Evans, Over & Handley, 2005; Johnson-Laird, 1997a, 1997b, 1999; Johnson-Laird, Byrne & Schaeken, 1994; O'Brien, Braine & Yang, 1994). Dieser erfolgt meist in der Form, dass solche experimentellen Befunde angeführt werden, die sich mit der einen Theorie erklären lassen und mit der anderen nicht. Die Erklärungskraft für empirische Befunde bildet also das „Herzstück“ der Argumentation für oder gegen eine Theorie. Dies soll daher auch die wichtigste Rolle bei der Auswahl der kognitionspsychologischen Theorie spielen, auf der die beabsichtigte Testkonstruktion letztlich aufgebaut wird. In Abschnitt 2.1.2 werden zwei empirische Befunde als besonders prägnant für die vier Schlussfiguren Konditionalen Schlussfolgerns vorgestellt: Zum einen, dass unter den validen Schlussfiguren Aufgaben zum Modus Ponens häufiger korrekt gelöst werden als Aufgaben zum Modus Tollens und zum anderen, dass Aufgaben zu den validen Schlussfiguren (Modus Ponens, Modus Tollens) generell häufiger korrekt gelöst werden als Aufgaben zu den invaliden Schlussfiguren (Negation des Antezedens, Bestätigung der Konsequenz). Es kann zusammenfassend festgehalten werden, dass die Logiktheorie Konditionalen Schlussfolgerns (Braine & O’Brien, 1991) beide Befunde explizit erklärt (vgl. Abschnitt 2.1.2.2), die Modelltheorie Konditionalen Schlussfolgerns (Johnson-Laird & Byrne, 2002) hingegen nur den ersten (vgl. Abschnitt 2.1.2.1). Dieser Punkt spricht für die Logiktheorie als Grundlage für eine Testkonstruktion zum Konditionalen Schlussfolgern. Die Stärke der Modelltheorie liegt zweifelsohne in der expliziten Thematisierung semantischer Aspekte. Sind diese im Rahmen der beabsichtigten Testkonstruktion von großem Interesse, bietet sich die Modelltheorie als Basis an. Sind semantische Aspekte jedoch nicht im Fokus des zu konstruierenden Tests, dann legt die Erklärungskraft für beide zentralen empirischen Befunde die Wahl der Logiktheorie als Grundlage für die beabsichtigte Testkonstruktion nahe.

2.1.3 Interindividuelle Unterschiede im Konditionalen Schlussfolgern

Allgemein- bzw. kognitionspsychologische Grundlagen des Konditionalen Schlussfolgerns sind nun dargelegt. Ein Test verfolgt jedoch in der Regel das Ziel, differenzialpsychologische Informationen über die damit untersuchten Personen zu erhalten (z.B. Rost, 2004). Daher soll an dieser Stelle ein Perspektivwechsel in Form einer Integration differenzialpsychologischer Befunde erfolgen. Bevor man sich allerdings wie Stanovich (1999) der differenzialpsychologisch interessanten Frage „*Who* is rational?“ zuwenden kann − also im Falle dieser Arbeit „*Wer* zieht korrekte Konditionale Schlüsse?“ − soll hier kurz rekapituliert werden, dass für die beabsichtigte Testkonstruktion normativ rationales Verhalten sensu Anderson (1990, 1991) als korrektes Verhalten festgelegt wurde (vgl. Abschnitt 2.1.1). Der folgende Abschnitt widmet sich daher interindividuellen Unterschieden im normativ-logischen, rationalen Schlussfolgern bei Konditionalaussagen. Am Ende soll die Frage beantwortet werden, ob die Betrachtung interindividueller Unterschiede für Konditionales Schlussfolgern überhaupt sinnvoll ist, da die Konstruktion eines entsprechenden Tests es andernfalls sicher nicht wäre.

Ein stabiler empirischer Befund ist, dass Menschen beim Konditionalen Schlussfolgern Fehler machen (siehe Evans et al., 1993 für einen Überblick). Beide vorgestellten kognitionspsychologischen Theorien (siehe Abschnitte 2.1.2.1 und 2.1.2.2) nehmen u.a. Limitationen des Arbeitsgedächtnisses als Quelle für diese Fehler an (Braine & O’Brien, 1991; Johnson-Laird & Byrne, 2002). Bezüglich Arbeitsgedächtniskapazität sind nach Kyllonens (1994) *Vier-Quellen-Modell* (siehe auch Ausführungen zu Intelligenstrukturtheorien später in Abschnitt 2.1.4.1) interindividuelle Unterschiede nahezu unstrittig. Dies rechtfertigt nicht nur eine performanzbedingte Interpretation dieser Fehler, sondern demnach auch differenzialpsychologische Betrachtungen zum Konditionalen Schlussfolgern. Seit etwa einer Dekade sind derartige Ansätze auch verstärkt zu beobachten, sowohl für Deduktives Denken im Allgemeinen (z.B. Newstead, Handley, Harley, Wright & Farrelly, 2004; Stanovich, 1999; Stanovich & West, 1998, 2000) als auch für Konditionales Schlussfolgern im Speziellen (z.B. Evans et al., 2007; Klaczynski & Daniel, 2005; Newstead et al., 2004; Oberauer, Geiger, Fischer & Weidenfeld, 2007). Dennoch scheint dieser „Blickwinkel“ vergleichsweise jung gegenüber der fast 100-jährigen allgemeinpsychologischen Betrachtung logischer Denkprozesse (vgl. z.B. Wilhelm, 2000). Gerade für Konditionales Schlussfolgern ist erst seit der Arbeit von Newstead et al. (2004) eine theoretische und gleichzeitig auf ausreichend großen Stich-

proben basierende empirische Auseinandersetzung mit interindividuellen Unterschieden zu verzeichnen (Evans et al., 2007).

Dass also die Betrachtung interindividueller Unterschiede beim Konditionalen Schlussfolgern sinnvoll ist, ist hinreichend argumentiert worden. Dass diese Unterschiede (zumindest zum Teil) auf intelligentes Verhalten zurückgeführt werden können, muss hingegen noch gezeigt werden. Stanovich (1999) kommt zu dem Schluss, dass Fehler beim logischen Schlussfolgern nicht unsystematisch sein können. Als Beleg werden Zusammenhänge mittlerer Stärke zwischen verschiedenen Aufgaben zum logischen Schlussfolgern präsentiert, die sich in einer Vielzahl von Studien zeigen (vgl. Stanovich, 1999). Mehr noch, es zeigen sich systematische Zusammenhänge zwischen Intelligenzleistungen (insbesondere im Schlussfolgernden Denken) und kognitiven Fähigkeiten (vgl. z.B. Stanovich & West, 1998). Intelligentere Personen verhalten sich „rationaler", wobei sich ein moderater Anteil der Varianz bspw. mit interindividuellen Unterschieden in der Verarbeitungskapazität (bzw. Verarbeitungslimitationen) erklären lässt (Stanovich, 1999).

Zusammenfassend kann festgehalten werden, dass sowohl die theorieübergreifende Annahme performanzbedingter Ursachen als auch zahlreiche empirische Befunde für die Betrachtung interindividueller Unterschiede beim Konditionalen Schlussfolgern sprechen. Des Weiteren scheint ein aktueller Trend für eine Testkonstruktion zum Konditionalen Schlussfolgern interessant: der Einsatz allgemeinpsychologischer Aufgaben[8] für die Beantwortung differenzialpsychologischer Fragestellungen. Wenngleich dieser Ansatz vergleichsweise jung ist, scheint er aufgrund einer Vielzahl empirischer Befunde (z.B. Evans et al., 2007; Evans, Handley & Over, 2003; Newstead et al., 2004; Oberauer et al., 2007; Oberauer & Wilhelm, 2003; Stanovich & West, 1998, 2000) äußerst vielversprechend. Der Vorteil allgemeinpsychologischer Aufgaben liegt darin, dass ihnen eine elaborierte Theorie über das Zustandekommen der Leistungen zugrunde liegt. Eine „Qualitätssicherung" des Transfers der allgemeinpsychologischen Aufgaben in einen differenzialpsychologischen Rahmen kann dadurch unterstützt werden, dass Vergleiche mit etablierten diagnostischen Verfahren bspw. zur Intelligenzmessung angestellt werden. Generell sind dabei Zusammenhänge mittlerer Stärke zwischen allgemeinpsychologischen Aufgaben und klassischen Intelligenztests zu erwarten, wie Arbeiten von Stanovich (zusammenfassend Stanovich, 1999; Stanovich & West, 2000) zeigen.

[8] Unter „allgemeinpsychologischen Aufgaben" sind hier Aufgaben zu verstehen, die sonst im Rahmen von Experimenten zur Bewertung allgemeinpsychologischer Theorien verwendet werden.

2.1.4 Diagnostik einer Fähigkeit zum Konditionalen Schlussfolgern

Nachdem Abschnitt 2.1.3 dafür spricht, dass interindividuelle Unterschiede im Konditionalen Schlussfolgern zumindest zum Teil auf intelligentes Verhalten zurückgeführt werden können, liegt es zunächst nahe, eine Fähigkeit zum Konditionalen Schlussfolgern als intellektuelle Fähigkeit zu betrachten. Es stellt sich die Frage nach einer solchen Fähigkeitskonzeption im Rahmen von *Intelligenzstrukturtheorien* sowie nach der Diagnostik einer solchen Fähigkeit. Entsprechend ist der folgende Abschnitt aufgebaut. Zuerst wird überprüft, ob im Rahmen (klassischer wie aktueller) Intelligenzstrukturtheorien eine Fähigkeit zum Konditionalen Schlussfolgern thematisiert wird, anschließend werden (ggf.) Testverfahren zur Erfassung einer solchen Fähigkeit vorgestellt.

2.1.4.1 Fähigkeit zum Konditionalen Schlussfolgern in Intelligenzstrukturtheorien

So vielfältig Definitionen von *Intelligenz* aktuell sind (für einige Beispiele siehe z.B. Süß, 2007; für eine aktuelle „Bestandsaufnahme“ zum Thema Intelligenz siehe Rost, 2009), so herausragend ist die praktische Bedeutung von Intelligenz, bspw. bei der Vorhersage von Berufserfolg (z.B. Schmidt & Hunter, 1998). Exemplarisch sei an dieser Stelle Wechslers Definition angeführt, der Intelligenz beschreibt als „… globale Fähigkeit des Individuums, zweckvoll zu handeln, vernünftig zu denken und sich mit seiner Umgebung wirkungsvoll auseinanderzusetzen“ (Wechslers, 1964, S. 13). Ebenso vielfältig wie die Definitionen von Intelligenz sind die Theorien zu ihrer Struktur. Die Frage nach einer Fähigkeitskonzeption Konditionalen Schlussfolgerns im Rahmen solcher Intelligenzstrukturtheorien lässt sich zunächst recht einfach beantworten: In keiner (klassischen wie aktuellen) Intelligenzstrukturtheorie existiert eine explizite Fähigkeit zum Konditionalen Schlussfolgern. Eine solche Fähigkeit müsste demnach konstruiert und − unter der Annahme, es handle sich um eine intellektuelle Leistung − zumindest theoretisch in aktuelle Intelligenzstrukturtheorien eingeordnet werden. Hierzu soll die Idee genutzt werden, die der Entwicklung vieler Intelligenzstrukturtheorien zugrunde liegt. Dabei wird eine möglichst umfassende Menge von Aufgabentypen zur Erfassung intellektueller Leistungen faktoren- oder facettenanalytisch untersucht. So entstehen Gruppen (Faktoren oder Facetten) von Aufgabentypen, die dann unter bestimmten Oberbegriffen zusammengefasst werden. Überträgt man diese Idee auf Aufgaben zum Konditionalen Schlussfolgern, stellt sich die Frage, mit welchen Aufgabentypen diese theoretisch gruppiert werden könnten. Hierzu kann auf den

aussagenlogischen Grundlagen aus Abschnitt 2.1.1 aufgebaut werden. Demnach zählen Konditionale Schlüsse zu den Propositionalen Schlüssen. Wilhelm (1995; siehe auch Wilhelm & Conrad, 1998) fasst das Lösen von Aufgaben zum Propositionalen Schlussfolgern, Syllogistischen Schlussfolgern, Relationalen Schlussfolgern, mehrfach quantifizierten sowie metadeduktiven Denken zur *Fähigkeit zum Lösen logischer Denkprobleme* zusammen. Diese Fähigkeit entspricht konzeptionell Deduktivem Denken, also der Fähigkeit, aus Prämissen notwendigerweise wahre Schlüsse abzuleiten. Deduktives Denken zählt durchaus zu den Kernbereichen menschlicher Intelligenz (Wilhelm, 1995; Wilhelm & Conrad, 1998) und ist in einer Vielzahl von Intelligenzstrukturtheorien repräsentiert. In der Folge sollen einige besonders populäre (klassische wie moderne) Intelligenzstrukturtheorien dahingehend skizziert werden, ob bzw. inwieweit Deduktives Denken als eigenständiger Faktor konzipiert wird. Wird Deduktives Denken nicht als eigenständig behandelt, soll zumindest die Rolle von *Reasoning*[9] innerhalb der betrachteten Intelligenzstrukturtheorien thematisiert werden. Reasoning umfasst Deduktives und Induktives Denken. Daher wird zunächst die Rolle von Reasoning innerhalb der jeweiligen Intelligenzstrukturtheorie betrachtet und − falls vorhanden − die Differenzierung in Deduktives und Induktives Denken thematisiert.

Die erste betrachtete Theorie ist die *Generalfaktor-Theorie* der Intelligenz von Spearman (1904), nach der ein latenter *g-Faktor* das Gemeinsame aller Leistungstestaufgaben erklärt. Interindividuelle Unterschiede werden in dieser Theorie auf g-bedingte und testspezifische Varianz zurückgeführt. Reasoning (und damit auch Deduktivem Denken) kommt dabei kein besonderer Status zu, wenngleich Spearman (1938, 1939) den Einsatz sog. *Reasoning-Tests* zur Bestimmung des g-Faktors empfiehlt. Der hohe g-bedingte Varianzanteil dieser Tests ist bis heute vielfach empirisch repliziert worden und gilt daher als unstrittig.

Als zweite wichtige Intelligenzstrukturtheorie soll die *Primärfaktorentheorie* der Intelligenz (Thurstone & Thurstone, 1941) erwähnt werden, deren Kern die Definition mehrerer Primärfaktoren der Intelligenz ist. Einen dieser Primärfaktoren stellt Reaso-

[9] Gleichwohl in dieser Arbeit – soweit möglich – deutsche Begriffe verwendet werden sollen, wird *Reasoning* als terminus technicus beibehalten und nicht die häufige deutsche Übersetzung *Schlussfolgerndes Denken* verwendet. Wirklich schlussfolgernd ist per Definition lediglich Deduktives Denken. Reasoning umfasst hingegen auch Induktives Denken und damit mehr als durch die Bezeichnung „Schlussfolgerndes Denken“ nahegelegt wird. Ergänzend sei erwähnt, dass sich in der mehr als 100-jährigen Auseinandersetzung mit dem psychologischen Konstrukt Reasoning auch andere Bezeichnungen wie *fluide Intelligenz*, *Verarbeitungskapazität* oder *Deduktives und Induktives Denken* für die damit assoziierten Denkleistungen finden (Wilhelm, 2000). Reasoning ist jedoch die populärste Bezeichnung und wird deshalb in der vorliegenden Arbeit verwendet.

ning dar, das Thurstone (1938) zunächst durch spezifische Reasoningfaktoren (u.a. Induktion und Deduktion) beschreibt, diese Unterteilung später aber aufhebt und sowohl Induktion als auch Deduktion anderen Faktoren zuordnet (Thurstone & Thurstone, 1941). Dennoch gehen die auch heute noch in der Differenziellen Psychologie häufig getroffene Unterscheidung in Induktion und Deduktion und insbesondere die entsprechenden Operationalisierungen maßgeblich auf Thurstone (1938) zurück.

Nahezu alle aktuellen Intelligenzstrukturtheorien bzw. -modelle sind Weiterentwicklungen dieser beiden Theorien. Besonders populär sind derzeit die *Drei-Stratum-Theorie* (Carroll, 1993, 1997, 2005) und die *erweiterte GfGc-Theorie*[10] (Horn & Blankson, 2005; Horn & Noll, 1997) sowie deren Zusammenführung zur *Cattell-Horn-Carroll*(*CHC*)-*Theorie* (McGrew, 2005, 2009). Reasoning ist in allen drei Theorien von zentraler Bedeutung und Deduktives Denken stellt stets einen eigenständigen Faktor dar.

Schließlich sollen noch zwei Intelligenzstrukturmodelle kurz beschrieben werden, auf die im weiteren Verlauf der vorliegenden Arbeit vermehrt Bezug genommen wird. Dabei handelt es sich um das Vier-Quellen-Modell (z.B. Kyllonen, 1994) und das *Berliner Intelligenzstrukturmodell* (Jäger, 1982).

Im Vier-Quellen-Modell von Kyllonen (z.B. Kyllonen, 1994) gilt Arbeitsgedächtniskapazität in allen Phasen des Fertigkeitserwerbs als wichtiger Prädiktor. Dabei wird Reasoning mit Arbeitsgedächtniskapazität nahezu gleichgesetzt (Kyllonen & Christal, 1990). Abweichungen werden lediglich auf Unterschiede in der Wissensbeteiligung bei den üblichen Aufgaben zu diesen Konstrukten zurückgeführt. Es wird angenommen, dass typische Reasoning-Aufgaben höhere Vorwissenanteile aufweisen als typische Arbeitsgedächtnisaufgaben (siehe dazu auch Wilhelm, 2000).

Beim Berliner Intelligenzstrukturmodell (Jäger, 1982) handelt es sich um ein bimodales Intelligenzstrukturmodell (Jäger, 1984). Es wird in die zwei Facetten *Inhalt* und *Operation* unterteilt. Als Kreuzprodukt der drei Inhaltsklassen (*verbal, numerisch, figural*) mit den vier Operationsklassen (*Verarbeitungskapazität, Einfallsreichtum, Merkfähigkeit, Bearbeitungsgeschwindigkeit*) ergeben sich 12 Zellen, über denen − quasi als Integral sämtlicher Leistungen − die allgemeine Intelligenz steht (siehe Abbildung 1).

[10] Dabei handelt es sich um eine Weiterentwicklung von Catells (1963) GfGc-Theorie.

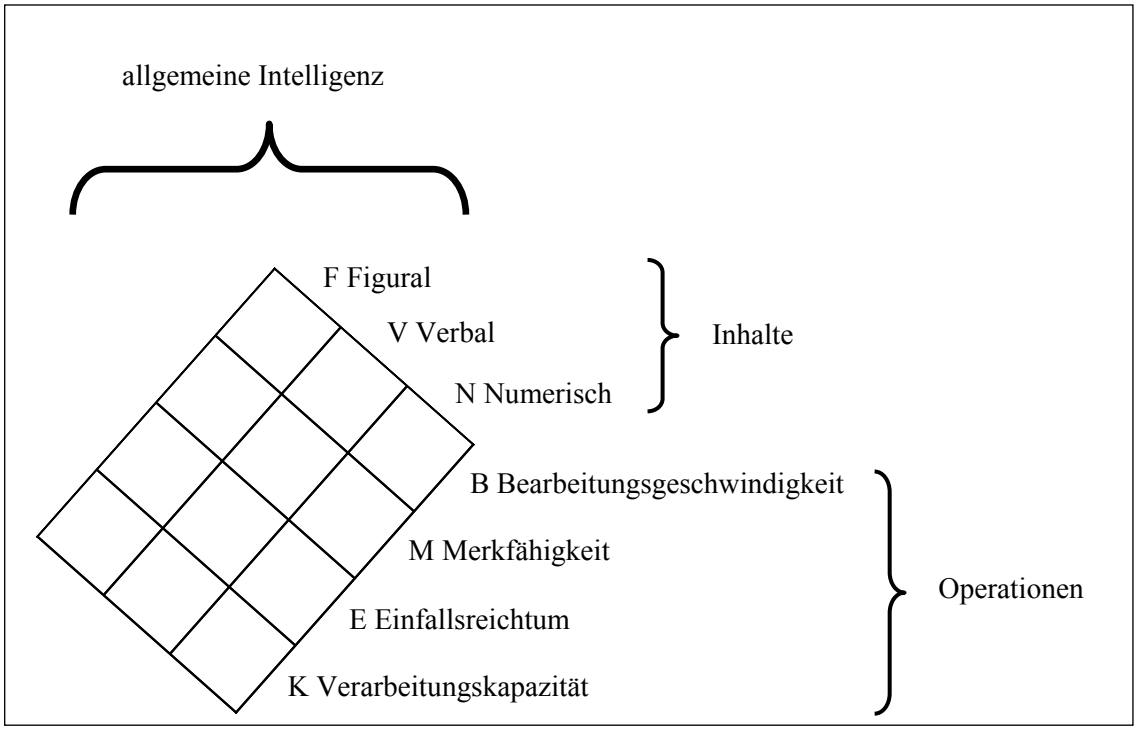


Abbildung 1: Das Berliner Intelligenzstrukturmodell nach Jäger

Für Reasoning − oder wie Jäger (1982, 1984) es bezeichnet „Verarbeitungskapazität, Urteilsfähigkeit und logisches Denken“ − ist in diesem Modell die Operationsklasse *K* (Verarbeitungskapazität) von zentraler Bedeutung. Ihr werden die Reasoning-assoziierten Denkleistungen zugeordnet.

Die Repräsentation und damit zentrale Bedeutung von Reasoning in klassischen wie modernen Intelligenzstrukturtheorien erweist sich nach diesen Ausführungen als unstrittig. Sie zeigt sich auch in weiteren facettentheoretischen Modellen intellektueller Leistungen (z.B. Guttman & Levy, 1991) sowie in der hohen prädiktiven Potenz von Reasoning (Wittmann & Süß, 1996). Weit weniger eindeutig ist jedoch, was in den betrachteten Theorien unter Reasoning verstanden wird. Beispielsweise wird es nicht in allen Intelligenzstrukturtheorien in Induktives und Deduktives Denken binnendifferenziert (Wilhelm, 2000), obwohl diese Unterscheidung definitorisch recht eindeutig ist. Betrachtet man jedoch den übergeordneten Prozess als Problemlösen, so verwischen diese definitorischen Unterschiede (Waldmann & Weinert, 1990), was sich auch darin widerspiegelt, dass sich Induktives und Deduktives Denken empirisch oft nicht klar voneinander trennen lassen (z.B. Wilhelm, 2000). Eine mögliche Erklärung hierfür wäre die bspw. in Kyllonens (1994) Vier-Quellen-Modell angenommene Limitation beider Aufgabenklassen durch die Arbeitsgedächtniskapazität. Eine zweite Erklärung könnte

sein, dass Induktive Schlüsse auch als Deduktive Schlüsse betrachtet werden können, bei denen eine oder mehrere Prämissen implizit „dazugedacht“ werden, obwohl sie nicht explizit formuliert sind (Wilhelm, 2000). Auch bei klassischen Reasoning-Tests verwischen die Grenzen. Zwar werden die meisten dieser Tests als Induktive Denktests bezeichnet, ihre Aufgabentypen umfassen jedoch fast immer (in der evaluativen Phase) auch die Regelanwendung und damit Deduktives Denken (Wilhelm, 2000). Ein Beispiel hierfür sind Zahlenreihen. Dabei muss unterschieden werden zwischen dem Erkennen der Regel und der folgerichtigen Fortsetzung, die sich bei Anwendung dieser erschlossenen Regel ergibt. Ersteres ist induktiv, zweiteres eher deduktiv. Induktives Denken und Reasoning gleichzusetzen, scheint daher unangebracht, wenn nicht gar definitorisch falsch. Vielmehr bildet Induktives Denken ebenso wie Deduktives Denken einen zentralen Bestandteil von Reasoning. Der Einsatz Deduktiver Denkprobleme in Intelligenzstrukturtests ist in der Tat seltener als der Einsatz Induktiver Denkprobleme (Beckmann & Guthke, 1999; Wilhelm, 2000). Betrachtet man jedoch Konstruktbeschreibungen zu Reasoning, zur Verarbeitungskapazität oder zur allgemeinen Intelligenz, so zählt die Fähigkeit, zu logisch zwingenden Schlussfolgerungen zu kommen oder notwendig richtige Schlüsse als solche zu erkennen, unverändert zu den zentralen Merkmalen dieser Konstrukte.

Abschließend kann festgehalten werden, dass eine Fähigkeit zum Ziehen logisch korrekter Inferenzen bei Konditionalaussagen zumindest theoretisch Deduktivem Denken zuzuordnen wäre und daher bspw. als eine „Teilfähigkeit“ Deduktiven Denkens aufgefasst werden kann. Deduktives Denken wird in aktuell besonders populären Intelligenzstrukturtheorien wie der Drei-Stratum-Theorie (Carroll, 1993, 1997, 2005), der erweiterten GfGc-Theorie (Horn & Blankson, 2005; Horn & Noll, 1997) oder der CHC-Theorie (McGrew, 2005, 2009) stets als eigenständiger Faktor repräsentiert und jeweils einem Faktor zugeordnet, der konzeptuell Reasoning entspricht. Zumindest theoretisch ist also die Einordnung einer Fähigkeit zum Konditionalen Schlussfolgern in Intelligenzstrukturtheorien möglich, sodass die Frage gerechtfertigt ist, inwieweit bereits Versuche unternommen wurden, eine solche Fähigkeit messbar zu machen. Dies ist Gegenstand des folgenden Abschnitts.

2.1.4.2 Testverfahren zur Erfassung einer Fähigkeit zum Konditionalen Schlussfolgern

Es steht außer Frage, dass ein Testverfahren zur Erfassung einer Fähigkeit zum Konditionalen Schlussfolgern aus Aufgaben zum Konditionalen Schlussfolgern bestehen sollte. Der folgende Abschnitt widmet sich solchen Testverfahren in Verbindung mit der Frage, ob eine eigene Testkonstruktion überhaupt notwendig ist und – gegeben sie ist es – welche Implikationen aus der Betrachtung bereits existierender Tests für die beabsichtigte Testkonstruktion abgeleitet werden können.

Die Ausführungen zu interindividuellen Unterschieden im Konditionalen Schlussfolgern (siehe Abschnitt 2.1.3) zeigen einen aktuellen Trend: die Verwendung allgemeinpsychologischer Aufgaben zur Beantwortung differenzialpsychologischer Fragestellungen. Dabei scheint zunächst naheliegend, noch einen Schritt weiterzugehen und aus solchen Aufgaben einen Test zur Erfassung interindividueller Unterschiede im Konditionalen Schlussfolgern zu konstruieren. Es existiert jedoch nur ein publizierter deutschsprachiger Test, der aus Aufgaben zum Konditionalen Schlussfolgern besteht. Dabei handelt es sich um den „Leistungsprofiltest Schlussfolgerndes Denken – Verbal“ (SDV; Spiel, Glück & Gößler, 2004), auf den später in diesem Abschnitt auch vertieft eingegangen wird. Doch bereits an dieser Stelle sei erwähnt, dass dieser Test in einem anderen Kontext entstanden ist, nämlich zur Analyse bzw. Beschreibung entwicklungspsychologischer Prozesse im Kindes- und Jugendalter (vgl. Spiel et al., 2004). Umso mehr stellt sich die Frage, warum dieser naheliegende Schritt bislang nicht unternommen wurde. Die Antwort liegt aus Sicht des Autors nicht allein im Konditionalen Schlussfolgern begründet. Betrachtet man die Vielzahl an existierenden Reasoning-Tests, so fällt auf, dass Tests zum Deduktiven Denken gegenüber Tests zum Induktiven Denken stark unterrepräsentiert sind (Beckmann & Guthke, 1999; Wilhelm, 2000). Wilhelm (2000) führt das darauf zurück, dass Tests bzw. Items zum Induktiven Denken üblicherweise sehr gute psychometrische Eigenschaften aufweisen. Möglicherweise reichen auch Tests zum Induktiven Denken aus, um Reasoning hinreichend zu erfassen, wie Beckmann und Guthke (1999) vermuten. Unabhängig von den Gründen – angesichts der langen Tradition allgemeinpsychologischer Forschung zum Deduktiven Denken scheint die Unterrepräsentation in der differenzialpsychologischen Forschung bzw. in der Diagnostik insgesamt unangemessen. Zumindest existieren Tests zu einzelnen Formen Deduktiven Denkens (zum Syllogistischen sowie Relationalen Schlussfolgern z.B. Wilhelm & Conrad, 1998). Doch zum Propositionalen Schlussfolgern (dem Kondi-

tionales Schlussfolgern formal zugeordnet wird; vgl. Abschnitt 2.1.1) existieren lediglich Aufgabensammlungen (z.B. O'Brien et al., 1994; Rips, 1983). Zwar enthalten diese auch Aufgaben zum Konditionalen Schlussfolgern, dennoch bleibt festzuhalten, dass zum Deduktiven Denken vergleichsweise wenige Testverfahren existieren und explizit zum Konditionalen Schlussfolgern überhaupt keine (deutschsprachigen) Tests bis auf den bereits erwähnten SDV (Spiel et al., 2004). Dieser soll daher in der Folge näher betrachtet werden.

Als erstes ist festzuhalten, dass der SDV nicht explizit auf die Messung einer Fähigkeit zum Konditionalen Schlussfolgern abzielt. Beim SDV werden Konditionale Syllogismen als Indikatoren für die Fähigkeit zum Schlussfolgernden Denken verwendet. Die theoretische Basis des SDV bildet Piagets (1971) Entwicklungstheorie, die etwa im 12. Lebensjahr einen Übergang vom konkret-operatorischen zum formal-operatorischen Denken annimmt. Die Annahme dieses Übergangs wird von Spiel et al. (2004) zunächst auf das Lösen von Aufgaben zum Konditionalen Schlussfolgern übertragen. Aufbauend auf Befunden der Performanz-Kompetenz-Forschung sowie eigenen Arbeiten (Spiel, Gittler, Sirsch & Glück, 1997; Spiel, Glück & Gößler, 2001) vermuten Spiel et al. (2004), dass sich dieser Übergang zum einen in kleineren Schritten vollzieht als von Piaget (1971) angenommen und zum anderen von verschiedenen Moderatorvariablen beeinflusst wird. Es werden zwei Übergangsstadien postuliert sowie moderierende Effekte des Aufgabeninhaltes und der Verwendung von Negationen im Antezedens der Hauptprämisse. Das Untersuchungsmaterial des SDV bilden 24 klassische Aufgaben zum Konditionalen Schlussfolgern, in denen die Moderatoren (Negation vs. keine Negation im Antezedens der Hauptprämisse sowie drei Inhaltsklassen) systematisch variiert und mit den vier Schlussfiguren des Konditionalen Schlussfolgerns (MP, NA, BK, MT; siehe Abschnitt 2.1.1) kombiniert werden. Unter Verwendung von *Mixed-Rasch-Modellen* (*MRM*; Rost, 1989, 1990; siehe auch Rost, 2004) werden die Daten von 418 Schülerinnen und Schülern der Klassenstufen 7-12 analysiert. Die Ergebnisse werden als konform mit den aufgestellten Hypothesen (siehe dazu Spiel et al., 2004) interpretiert und Befunde zu Reliabilität, Validität und Objektivität des SDV angeführt.

Gegen einen Einsatz des SDV im Rahmen des in Kapitel 1 skizzierten Forschungsprojektes spricht jedoch, dass die Zielgruppe des SDV Jugendliche sind, während für den zu konstruierenden Test Erwachsene die Zielgruppe darstellen, konkret Chipdesigner (siehe Ausführungen in Kapitel 1). Bei Chipdesignern handelt es sich

größtenteils um Hochschulabsolventen, für die weder theoretisch noch praktisch die Vorgabe eines „Tests für Jugendliche“ angezeigt wäre. Dennoch stellt sich die Frage, welche Implikationen aus den Entwicklungsarbeiten des SDV für die Eigenkonstruktion eines Tests zum Konditionalen Schlussfolgern abgeleitet werden können. Letztlich sprechen jedoch methodische Gründe wie auch empirische Befunde zum SDV für eine grundständige Neukonstruktion. Da der zu entwickelnde Test im Rahmen der Betrachtungen zum Testgütekriterium „Nützlichkeit“ (siehe Abschnitt 3.7) ohnehin einem konzeptuellen wie methodischen Vergleich mit dem SDV standhalten muss, werden diese Punkte jedoch erst später bei den Ausführungen zu diesem Testgütekriterium ausführlich thematisiert (siehe dazu Abschnitt 3.7).

Zwei weitere Testverfahren sollen an dieser Stelle ebenfalls kurz erwähnt werden, wenngleich sie nicht aus Aufgaben zum Konditionalen Schlussfolgern bestehen. Sie erfassen aber zumindest andere Formen Deduktiven Denkens und basieren auf einer der beiden in Abschnitt 2.1.2 vorgestellten kognitionspsychologischen Theorien, nämlich der Theorie mentaler Modelle. Es handelt sich um zwei Messinstrumente zur Erfassung der Fähigkeit zum Lösen logischer Denkprobleme (Wilhelm, 1995; siehe auch Wilhelm & Conrad, 1998). Bei beiden Instrumenten wird angenommen, dass das Lösen logischer Denkaufgaben auf der Grundlage der Erzeugung, Aufrechterhaltung und Bearbeitung mentaler Modelle erfolgt. Die konstruierten Items beschränken sich auf die Bereiche des Syllogistischen und (räumlich) Relationalen Schlussfolgerns. Aus hohen Korrelationen der beiden Tests mit Verarbeitungskapazität ($r = .62$ bzw. $r = .63$) wird geschlussfolgert, die Fähigkeit zum Logischen Denken könne als Teilbereich der Verarbeitungskapazität aufgefasst werden. Entgegen den Erwartungen von Wilhelm (1995; siehe auch Wilhelm & Conrad, 1998) fällt die Interkorrelation der beiden Tests zum Logischen Denken mit $r = .42$ vergleichsweise niedrig aus. Die Itemschwierigkeiten des Syllogistischen Tests sind theoriegeleitet gut vorhersagbar, dafür ist er psychometrisch auffällig (entgegen den Erwartungen sind die Items nicht homogen). Im Gegensatz dazu erweisen sich die Items des Relationalen Tests als psychometrisch homogen. Hier ist jedoch die theoriegeleitete Vorhersage der Itemschwierigkeiten nicht gewährleistet. Dennoch scheint die Theorie mentaler Modelle als Grundlage für die Konstruktion von Tests zum Deduktiven Denken ein möglicher Ansatz. Zu einem ähnlichen Ergebnis gelangen Beckmann und Guthke (1999), die besonders die modelltheoretisch abgesicherte Konstruktion der beiden Tests von Wilhelm (1995; siehe auch Wilhelm & Conrad, 1998) positiv hervorheben.

Zusammenfassend kann festgehalten werden, dass die Eigenkonstruktion eines Tests zum Konditionalen Schlussfolgern prinzipiell gerechtfertigt scheint, da aktuell keine deutschsprachigen Tests mit Aufgaben zum Konditionalen Schlussfolgern existieren, die den Anforderungen des in der Einleitung beschriebenen Forschungsprojektes (vgl. Kapitel 1) genügen. Allerdings existieren Tests zum Deduktiven Denken mit anderen Aufgabentypen, die zwar zum Teil Einschränkungen bezüglich ihrer psychometrischen Kriterien aufweisen, zumindest aber auf einer kognitionspsychologischen Theorie, der Theorie mentaler Modelle, basieren.

2.2 Ableiten eines kognitionspsychologisch fundierten Modells als Grundlage für die Testkonstruktion

Im folgenden Abschnitt soll aus den bisherigen Betrachtungen ein kognitionspsychologisch fundiertes Modell abgeleitet werden, das als Grundlage für die Testkonstruktion dient. Dieses Modell sollte aufgrund der Ausführungen zu den beiden zentralen kognitionspsychologischen Theorien zum Konditionalen Schlussfolgern (Modelltheorie, siehe Abschnitt 2.1.2.1; Logiktheorie, siehe Abschnitt 2.1.2.2) auf der Logiktheorie basieren, da diese bezüglich der vorgestellten empirischen Befunde die größere Erklärungskraft besitzt (vgl. Abschnitt 2.1.2.3). Allerdings sprechen die Ausführungen zu bereits existierenden Testverfahren (siehe Abschnitt 2.1.4.2) für die Modelltheorie, da Tests zu anderen Formen Deduktiven Denkens (Wilhelm, 1995; Wilhelm & Conrad, 1998) auf Basis der Modelltheorie konstruiert werden können (vgl. Abschnitt 2.1.4.2). Daher soll zunächst überprüft werden, inwieweit eine Übertragbarkeit der Ergebnisse von Wilhelm (1995; siehe auch Wilhelm & Conrad, 1998) auf Konditionales Schlussfolgern möglich ist.

Neben den Aufgaben zum Syllogistischen und Relationalen Schussfolgern werden von Wilhelm (1995; siehe auch Wilhelm & Conrad, 1998) in der dazu durchgeführten Studie auch einige Items zum Konditionalen Schlussfolgern vorgegeben (Wilhelm, 2000). Ein Großteil dieser Items weist – im Gegensatz zu denen zum Syllogistischen bzw. Relationalen Schlussfolgern – keine adäquaten psychometrischen Eigenschaften auf und muss daher aus weiteren Analysen ausgeschlossen werden (Wilhelm, 2000). So erweist sich die Modelltheorie für die Konstruktion von Aufgaben zum Syllogistischen und Relationalen Schlussfolgern offenbar als gut geeignet, scheint jedoch als Basis für

die Konstruktion eines Tests zum Konditionalen Schlussfolgern eher ungeeignet zu sein[11]. Damit überwiegen die Argumente für die Logiktheorie als Basis für die Testkonstruktion[12] und der nächste Schritt ist, ein Modell abzuleiten bzw. zu entwickeln, welches differenzialpsychologische Aussagen erlaubt und somit die Grundlage für die Testkonstruktion bildet. Gegenstand dieses Abschnitts ist die Entwicklung eines solchen Modells, inklusive dessen (theoretischer) Erklärungskraft für bereits berichtete (siehe Abschnitt 2.1.2) wie auch weitere empirische Befunde.

2.2.1 Der Weg zu einem differenzialpsychologischen Modell Konditionalen Schlussfolgerns

Es steht außer Frage, dass ein Test mit dem Anspruch, eine Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern zu erfassen, alle vier Schlussfiguren (siehe Abschnitt 2.1.1) berücksichtigen muss: Modus Ponens (MP), Modus Tollens (MT), Negation des Antezedens (NA) und Bestätigung der Konsequenz (BK). Folglich müssen auch alle vier Schlussfiguren in dem der Testkonstruktion zugrundeliegenden Modell repräsentiert sein. Ebenso steht außer Frage, dass dieses Modell mit den bereits mehrfach erwähnten empirischen Befunden zum Konditionalen Schlussfolgern (siehe Abschnitt 2.1.2) vereinbar sein muss. Aus allgemeinpsychologischer Perspektive müssen mit dem zu entwickelnden Modell die Befunde erklärbar sein, dass Modus-Ponens-Aufgaben häufiger korrekt gelöst werden als Modus-Tollens-Aufgaben und dass Aufgaben zu den validen Schlussfiguren (MP, MT) häufiger korrekt gelöst werden als Aufgaben zu den invaliden (NA, BK). Da die Testkonstruktion auf der Logiktheorie basieren soll, ist beides gegeben (vgl. Abschnitt 2.1.2.2). Unter Berücksichtigung dieser Punkte ist nun ein Modell zu entwickeln, das zudem differenzialpsychologische Aussagen erlaubt. Dies ist Gegenstand des folgenden Abschnitts.

Ausgangspunkt ist eine Idee von Rumain, Connell und Braine (1983), die in ihrer Arbeit der Frage nachgehen, ob die Fehler, die Kinder beim Konditionalen Schlussfolgern machen, die Folge eines fehlerhaften lexikalischen Eintrags des Wortes „wenn“ sind. Ist dieser Eintrag bei Kindern einer bestimmten Altersstufe bikonditional (statt korrekt: konditional), wie bspw. auch Johnson-Laird (2001; Johnson-Laird & Byrne,

[11] Mögliche Erklärungen hierfür ergeben sich im weiteren Verlauf der vorliegenden Arbeit und werden in Kapitel 4 ausführlich diskutiert.
[12] Aufgrund der immensen Bedeutung der Theorie mentaler Modelle für Deduktive Denkprozesse (siehe dazu z.B. Wilhelm, 2000) sollen jedoch die angestellten Überlegungen an entscheidenden Stellen immer auch aus Perspektive der Theorie mentaler Modelle betrachtet werden.

2002) annimmt, dann wäre eine Reihe von Fehlern beim Konditionalen Schlussfolgern darauf zurückführbar. Rumain et al. (1983) bevorzugen als Erklärung allerdings das auf Geis und Zwicky (1971) zurückgehende Konzept der einladenden Inferenzen (siehe auch Abschnitt 2.1.2.2), welches sich dem Konditionalen Schlussfolgern aus einer linguistischen Perspektive nähert[13]. Zur Erinnerung sei folgendes Beispiel zur einladenden Inferenz erneut angeführt (siehe auch Abschnitt 2.1.2.2): „*Wenn Du den Rasen mähst, gebe ich Dir 5 Dollar*". Die entsprechende einladende Inferenz wäre dann „*Wenn Du den Rasen nicht mähst, gebe ich Dir keine 5 Dollar*". Durch das Akzeptieren dieser einladenden Inferenz wird eine eigentlich konditionale Aussage ($p \rightarrow q$) als bikonditional interpretiert und es werden auch für die Schlussfiguren Negation des Antezedens und Bestätigung der Konsequenz eindeutige Inferenzen gezogen (bei NA: $\neg q$, bei BK: p). Rumain et al. (1983) betonen dabei, dass die bikonditionale Interpretation nicht die Folge eines fehlerhaften lexikalischen Eintrags für „wenn" ist, sondern Folge des Akzeptierens der einladenden Inferenz. Zur Überprüfung dieser Idee ergänzen Rumain et al. (1983) zusätzliche Prämissen, die die einladende Inferenz blockieren sollen. Folgendes Beispiel soll das illustrieren:

Hauptprämisse: „Wenn Du den Rasen mähst, gebe ich Dir 5 Dollar."
Zusatzprämisse: „Wenn Du abwäschst, gebe ich Dir 5 Dollar."
Nebenprämisse 1 (Negation des Antezedens): „Die Person mäht nicht den Rasen."
Nebenprämisse 2 (Bestätigung der Konsequenz): „Die Person erhält 5 Dollar."
Nebenprämisse 3 (Modus Tollens): „Die Person erhält keine 5 Dollar."

Es zeigt sich − wie von Rumain et al. (1983) erwartet − ein Anstieg der für die konditionale Interpretation korrekten „keine Aussage möglich (k.A.m.)"-Reaktionen im Falle von NA und BK, jedoch steigt ebenso die Zahl der „k.A.m."-Reaktionen für den Modus Tollens (Rumain et al., 1983). Rumain et al. (1983) erklären das damit, dass auch ein Modus-Tollens-Schluss auf der einladenden Inferenz und damit auf einer bikonditionalen Interpretation basieren kann. Durch das Blockieren der einladenden Inferenz wird dieser Modus-Tollens-Schluss nun ebenfalls blockiert. Zusätzlich zu einem solchen unsophistizierten, auf einer bikonditionalen Interpretation basierenden Modus-Tollens-Schluss existiert noch ein sophistizierter Modus-Tollens-Schluss, der bei einer konditionalen und damit korrekten Interpretation ganz im Sinne der Logiktheorie (Braine &

[13] Die große Bedeutung linguistischer Prozesse beim Schlussfolgernden Denken ist unter Kognitionspsychologen überdies seit Langem unumstritten (z.B. Evans, 1977a).

O'Brien, 1991) gezogen wird. Für diesen ist die Aneinanderreihung vier verschiedener Inferenzschemata notwendig: einer Annahme (p), des Modus Ponens (q), einer Kontradiktion (Inkompatibilität) sowie der reductio ad absurdum ($\neg p$). Dieser sophistizierte Modus-Tollens-Schluss ist nicht Folge des Akzeptierens der einladenden Inferenz und sollte daher unabhängig davon gezogen werden.

Sämtliche beschriebene Effekte zeigen sich nicht nur bei Kindern, sondern in nicht unerheblichem Maße auch bei Erwachsenen. Rumain et al. (1983) schlussfolgern daraus, der lexikalische Eintrag für „wenn" sei bei Kindern der gleiche wie bei Erwachsenen. Fehler, die Erwachsene bei der Negation des Antezedens und bei der Bestätigung der Konsequenz machen, seien wie auch bei Kindern auf das Akzeptieren der einladenden Inferenz beim Konditionalen Schlussfolgern zurückzuführen bzw. beim Modus Tollens auf deren Blockieren (Rumain et al., 1983). Ein entscheidender Punkt in der Argumentation von Rumain et al. (1983) ist, dass einladenden Inferenzen zu „widerstehen" und damit die Aussage als konditional zu interpretieren als (kognitive) Leistung interpretiert und als „*Sophistiziertheits-Stufe*" bezeichnet wird. Als Konsequenz für differenzialpsychologische Überlegungen bleibt festzuhalten, dass nach Rumain et al. (1983) bezüglich des logischen Schlussfolgerns bei Konditionalaussagen zwei Gruppen von Personen existieren: eine, die die einladende Inferenz akzeptiert, und eine, die ihr widersteht. Dieses „Widerstehen" kann als Leistung interpretiert werden. Des Weiteren gibt es zwei unterschiedliche Möglichkeiten einen formal korrekten Modus-Tollens-Schluss zu ziehen: eine unsophistizierte infolge des Akzeptierens der einladenden Inferenz und eine sophistizierte im Sinne der Logiktheorie. Beide Überlegungen werden von Rijmen und De Boeck (2003) zu einem *Stufen-Modell* zusammengefasst.

2.2.2 Das Stufen-Modell zum Konditionalen Schlussfolgern

Rijmen und De Boeck (2003) postulieren drei Stufen. Auf einer ersten, sog. *Unsophistizierten*-Stufe akzeptieren Personen die einladende Inferenz ($\neg p \rightarrow \neg q$) einer Konditionalaussage ($p \rightarrow q$), interpretieren die Aussage daher als bikonditional und ziehen für alle vier Schlussfiguren eine (eindeutige) Inferenz. Auf einer *Fortgeschrittenen*-Stufe gelingt es Personen, die einladende Inferenz zu blockieren und folglich nur bei Modus-Ponens-Problemen eine eindeutige Inferenz zu ziehen. Für die anderen drei Schlussfiguren (NA, BK, MT) wird die „k.A.m."-Reaktion gewählt, die zumindest für NA und BK auch korrekt ist. Eine dritte Stufe baut auf der Idee eines sophistizierten Modus-Tollens-

Schlusses (siehe Abschnitt 2.2.1) im Sinne der Logiktheorie (Braine & O'Brien, 1991) auf. Auf dieser *Sophistizierten*-Stufe gelingt es Personen, sowohl die einladende Inferenz zu blockieren als auch einen sophistizierten Modus-Tollens-Schluss zu ziehen. Personen auf dieser Stufe ziehen also bei Modus-Ponens- und bei Modus-Tollens-Problemen eindeutige Inferenzen, entscheiden sich bei NA und BK allerdings für die „k.A.m."-Reaktion. In der von Rijmen und De Boeck (2003) berichteten Studie lassen sich Personen auf der Unsophistizierten-Stufe und Personen auf der Fortgeschrittenen-Stufe identifizieren. Empirische Evidenz für die Sophistizierten-Stufe ergibt sich hingegen nicht. Da es jedoch unplausibel ist, dass generell niemand in der Lage sein soll, sämtliche vier Schlussfiguren zum Konditionalen Schlussfolgern korrekt zu lösen, scheinen für diesen Befund andere Ursachen (z.B. Stichprobeneffekte, methodische Probleme) wahrscheinlich. Die Stufen-Überlegungen von Rijmen und De Boeck (2003) werden hinsichtlich des korrekten Lösens der vier Schlussfiguren von Konditionalaussagen in Tabelle 4 zusammengefasst.

Tabelle 4: Löseverhalten für die vier Schlussfiguren Konditionalen Schlussfolgerns auf Basis der Überlegungen des Stufen-Modells

	Schlussfiguren bei der Konditionalaussage $p \rightarrow q$			
Sophistiziertheits-Stufe	MP (p)	NA ($\neg p$)	BK (q)	MT ($\neg q$)
Lösung auf Stufe 1: (Unsophistiziert)	q (korrekt)	$\neg q$ (Fehler)	p (Fehler)	$\neg p$ (korrekt)
Lösung auf Stufe 2: (Fortgeschritten)	q (korrekt)	k.A.m. (korrekt)	k.A.m. (korrekt)	k.A.m. (Fehler)
Lösung auf Stufe 3: (Sophistiziert)	q (korrekt)	k.A.m. (korrekt)	k.A.m. (korrekt)	$\neg p$ (korrekt)

Anmerkung. In Klammern ist die normativ-logische Korrektheit der Lösung angegeben. MP … Modus Ponens, NA … Negation des Antezedens, BK … Bestätigung der Konsequenz, MT … Modus Tollens, k.A.m. … keine Aussage möglich.

Aus kognitionspsychologischen Theorien lässt sich also ein Modell ableiten, mit dem interindividuelle Unterschiede im Konditionalen Schlussfolgern erklärt und unterschiedliche Ausprägungen im Sinne einer kognitiven Leistung interpretiert werden können. Allerdings handelt es sich dabei nicht um eine dimensionale Fähigkeitskonzeption Konditionalen Schlussfolgerns (als „Teilfähigkeit" Deduktiven Denkens wie in Abschnitt 2.1.4.1 beschrieben), sondern vielmehr um eine Art (Kompetenz-)Stufen-

Modell. In der Folge soll überprüft werden, inwieweit dieses Stufen-Modell in der Lage ist, empirische Befunde zu erklären.

2.2.3 Erklärungskraft des Stufen-Modells für weitere empirische Befunde

Obwohl das Stufen-Modell in der postulierten Form noch nicht vollständig empirisch repliziert werden kann (vgl. Rijmen & De Boeck, 2003), spricht für dieses Modell, dass es einige bislang eher kontraintuitive empirische Befunde erklären kann. Ein besonders auffälliger ist der *negative* Zusammenhang zwischen dem Lösen von Modus-Tollens-Problemen und den Ergebnissen klassischer Reasoning-Tests (Evans et al., 2007[14]; Newstead et al., 2004). Zum Lösen eines Modus-Tollens-Problems sind nach der Logiktheorie (Braine & O'Brien, 1991) insgesamt vier Inferenzschemata notwendig (vgl. Abschnitt 2.2.1), nach der Modelltheorie (Johnson-Laird & Byrne, 2002) drei vollständig explizite Modelle (vgl. Abschnitt 2.1.2.1). Sowohl die Zahl der verfügbaren Inferenzschemata als auch die Zahl der vollständig expliziten Modelle werden als abhängig von der Arbeitsgedächtniskapazität angenommen (Braine & O'Brien, 1991; Johnson-Laird & Byrne, 2002). In Kyllonens (1994) Vier-Quellen-Modell gelten nun Arbeitsgedächtniskapazität und Reasoning als nahezu identisch (Kyllonen & Christal, 1990), was durch sehr hohe empirische Zusammenhänge (z.B. Süß, Oberauer, Wittmann, Wilhelm & Schulze, 2002) gestützt wird. Demnach wäre nach beiden Theorien ein positiver Zusammenhang zwischen dem Lösen von Modus-Tollens-Problemen und den Ergebnissen klassischer Reasoning-Tests zu erwarten. Nach dem Stufen-Modell lösen nun sowohl Unsophistizierte als auch Sophistizierte ein Modus-Tollens-Problem korrekt. Lediglich auf der Fortgeschrittenen-Stufe wird ein Modus-Tollens-Problem nicht korrekt gelöst. Angenommen, man würde die Ordnung der drei Stufen auf Reasoning[15] übertragen, dann würden Personen mit niedrigen Reasoning-Ausprägungen ein Modus-Tollens-Problem ebenfalls korrekt lösen, wenn auch aufgrund einer bikonditionalen Interpretation infolge des Akzeptierens der einladenden Inferenz. Personen mit mittleren Reasoning-Ausprägungen lösen ein Modus-Tollens-Problem nicht korrekt, da sie zwar der einladenden Inferenz widerstehen, jedoch bspw. nicht über ausreichend Arbeitsgedächtniskapazität verfügen, um die Inferenzschemata für eine sophistizierte

[14] Bei Evans et al. (2007) zeigt sich der negative Zusammenhang tendenziell, jedoch nicht signifikant.
[15] In der Folge bezieht sich der Begriff „Reasoning" stets auf eine (dimensionale) Fähigkeitskonzeption, also die „Fähigkeit zum Reasoning".

Lösung eines Modus-Tollens-Problems anzuwenden. Personen mit hohen Reasoning-Ausprägungen gelingt letztlich beides – sowohl das Blockieren der einladenden Inferenz als auch das sophistizierte Lösen des Modus-Tollens-Problems. Der Zusammenhang zwischen dem Lösen von Modus-Tollens-Problemen und Reasoning wäre nach dem Stufen-Modell also allenfalls durch eine quadratische Funktion (zunächst fallend, dann wieder steigend) beschreibbar und könnte demnach durchaus zu einer Nullkorrelation führen[16]. Geht man nun noch davon aus, dass Personen auf der Unsophistizierten-Stufe (mit niedrigen Reasoning-Ausprägungen) gegenüber Personen auf der Sophistizierten-Stufe (mit hohen Reasoning-Ausprägungen) stark überrepräsentiert sind, ergibt sich der beschriebene negative (lineare) Zusammenhang. Für die Überrepräsentation von Personen auf der Unsophistizierten-Stufe sprechen sowohl die Ergebnisse von Rijmen und De Boeck (2003), die überhaupt keine Personen auf der Sophistizierten-Stufe identifizieren können[17], als auch die empirischen Befunde zu Beginn von Abschnitt 2.1.2, was in der Folge verdeutlicht werden soll.

Prüfung der Behauptung „Die Unsophistizierten-Stufe ist die häufigste."

Ausgangspunkt der folgenden Argumentation ist die Annahme, dass das Stufen-Modell gilt. Demnach sind Personen auf der Unsophistizierten-Stufe daran zu erkennen, dass sie bei Negation des Antezedens (NA) und bei Bestätigung der Konsequenz (BK) eine (logisch falsche) Inferenz ziehen (vgl. Tabelle 4 in Abschnitt 2.2.2). Das unterscheidet sie von Personen auf den anderen beiden Stufen. Als Datengrundlage werden die bei Kleinbeck (2005) berichteten, aggregierten Daten verwendet, da diese auf der größten, dem Autor bekannten Stichprobe ($N = 325$) zu den vier Schlussfiguren Konditionalen Schlussfolgerns basieren. Es ist jedoch anzumerken, dass weder bei Kleinbeck (2005) noch in einer der anderen betrachteten allgemeinpsychologischen Studien (siehe Evans et al., 1993 für einen Überblick) abhängige Antwortmuster auf die vier Schlussfiguren berichtet werden. Demnach sind sämtliche Schätzungen provisorisch, da sie lediglich auf den Prozentangaben der einzelnen Schlussfiguren basieren und zudem keinerlei Messfehler berücksichtigt werden. Bei Kleinbeck (2005) zeigen sich jedenfalls folgende prozentualen Häufigkeiten für gezogene eindeutige Inferenzen bei den vier Schlussfiguren:

[16] Diese theoretisch mögliche Nullkorrelation erklärt möglicherweise die Befunde von Evans et al. (2007).

[17] Hierfür werden allerdings noch andere Gründe vermutet, die jedoch erst in Abschnitt 2.3.1 angeführt werden.

MP: 95% NA: 60% BK: 68% MT: 67%

Es ziehen also 60% (bei NA) bzw. 68% (bei BK) der Personen logisch falsche Inferenzen. Unterstellt man, dass die drei Stufen des Stufen-Modells die Stichprobe erschöpfend charakterisieren, kann man davon ausgehen, dass Personen auf der Unsophistizierten-Stufe den größten Teil der betrachteten Stichprobe bilden. Dies spricht für die aufgestellte Behauptung („*Die Unsophistizierten-Stufe ist die häufigste*"). Als Schätzung für die Häufigkeit der Unsophistizierten-Stufe soll vorerst der Durchschnitt aus den prozentualen Häufigkeiten für NA und BK dienen. Demnach wären 64% der betrachteten Stichprobe dieser Stufe zuzuordnen, wobei nochmals betont werden soll, dass dies aus den bereits genannten Gründen lediglich eine provisorische Schätzung darstellt. Auch für die anderen beiden Stufen lassen sich aus den Daten Auftretenshäufigkeiten schätzen. So wären nach dem Stufen-Modell sämtliche Personen, die beim Modus Tollens keine Inferenz ziehen, der Fortgeschrittenen-Stufe zuzuordnen, da auf den anderen beiden Stufen die korrekte Modus-Tollens-Inferenz gezogen wird. Für die bei Kleinbeck (2005) berichteten Daten sind das 33%. Abschließend sind aus den Personen, die beim Modus Tollens eine Inferenz ziehen (67%), diejenigen herauszufiltern, die der Sophistizierten-Stufe zuzuordnen sind. Nach Abzug der 64% Unsophistizierten wären das also lediglich 3%. Tabelle 5 fasst diese provisorischen Schätzungen zusammen.

Tabelle 5: Provisorische Schätzung der Auftretenshäufigkeit der drei Stufen des Stufen-Modells anhand der bei Kleinbeck (2005) berichteten Daten

Sophistiziertheits-Stufe	Schätzung anhand der bei Kleinbeck (2005) berichteten Daten
Unsophistiziert	64%
Fortgeschritten	33%
Sophistiziert	3%

Doch das Stufen-Modell wird noch durch weitere empirische Befunde gestützt. Im Gegensatz zum Modus Tollens finden sich für das korrekte Lösen von NA- und BK-Problemen positive Zusammenhänge mit der Arbeitsgedächtniskapazität (De Neys, Schaeken & d'Ydewalle, 2005; Newstead et al., 2004) und damit mit Reasoning[18]. Nach dem Stufen-Modell lösen nun Personen auf der Unsophistizierten-Stufe Aufgaben zu

[18] Ergänzend sei erwähnt, dass es sich bei den Ergebnissen von De Neys et al. (2005) sowie Newstead et al. (2004) um negative Zusammenhänge zwischen der Arbeitsgedächtniskapazität und dem Ziehen der falschen Inferenzen bei NA bzw. BK handelt, die vice versa interpretiert werden.

NA und BK nicht korrekt, wohl aber Personen auf der Fortgeschrittenen- oder Sophistizierten-Stufe. Diese Befunde unterstützen also die Annahme, dass es sich bei den drei Stufen tatsächlich um eine Art (geordnete) Kompetenz-Stufen handelt. Höhere Stufen gehen mit höheren Ausprägungen der Arbeitsgedächtniskapazität und damit höheren Ergebnissen in klassischen Reasoning-Tests einher. Abbildung 2 fasst die vermutete Beziehung zwischen den Sophistiziertheits-Stufen Konditionalen Schlussfolgerns (basierend auf dem Stufen-Modell) und Reasoning zusammen.

	Löseverhalten bezüglich der Schlussfiguren bei einer Konditionalaussage $p \rightarrow q$		
Sophistiziertheits-Stufe nach dem Stufen-Modell	NA ($\neg p$)	BK (q)	MT ($\neg q$)
Lösung auf Stufe 1: (Unsophistiziert)	$\neg q$ (Fehler)	p (Fehler)	$\neg p$ (korrekt)
Lösung auf Stufe 2: (Fortgeschritten)	k.A.m. (korrekt)	k.A.m. (korrekt)	k.A.m. (Fehler)
Lösung auf Stufe 3: (Sophistiziert)	k.A.m. (korrekt)	k.A.m. (korrekt)	$\neg p$ (korrekt)

Reasoning ↓

Legende. NA … Negation des Antezedens, BK … Bestätigung der Konsequenz, MT … Modus Tollens, k.A.m. … keine Aussage möglich.
Anmerkung. Auf die Darstellung der Schlussfigur Modus Ponens wird verzichtet, da diese auf allen drei Stufen korrekt gelöst wird (siehe Tabelle 4 in Abschnitt 2.2.2). In Klammern ist jeweils die normativ-logische Korrektheit der Lösung angegeben.

Abbildung 2: Vermutete Beziehung zwischen Reasoning und den Sophistiziertheits-Stufen Konditionalen Schlussfolgerns nach dem Stufen-Modell

Das Stufen-Modell ist also in der Lage, drei qualitativ unterschiedliche Antwortmuster in eine Rangreihe zu bringen, die − zumindest theoretisch − positiv mit Reasoning zusammenhängen sollte. Personen unterscheiden sich demnach in ihrer Sophistiziertheits-Stufe Konditionalen Schlussfolgerns. Auf Basis des Stufen-Modells sind also differenzialpsychologische Aussagen möglich, womit es den eingangs formulierten Anforderungen (siehe Abschnitt 2.2) genügt.

Es sei an dieser Stelle erwähnt, dass die Idee, qualitativ unterschiedliches Antwortverhalten beim Konditionalen Schlussfolgern mit Reasoning in Beziehung zu setzen, auch von Evans et al. (2007) verfolgt wird. Im Gegensatz zum Stufen-Modell wird dort jedoch anstelle einer bikonditionalen Interpretation (inklusive eines entsprechenden

Antwortmusters) eine *konjunktive Antwortwahrscheinlichkeit* postuliert. Diese entsteht, weil bedingte Wahrscheinlichkeiten [$P(A \mid B)$] von einigen Personen fälschlicherweise als Verbundwahrscheinlichkeiten [$P(A \cap B)$] geschätzt werden. Die in dieser Studie eingesetzten „*probability-of-conditional*“ *tasks* (Evans et al., 2003) stützen zumindest für abstrakte Konditionalaussagen die Idee qualitativ unterschiedlichen Antwortverhaltens bei verschiedenen Personengruppen (Evans et al., 2003; Oberauer & Wilhelm, 2003). Eine Minderheit scheint dabei ein prinzipiell anderes Verständnis von Konditionalaussagen zu haben. Personen, die fälschlicherweise Verbundwahrscheinlichkeiten schätzen, zeigen insgesamt niedrigere Ergebnisse in einem Reasoning-Test als Personen, die bedingte Wahrscheinlichkeiten schätzen (Evans et al., 2007). Es kann also festgehalten werden, dass die Idee, qualitativ unterschiedliches Antwortverhalten beim Konditionalen Schlussfolgern mit Reasoning in Beziehung zu setzen, auch andernorts (z.B. bei Evans et al., 2007) theoretische wie empirische Evidenz findet.

Abschließend soll noch ein dritter − wieder eher kontraintuitiver − empirischer Befund angeführt werden, den das Stufen-Modell erklären kann: Kinder lösen mitunter mehr Modus-Tollens-Probleme korrekt als Erwachsene (O’Brien & Overton, 1982; Rumain et al., 1983; Wildman & Fletcher, 1977). Überträgt man die Ideen entwicklungspsychologischer Theorien, dass Logisches Denken sich im Laufe des Lebens entwickelt (z.B. Piaget, 1971), auf die Sophistiziertheit Konditionalen Schlussfolgerns, liegt es nahe, dass bei Kindern die Unsophistizierten-Stufe häufiger vorkommt als bei Erwachsenen und bei diesen wiederum die Fortgeschrittenen-Stufe häufiger als bei Kindern. Folglich kann es durchaus sein, dass Kinder mehr Modus-Tollens-Probleme lösen, allerdings infolge der unter ihnen häufiger auftretenden Unsophistizierten-Stufe.

Zusammenfassend kann festgehalten werden, dass das Stufen-Modell Konditionalen Schlussfolgerns einen vielversprechenden Zugang zu interindividuellen Unterschieden (hier: Stufen-Zuordnungen) im Konditionalen Schlussfolgern darstellt und damit differenzialpsychologische Aussagen erlaubt. Es ist mit empirischen Befunden vereinbar und erklärt darüber hinaus sogar solche, die bislang kontraintuitiv schienen. Allerdings bedarf es weiterer empirischer Arbeit, da keine Studie existiert, in der alle drei Sophistiziertheits-Stufen identifiziert werden können. Des Weiteren ist auch der hier theoretisch hergeleitete Zusammenhang mit Reasoning empirisch zu überprüfen. Um also die Testkonstruktion auf dem Stufen-Modell aufbauen zu können, sind in jedem Falle folgende zwei Basishypothesen bezüglich des Stufen-Modells zu prüfen:

Basishypothese 1: Für Aufgaben zum Konditionalen Schlussfolgern zeigen sich die drei Stufen des Stufen-Modells.

Basishypothese 2: Gegeben die drei Stufen können identifiziert werden (Basishypothese 1), sind sie entsprechend den Annahmen des Stufen-Modells in eine Rangreihe zu bringen (Unsophistizierte - Rang 1, Fortgeschrittene - Rang 2, Sophistizierte - Rang 3). Die (dann ordinale) Variable *Sophistiziertheit Konditionalen Schlussfolgerns* (*SKS*) korreliert positiv mit Reasoning.

Zwei Dinge sind dabei anzumerken: Erstens ist *Basishypothese 1* notwendige Voraussetzung, um *Basishypothese 2* überhaupt prüfen zu können. Zweitens werden sich nach der konkreten Aufgabenkonstruktion weitere Differenzierungen dieser Basishypothesen ergeben.

2.3 Entwicklung von Items zum Konditionalen Schlussfolgern

Mit dem Stufen-Modell zum Konditionalen Schlussfolgern ist nun sowohl die Forderung der kognitionspsychologischen Fundierung der Testkonstruktion (siehe Ausführungen zu Beginn von Kapitel 2) erfüllt als auch die Möglichkeit differenzialpsychologischer Aussagen gegeben, sodass nun die konkrete Umsetzung der Testkonstruktion thematisiert werden kann.

Zentrales Merkmal eines (psychologischen) Tests ist die mehrfache Messung des interessierenden Konstruktes, bspw. durch verschiedene Items[19]. Es stellt sich also nicht nur die Frage der konkreten Umsetzung des Stufen-Modells bei der Itemkonstruktion, sondern auch die Frage, welche Variationen von Items möglich sind und welche Konsequenzen für die Itemparameter sich aus diesen Variationen (theoretisch) ergeben. Beides ist Gegenstand des folgenden Abschnitts. Bezüglich der Variationsmöglichkeiten werden weitere kognitionspsychologische Befunde zu Konditionalaussagen vorgestellt und hinsichtlich ihrer Relevanz für die Itemkonstruktion bewertet. Als Ergebnis werden am Ende dieses Abschnitts die 16 Testaufgaben präsentiert.

[19] Erst durch die mehrfache Messung ist es möglich, die Messgenauigkeit (Reliabilität) eines Tests überhaupt bestimmen zu können (zu Reliabilität siehe Abschnitt 3.2).

2.3.1 Möglichkeiten der Variation von Items zum Konditionalen Schlussfolgern

Wie bereits erwähnt, ist die mehrfache Messung beim Testen essentiell. Nicht nur im Rahmen des Stufen-Modells, sondern generell bilden die vier Schlussfiguren das „Herzstück" Konditionalen Schlussfolgerns. Mehrfache Messung beim Konditionalen Schlussfolgern bedeutet also mehrfache Messung der vier Schlussfiguren. Für die Variation von Aufgaben zu den vier Schlussfiguren existieren prinzipiell zwei Möglichkeiten: die Variation der aussagenlogischen Form und die Variation des Inhalts (Kleinbeck, 2005). Das Zusammenspiel beider ist entscheidend für die Interpretation von Konditionalaussagen (Kleinbeck, 2005) und kann bspw. mithilfe des *Dual-Source-Ansatzes* (Beller 1997; Beller & Spada, 2003) erklärt werden. Zunächst erfolgen Betrachtungen zur Variation der Form von Konditionalaussagen. Eine Möglichkeit ist die Integration weiterer aussagenlogischer Junktoren. So kombinieren bspw. Rijmen und De Boeck (2003) bei ihren Aufgaben zum Konditionalen Schlussfolgern die vier Schlussfiguren mit verschiedenen Konjunktionen sowie Disjunktionen[20]. Auf diese Weise sollen Decken-Effekte vermieden werden (Rijmen & De Boeck, 2003). Die Verwendung von Konjunktionen und Disjunktionen scheint jedoch problematisch, was folgende Argumentation verdeutlichen soll:

Die korrekte Lösung eines Modus-Tollens-Problems ist die Negation des Antezedens der Hauptprämisse (vgl. Abschnitt 2.1.1). Angenommen das Antezedens ist nun keine atomare Aussage, sondern eine Konjunktion (oder Disjunktion) zweier atomarer Aussagen wie bei Rijmen und De Boeck (2003), dann ist für das korrekte Lösen der Aufgabe also diese Konjunktion (oder Disjunktion) zu negieren. Empirische Befunde zeigen nun aber, dass derartige Negationen für Probanden schwierig sind (vgl. z.B. Byrne & Handley, 1992 für die Negation von Konjunktionen sowie z.B. Roberge, 1976 für die Negation von Disjunktionen). Das heißt, das Lösen eines Modus-Tollens-Problems ist nicht mehr nur von der Sophistiziertheit Konditionalen Schlussfolgerns einer Person abhängig, sondern auch davon, inwieweit die Person in der Lage ist, Konjunktionen oder Disjunktionen zu negieren. Letzteres muss keineswegs das Gleiche sein wie die Sophistiziertheit Konditionalen Schlussfolgerns. Rijmen und De Boeck (2003) berücksichtigen das weder theoretisch noch praktisch, sondern gehen offenbar von einer Steigerung der Itemschwierigkeiten auf einer gemeinsamen Dimension aus. Darin liegt

[20] zur Beschreibung dieser Junktoren siehe Abschnitt 2.1.1

möglicherweise eine der Ursachen dafür, dass bei Rijmen und De Boeck (2003) keine Personen auf der Sophistizierten-Stufe identifiziert werden können. Um das zu verdeutlichen, wird zunächst angenommen, es existieren Personen in der Stichprobe, die Modus-Tollens-Probleme sophistiziert lösen können, nicht jedoch in der Lage sind, Konjunktionen (oder Disjunktionen) korrekt zu negieren. Das ist durchaus naheliegend, wenn man folgendes Modus-Tollens-Problem betrachtet, das eine Konjunktion beinhaltet:

$(p \text{ und } q) \rightarrow r$
$\neg r$

$\neg(p \text{ und } q)$

Es gilt (aussagenlogisch) nach der sog. *de Morgan'schen Regel*:

$\neg(p \text{ und } q) \leftrightarrow \neg p \text{ oder } \neg q$[21]

Dies wäre dann in den Antwortoptionen zu verbalisieren. Um diese Aufgabe korrekt zu lösen, reicht es also nicht aus, sich bezüglich des Modus Tollens auf der Sophistizierten-Stufe befinden; es muss zudem die de Morgan'schen Regel beherrscht werden. Da diese nach den Theorien einer mentalen Logik ebensowenig Teil des Repertoires von Personen ist wie der Modus Tollens (Braine & O'Brien, 1991, 1998; Rips, 1994), wird die benötigte Arbeitsgedächtniskapazität größer, die Aufgabe fehleranfälliger und damit schwerer (Rips, 1994).

Möglicherweise existieren also auch in der untersuchten Stichprobe von Rijmen und De Boeck (2003) Personen auf der Sophistizierten-Stufe, zumindest in Bezug auf einfache Konditionalaussagen. Haben diese Personen jedoch Schwierigkeiten bei der Negation von Konjunktionen (oder Disjunktionen), können sie solche Modus-Tollens-Probleme nicht lösen und werden (zumindest bei diesen Aufgaben) lediglich der Fortgeschrittenen-Stufe zugeordnet.

Um ähnliche Effekte von vornherein auszuschließen, soll bei dem in dieser Arbeit vorgestellten Test für die mehrfache Messung lediglich eine weitere (aussagenlogische) Variation der Aufgaben durchgeführt werden. Diese soll zudem theoretisch fundiert und empirisch überprüft werden. Aus mehreren Gründen bietet sich dafür die *Negation* als weiterer Junktor an. Erstens handelt es sich bei der Verwendung von Ne-

[21] Das Symbol ↔ steht in der Aussagenlogik für die Äquivalenz („wenn…, dann und nur genau dann"; siehe auch Abschnitt 2.1.1)

gationen in Konditionalaussagen um ein − zumindest allgemeinpsychologisch − vielfach untersuchtes Phänomen (z.B. Evans & Handley, 1999; Kleinbeck, 2005; Roberge, 1971; Schroyens, Verschueren, Schaeken, & d'Ydewalle, 2000). Zweitens existieren auch vielfältige empirische Befunde zur Kombination von Konditionalaussagen und Negationen (für einen Überblick siehe z.B. Evans et al., 1993). Der dritte und wohl wichtigste Grund liegt in den Schlussfiguren des Konditionalen Schlussfolgerns selbst. Negationen sind bereits integraler Bestandteil sowohl des Modus Tollens als auch der Negation des Antezedens. Damit kommt kein qualitativ neuer Junktor hinzu. Folglich spricht (zumindest theoretisch) nichts dagegen, davon auszugehen, dass durch Negationen variierte Aufgaben die gleiche Kompetenz messen wie die vier „Basisaufgaben" zu den vier Schlussfiguren. Die Variation der Testaufgaben durch Verwendung von Negationen soll daher in der Folge näher betrachtet werden.

2.3.2 Variation der Testaufgaben durch zusätzliche Negationen – Das Negationsparadigma

Wie bereits argumentiert, sollen zur Variation der aussagenlogischen Form von Items in dem zu konstruierenden Test zusätzliche Negationen eingesetzt werden. Hierzu wird zunächst ein kurzer Überblick über die Bedeutung von Negationen beim Konditionalen Schlussfolgern gegeben. Anschließend werden die Ergebnisse dieser Ausführungen in das Stufen-Modell integriert. Ziel dieses Abschnitts ist die Entwicklung eines erweiterten Stufen-Modells, das auch negierte Komponenten von Konditionalaussagen berücksichtigt und darauf aufbauend noch differenziertere Aussagen zur Sophistiziertheit Konditionalen Schlussfolgerns von Personen erlaubt.

2.3.2.1 Zur Bedeutung von Negationen beim Konditionalen Schlussfolgern

Die Bedeutung von Negationen ist ein fester Bestandteil der Forschung zum Logischen Denken (z.B. Lea & Mulligan, 2002; Roberge, 1976). Für Konditionales Schlussfolgern wird davon ausgegangen, dass bei abstrakten Konditionalaussagen mit negierten Komponenten Interpretationsschwierigkeiten auftreten können (für einen Überblick siehe z.B. Manktelow, 1999). Als Ursache vermuten Lüer und Spada (1992), dass Menschen nicht daran gewöhnt sind, mit „verneinten" Aussagen umzugehen und aus negierten Komponenten einer Konditionalaussage Schlussfolgerungen zu ziehen. Die häufigste Methode, die Bedeutung von Negationen beim Konditionalen Schlussfolgern zu unter-

suchen, ist deren systematische Integration in eine Konditionalaussage ($p \rightarrow q$). Dabei werden zunächst jeweils Antezedens und Konsequenz negiert und schließlich beide gemeinsam. Auf diese Weise ergeben sich vier verschiedene Hauptprämissen:

$$p \rightarrow q$$
$$p \rightarrow \neg q$$
$$\neg p \rightarrow q$$
$$\neg p \rightarrow \neg q$$

Werden für diese vier Hauptprämissen nun jeweils Nebenprämissen zur Generierung der vier Schlussfiguren (MP, NA, BK, MT) aufgestellt, ergeben sich exakt 16 Aufgaben, die Oaksford und Stenning (1992) als *Negationsparadigma* bezeichnen. Eine Vielzahl allgemeinpsychologischer Experimente (z.B. Evans, 1977b; Evans, Clibbens & Rood, 1995; Pollard & Evans, 1980; Wildman & Fletcher, 1977) basiert auf dem Negationsparadigma, wobei meist Vergleiche zwischen den vier verschiedenen Hauptprämissen hinsichtlich der vier Schlussfiguren betrachtet werden. Allerdings weisen die Ergebnisse eine sehr große Varianz auf (Evans et al., 1993), sodass kaum Effekte von Negationen in Konditionalaussagen existieren, die als allgemeingültig gelten. Effekte, die sich zumindest einigermaßen stabil zeigen, werden zumeist als *Biases*, also systematische Verzerrungen bzw. Antworttendenzen interpretiert. Als weitgehend unumstritten gilt dabei der sog. *Negative Conclusion Bias* (*NCB*; Evans, 1977a; Evans, 1982; Pollard & Evans, 1980). Andere gefundene Effekte wie der *Affirmative Premise Bias* oder der *Negative Categorical Premise Bias* (für Beschreibungen siehe z.B. Schaeken & Schroyens, 2000) zeigen sich deutlich weniger stabil und sind insgesamt von geringerer Bedeutung. Daher soll in der Folge lediglich der NCB näher betrachtet werden. Dieser besagt, dass Personen häufiger eine Inferenz ziehen, wenn die Konklusion eines Konditionalen Schlusses negativ ist. Einigkeit herrscht darüber, dass der NCB beim Modus Ponens nicht auftritt (Evans et al., 1995; Kleinbeck, 2005). Dieser Effekt ist vor allem bei den Schlussfiguren Negation des Antezedens und Modus Tollens zu beobachten (z.B. Evans et al., 1995; Schaeken & Schroyens, 2000), was sich wie folgt äußert:

1. Bei der Negation des Antezedens werden mehr Inferenzen gezogen, wenn die Konsequenz der Hauptprämisse positiv ist, also für $p \rightarrow q$ und $\neg p \rightarrow q$.
   (Da eine korrekte Lösung der Schlussfigur Negation des Antezedens eine „k.A.m."-Reaktion wäre, werden in diesen Fällen also *mehr Fehler* gemacht.)

2. Beim Modus Tollens werden mehr Inferenzen gezogen, wenn das Antezedens der Hauptprämisse positiv ist, also für $p \rightarrow q$ und $p \rightarrow \neg q$.
   (Da eine korrekte Lösung der Schlussfigur Modus Tollens eine eindeutige Inferenz, nämlich die Negation des Antezedens ist, werden in diesen Fällen also *weniger Fehler* gemacht.)

Uneinigkeit herrscht darüber, ob der NCB bei der Schlussfigur Bestätigung der Konsequenz lediglich in geringerem Maße (z.B. Schaeken & Schroyens, 2000; Schroyens, Schaeken & d'Ydewalle, 2001) oder überhaupt nicht auftritt (z.B. Evans et al., 1995). Hinweise liefern möglicherweise die theoretischen Erklärungen für den NCB. Evans et al. (1995) gehen davon aus, dass der NCB die Folge von Schwierigkeiten im Umgang mit doppelten Negationen ist. Nach den Theorien einer mentalen Logik ist ein zusätzliches Inferenzschema (Braine & O'Brien, 1998; O'Brien, 2004) bzw. eine zusätzliche Regel (Rips, 1994) notwendig, um eine doppelte Negation aufzulösen. Das macht eine solche Aufgabe schwieriger und führt zu weniger Inferenzen. Im Negationsparadigma müssen derartige doppelte Negationen für die Schlussfigur Negation des Antezedens bei den Hauptprämissen $p \rightarrow \neg q$ und $\neg p \rightarrow \neg q$ und für die Schlussfigur Modus Tollens bei den Hauptprämissen $\neg p \rightarrow q$ und $\neg p \rightarrow \neg q$ aufgelöst werden. Dies sei am Beispiel des Modus Tollens bei der Hauptprämisse $\neg p \rightarrow q$ kurz verdeutlicht: Die Nebenprämisse lautet entsprechend $\neg q$, die Konklusion $\neg\neg p$, also nach Auflösen der doppelten Negation p. Bei den vier angeführten Fällen handelt es sich genau um diejenigen, bei denen der NCB nicht auftritt. Nach diesem Erklärungsansatz werden deswegen mehr Inferenzen gezogen, wenn die Konklusion negativ ist, weil dazu keine doppelte Negation aufgelöst werden muss. Bei der Schlussfigur Bestätigung der Konsequenz braucht hingegen keinerlei doppelte Negation aufgelöst werden, sodass nach diesem Erklärungsansatz der NCB bei dieser Schlussfigur nicht auftreten sollte. Bei konsequenter Verfolgung des Erklärungsansatzes von Evans et al. (1995) ergeben sich jedoch zwei weitere (der 16) Fälle des Negationsparadigmas, in denen eine doppelte Negation aufgelöst werden muss, und zwar:

1. beim Modus Tollens und der Hauptprämisse $p \rightarrow \neg q$ sowie
2. bei der Negation des Antezedens und der Hauptprämisse $\neg p \rightarrow q$.

Im letzteren Fall bspw. lautet die Nebenprämisse p. Um die Schlussfigur Negation des Antezedens anzuwenden, muss nun das Antezedens der Hauptprämisse negiert werden.

Dabei ist zumindest zu enkodieren, dass *p* der Negation des Antezedens entspricht, also ¬¬*p*. Diese doppelte Negation sollte deutlich leichter aufzulösen sein als die vier angeführten „klassischen" Beispiele, da sie nicht aktiv gebildet, sondern lediglich enkodiert werden muss. Dennoch sollten auch hier weniger Inferenzen gezogen werden. Gleiches gilt für die Nebenprämisse *q* (¬¬*q*) im ersten angeführten Fall. Diese Überlegung deckt sich bspw. mit den Befunden von Evans (1977b), die in Tabelle 6 dargestellt sind. Dabei zeigt sich, dass in den beiden betrachteten Fällen tatsächlich weniger Inferenzen gezogen werden als bei der Hauptprämisse ohne jegliche Negation.

Tabelle 6: Prozentuale Häufigkeit gezogener Inferenzen bei den Schlussfiguren Negation des Antezedens (NA) und Modus Tollens (MT) für die vier Hauptprämissen des Negationsparadigmas bei Evans (1977b)

	Hauptprämisse			
Schlussfigur	$p \rightarrow q$	$p \rightarrow \neg q$	$\neg p \rightarrow q$	$\neg p \rightarrow \neg q$
NA	69	12	50*	19
MT	75	56*	12	25

Anmerkung. Bei den mit * gekennzeichneten Fällen werden weniger Inferenzen gezogen als bei der entsprechenden Hauptprämisse ohne jegliche Negation. Als Ursache wird das Enkodieren einer doppelten Negation vermutet.

Für die beiden zusätzlich betrachteten Fälle ist die Häufigkeit gezogener Inferenzen also höher als in den „klassischen" NCB-Fällen, jedoch niedriger als für die Konditionalaussage ohne jegliche Negationen. Dies kann zusammenfassend so interpretiert werden, dass das Vorhandensein mindestens einer Negation in der Hauptprämisse das Ziehen von Inferenzen prinzipiell erschwert. Dass jegliche Negation in der Hauptprämisse einen Effekt hat, zeigt sich auch anhand anderer Phänomene beim Konditionalen Schlussfolgern, wie bspw. dem sog. *Matching Bias* (für einen Überblick siehe Evans, 1998). Darunter versteht Evans (1993) eine Tendenz, als Grundlage für Schlussfolgerungen die in der Prämisse enthaltenen Größen statt deren Negationen zu wählen. Der Matching Bias gilt als weitgehend unumstritten und wird insbesondere anhand der *Wason Selection Task* (*WST*, Wason, 1966; siehe auch Abschnitt 4.3.1) untersucht. Es existieren aber auch Untersuchungen, in denen die Ursachen des Matching Bias anhand einfacher Konditionalaussagen und deren systematischer Negation untersucht werden (z.B. Schroyens et al., 2000). Weitere Erklärungsansätze für den NCB basieren auf der *Probabilistischen Theorie des Deduktiven Schließens* (Chater & Oaksford, 1999) und

der *Bildung von Kontrastmengen* (Schaeken & Schroyens, 2000). Diese sollen jedoch nicht weiter vertieft werden.

Festzuhalten bleibt, dass als relativ robuster Effekt von Negationen in Konditionalaussagen der Negative Conclusion Bias gelten kann. Dieser sollte (zumindest auf Basis des hier vorgestellten Erklärungsansatzes) lediglich bei den Schlussfiguren Negation des Antezedens und Modus Tollens auftreten. Zusätzlich lässt sich die vermutete Ursache „Erschweren einer Inferenz durch Auflösen einer doppelten Negation" auch auf zwei Fälle übertragen, in denen die Konklusion positiv ist, jedoch die Nebenprämisse als doppelte Negation enkodiert werden muss. Auch in diesen Fällen liegt es nahe, dass das Ziehen einer Inferenz erschwert wird, allerdings sollte sich der Effekt schwächer zeigen als bei den „klassischen" NCB-Fällen.

2.3.2.2 Erweiterung des Stufen-Modells auf das Negationsparadigma

Das Stufen-Modell Konditionalen Schlussfolgerns (siehe Abschnitt 2.2.2) erlaubt differenzialpsychologische Aussagen zur Sophistiziertheit Konditionalen Schlussfolgerns von Personen. Nun stellt sich die Frage, inwieweit die bisherigen Ausführungen zu Negationen in Konditionalaussagen (siehe Abschnitt 2.3.2.1) differenzialpsychologisch betrachtet werden können und ob sie in das Stufen-Modell integrierbar sind. Allerdings existieren − zumindest nach Kenntnis des Autors − keine differenzialpsychologischen Befunde zum Negationsparadigma, obwohl es durchaus in Untersuchungen zu interindividuellen Unterschieden eingesetzt wird (z.B. Evans, et al., 2007; Oberauer et al., 2007). Die Frage, bei welchen Personen die Effekte von Negationen in Konditionalaussagen wirken, scheint prinzipiell gerechtfertigt, da sich durch die Negationen stets nur das Antwortverhalten eines Teils der Probanden ändert (vgl. z.B. Befunde von Evans, 1977b; Evans et al., 1995; Pollard & Evans, 1980; Wildman & Fletcher, 1977). Dies spricht dafür, dass sich für einen Teil der Personen auch die Stufen des Stufen-Modells stabil zeigen sollten – unabhängig davon, ob in der Konditionalaussage Negationen enthalten sind oder nicht. Auch bei Rijmen und De Boeck (2003) zeigen sich die postulierten Stufen relativ robust, trotz Manipulation der Aufgaben durch verschiedene andere Junktoren (Konjunktionen, Disjunktionen).

Ein anderer Teil der Personen reagiert offenbar auf Negationen in der Konditionalaussage. Eine mögliche Erklärung ist die Anforderung „Auflösen einer doppelten Negation" bei insgesamt sechs der 16 Aufgaben des Negationsparadigmas (vgl. Abschnitt 2.3.2.1). Dabei handelt es sich jeweils um die drei Aufgaben zur Negation des

Antezedens und zum Modus Tollens, bei denen mindestens eine Negation in der Hauptprämisse enthalten ist (vgl. Abschnitt 2.3.2.1). Bei allen sechs Aufgaben wird von der gleichen zusätzlichen Anforderung ausgegangen, wenn auch bei zwei Aufgaben in deutlich geringerer Intensität (vgl. Abschnitt 2.3.2.1). Es spricht also (theoretisch) nichts dagegen, dass für bestimmte Personen dieser Effekt bei allen sechs Aufgaben auftritt. Der beschriebene Effekt ist nun in das Stufen-Modell zu integrieren. Betrachtet man das Antwortverhalten bei mindestens einer Negation in der Hautprämisse, so werden bei NA und MT weniger Inferenzen gezogen (vgl. Abschnitt 2.3.2.1). Entsprechend wird häufiger die „k.A.m."-Reaktion gewählt. Das dann gezeigte Antwortverhalten (bei NA und MT) entspricht dem von Personen auf der Fortgeschrittenen-Stufe des Stufen-Modells. Die Frage ist nun, welcher Stufe diese Personen zugeordnet werden, wenn keine Negation in der Hauptprämisse enthalten ist. Für Negation des Antezedens ist dies die Unsophistizierten-Stufe, da Personen auf der Sophistizierten-Stufe bei dieser Schlussfigur ebenfalls die „k.A.m."-Reaktion wählen. Für die Modus-Tollens-Aufgaben wären zunächst beide Stufen möglich, da sowohl Personen auf der Unsophistizierten-Stufe als auch Personen auf der Sophistizierten-Stufe Inferenzen ziehen. Allerdings sprechen Befunde von Evans et al. (2007) dafür, dass Personen auf höheren Stufen weniger anfällig für Biases sind als Personen auf niedrigeren Stufen. Das legt nahe, dass es sich um Personen handelt, die bei Konditionalaussagen ohne Negationen der Unsophistizierten-Stufe zugeordnet werden. Dafür spricht auch, dass für einen (sophistizierten) Modus-Tollens-Schluss, wie er lediglich auf der Sophistizierten-Stufe gezogen wird, nach der Logiktheorie (Braine & O'Brien, 1991; siehe Abschnitt 2.1.2.2) bereits vier Inferenzschemata notwendig sind (vgl. Abschnitt 2.2.1) bzw. drei vollständig explizite Modelle nach der Modelltheorie (Johnson-Laird & Byrne, 2002; vgl. Abschnitt 2.1.2.1). Es scheint unplausibel, dass Personen, die diese Anforderungen bewältigen, anfällig für derartige Biases sein sollten. Die bisher angestellten Überlegungen lassen sich zu folgender (vorläufiger) Behauptung zusammenfassen:

Bei zusätzlichen Negationen in der Hauptprämisse steigt der Anteil an Personen, die bei den Schlussfiguren NA und MT das Antwortverhalten der Fortgeschrittenen-Stufe zeigen. Bei einer Hauptprämisse ohne Negationen zeigen diese Personen bei diesen Schlussfiguren das Antwortverhalten der Unsophistizierten-Stufe.

Berücksichtigt man nun die Ordnung der drei Stufen des Stufen-Modells entlang der Dimension „Reasoning" (vgl. Abbildung 2 in Abschnitt 2.2.3), dann handelt es sich

offenbar um Personen auf einer „Zwischen-Stufe“ zwischen der Unsophistizierten- und der Fortgeschrittenen-Stufe. Nach der dem Stufen-Modell zugrundeliegenden Logiktheorie unterscheiden sich Personen auf diesen beiden Stufen darin, ob es ihnen gelingt, die einladende Inferenz zu blockieren (vgl. Abschnitt 2.2.2). Folglich unterstützen die Negationen in der Hauptprämisse bei den Schlussfiguren NA und MT das Blockieren der einladenden Inferenzen und ermöglichen so (bestimmten Personen) den „Sprung“ von der Unsophistizierten- auf die Fortgeschrittenen-Stufe. Das Antwortverhalten von Personen auf dieser Zwischen-Stufe entspricht also prinzipiell dem von Personen auf der Unsophistizierten-Stufe, ist aber außerdem dadurch charakterisiert, dass

1. der einladenden Inferenz bei der Schlussfigur Negation des Antezedens im Falle einer oder mehrerer Negationen in der Hauptprämisse leichter „widerstanden“ wird, was zu weniger Inferenzen und damit einer deutlich *höheren* Lösungswahrscheinlichkeit führt und

2. der einladenden Inferenz bei der Schlussfigur Modus Tollens im Falle einer oder mehrerer Negationen in der Hauptprämisse leichter „widerstanden“ wird, was zu weniger Inferenzen und damit einer deutlich *niedrigeren* Lösungswahrscheinlichkeit führt.

Zur Erläuterung von Punkt 2 sei daran erinnert, dass Modus-Tollens-Items auf der Unsophistizierten-Stufe korrekt gelöst werden, allerdings aufgrund einer bikonditionalen Interpretation infolge des Akzeptierens der einladenden Inferenz (vgl. Abschnitt 2.2.2). Auf der nächsthöheren Stufe (Fortgeschrittene) werden die Modus-Tollens-Items dann nicht mehr korrekt gelöst, weshalb die niedrigere Lösungswahrscheinlichkeit unter Punkt 2 dennoch einen „(Kompetenz-)Sprung“ darstellt, und zwar von der Unsophistizierten- auf die Fortgeschrittenen-Stufe. Da jedoch bei einer Hauptprämisse ohne Negation das Antwortverhalten bei NA und MT immer noch der Unsophistizierten-Stufe entspricht, ebenso wie bei sämtlichen BK-Inferenzen, wird diese Zwischen-Stufe in der Folge „*Unsophistizierte II*“ genannt. Zusammenfassend ergibt sich also folgende Ordnung für die nunmehr vier Sophistiziertheits-Stufen im Konditionalen Schlussfolgern:

Stufe 1: Unsophistizierte (wie bei Rijmen & De Boeck, 2003)
Stufe 2: Unsophistizierte II (bei mindestens einer Negation in der Hauptprämisse: „Sprung“ auf die Fortgeschrittenen-Stufe, allerdings nur bei NA und MT)
Stufe 3: Fortgeschrittene (wie bei Rijmen & De Boeck, 2003)
Stufe 4: Sophistizierte (wie bei Rijmen & De Boeck, 2003)

Damit lassen sich die beiden Basishypothesen dieser Arbeit (siehe Abschnitt 2.2.3) zu folgenden Arbeitshypothesen ausdifferenzieren:

Arbeitshypothese 1: Für die 16 Aufgaben des Negationsparadigmas zum Konditionalen Schlussfolgern zeigen sich die vier Stufen des erweiterten Stufen-Modells, die folgende Charakteristika bezüglich des Löseverhaltens aufweisen:

Arbeitshypothese 1a: Es zeigt sich eine Gruppe von Personen, die (bis auf zufällige Fehler) Modus-Ponens-Aufgaben korrekt löst, Aufgaben zur Negation des Antezedens wie zur Bestätigung der Konsequenz nicht korrekt löst sowie Modus-Tollens-Aufgaben korrekt löst. (Dabei handelt es sich um Personen auf der Unsophistizierten-Stufe.)

Arbeitshypothese 1b: Es zeigt sich eine Gruppe von Personen, die (bis auf zufällige Fehler) Modus-Ponens-Aufgaben korrekt löst und Aufgaben zur Bestätigung der Konsequenz nicht korrekt löst. Bei Aufgaben zur Negation des Antezedens zeigen diese Personen eine deutlich höhere Lösungswahrscheinlichkeit, wenn die Hauptprämisse mindestens eine Negation enthält. Bei Modus-Tollens-Aufgaben zeigen diese Personen eine deutlich niedrigere Lösungswahrscheinlichkeit, wenn die Hauptprämisse mindestens eine Negation enthält. (Dabei handelt es sich um Personen auf der Unsophistizierten-Stufe II.)

Arbeitshypothese 1c: Es zeigt sich eine Gruppe von Personen, die (bis auf zufällige Fehler) Modus-Ponens-Aufgaben korrekt löst, Aufgaben zur Negation des Antezedens wie zur Bestätigung der Konsequenz ebenfalls korrekt löst sowie Modus-Tollens-Aufgaben nicht korrekt löst. (Dabei handelt es sich um Personen auf der Fortgeschrittenen-Stufe.)

Arbeitshypothese 1d: Es zeigt sich eine Gruppe von Personen, die (bis auf zufällige Fehler) sowohl Modus-Ponens-Aufgaben, Aufgaben zur Negation des Antezedens, Aufgaben zur Bestätigung der Konsequenz sowie Modus-Tollens-Aufgaben korrekt löst. (Dabei handelt es sich um Personen auf der Sophistizierten-Stufe.)

Arbeitshypothese 2: Gegeben die vier Stufen können identifiziert werden (Arbeitshypothese 1), sind sie entsprechend den Annahmen des erweiterten Stufen-Modells in eine Rangreihe zu bringen (Unsophistizierte - Rang 1, Unsophistizierte II - Rang 2, Fortgeschrittene - Rang 3, Sophistizierte - Rang 4). Die (dann ordinale) Variable *Sophistiziertheit Konditionalen Schlussfolgerns* (*SKS*) korreliert positiv (bei einem mittleren Effekt[22]) mit Reasoning.

Vor dem Hintergrund einer Testentwicklung zum Konditionalen Schlussfolgern ist das Negationsparadigma in mehrfacher Hinsicht attraktiv. Es hat sich empirisch vielfach bewährt und ist zudem sehr ökonomisch, da durch Vorgabe von lediglich 16 vergleichsweise kurzen Items die suffiziente Betrachtung eines kompletten Paradigmas möglich ist. Zudem lassen sich allgemeinpsychologische Befunde zum Negationsparadigma in das Stufen-Modell integrieren und erlauben durch eine zusätzliche „Negationen-sensitive“ Stufe noch differenziertere Aussagen als das ursprüngliche Stufen-Modell.

2.3.3 Variation der Testaufgaben durch verschiedene Inhalte

Neben der Variation der aussagenlogischen Form von Konditionalaussagen (bspw. durch Negationen; siehe Abschnitt 2.3.2) kann auch deren Inhalt variiert werden. Inhaltseffekte gelten nicht nur als unumstritten, sondern sind auch essentieller Bestandteil der Modelltheorie (vgl. Abschnitt 2.1.2.1). Generell wird davon ausgegangen, dass sowohl beim Induktiven als auch beim Deduktiven Denken das Ziehen glaubwürdiger Schlussfolgerungen begünstigt wird (Wilhelm, 2000). Dies kann demnach auch für Konditionales Schlussfolgern angenommen werden. Das Prinzip der semantischen Modulation im Rahmen der Modelltheorie (siehe Abschnitt 2.1.2.1) besagt hierzu, dass die semantisch sinnvolle Verknüpfbarkeit von Antezedens und Konsequenz die Bildung vollständig expliziter Modelle und folglich auch das korrekte Schlussfolgern erleichtert. Allerdings zeigen sich Zusammenhänge mit Intelligenz für abstrakte Aufgaben deutlicher als für Aufgaben mit konkreten Inhalten (Wilhelm, 2000). Als Ursache wird angenommen, dass bei letzteren das korrekte Lösen zu stark von Erfahrungen mit den entsprechenden Inhalten abhängig ist (Wilhelm, 2000). Inhalte von Aufgaben gezielt zu

[22] Der Effekt wird in Anlehnung an die berichteten Effekten in den Arbeiten von Stanovich (zusammenfassend Stanovich, 1999; Stanovich & West, 2000; siehe auch Abschnitt 2.1.4.1) im mittleren Bereich vermutet.

manipulieren, scheint extrem schwierig, sie zumindest zu kontrollieren indes recht einfach. Wird der Inhalt über die Aufgaben hinweg konstant gelassen, dann gilt er über sämtliche Aufgaben hinweg und Varianzquelle sind lediglich aussagenlogische Veränderungen wie die für diese Testentwicklung beschriebenen Negationen (siehe Abschnitt 2.3.2). In engem Zusammenhang mit Inhaltseffekten steht die Bedeutung von Vorwissen. Dies spiegelt sich im Prinzip der pragmatischen Modulation der Modelltheorie (siehe Abschnitt 2.1.2.1) wider. Nach diesem Prinzip kann auch der Kontext einer Konditionalaussage in Abhängigkeit vom allgemeinen wie spezifischen Vorwissen der Person die Bildung vollständig expliziter Modelle und damit auch das korrekte Schlussfolgern erleichtern. Allerdings sind Unterschiede im Vorwissen von Personen schwer zu erheben. Unumstritten ist jedenfalls, dass für deduktive Aufgaben nicht-deduktive Varianzanteile existieren, insbesondere wissensbasierte, die nur schwer vollständig ausschließbar sind (Wilhelm, 2000). Damit stellt sich die Frage, inwieweit solche Effekte bei dieser Testkonstruktion berücksichtigt werden können.

Wie in Kapitel 1 bereits ankündigt, sind in verschiedenen Phasen der Testkonstruktion pragmatische Entscheidungen aufgrund bestimmter Vorgaben (in diesem konkreten Fall Vorgaben des Auftraggebers) zu treffen. Die Herausforderung besteht dabei vor allem darin, wissenschaftlichen Anspruch und ökonomische Pragmatik in Einklang zu bringen. Die Wahl des Inhaltes ist bei dieser Testkonstruktion sicher das prägnanteste Beispiel dafür: Die Zielgruppe für den zu konstruierenden Test sind − zumindest im Rahmen der Testkonstruktion − Chipdesigner (vgl. Kapitel 1). Die Aufgaben sind folglich so zu konstruieren, dass sie inhaltlich für diese Zielgruppe attraktiv sind (Testgütekriterium *Attraktivität*; siehe Abschnitt 3.11). Darüber hinaus müssen sie von dieser Zielgruppe inhaltlich akzeptiert werden und gleichzeitig abstrakt genug sein, um nicht zu stark durch die beschriebenen Inhaltseffekte konfundiert zu werden. Für den beabsichtigten Test ist der gewählte Inhalt der Aufgaben das Funktionieren elektrischer Schaltungen. Dieser Aufgaben-Inhalt scheint ausreichend abstrakt zu sein, da das Funktionieren einer Schaltung einer (abstrakten) Aussage p sowie das Nicht-Funktionieren einer (abstrakten) Aussage $\neg p$ entspricht. Gleichzeitig ist dieser Inhalt für die Zielgruppe akzeptabel, wie die Ergebnisse eines Workshops zeigen, in dem dieser Inhalt mit Experten für den Chipdesign-Prozess abgestimmt wurde (siehe dazu Böhme & Steyer, 2008). Wird dieser Inhalt nun über sämtliche Testaufgaben hinweg konstant gelassen, können Inhaltseffekte zwar nicht ausgeschlossen, aber durch das Konstanthalten zumindest kontrolliert werden.

2.3.4 Die 16 Testaufgaben

Nachdem das Negationsparadigma die Basis für die aussagenlogische Form der 16 Testaufgaben bildet und das Funktionieren elektrischer Schaltungen den Inhalt, sind die 16 Testaufgaben nun eindeutig determiniert. Die Vorstellung der 16 Items soll auf Ebene 1 anhand der vier Schlussfiguren und auf Ebene 2 anhand der vier verschiedenen Hauptprämissen des Negationsparadigmas erfolgen. Für die vier Schlussfiguren wird folgende Ordnung festgelegt:

Modus Ponens (MP)
Negation des Antezedens (NA)
Bestätigung der Konsequenz (BK)
Modus Tollens (MT)

Stehe *p* für die Aussage „Schaltung 1 funktioniert" und *q* für die Aussage „Schaltung 2 funktioniert", dann wird für die vier Hauptprämissen des Negationsparadigmas im weiteren Verlauf folgende Bezeichnung und damit Ordnung festgelegt:

1: $p \rightarrow q$
2: $p \rightarrow \neg q$
3: $\neg p \rightarrow q$
4: $\neg p \rightarrow \neg q$

Die konkrete Nebenprämisse eines Items richtet sich nach der jeweiligen Hauptprämisse und bezieht sich inhaltlich ebenfalls auf das Funktionieren bzw. Nicht-Funktionieren von Schaltung 1 bzw. Schaltung 2. Zur Illustration sei das (willkürlich gewählte) Beispiel BK3 angeführt. Es handelt sich dabei entsprechend obiger Ordnung also um die Schlussfigur Bestätigung der Konsequenz (BK) und die Hauptprämisse mit einer Negation im Antezedens (3). Konkret sieht das Item BK3 wie folgt aus:

<u>Item BK3:</u>
Hauptprämisse: Wenn Schaltung 1 *nicht* funktioniert, dann funktioniert Schaltung 2.
Nebenprämisse: Schaltung 2 funktioniert.
Korrekter (normativ logischer) Schluss: Schaltung 1 funktioniert oder Schaltung 1 funktioniert nicht. Es ist also keine eindeutige Aussage über das Funktionieren von Schaltung 1 ableitbar („k.A.m."-Reaktion).

Jedes der 16 Items ist damit eindeutig determiniert und kann entsprechend obiger Nomenklatur [{MP, NA, BK, MT} x {1, 2, 3, 4}] bezeichnet werden. Die 16 Testaufgaben finden sich im Anhang A.2 dieser Arbeit[23]. Mit dem erweiterten Stufen-Modell existiert nun ein kognitionspsychologisch fundiertes Modell, das differenzialpsychologische Aussagen erlaubt. Weiterhin existieren konkrete Arbeitshypothesen zur Anwendung des erweiterten Stufen-Modells auf 16 Items, die in der Folge als *Kurztest zum Konditionalen Schlussfolgern* (*KKS*) bezeichnet werden[24]. Bevor tatsächlich von einem Test gesprochen werden kann, sind die Arbeitshypothesen zunächst jedoch empirisch zu überprüfen. Dies ist Gegenstand des folgenden Abschnitts.

2.4 Eine erste empirische Überprüfung des erweiterten Stufen-Modells und damit des KKS

Bevor weitere theoretische Annahmen und insbesondere Testgütekriterien des KKS systematisch überprüft werden, soll das postulierte Stufen-Modell und damit implizit auch der KKS einer ersten empirischen Erprobung unterzogen werden. In diesem Vortest sind dazu die sechs Arbeitshypothesen (siehe Abschnitt 2.3.2.2) zu überprüfen. Nur wenn dieser Vortest positiv ausfällt, scheint eine Überprüfung weiterer Annahmen sowie die Betrachtung von Testgütekriterien überhaupt sinnvoll. Ein erfolgreicher Vortest ist also eine notwendige, jedoch keineswegs hinreichende Bedingung für den Einsatz des KKS zur Diagnostik der Sophistiziertheit Konditionalen Schlussfolgerns. Die theoretische Basis für die in der Folge vorgestellte Studie bilden sämtliche bislang angestellten Überlegungen. Es folgt ein Überblick über die verwendeten Methoden dieser ersten empirischen Überprüfung, eine Darstellung der Ergebnisse der Hypothesenprüfung sowie deren Diskussion vor dem theoretischen Hintergrund. Nach einer kritischen Würdigung der Methoden werden abschließend Konsequenzen für die weitere empirische Erprobung des KKS und dabei insbesondere für die Bestimmung seiner Testgütekriterien abgeleitet.

[23] Die Reihenfolge für die konkrete Vorgabe der Items im Test kann natürlich von dieser Nomenklatur abweichen. Sie wird in Abschnitt 2.4.1.2 vorgestellt.

[24] Die Bezeichnung „Kurztest“ ist aufgrund der geringen Aufgabenzahl durchaus angemessen, bezieht sich aber vor allem auf die kurze Bearbeitungszeit von lediglich 5-10 Minuten (siehe Testgütekriterium *Testökonomie* in Abschnitt 3.6).

2.4.1 Methoden

Gegenstand dieses Abschnitts sind die verwendeten Methoden zur Prüfung der sechs Arbeitshypothesen. Dabei wird auf das allgemeine Untersuchungsdesign, die verwendeten Erhebungsinstrumente, die konkrete Durchführung der Untersuchung, die Untersuchungsstichprobe und die Auswertungsmethoden eingegangen. Doch zunächst soll eine zweite prinzipielle Forderung an die vorliegende Arbeit gestellt werden. Die erste ist die kognitionspsychologische Fundierung der Testkonstruktion (vgl. Ausführungen zu Beginn von Kapitel 2), die zweite eine Prüfung der Arbeitshypothesen auf Basis von *Latente-Variablen-Modellen*, um so das *Messfehler-Problem* (z.B. Steyer & Eid, 2001) von vornherein zu berücksichtigen.

2.4.1.1 Allgemeines Untersuchungsdesign

Wie für Leistungstests allgemein üblich wurde ein quantitatives Untersuchungsdesign umgesetzt. Da für die Umsetzung der zweiten Forderung an die vorliegende Arbeit (Einsatz von Latente-Variablen-Modellen zur empirischen Hypothesenprüfung) meist umfangreiche Untersuchungsstichproben notwendig sind, bot sich eine Online-Erhebung an, mit der vergleichsweise einfach die Rekrutierung einer großen Untersuchungsstichprobe möglich ist (Funke & Reips, 2007; Reips, 2002). Nähere Ausführungen hierzu finden sich im Abschnitt zur Durchführung der Untersuchung (siehe Abschnitt 2.4.1.3).

2.4.1.2 Erhebungsinstrumente

Zur Überprüfung des *Arbeitshypothesenkomplexes 1* wurden die 16 Items des KKS verwendet, die bereits in Abschnitt 2.3.4 ausführlich vorgestellt werden. Allerdings fehlen aktuell noch Angaben zur Reihenfolge dieser 16 Items im KKS sowie zum verwendeten Paradigma bezüglich der Antworten. Beides wird in diesem Abschnitt thematisiert. Gegenstand von *Arbeitshypothese 2* ist der Zusammenhang des Konstruktes Sophistiziertheit Konditionalen Schlussfolgerns mit Reasoning. Nachdem erstere mit dem KKS erhoben wurde, bleibt noch die Vorstellung des Erhebungsinstruments für Reasoning als Abschluss dieses Abschnitts.

Kurztest zum Konditionalen Schlussfolgern (*KKS*)

Die 16 Items des KKS (siehe auch Anhang A.2) werden bereits in Abschnitt 2.3.4 ausführlich vorgestellt. Allerdings stellt sich die Frage nach dem konkreten Paradigma zur

Präsentation und damit auch nach der Auswertung der Antworten auf die Items. Prinzipiell existieren zwei etablierte Paradigmen für die Bearbeitung von Aufgaben mit Konditionalaussagen: das Generierungsparadigma und das Verifikationsparadigma (vgl. z.B. Knauff, 2006).

Generierungsparadigma

Beim Generierungsparadigma werden den Probanden Prämissen vorgegeben und es wird nach den logischen Folgerungen gefragt. Die Probanden müssen die Konklusion selbst ableiten. Die Variation der Schwierigkeit der Aufgaben erfolgt anhand der Anzahl oder Reihenfolge gegebener Prämissen. Aufgaben für ein Generierungsparadigma sind so zu konstruieren, dass die Probanden unterschiedliche Folgerungsbeziehungen anwenden müssen, um zu einer validen Konklusion zu gelangen. Als Ergebnisse gelten sämtliche von den Probanden generierten Konklusionen. Als Maß für die Schwierigkeit einer Inferenz wird das Verhältnis zwischen generierten gültigen und ungültigen Konklusionen betrachtet. Die relative Häufigkeit richtiger und falscher Reaktionen wird dabei berücksichtigt. Zudem erfolgt in einem Generierungsparadigma häufig zusätzlich eine Zeitmessung, wie viel Zeit der Proband bis zur Konklusion benötigt. Es lässt sich außerdem noch zwischen partiellen und generellen Generierungsaufgaben unterscheiden (Knauff, Rauh & Schlieder, 1995), worauf allerdings nicht weiter eingegangen werden soll.

Verifikationsparadigma

Beim Verifikationsparadigma werden den Probanden Prämissen und mindestens eine Konklusion präsentiert. Die Aufgabe der Probanden besteht darin, zu beurteilen, ob die Konklusion logisch aus den Prämissen folgt. Die Zeitmessung fungiert dabei als Maß für die Schwierigkeit der Aufgabe. Es werden entsprechend der Signal-Detektions-Theorie (Green & Swets, 1966; siehe auch Velden, 1982) vier mögliche Reaktionen unterschieden: Treffer, Verpasser, Falscher Alarm und Korrekte Zurückweisung, die in Tabelle 7 zusammengefasst werden.

Tabelle 7: Überblick über die vier möglichen Reaktionen eines Probanden im Rahmen des Verifikationsparadigmas

	Urteil des Probanden	
Vorgegebene Konklusion	„valide“	„nicht valide“
valide	**Treffer**	Verpasser
nicht valide	Falscher Alarm	**Korrekte Zurückweisung**

Anmerkung. Logisch korrekte Urteile des Probanden sind fettgedruckt.

Wendet man die Terminologie des Verifikationsparadigmas auf die bereits mehrfach berichteten empirischen Befunde zum Konditionalen Schlussfolgern (siehe z.B. Abschnitt 2.1.2) an, dann ist die Zahl der Treffer für Modus-Ponens-Aufgaben sehr hoch. Für Modus-Tollens-Aufgaben ist die Zahl der Treffer etwas niedriger. Aufgaben zur Negation des Antezedens und zur Bestätigung der Konsequenz zeichnen sich durch eine hohe Zahl falscher Alarme aus, wenn die in Abschnitt 2.1.2 beschriebenen Konklusionen vorgegeben werden. Das heißt, werden diese invaliden Konklusionen vorgegeben, werden sie von Probanden häufig (jedoch fälschlicherweise) als „valide“ bezeichnet (vgl. z.B. Knauff, 2006).

Für den KKS bedurfte es nun einer möglichst ökonomischen Messung der 16 Items, da die beabsichtigten Latente-Variablen-Modelle meist große Stichproben erfordern und folglich Daten einer entsprechend großen Untersuchungsstichprobe auszuwerten sein würden (vgl. Abschnitt 2.4.1.1). Offene Antworten wie bspw. im beschriebenen Generierungsparadigma schienen daher ungeeignet. Stattdessen wurden in Anlehnung an klassische Multiple-Choice-Tests verschiedene Antworten vorgegeben, wobei der Proband diejenige auswählen sollte, die seines Erachtens die richtige Lösung der Aufgabe darstellte. Durch die vollständig determinierte Itemmenge (siehe Abschnitt 2.3.4) ließ sich auch die Menge möglicher Antwortalternativen für jedes Item eindeutig festlegen, was im Folgenden verdeutlicht werden soll.

Gegeben sei die Hauptprämisse 1 des KKS: *Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2*. Je nach Schlussfigur bezieht sich die Nebenprämisse nun entweder auf Schaltung 1 (bei Modus Ponens und Negation des Antezedens) oder auf Schaltung 2 (bei Bestätigung der Konsequenz und Modus Tollens). Folglich können sich mögliche Konklusionen lediglich auf die jeweils andere Schaltung beziehen. Es

ergeben sich genau drei Möglichkeiten für das Funktionieren der Schaltung, auf die sich die Konklusion bezieht:

1. Die Schaltung funktioniert.
2. Die Schaltung funktioniert nicht.
3. Es ist keine eindeutige Aussage über das Funktionieren der Schaltung ableitbar.

Letztere Möglichkeit kann als inhaltlich äquivalent zu der Aussage „*Die Schaltung funktioniert oder sie funktioniert nicht*“ betrachtet werden, die (formal-logisch) ebenfalls eine exklusive Disjunktion darstellt. Die gewählte Formulierung für Möglichkeit 3 scheint jedoch sprachlich besser und wird daher verwendet. Bei gemeinsamer Vorgabe mit den beiden anderen Antwortoptionen können Missverständnisse bei der dritten Antwortoption nach Ansicht des Autors ausgeschlossen werden. Um die Wahl der dritten Antwortalternative nicht artifiziell zu erschweren, wurde in der Instruktion (siehe Anhang A.1) explizit darauf hingewiesen, dass diese Antwortalternative für einige Aufgaben durchaus korrekt sein kann.

Die drei „logischen“ Antwortoptionen (je nach Schlussfigur und Hauptprämisse bezogen auf Schaltung 1 oder Schaltung 2) stellen also für jede der 16 Aufgaben des KKS eine erschöpfende Antwortmenge dar. Es ist für jede Aufgabe eine Antwort zu wählen, wobei auch maximal nur eine Antwort pro Aufgabe gewählt werden kann. Der Terminologie der etablierten Paradigmen für Aufgaben zum Konditionalen Schlussfolgern lässt sich diese Art der Itemvorgabe nur schwer zuordnen. Ausgehend von klassischen Multiple-Choice-Tests könnte man von einem *Multiple-Choice-Generierungsparadigma* sprechen, da aus sämtlichen „denkbaren“ Konklusionen diejenige ausgewählt werden soll, die logisch korrekt ist. Ähnlich wie bei klassischen Multiple-Choice-Tests scheint durch die Vorgabe von Antworten allerdings auch hier eine Minderung der Aufgabenschwierigkeit wahrscheinlich. So ist es durchaus möglich, dass Probanden die selbstständige Generierung von Antwortoptionen wie „*Es ist keine eindeutige Aussage über das Funktionieren der Schaltung ableitbar*“ schwerfallen würde. Bei diesem Aufgabentyp finden sich aber durchaus auch Elemente des Verifikationsparadigmas. So ist bspw. für jede Antwortoption durch die Probanden zu entscheiden, ob sie für die gegebenen Prämissen valide ist oder nicht. Durch die Auswahlmöglichkeit lediglich einer Antwort ist jedoch keine Unabhängigkeit zwischen den Antworten innerhalb eines Items gegeben. Insgesamt beschreibt die Bezeichnung Multiple-Choice-Generierungsparadigma den gewählten Aufgabentyp aus Sicht des Autors am besten, da

sämtliche Charakteristika (Wahlmöglichkeit einer Antwortalternative, Generierung der logisch korrekten Antwort aus der Menge aller möglichen Antworten) enthalten sind. Fortan soll daher diese Bezeichnung verwendet werden.

Reihenfolge der Items im KKS

Die Items wurden systematisch vorgegeben. Die Konditionalaussage (Hauptprämisse) blieb für jeweils vier Aufgaben (die vier Nebenprämissen) gleich. Die Vorgabe der Nebenprämissen und damit der Schlussfiguren erfolgte jeweils in der Reihenfolge MP - NA - BK - MT. Nacheinander wurden so die vier Hauptprämissen des Negationsparadigmas vorgegeben, zuerst die ohne jegliche Negation, dann folgte die Hauptprämisse mit einer Negation in der Konsequenz, anschließend die mit einer Negation im Antezedens und letztlich die mit zwei Negationen (siehe auch Nomenklatur 1 bis 4 in Abschnitt 2.3.4). Wie auch bei Rumain et al. (1983) oder Evans et al. (2007) wurde in der Instruktion ein einfaches Modus-Ponens-Problem (hier MP1, also das ohne jegliche Negationen in der Hauptprämisse) verwendet, um die instruktionsgemäße Aufgabenbearbeitung zu illustrieren.

Reasoning

In *Arbeitshypothese 2* wird ein positiver Zusammenhang zwischen (der Stufe) der Sophistiziertheit Konditionalen Schlussfolgerns und Reasoning postuliert. Als bester Marker für Reasoning gelten klassische (Raven-)Matrizenaufgaben (Carpenter, Just & Shell, 1990; Carroll, 1993). Daher wurden zur Erfassung von Reasoning 10 Matrizenaufgaben vom Raven-Typ verwendet. Diese Matrizenitems wurden mit dem Programm ITEMGENERATOR[25] (Ihme, 2007) konstruiert. Aufgrund der regelgeleiteten Itemkonstruktion, die auf dem aktuellen Stand der Forschung aufbaut, können diese 10 Items als kontentvalide (im Sinne der Kontentvalidität nach Klauer, 1978, 1984) angenommen werden. Angaben zur Modellgüte wie zur Reliabilität folgen zu Beginn der Darstellung der Ergebnisse dieser Studie (siehe Abschnitt 2.4.2). Die Probanden konnten bei jeder der 10 Aufgaben eine von acht Alternativen zur Vervollständigung einer klassischen 3x3-Matrix wählen, wobei eine Antwortalternative gewählt werden musste, um zum nächsten Item zu gelangen. Ein Überblick über die 10 vorgegebenen Matrizenitems findet sich in Anhang B dieser Arbeit.

[25] Das Programm ITEMGENERATOR (Ihme, 2007) ermöglicht die Konstruktion von Items auf der Basis vorher definierter Regeln. Für diese Studie wird ein Subset (Ihme, 2007) klassischer Konstruktionsregeln für Matrizenaufgaben (für einen Überblick siehe z.B. Preckel, 2003) definiert.

2.4.1.3 Durchführung

Die Untersuchung fand in Form einer Online-Erhebung statt (vgl. Abschnitt 2.4.1.1). Hierzu wurden die vorgestellten Erhebungsinstrumente (siehe Abschnitt 2.4.1.2) unter Verwendung von PHP, flexSURVEY (Hartenstein, 2007) und MySQL als Online-Test programmiert. Die beiden Erhebungsinstrumente wurden noch ergänzt um Angaben zur Demographie, einzelne Übergangsseiten sowie weitere Aufgaben, die jedoch im Rahmen dieser Arbeit nicht ausgewertet werden. Der Online-Test existierte in zwei Varianten, die sich hinsichtlich der Reihenfolge der Aufgaben unterschieden. Die Reihenfolge der Items innerhalb einer Aufgabe war jedoch in beiden Varianten gleich. Die Testteilnehmer wurden randomisiert einer der beiden Varianten zugewiesen. Tabelle 8 zeigt den Aufbau beider Varianten.

Tabelle 8: Aufbau der beiden Varianten des Online-Tests

Variante 1	Variante 2
Codeabfrage	Codeabfrage
allgemeine Einleitung	allgemeine Einleitung
16 Items des KKS	Wason Selection Task *
16 weitere Items zum Konditionalen Schlussfolgern *	Demographie
Demographie	10 Matrizenitems
10 Matrizenitems	Pause
Pause	16 Items des KKS
Wason Selection Task *	16 weitere Items zum Konditionalen Schlussfolgern *
Verabschiedung	Verabschiedung

Anmerkung. Die mit * gekennzeichneten Items bzw. Aufgaben werden im Rahmen der vorliegenden Arbeit nicht ausgewertet.

Da in Variante 1 zuerst die 16 Items des KKS und dann die 10 Matrizenitems dargeboten wurden und in Variante 2 die Reihenfolge umgekehrt war, konnten durch Vergleich der Ergebnisse beider Varianten Reihenfolgeeffekte überprüft werden. Für die Durchführung des (gesamten) Online-Tests wurde keine Zeitbeschränkung festgelegt, lediglich eine Angabe des voraussichtlichen Zeitbedarfs (ca. 45 Minuten) erfolgte in der allgemeinen Einleitung.

2.4.1.4 Untersuchungsstichprobe

Die Rekrutierung der Untersuchungsstichprobe erfolgte über eine Art Schneeballsystem. Dabei wurden Psychologiestudierende zweier Seminare gebeten, an jeweils 10 bis 20 Bekannte den Link zur Online-Studie in Verbindung mit der Bitte um Teilnahme zu versenden. 905 der auf diese Weise rekrutierten Personen bearbeiteten den KKS vollständig und stellten damit die Untersuchungsstichprobe dar. Der überwiegende Teil dieser Personen machte soziodemographische Angaben zur eigenen Person (jeweils über 90%). Diese Angaben werden in der Folge kurz zusammengefasst:

51,6% der Teilnehmer waren Frauen, entsprechend 48,4% Männer. Das Geschlechterverhältnis der Untersuchungsstichprobe ist also nahezu ausgeglichen. Das Durchschnittsalter lag bei 25 Jahren (Sd = 7,6 Jahre). Als höchsten Bildungsabschluss gaben 7,9% Realschule oder niedriger an, 10,6% eine abgeschlossene Berufsausbildung, 60,9% Abitur und 18,4% einen Hochschulabschluss[26]. 70,3% befanden sich zum Untersuchungszeitpunkt in Ausbildung (auch Studium), 17,4% waren Angestellte[27]. Da der Altersdurchschnitt für eine Stichprobe junger Erwachsener spricht, kann aufgrund der Angaben zur aktuellen Tätigkeit und wegen des vergleichsweise hohen Bildungsniveaus davon ausgegangen werden, dass Studierende den überwiegenden Teil der Stichprobe darstellen. Außerdem stammten 52% der Teilnehmer aus Thüringen. Dieses Bundesland ist damit stark überrepräsentiert.

2.4.1.5 Auswertungsmethoden

Zentraler Punkt der Betrachtungen zu den verwendeten Auswertungsmethoden ist die Auswahl eines geeigneten Analyseverfahrens für die 16 Items des KKS, also zur Überprüfung des *Arbeitshypothesenkomplexes 1* (siehe Abschnitt 2.3.2.2). Eine solche Auswahl richtet sich üblicherweise nach dem Skalenniveau der manifesten sowie der latenten Variablen. Zunächst soll dazu das Antwortformat der Items festgelegt werden. Wenngleich für differenzialdiagnostische Analysen ein 0,1-Format (0…Item nicht gelöst, 1…Item gelöst) naheliegt, ist das für Aufgaben zum Konditionalen Schlussfolgern keineswegs selbstverständlich. In allgemeinpsychologischen Experimenten bspw. werden häufig bei den Schlussfiguren Negation des Antezedens und Bestätigung der Konsequenz die formal-logisch falschen Inferenzen, anstelle korrekter Lösungen analy-

[26] Außerdem gaben 1,3% „Promotion“ an sowie 1,0% „Sonstiges“.
[27] Außerdem gaben 2,5% „Selbstständigkeit“, 2,1% „Hausfrau/-mann“, 1,0% „Leitende(r) Angestellte(r)“ sowie 6,8% „Sonstiges“ an.

siert (vgl. Abschnitt 2.1.2). Daher sei an dieser Stelle nochmals explizit festgehalten, dass für diese (differenzialdiagnostische) Analyse, aufbauend auf den theoretischen Überlegungen und den konkreten Arbeitshypothesen, lediglich betrachtet wird, ob ein Item gelöst wird oder nicht. Ein Item gilt dabei als gelöst, wenn die korrekte logische Schlussfolgerung aus den drei Antwortoptionen des Multiple-Choice-Generierungsparadigmas (siehe Abschnitt 2.4.1.2) gewählt wird. Es gilt als nicht gelöst, wenn eine der beiden formal-logisch falschen Antwortoptionen gewählt wird. Das 0,1-Format der Items und damit das Skalenniveau der manifesten Variablen ist folglich *geordnet kategorial*.

Als nächstes interessiert das Skalenniveau der latenten Variable, also hier der Sophistiziertheit Konditionalen Schlussfolgerns (SKS). Der häufigste psychometrische Ansatz bei der Analyse von Items, die ein gemeinsames Konstrukt messen sollen, ist die Annahme einer kontinuierlichen latenten Variable – so umgesetzt bspw. im *Rasch*- oder im *Birnbaum-Modell* (für eine Beschreibung dieser beiden Modelle siehe z.B. Rost, 2004). Für die theoretisch postulierten Sophistiziertheits-Stufen Konditionalen Schlussfolgerns scheint dieser Ansatz allerdings ungeeignet, was am Beispiel der Modus-Tollens-Aufgaben illustriert werden soll. Diese werden von Personen auf der niedrigsten und der höchsten Sophistiziertheits-Stufe gelöst, nicht jedoch oder lediglich zum Teil von Personen auf den beiden mittleren Stufen (Unsophistizierte II, Fortgeschrittene). Das macht die Modellierung einer kontinuierlichen latenten Variable auf Basis des Löseverhaltens bei den 4 Modus-Tollens-Aufgaben sehr schwierig, insbesondere da für die anderen drei Schlussfiguren (MP, NA, BK) diese theoretische Annahme nicht gilt. Das bedeutet: Auch wenn die vorgestellten Sophistiziertheits-Stufen eine Ordnungsrelation implizieren, kann keine kontinuierliche latente Variable konstruiert werden, die für alle vier Schlussfiguren monotone Response-Funktionen aufweist. Indikativ für die Ausprägung der latenten Variable ist folglich das spezifische Antwortmuster. Es werden sowohl durch das Stufen-Modell (Rijmen & De Boeck, 2003; siehe auch Abschnitt 2.2.2) als auch durch dessen Erweiterung auf das Negationsparadigma (siehe Abschnitt 2.3.2.2) spezifische Antwortmuster für die einzelnen Stufen postuliert. Es wird also eine *Typologie* von Personen beim Konditionalen Schlussfolgern entwickelt: Unsophistizierte, Unsophistizierte II, Fortgeschrittene und Sophistizierte, wobei sich diese vier Typen in ihren Antwortmustern unterscheiden. Das Datenniveau der latenten Variable SKS ist also zunächst lediglich nominal. Da sich die vier Typen (repräsentiert durch das spezifische Antwortmuster) theoriegeleitet ordnen lassen, wird die latente Variable im Weite-

ren als *geordnet kategorial*, also *ordinal* betrachtet. Dabei soll jedoch nochmals betont werden, dass dies nicht aus der Konstruktion der latenten Variable folgt, sondern aus theoretischen Überlegungen.

Als Verfahren zur Identifikation einer latenten nominalen Variable mithilfe kategorialer Items bietet sich eine *Analyse latenter Klassen* an. Dieses Verfahren sowie dessen Anwendung im Falle der latenten Variable SKS werden in der Folge kurz beschrieben.

Analyse latenter Klassen

Die Analyse latenter Klassen (*LCA*[28]; Formann, 1984; Goodman, 1974; Lazarsfeld & Henry, 1968; McCutcheon, 1987) ist ein probabilistisches Verfahren zur Identifikation nominaler latenter Variablen. Die Ausprägungen einer solchen nominalen latenten Variable werden bei der LCA als *Klassen* bezeichnet. Die Grundideen eines solchen *Latente-Klassen-Modells* (*LCM*[29]) sind folgende:

1. Jede Person gehört zu einer und nur zu einer latenten Klasse. Die Klassen sind also paarweise disjunkt und exhaustiv.
2. Innerhalb jeder latenten Klasse ist die Antwortwahrscheinlichkeit für jede Antwortkategorie der betrachteten Items konstant. Zwischen den latenten Klassen unterscheiden sich diese Antwortwahrscheinlichkeiten.
3. Innerhalb jeder latenten Klasse gilt (lokale) stochastische Unabhängigkeit der Antworten auf die Items.

(siehe z.B. Gollwitzer, 2007; Rost, 2004)

Diese Grundideen lassen sich theoretisch auf die latente Variable SKS übertragen: Personen sind genau einer der Sophistiziertheits-Stufen zuordenbar. Alle Personen auf derselben Stufe haben infolge der postulierten stufenspezifischen „typischen" Antwortmuster bezüglich der 16 Items die gleichen Antwort- genauer Lösungswahrscheinlichkeiten. Zwischen den Stufen unterscheiden sich diese Lösungswahrscheinlichkeiten. Eine LCA ist demnach mehr als nur ein adäquates Verfahren zur Identifikation der vier Sophistiziertheits-Stufen Konditionalen Schlussfolgerns mithilfe der 16 Items des KKS. Ein LCM mit vier Klassen ist eine Präzisierung des erweiterten Stufen-Modells in der Sprache der Wahrscheinlichkeitstheorie.

[28] Abk. für die englische Bezeichnung *Latent Class Analysis*
[29] Abk. für die englische Bezeichnung *Latent Class Model*

Noch angemessener wäre eine Präzisierung des erweiterten Stufen-Modells, bei der auch die Ordnung der Stufen bereits durch das Modell berücksichtigt würde. Eine solche Modellierung als *ordinale latente Klassenvariable*, wie sie bspw. Croon (1990) vorschlägt, ist jedoch nicht möglich, da dabei stets die Annahme monotoner Response-Funktionen über die geordneten Klassen hinweg gemacht werden muss. Diese Annahme ist für die 16 Items des KKS jedoch (theoriegeleitet) unplausibel, da bspw. Modus-Tollens-Items von Personen auf der niedrigsten Sophistiziertheits-Stufe korrekt gelöst werden, ebenso von Personen auf der höchsten, nicht jedoch von Personen auf der Fortgeschrittenen-(Zwischen-)Stufe (vgl. z.B. Abschnitt 2.2.2).

Bei einer LCA ist lediglich ein Parameter notwendigerweise im Vorhinein festzulegen: die Anzahl der Klassen. Aufgrund der vier postulierten Stufen wären also vier Klassen vorzugeben. Angenommen diese vier Klassen würden die Stichprobe − zumindest in Bezug auf Sophistiziertheit Konditionalen Schlussfolgerns − erschöpfend charakterisieren, so wäre bei Vorgabe von vier Klassen jedoch keinerlei „Raum" für anderweitiges stereotypes Antwortverhalten, bspw. infolge bestimmter Antworttendenzen (zu Antworttendenzen siehe z.B. Podsakoff, MacKenzie, Lee & Podsakoff, 2003). Derartiges Antwortverhalten kann sich bei einer LCA allerdings in eigenständigen Klassen manifestieren (Gollwitzer, 2007). Daher kann es durchaus sinnvoll sein, „Restklassen" zuzulassen, also Klassen, denen Personen zugeordnet werden, die keiner der postulierten Klassen zugeordnet werden können. Dies scheint umso angebrachter, da es sich um eine erste empirische Überprüfung des erweiterten Stufen-Modells (siehe Abschnitt 2.3.2.2) handelt.

Die bestpassende Klassen-Lösung, also das bestpassende LCM, wird bei einer LCA durch ein eher exploratives Vorgehen bestimmt (siehe dazu z.B. Clogg, 1995). Beginnend mit der Vorgabe einer latenten Klasse[30] wird die Zahl der Klassen schrittweise gesteigert, bis das theoretische Maximum erreicht ist (Clogg, 1995). Wenngleich sich durch Hinzunahme von Klassen der (absolute) Modellfit verbessert, erhöht sich gleichzeitig auch die Modellkomplexität, was wiederum die Qualität der Vorhersage senkt (Clogg, 1995). Daher werden ausgehend von der streng hypothesenkonformen Vier-Klassen-Lösung zusätzlich Modelle mit 1, 2, 3, 5, 6 und 7 Klassen analysiert. Zum Vergleich dieser Klassen-Lösungen werden sog. *Informationskriterien* genutzt, wie

[30] Wird lediglich eine latente Klasse angenommen, spricht man vom sog. *Unabhängigkeitsmodell* (z.B. Clogg, 1995). In der einen latenten Klasse gilt entsprechend den Annahmen eines LCMs lokale stochastische Unabhängigkeit der Items. Passt dieses Modell, „messen" die Items also nicht (Clogg, 1995).

bspw. das *Bayesian Information Criterion* (*BIC*) oder *Akaikes Information Criterion* (*AIC*). Generell gilt für den Vergleich verschiedener Latente-Klassen-Modelle der BIC als das Informationskriterium der Wahl (Nylund, Asparouhov & Muthén, 2007), wobei die Klassen-Lösung mit dem niedrigsten BIC-Wert die bestpassende Klassen-Lösung indiziert. Legt man *Arbeitshypothese 1* liberaler aus, dann wäre es ausreichend, wenn die vier postulierten Klassen (Stufen) des erweiterten Stufen-Modells eine Teilmenge der im bestpassenden LCM identifizierten Klassen bilden würden. Auch wenn sich eine Fünf-Klassen-Lösung (oder sogar eine Lösung mit mehr als fünf Klassen) als die bestpassende herausstellt, kann *Arbeitshypothese 1* demnach beibehalten werden, vorausgesetzt, vier der Klassen sind im Sinne des erweiterten Stufen-Modells interpretierbar. Bei der fünften Klasse (ggf. auch weiteren) könnte es sich bspw. um eine „Restklasse" im beschriebenen Sinne handeln. Dies wäre anhand des Musters der Antwort- bzw. Lösungswahrscheinlichkeiten in dieser Klasse zu überprüfen. In einem solchen Fall würde der Prozentsatz derjenigen Personen, die den vier postulierten Klassen des erweiterten Stufen-Modells zugeordnet werden, eine Art Gütekriterium der 16 Items des KKS darstellen.

Doch alleine die vergleichende Bewertung der BIC-Werte reicht nicht aus, um von einer guten Modellpassung sprechen zu können. Hierfür wird üblicherweise die *Log-Likelihood* der Daten mit der des theoretisch postulierten Modells verglichen. Die statistische Prüfung dieser Differenz erfolgt meist mittels *Likelihood-Ratio*- oder mittels *Chi-Quadrat-Test*. Allerdings ergeben sich für die 16 Items des KKS insgesamt 65.536 ($= 2^{16}$) mögliche Antwortmuster, die trotz der vergleichsweise umfangreichen Stichprobe ($N = 905$; siehe Abschnitt 2.4.1.4) unmöglich alle auftreten können. Da aufgrund dieses sog. *sparse-data-Problems* sowohl für den Likelihood-Ratio- als auch den Chi-Quadrat-Test sowie auch für die übrigen Prüfgrößen der Power-Divergence-Familie (siehe dazu z.B. Davier, 1997) die Verteilungen unbekannt sind, soll auf den Modelltest mittels *parametrischen Bootstraps* (für eine Einführung siehe Efron & Tibshirani, 1993) zurückgegriffen werden. Dabei werden basierend auf den geschätzten Parametern des berechneten LCMs neue Stichprobendaten simuliert, für die das jeweils spezifizierte LCM gilt. Als Prüfgrößen empfiehlt Davier (1997) Chi-Quadrat und *Cressie Read* (Cressie & Read, 1984), da sich diese auch im „sparse data"-Fall als recht robust erweisen. Die Werte für Chi-Quadrat und Cressie Read der Originalstichprobe werden dann mit denen der Bootstrap-Stichproben verglichen. Wenngleich für einen Modelltest bereits 40 derartige Bootstrap-Stichproben ausreichen (Davier, 1997), ist eine deutlich

größere Zahl notwendig, um die Verteilung der Cressie-Read- bzw. Chi-Quadrat-Werte ausreichend gut schätzen zu können (Efron & Tibshirani, 1993). Daher werden für diese Untersuchung 200 Bootstrap-Stichproben zur Prüfung des Modellfits gezogen. Ist die Wahrscheinlichkeit für den Chi-Quadrat-Wert der Originalstichprobe oder einen größeren, gegeben ein bestimmtes LCM gilt, größer-gleich .05 (konventionelle Festlegung des alpha-Niveaus; vgl. z.B. Bonnefon, Eid, Vautier & Jmel, 2008; Gollwitzer, 2007), braucht dass jeweilige Modell aufgrund der Daten nicht verworfen zu werden. Gleiches gilt für den Cressie-Read-Wert.

Es sei an dieser Stelle erwähnt, dass neben der Anzahl der Klassen auch andere Parameter bei einer LCA im Vorhinein festgelegt werden können. Es gibt zahlreiche Möglichkeiten, weitere Parameter zu restringieren (für einen Überblick siehe z.B. Rost, 2004) und den Modellfit dann − gegeben dieser Restriktionen − zu interpretieren. Da es sich jedoch um eine erste empirische Überprüfung des erweiterten Stufen-Modells handelt, soll zunächst auf jegliche zusätzliche Restriktionen verzichtet werden. Man spricht in diesem Fall von einer *unrestringierten LCA*. In der Folge werden die Parameter vorgestellt, die bei einer solchen LCA geschätzt werden.

Für jede der vorgegebenen Klassen wird die *Klassengröße* geschätzt. Dabei handelt es sich um die Zuweisungswahrscheinlichkeit P_C zu einer bestimmten Klasse C gegeben eine beliebige Person. Des Weiteren werden in einer LCA die *klassenbedingten Antwortwahrscheinlichkeiten* $P(Y_i = y \mid C = c)$ bzw. *klassenbedingten Lösungswahrscheinlichkeiten* $P(Y_i = 1 \mid C = c)$ für jedes Item Y_i geschätzt. Da diese das „typische" Antwort- bzw. Löseverhalten für jede Klasse C repräsentieren, stellen sie die entscheidende Quelle zur Überprüfung der *Arbeitshypothesen 1a* bis *1d* dar. Schließlich werden in einer LCA noch die *bedingten Klassenzuordnungswahrscheinlichkeiten* $P(C = c \mid \mathbf{y} = \mathbf{y})$ geschätzt. Dabei handelt es sich um die jeweiligen Zuordnungswahrscheinlichkeiten eines spezifischen Antwortmusters **y** zu sämtlichen Klassen C. Damit ergibt sich für jedes Antwortmuster auch eine Klasse, für die diese Zuordnungswahrscheinlichkeit maximal ist[31]. Diese Klasse ist dann für eine Person mit diesem Antwortmuster am wahrscheinlichsten, woraufhin diese Person (mit eben dieser Wahrscheinlichkeit) dieser Klasse zugeordnet wird. Das arithmetische Mittel sämtlicher Zuordnungswahrscheinlichkeiten von Personen, die einer Klasse C zugeordnet werden, wird als *Treffsicherheit* T_C für diese Klasse bezeichnet (Gollwitzer, 2007; Rost, 2004).

[31] Ein Antwortmuster kann allerdings auch gleichgroße (maximale) Zuordnungswahrscheinlichkeiten zu mehreren Klassen haben.

Das arithmetische Mittel der Zuordnungswahrscheinlichkeiten aller Personen über die Klassen hinweg, denen sie zugeordnet werden, ist die Treffsicherheit T des LCMs. Sowohl die klassenspezifischen Treffsicherheiten T_C als auch die Treffsicherheit T des gesamten LCMs sind nach Rost (2004) Maße für die Messgenauigkeit eines LCMs, für deren Höhe ähnliche Konventionen wie für Reliabilitätskoeffizienten gelten ($T > .85$: relativ hohe Treffsicherheit, $T > .90$: hohe Treffsicherheit). Ob es sich bei der Treffsicherheit tatsächlich um einen Indikator für Messgenauigkeit handelt, wird bei der Thematisierung des entsprechenden Testgütekriteriums (vgl. Abschnitt 3.2.1) erörtert. In jedem Falle handelt es sich bei der Treffsicherheit um ein Gütekriterium eines LCMs.

Ein weiteres, besonders wichtiges Gütekriterium einer LCA stellt die Interpretierbarkeit der Ergebnisse dar (Clogg, 1995). Dies gilt folglich insbesondere für die Überprüfung der *Arbeitshypothesen 1a* bis *1d*. Die Frage lautet, ob sich das postulierte klassenspezifische Antwort- bzw. Löseverhalten für die 16 Items des KKS in den klassenbedingten Lösungswahrscheinlichkeiten widerspiegelt. Das postulierte Löseverhalten in den vier vermuteten Klassen ist allerdings so spezifisch, dass anhand des Musters der Lösungswahrscheinlichkeiten bereits „prima facie“ erkennbar sein sollte, um welche Klasse es sich handelt. Die Prüfung klassenbedingter Antwort- oder Lösungswahrscheinlichkeiten per Augenschein ist im Falle einer unrestringierten LCA ohnehin die Analysemethode der Wahl (vgl. z.B. Gollwitzer, 2007). Für diese erste empirische Analyse soll die Prüfung der *Arbeitshypothesen 1a* bis *1d* daher auch auf diese Weise erfolgen. Ist nicht sofort ersichtlich, dass es sich um eine der postulierten Klassen handelt, so können zusätzliche Restriktionen bezüglich der Lösungswahrscheinlichkeiten in den einzelnen Klassen eingeführt und deren Auswirkung auf den Modellfit überprüft werden. Theoriegeleitet könnten in einer Klasse bspw. die Lösungswahrscheinlichkeiten für Aufgaben zu den Schlussfiguren MP, NA und BK nahe eins sowie für Aufgaben zum Modus Tollens nahe null fixiert werden. Dies entspräche der Klasse „Fortgeschrittene“ des erweiterten Stufen-Modells. Ähnlich könnte für die Lösungswahrscheinlichkeiten in anderen Klassen verfahren werden. Das so entstehende restringierte Modell könnte dann mit dem unrestringierten verglichen werden, bspw. mittels Likelihood-Ratio-Test. Es wird aufgrund der Eindeutigkeit des postulierten Antwort- bzw. Löseverhaltens in den vermuteten Klassen allerdings davon ausgegangen, dass solche zusätzlichen Tests unnötig sind. Durch die Präzisierung des erweiterten Stufen-Modells als Latente-Klassen-Modell lassen sich nun auch die Arbeitshypothesen aus *Arbeitshypothesenkomplex 1* (siehe Abschnitt 2.3.2.2) wie folgt präzisieren:

Arbeitshypothese 1: Bei Berechnung einer LCA für die 16 Aufgaben des KKS zeigen sich vier latente Klassen, die folgende Charakteristika bezüglich der klassenbedingten Lösungswahrscheinlichkeiten aufweisen:

Arbeitshypothese 1a: Es zeigt sich eine latente Klasse mit hohen Lösungswahrscheinlichkeiten bei Modus-Ponens-Aufgaben und Modus-Tollens-Aufgaben sowie niedrigen Lösungswahrscheinlichkeiten bei Aufgaben zur Negation des Antezedens und Aufgaben zur Bestätigung der Konsequenz. (Klasse „Unsophistizierte“)

Arbeitshypothese 1b: Es zeigt sich eine latente Klasse mit hohen Lösungswahrscheinlichkeiten bei Modus-Ponens-Aufgaben und niedrigen Lösungswahrscheinlichkeiten bei Aufgaben zur Bestätigung der Konsequenz. Bei Aufgaben zur Negation des Antezedens zeigt sich eine deutlich höhere Lösungswahrscheinlichkeit, wenn die Hauptprämisse mindestens eine Negation enthält. Bei Modus-Tollens-Aufgaben zeigt sich eine deutlich niedrigere Lösungswahrscheinlichkeit, wenn die Hauptprämisse mindestens eine Negation enthält. (Klasse „Unsophistizierte II“)

Arbeitshypothese 1c: Es zeigt sich eine latente Klasse mit hohen Lösungswahrscheinlichkeiten bei Modus-Ponens-Aufgaben, Aufgaben zur Negation des Antezedens und Aufgaben zur Bestätigung der Konsequenz sowie niedrigen Lösungswahrscheinlichkeiten bei Modus-Tollens-Aufgaben. (Klasse „Fortgeschrittene“)

Arbeitshypothese 1d: Es zeigt sich eine latente Klasse mit hohen Lösungswahrscheinlichkeiten bei Modus-Ponens-Aufgaben, Aufgaben zur Negation des Antezedens, Aufgaben zur Bestätigung der Konsequenz und Modus-Tollens-Aufgaben. (Klasse „Sophistizierte“)

Können nach der empirischen Überprüfung die vier *Arbeitshypothesen 1a* bis *1d* beibehalten werden, braucht folglich auch *Arbeitshypothese 1* nicht verworfen werden, die die Identifikation aller vier Klassen postuliert.

Zusammenhang zwischen SKS und Reasoning

Die Bestimmung der nominalen latenten Variable SKS ist mit den Ausführungen zur LCA (bzw. zum entsprechenden LCM) hinreichend behandelt. Für die Matrizenitems wird − wie für derartige Tests üblich – von einer kontinuierlichen latenten Variable ausgegangen, die Reasoning repräsentieren soll und daher in der Folge auch so bezeichnet wird. Als Messmodell für die 10 Matrizenitems wird ein dichotomes Rasch-Modell

(siehe z.B. Rost, 2004; Steyer & Eid, 2001) angenommen. Die Überprüfung des Modellfits erfolgt wiederum mittels parametrischen Bootstraps, da auch für die 10 Matrizenitems mit 1.024 ($= 2^{10}$) möglichen Antwortmustern wieder ein sparse-data-Problem vorliegt. Das Prinzip des parametrischen Bootstraps ist das gleiche wie bei der soeben beschriebenen LCA mit dem Unterschied, dass bei der Überprüfung eines Rasch-Modells parametrische Stichproben simuliert werden, in denen ein Rasch-Modell gilt. Wiederum wird das alpha-Niveau konventionell festgelegt ($\alpha = .05$). Vorausgesetzt, aufgrund dieser Modellprüfung braucht die Annahme eines Rasch-Modells nicht verworfen zu werden, können zusätzlich die *Z*-Werte der Q-Indices überprüft werden. Diese reagieren auf Abweichungen des geschätzten Diskriminationsparameters β_i eines Items von eins. Wie für *Z*-Werte üblich wird dies durch *Z*-Werte $Z < -1{,}96$ bzw. $Z > 1{,}96$ indiziert. Da in einem Rasch-Modell für alle Items gleichgroße Diskriminationsparameter angenommen werden (Annahme der Rasch-Homogenität; siehe z.B. Steyer & Eid, 2001), können so Items identifiziert werden, die möglicherweise nur einem Birnbaum-Modell (siehe z.B. Rost, 2004) genügen, also einem Modell mit unterschiedlichen Diskriminationsparametern. Hinweise darauf liefern − wie eben erläutert − von eins verschiedene Diskriminationsparameter. Als Indikator für die Messgenauigkeit wird *Andrichs Reliabilität* (Andrich, 1988) für die Personenparameterschätzung angegeben, zu deren Interpretation jedoch angemerkt werden muss, dass die Reliabilität im Rahmen der *Item-Response-Theorie* (*IRT*) abhängig vom Personenparameter ist. Das heißt, die Reliabilität kann für verschiedene Ausprägungen der latenten Variable (geschätzt durch den Personenparameter) verschieden hoch sein. Andrichs Reliabilität ist also im Prinzip nur ein „durchschnittlicher“ Wert für die Reliabilität der einzelnen Personenparameterschätzungen, was bei der Interpretation stets berücksichtigt werden sollte.

Nachdem nun die Methoden zur Überprüfung der Messmodelle für beide Konstrukte festgelegt sind, stellt sich die Frage, auf welche Weise der Zusammenhang zwischen einer kontinuierlichen Variable und einer nominalen Variable bestimmt werden kann, speziell auf latenter Ebene, gemäß der zweiten Forderung an die vorliegende Arbeit (siehe Abschnitt 2.4.1). Ein geeignetes Analyseverfahren stellt eine *multinomiale logistische Regression für latente Variablen* dar. Die Grundidee der multinomialen logistischen Regression ist die Vorhersage eines nominalskalierten Regressanden (hier der latenten Klassenvariable SKS) durch einen oder mehrere (kategoriale oder kontinuierliche) Regressoren. Für die beabsichtigte Analyse ist dieser Regressor die kontinuier-

liche latente Variable Reasoning, gemessen durch die 10 Matrizenitems. Auf eine ausführliche Beschreibung der multinomialen logistischen Regression sowie eine Darstellung der Modellgleichung soll verzichtet werden (siehe dazu bspw. Fahrmeir, Kneib & Lang, 2007[32]). Lediglich das Grundprinzip wird kurz beschrieben: Eine der Kategorien des nominalen Regressanden bildet die sog. *Referenzkategorie*. Jede weitere Kategorie wird zu dieser Referenzkategorie in Beziehung gesetzt, indem jeweils eine *binär-logistische Regression* (siehe z.B. Fahrmeir et al., 2007) berechnet wird. Im Falle der vier postulierten Klassen (Stufen) der Sophistiziertheit Konditionalen Schlussfolgerns hätte der nominale Regressand also vier Kategorien. Bei einer Referenzkategorie würden entsprechend drei binär-logistische Regressionen gerechnet und es ergäben sich drei Anstiegskoeffizienten. Dabei bedeutet ein Anstiegskoeffizient größer null, dass mit steigender Ausprägung des Regressors die Wahrscheinlichkeit steigt, der mit der Referenzkategorie verglichenen Kategorie zugeordnet zu werden, ein Anstiegskoeffizient kleiner null entsprechend das Gegenteil. Zur Hypothesentestung wird der Likelihood-Ratio-Test (*LR-Test*) verwendet. Dessen Nullhypothese besagt, dass die logit-transformierte latente Klassenvariable linear regressiv unabhängig ist vom (bspw. kontinuierlichen) latenten Regressor, in diesem Fall Reasoning. Die Alternativhypothese besagt entsprechend, dass eine lineare regressive Abhängigkeit besteht. Beim LR-Test werden die Log-Likelihoods zweier genesteter Modelle ins Verhältnis gesetzt. Im restriktiveren Modell, welches die Nullhypothese repräsentiert, werden sämtliche Anstiegskoeffizienten auf null fixiert. Im liberaleren Modell, das die Alternativhypothese repräsentiert, werden die Anstiegskoeffizienten frei geschätzt. Durch Multiplikation der erhaltenen Log-Likelihood-Differenz[33] mit dem Faktor −2 erhält man eine Chi-Quadrat-verteilte Prüfgröße (Tutz, 2000). Im Falle eines LCMs mit vier Klassen werden also drei Klassen mit der Referenzklasse (Referenzkategorie) verglichen, das heißt, im restriktiveren Modell werden drei Anstiegskoeffizienten auf 0 fixiert. Gegenüber dem liberaleren Modell entstehen also 3 Freiheitsgrade. Es wird folgende Nullhypothese geprüft:

H_0: $\beta_{10} = \beta_{20} = \beta_{30} = 0$,

wobei der Index 0 jeweils die Referenzklasse indiziert und die Indizes 1 bis 3 die verbliebenen 3 Klassen[34]. Die Prüfung der Nullhypothese erfolgt also durch die Chi-

[32] Bei Fahrmeier et al. (2007) wird das zugrunde liegende Modell als *Mehrkategoriales Logit-Modell* bezeichnet.
[33] Die Log-Likelihood-Differenz ist die Differenz zwischen den Log-Likelihoods beider Regressionen.
[34] Dies impliziert gleichzeitig: $\beta_{21} = \beta_{31} = \beta_{32} = 0$.

Quadrat-verteilte Prüfgröße bei drei Freiheitsgraden. Vorzeichen und Betrag der Anstiegskoeffizienten indizieren zudem die Ordnung der Klassen in Bezug zur Referenzklasse. Dadurch sind Rückschlüsse auf die Ordnung der vier Klassen in Bezug auf Reasoning möglich, wie im zweiten Teil von *Arbeitshypothese 2* postuliert.

Zusätzlich zur Prüfung der Nullhypothese bedarf es einer Effektgröße, die die Stärke der regressiven Abhängigkeit angibt, quasi einer Art Determinationskoeffizient für die multinomiale logistische Regression. Einen dem Determinationskoeffizienten *R-Quadrat* der „klassischen" Regression (Rao, 1973; siehe z.B. auch Steyer, 2003) vergleichbaren Koeffizienten stellt *Nagelkerkes R-Quadrat* (Nagelkerke, 1991) dar (für die Berechnungsformel von Nagelkerkes R-Quadrat siehe Anhang C). Gegenüber anderen bedeutungsähnlichen Koeffizienten aus dieser sog. *Pseudo-R-Quadrat-Familie* wie bspw. dem von Cox und Snell (1989) besitzt Nagelkerkes R-Quadrat den entscheidenden Vorteil der Standardisierung und kann gerade deshalb durchaus mit dem R-Quadrat der „klassischen" Regression verglichen werden (Nagelkerke, 1991). Die üblichen Konventionen (z.B. Cohen, 1988) für die Einschätzung der Stärke des Effektes anhand des Determinationskoeffizienten werden entsprechend auf Nagelkerkes R-Quadrat übertragen (geringer Effekt: .02, mittlerer Effekt: .13, starker Effekt: .26; vgl. Bortz, 1999, S. 449). Es ist allerdings anzumerken, dass bei logistischen Regressionen prinzipiell niedrige (Pseudo-)R-Quadrat-Werte die Norm sind (Hosmer & Lemeshow, 2000). Bei der Bewertung der Höhe von Nagelkerkes R-Quadrat sollte das berücksichtigt werden. Damit lässt sich nun *Arbeitshypothese 2* für das gewählte Analyseverfahren (multinomiale logistische Regression für latente Variablen) präzisieren:

Arbeitshypothese 2a[35]*:* Es besteht eine multinomiale logistische regressive Abhängigkeit der latenten Variable SKS von der latenten Variable Reasoning. (Die latente Variable SKS wird dabei durch das LCM der 16 Items des KKS konstruiert, die latente Variable Reasoning durch ein Rasch-Modell der 10 Matrizenitems.) Nagelkerkes R-Quadrat liegt im mittleren Bereich. Die Anstiegskoeffizienten der multinomialen logistischen Regression zeigen (hinsichtlich der latenten Variable Reasoning) die durch das erweiterte Stufen-Modell postulierte Ordnung: Unsophistizierte - Rang 1, Unsophistizierte II – Rang 2, Fortgeschrittene - Rang 3, Sophistizierte - Rang 4.

[35] Da für *Arbeitshypothese 2* noch eine zweite Präzisierung folgt, wird diese erste Präzisierung durch den Zusatz „a" gekennzeichnet.

Da die Ergebnisse einer multinomialen logistischen Regression wenig anschaulich und schlecht mit den Ergebnissen häufiger verwendeter Verfahren wie bspw. Korrelationen vergleichbar sind, soll noch ein zweites Verfahren vorgestellt werden, mit dem *Arbeitshypothese 2* anschaulicher überprüft werden kann. Allerdings werden dabei nicht latente Variablen sondern manifeste Variablen zueinander in Beziehung gesetzt und daher das Messfehlerproblem nicht berücksichtigt. Die Grundidee ist, dass sich die Sophistiziertheits-Stufen Konditionalen Schlussfolgerns theoriegeleitet ordnen lassen, vorausgesetzt, Personen lassen sich aufgrund ihrer Testleistung hinreichend genau den postulierten Stufen zuordnen. Unter Berücksichtigung der postulierten Annahmen über die Ordnung der Stufen weist die zugrundeliegende latente Variable also ein ordinales Skalenniveau auf (siehe dazu auch Ausführungen zu Beginn dieses Abschnitts). Die Werte dieser ordinalen Variable bilden die Zuordnungen der Probanden zu der für sie wahrscheinlichsten Klasse im berechneten LCM. Die Klassen werden aufgrund der postulierten Ordnung wie folgt festgelegt:

1 (Rang 1) … Unsophistizierte
2 (Rang 2) … Unsophistizierte II
3 (Rang 3) … Fortgeschrittene
4 (Rang 4) … Sophistizierte

Jede Person hat damit einen Wert zwischen 1 und 4 auf der (manifesten) Variable SKS. Die Wahrscheinlichkeit dieser Klassenzuordnung wird nun jedoch nicht mehr berücksichtigt. Für die 10 Matrizenitems bildet der Summenwert eine suffiziente Statistik, vorausgesetzt, sie genügen einem Rasch-Modell (Steyer & Eid, 2001). Zur Bestimmung des Zusammenhangs zwischen der (manifesten) ordinalen Variable SKS und dem Summenwert der Matrizenitems werden einfache Rangkorrelationen berechnet und Spearmans Rho sowie Kendalls Tau (siehe z.B. Bortz, Lienert & Boehnke, 2000) als Koeffizienten angegeben. Da eine positive Korrelation postuliert wird, erfolgt die Signifikanzprüfung einseitig. Die Stärke der Effekte wird gemäß den Konventionen von Cohen (1990, 1992) für Korrelationskoeffizienten beurteilt: .10 – geringer Effekt, .30 − mittlerer Effekt, .50 – hoher Effekt (vgl. Bortz & Döring, 1995). *Arbeitshypothese 2* lässt sich demnach noch für ein zweites Analyseverfahren präzisieren:

Arbeitshypothese 2b: Nachdem die manifesten Klassenzuordnungen entsprechend den Annahmen des erweiterten Stufen-Modells in eine Rangreihe gebracht wurden (Unsophistizierte - Rang 1, Unsophistizierte II - Rang 2, Fortgeschrittene - Rang 3, Sophistizierte - Rang 4), zeigt sich für die so konstruierte ordinale manifeste Variable *SKS_manifest* eine positive Rangkorrelation (Spearmans Rho, Kendalls Tau) mit dem Summenwert der 10 Matrizenitems. Die Effektstärke des Rangkorrelationskoeffizienten liegt jeweils im mittleren Bereich.

Überprüfung von Reihenfolgeeffekten

Um Reihenfolgeeffekte im Testaufbau zu überprüfen, können beide Varianten des Online-Tests (Variante 1: KKS vor Matrizenitems, Variante 2: KKS nach Matrizenitems; vgl. Abschnitt 2.4.1.3) hinsichtlich der Ergebnisse des KKS verglichen werden. Hierfür bietet sich der Chi-Quadrat-Test einer Kreuztabelle „Klasse im KKS x Variante des Online-Tests“ an, da bei randomisierter Zuweisung zu den beiden Reihenfolge-Varianten die Verteilung der latenten Klassenvariable in beiden Varianten gleich sein sollte. Ein nicht-signifikantes Ergebnis spricht dafür, die Nullhypothese (keine Reihenfolgeeffekte in Bezug auf die Ergebnisse des KKS) beizubehalten. Für den Summenwert der 10 Matrizenitems kann ein solcher Reihenfolgeeffekt unter Verwendung eines zweiseitigen *t*-Tests überprüft werden (abhängige Variable: Summenwert der 10 Matrizenitems, unabhängige Variable: Variante des Online-Tests). Wiederum kann im Falle eines nicht-signifikanten Ergebnisses die Nullhypothese (keine Reihenfolgeeffekte in Bezug auf die Ergebnisse der 10 Matrizenitems) beibehalten werden.

Verwendete Software

Zur Bestimmung des Modellfits wird bei sämtlichen LCMs das Programm Winmira 2001 (Davier, 2001) verwendet, in welches die beschriebenen Bootstrap-Prozeduren integriert sind. Da die multinomiale logistische Regression für latente Variablen nicht in Winmira 2001 implementiert ist, wird für deren Berechnung die Software M*plus* 5 (Muthén & Muthén, 1998-2007) genutzt. Allerdings werden sämtliche Messmodelle (LCM für die 16 Items des KKS, Rasch-Modell für die 10 Matrizenitems) vorher mit Winmira 2001 und den integrierten Bootstrap-Prozeduren überprüft, da diese wiederum nicht in M*plus*[36] integriert sind[37]. Signifikanz sowie Stärke der multinomialen logisti-

[36] Wird in der Folge die Software M*plus* angeführt, ist damit stets die Version M*plus* 5 (Muthén & Muthén, 1998-2007) gemeint.

schen regressiven Abhängigkeit werden entsprechend dem beschriebenen Vorgehen (siehe dort) aus den jeweiligen Log-Likelihoods „von Hand" berechnet. Analysen zu demographischen Daten, die Berechnung der Rangkorrelationskoeffizienten sowie die Prüfung der Reihenfolgeeffekte erfolgt unter Verwendung der Software SPSS.

Zum Abschluss der Ausführungen zu Auswertungsmethoden soll noch ein Punkt betrachtet werden, der bei Online-Erhebungen dem Vorteil der vergleichsweise ökonomischen Rekrutierung großer Stichproben gegenübersteht: das Problem der Sicherstellung der Datenqualität (Funke & Reips, 2007; vgl. auch Reips, 2002). So ist meist ein effektives Datenscreening nötig, um bspw. Probanden zu identifizieren, die „einfach nur durchgeklickt haben" oder Ähnliches (für einen Überblick solch unerwünschter Verhaltensweisen der Teilnehmer von Online-Tests siehe z.B. Reips, 2001). Auch für die im Rahmen dieser Arbeit durchgeführten Online-Erhebungen[38] stellt sich das Problem der Sicherung der Datenqualität, insbesondere da die Stichprobe (ebenso wie deren Antwortverhalten) aufgrund des Schneeballsystems (siehe dazu Abschnitt 2.4.1.4) schwer kontrolliert werden kann. Daher wird in der Folge die Entwicklung einer eigens für den KKS konzipierten Methode zur Sicherung der Datenqualität vorgestellt.

Entwicklung einer Methode zur Sicherung der Datenqualität

Ausgangspunkt ist ein zentraler Befund aus Studien zum Negationsparadigma (siehe Abschnitt 2.3.2.1), auf dem auch die 16 Items des KKS basieren. In einer Vielzahl von Studien (für einen Überblick siehe z.B. Evans et al., 1993) zeigt sich dabei, dass unabhängig von der Verwendung von Negationen zwischen 95% und 100% der Probanden bei Modus-Ponens-Aufgaben den logisch korrekten Schluss ziehen (Evans et al., 1993). Diese Aufgaben können also als „extrem leicht" bezeichnet werden, was auf Basis beider kognitionspsychologischer Theorien zum Konditionalen Schlussfolgern auch völlig plausibel ist (vgl. Abschnitte 2.1.2.1 und 2.1.2.2). Im Rahmen des Negationsparadigmas werden genau vier Modus-Ponens-Aufgaben, also vier extrem leichte Aufgaben vorgegeben. Diese vier Aufgaben sollten also (bis auf zufällige Fehler) von allen Probanden korrekt gelöst werden – zumindest von denen, die den Test ernsthaft bearbeiten.

[37] In M*plus* sind lediglich Bootstrapverfahren zur Bestimmung der Parameter integriert, nicht jedoch für die verwendeten Prüfgrößen der Modellgeltungskontrolle (B. Muthén, 2009, persönliche Kommunikation).

[38] Neben der hier vorgestellten Studie folgen im Rahmen der Betrachtung von Testgütekriterien noch weitere Studien zum KKS (siehe Abschnitte 3.2 und 3.3).

Das in diesem Online-Test verwendete Multiple-Choice-Generierungsparadigma (siehe Abschnitt 2.4.1.2) ist so aufgebaut, dass die korrekte Antwortoption in ihrer Position variiert, sodass bspw. bei systematischem Anklicken jeweils der ersten Antwortoption lediglich zwei der vier Modus-Ponens-Aufgaben richtig gelöst würden. Für eine Vielzahl anderer „Durchklicker-Strategien" (z.B. wechselnd die erste, zweite oder dritte Antwortoption oder ähnliche) ergäben sich ähnlich zufällig wirkende Antwortmuster.

In der Konsequenz sollte es also zwei Klassen von Probanden geben, nämlich eine Klasse von Probanden, die den Test ernsthaft bearbeiten und folglich (bis auf zufällige Fehler) alle vier Modus-Ponens-Aufgaben korrekt lösen, und eine Klasse von Probanden, die zufällig anklicken oder auch systematisch durchklicken und entsprechend deutlich von eins verschiedene Lösungswahrscheinlichkeiten für die vier Modus-Ponens-Aufgaben aufweisen. Angesichts von jeweils drei Antwortalternativen pro Item (vgl. Abschnitt 2.4.1.2) sollten die erwarteten Lösungswahrscheinlichkeiten bei zufälligem Anklicken um .33 für jedes Item liegen. Zur Überprüfung dieser Hypothese und Identifikation dieser beiden Klassen bietet sich wiederum eine LCA an. Unter der Voraussetzung, dass ein Zwei-(latente-)Klassen-Modell auf die Daten der vier Modus-Ponens-Aufgaben passt und die Klassen in obigem Sinne interpretierbar sind, kann eine Reduktion der Gesamtdaten auf die Daten derjenigen Probanden, die der ersten Klasse zugeordnet werden können, die Datenqualität deutlich verbessern. Diese Technik wird im Folgenden als *Modus-Ponens-Reduktion* der Daten (bzw. der Stichprobe) bezeichnet. Zur Überprüfung ihrer Angemessenheit wird folgende Hypothese formuliert:

Hypothese MP-Reduktion_1: Es existieren zwei Klassen von Personen: eine umfangreiche Klasse ($C = 1$) von Personen, die den Test ernsthaft bearbeiten und entsprechend hohe Lösewahrscheinlichkeiten bezüglich aller vier Modus-Ponens-Aufgaben aufweisen [$P(MP_i = 1 \mid C = 1) \approx 1$, $i = 1, \ldots, 4$], und eine zweite weniger umfangreiche Klasse ($C = 2$) von Personen, die die Aufgaben nicht ernsthaft bearbeiten und deren Lösewahrscheinlichkeiten bezüglich der vier Modus-Ponens-Aufgaben im Zufallsbereich liegen [$P(MP_i = 1 \mid C = 2) \approx .33$, $i = 1, \ldots, 4$].

Die erste Klasse ($C = 1$) kann dann als „Modus-Ponens-Löser", die zweite Klasse ($C = 2$) als „Durchklicker" interpretiert werden. Zur Prüfung dieser Hypothese wird zunächst die bestpassende Klassen-Lösung mittels Vergleiches der BIC-Werte ausgewählt. Hierzu werden 1, 2 und 3 Klassen vorgegeben, wobei die Vorgabe von drei

Klassen in diesem Fall die theoretisch maximal mögliche Klassenzahl darstellt[39]. Im Anschluss wird mittels parametrischen Bootstrapverfahrens überprüft, ob die bestpassende Klassen-Lösung einen adäquaten Modellfit aufweist. Als Prüfgrößen werden Chi-Quadrat und Cressie Read analysiert. Zeigt sich ein adäquater Modellfit, werden die klassenbedingten Lösungswahrscheinlichkeiten per Augenschein mit den postulierten verglichen.

Wenngleich dieses Vorgehen intuitiv ist und hypothesenkonforme Ergebnisse zur Rechfertigung der Modus-Ponens-Reduktion ausreichen sollten, scheint dennoch eine Beziehung zu einem externen Kriterium sinnvoll. Auf diese Weise könnte eine Art Kriteriumsvalidierung der Modus-Ponens-Reduktion vorgenommen werden. Als solches Kriterium wird eine von Batinic (2006) vorgeschlagene Frage zur Sicherung der Datenqualität gewählt: „*Haben Sie den Test ernsthaft bearbeitet?*“ (vgl. z.B. auch Reips, 2000). Die Antwort („*ja*“/„*nein*“) auf diese Frage klärt einen überraschend hohen Anteil der Varianz von Testergebnissen auf (Batinic, 2006). Vorausgesetzt, die strukturellen Überlegungen zur Modus-Ponens-Reduktion bestätigen sich, soll eine Kreuztabelle angegeben werden, in der die zwei Klassen der Modus-Ponens-Reduktion mit der Angabe der ernsthaften Testbearbeitung („*ja*“/„*nein*“) gekreuzt werden. Die entsprechende Hypothese zur Kriteriumsvalidierung der Modus-Ponens-Reduktion lautet:

Hypothese MP-Reduktion_2: Unter den „Durchklickern“ ist der Anteil an Personen, die angeben, sie hätten den Test *nicht* ernsthaft bearbeitet, höher als unter den „Modus-Ponens-Lösern“.

Es bleibt festzuhalten, dass eine positive Prüfung von *Hypothese MP-Reduktion_1* notwendige Voraussetzung für die Überprüfbarkeit von *Hypothese MP-Reduktion_2* ist. Die Überprüfung von *Hypothese MP-Reduktion_2* erfolgt dann mittels Chi-Quadrat-Test. Bestätigen sich die *Hypothesen MP-Reduktion_1* und *MP-Reduktion_2*, dann werden zur Sicherung der Datenqualität in der hier beschriebenen Online-Erhebung nur Daten von Personen analysiert, die mit hoher Wahrscheinlichkeit (größer .85; nach Rost, 2004) der latenten Klasse der „Modus-Ponens-Löser“ zugeordnet werden können.

Damit ist die Darstellung der verwendeten Methoden zur Prüfung der Arbeitshypothesen abgeschlossen. Es folgt die Darstellung der Ergebnisse.

[39] Im Falle von vier Items hat eine (unrestringierte) LCA unter Vorgabe von drei Klassen genau einen Freiheitsgrad ($df = 1$) und ist damit identifizier- und testbar. Bei einer LCA unter Vorgabe von vier Klassen ist die Anzahl der zu schätzenden Parameter jedoch größer als die Anzahl der gegebenen ($df = -4$) und das Modell demnach weder identifizier- noch testbar.

2.4.2 Ergebnisse

In diesem Abschnitt werden die Ergebnisse zur Überprüfung der sechs Arbeitshypothesen dargestellt. Zunächst erfolgt jedoch die Sicherung der Datenqualität, im Rahmen derer die Annahmen zur Modus-Ponens-Reduktion (*Hypothesen MP-Reduktion_1* und *MP-Reduktion_2*) überprüft werden, um etwaige Konsequenzen bei den weiteren Analysen berücksichtigen zu können. Anschließend folgt die Prüfung des erweiterten Stufen-Modells anhand der 16 Items des KKS (*Arbeitshypothesen 1*, *1a* bis *1d*) und schließlich die Überprüfung der Zusammenhangshypothese zwischen Sophistiziertheit Konditionalen Schlussfolgerns und Reasoning (*Arbeitshypothese 2*). Zum Abschluss werden die Ergebnisse zur Prüfung auf Reihenfolgeeffekte präsentiert.

Sicherung der Datenqualität

Im Folgenden werden die Ergebnisse zur Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5) für diese Studie vorgestellt. Die drei überprüften Klassen-Lösungen für die vier Modus-Ponens-Aufgaben weisen folgende BIC-Werte auf:

1-Klassen-Lösung: BIC = 1464,12
2-Klassen-Lösung: BIC = 1258,29
3-Klassen-Lösung: BIC = 1288,34

Die Zwei-Klassen-Lösung erweist sich demnach als die bestpassende. Für die Prüfung des Modellfits der Zwei-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergeben sich folgende *p*-Werte für Chi-Quadrat und Cressie Read:

Chi-Quadrat: $p = .22$
Cressie Read: $p = .25$

Die Zwei-Klassen-Lösung der vier Modus-Ponens-Aufgaben weist also einen guten Modellfit auf. Die Treffsicherheiten der zwei resultierenden Klassen liegen bei .99 beziehungsweise .91 und sind damit vergleichsweise hoch. Für Klasse 1 (Klassengröße $P_{C=1} = .96$, Treffsicherheit $T_1 = .99$) ergeben sich folgende Lösungswahrscheinlichkeiten für die vier Modus-Ponens-Aufgaben (MP1, …, MP4[40]):

[40] zur ausführlichen Beschreibung der mit MP1, …, MP4 bezeichneten Items siehe Abschnitt 2.3.4

$P(MP1 = 1 \mid C = 1) = .96$

$P(MP2 = 1 \mid C = 1) = .99$

$P(MP3 = 1 \mid C = 1) = .97$

$P(MP4 = 1 \mid C = 1) = .99$

In Klasse 2 (Klassengröße $P_{C=2} = .04$, Treffsicherheit $T_2 = .91$) ergeben sich folgende Lösungswahrscheinlichkeiten für die vier Modus-Ponens-Aufgaben:

$P(MP1 = 1 \mid C = 2) = .39$

$P(MP2 = 1 \mid C = 2) = .32$

$P(MP3 = 1 \mid C = 2) = .33$

$P(MP4 = 1 \mid C = 2) = .33$

Innerhalb der ersten Klasse (C = 1) liegen die Lösungswahrscheinlichkeiten aller vier Modus-Ponens-Aufgaben also nahe eins, innerhalb der zweiten Klassen (C = 2) nahe .33 (*Hypothese MP-Reduktion_1*). Eine Interpretation der ersten latenten Klasse als „Modus-Ponens-Löser" und der zweiten latenten Klasse als „Durchklicker" scheint aufgrund der geschätzten klassenbedingten Lösungswahrscheinlichkeiten durchaus angemessen.

Da *Hypothese MP-Reduktion_1* beibehalten werden kann, ist die Überprüfung von *Hypothese MP-Reduktion_2* möglich. Dazu wird eine Kreuztabelle analysiert, in der die Klasse der Modus-Ponens-Reduktion mit der Angabe der ernsthaften Testbearbeitung („*ja*"/„*nein*") gekreuzt wird. Diese Kreuztabelle ist in Tabelle 9 dargestellt.

Tabelle 9: Kreuztabelle „Klasse Modus-Ponens-Reduktion gekreuzt mit ernsthafter Teilnahme"

		ernsthafte Teilnahme		
Klasse Modus-Ponens-Reduktion		ja	nein	gesamt
„Modus-Ponens-Löser"	Anzahl	661	26	687
	erwartete Anzahl	655,5	31,5	
„Durchklicker"	Anzahl	25	7	32
	erwartete Anzahl	30,5	1,5	
gesamt		686	33	719 *

Anmerkung. Der mit * gekennzeichnete reduzierte Stichprobenumfang resultiert daher, dass die Frage nach ernsthafter Testbearbeitung lediglich von 719 der 905 Personen beantwortet wurde.

Der Chi-Quadrat-Test auf Unabhängigkeit der beiden Variablen wird mit einem Chi-Quadrat-Wert von $\chi^2 = 22{,}85$ bei einem Freiheitsgrad ($df = 1$) signifikant ($p < .001$), sodass auch die *Hypothese MP-Reduktion_2* bezüglich der Modus-Ponens-Reduktion beibehalten kann.

Die Modus-Ponens-Reduktion scheint damit als Verfahren anwendbar sowie zur Sicherung der Datenqualität geeignet. Für weitere Analysen werden daher lediglich Personen berücksichtigt, deren Zuordnungswahrscheinlichkeit zur ersten Klasse (C = 1, „Modus-Ponens-Löser“) mindestens .85 beträgt. Dadurch reduziert sich die Untersuchungsstichprobe (ursprünglich $N = 905$; vgl. Abschnitt 2.4.1.4) um 38 Personen. Der Stichprobenumfang für weitere Analysen beträgt damit $N = 867$.

LCA der 16 Items des KKS

Für diese 867 Personen werden die Antworten auf die 16 Items des KKS unter Verwendung einer LCA analysiert. Als bestpassendes Modell wird dasjenige mit dem niedrigsten BIC-Wert ausgewählt. Die BIC-Werte der berechneten Klassen-Lösungen sind im Folgenden angegeben:

1-Klassen-Lösung: BIC = 14602,97 (Unabhängigkeitsmodell)
2-Klassen-Lösung: BIC = 12346,91
3-Klassen-Lösung: BIC = 11364,28
4-Klassen-Lösung: BIC = 10844,42
5-Klassen-Lösung: BIC = 10780,56
6-Klassen-Lösung: BIC = 10787,57
7-Klassen-Lösung: BIC = 10808,96

Die bestpassende Lösung für die 16 Items des KKS ist eine Fünf-Klassen-Lösung[41]. Für die Prüfung des Modellfits der Fünf-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergeben sich folgende *p*-Werte für Chi-Quadrat und Cressie Read:

Chi-Quadrat: $p = .33$
Cressie Read: $p = .13$

[41] Den höchsten BIC-Wert und damit die schlechteste Passung weist das Unabhängigkeitsmodell auf. Dabei handelt es sich um das Modell, welches impliziert, dass die Items nicht messen (Clogg, 1995). Dieses Ergebnis spricht dafür, dass die 16 Items des KKS nicht stochastisch unabhängig voneinander sind, also in jedem Falle (etwas) messen.

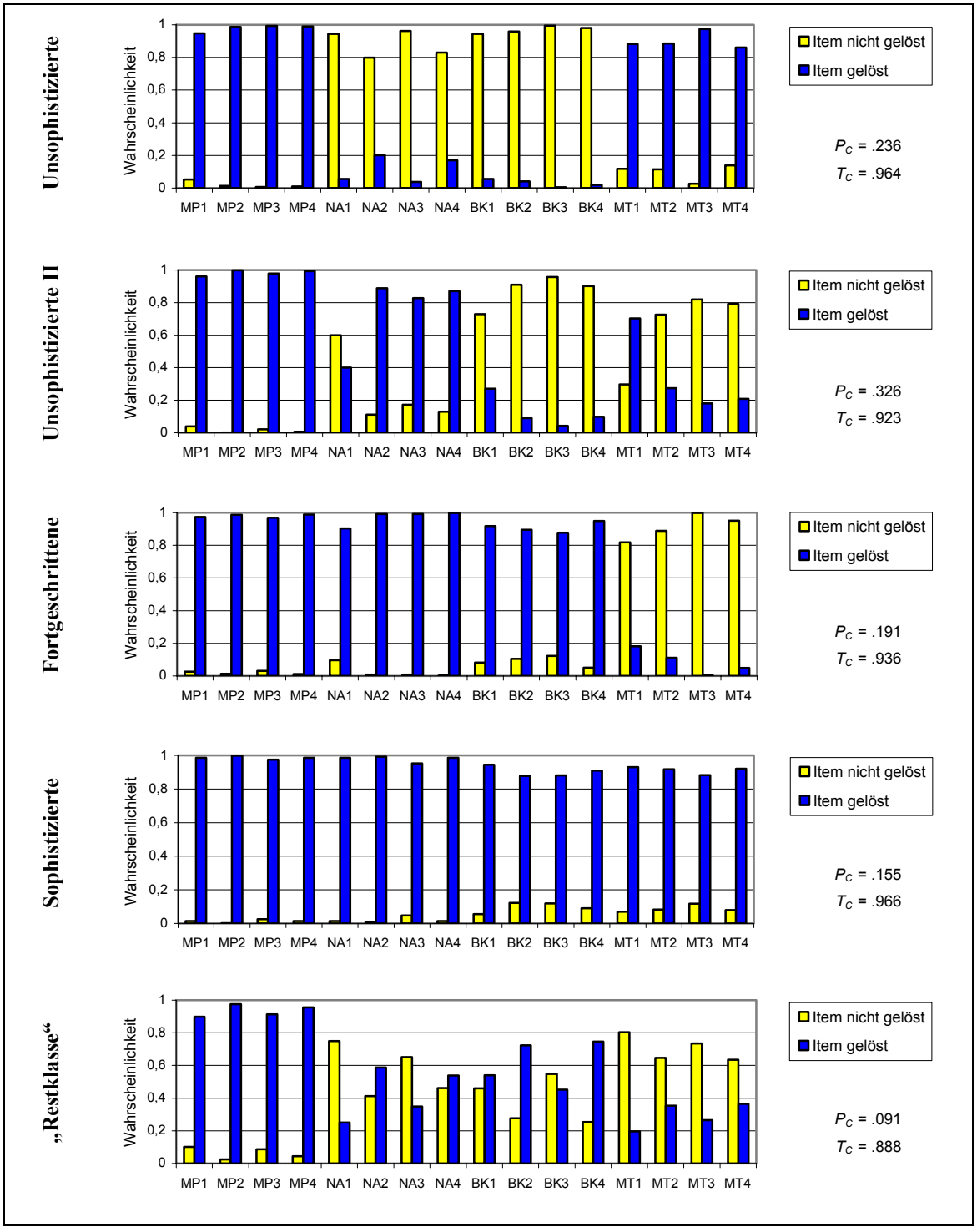
Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Fünf-Klassen-Lösung. Die Treffsicherheiten der fünf resultierenden Klassen liegen zwischen .888 und .966 sowie für die gesamte Fünf-Klassen-Lösung bei .939 und indizieren damit eine relativ hohe bis hohe Treffsicherheit, wie Tabelle 10 illustriert.

Tabelle 10: Geschätzte Klassengrößen (P_C) und Treffsicherheiten (T_C) der Fünf-Klassen-Lösung der 16 Items des KKS

Geschätzte Klassengrößen P_C	Treffsicherheit T_C
$P_{C=1} = .326$	$T_1 = .923$
$P_{C=2} = .236$	$T_2 = .964$
$P_{C=3} = .191$	$T_3 = .936$
$P_{C=4} = .155$	$T_4 = .966$
$P_{C=5} = .091$	$T_5 = .888$
Treffsicherheit der Fünf-Klassen-Lösung	$T = .939$

Anmerkung. Die Ordnung der Klassen erfolgt zunächst nach der geschätzten Klassengröße.

Sämtliche identifizierte Klassen weisen eine substanzielle Klassengröße auf und die Personen können mit hinreichend hoher Treffsicherheit den Klassen zugeordnet werden. Abbildung 3 zeigt die klassenbedingten Lösungswahrscheinlichkeiten der Fünf-Klassen-Lösung für die 16 Items des KKS. Diese indizieren für jede Klasse das „typische“ Muster der (klassenbedingten) Lösungswahrscheinlichkeiten und bilden damit die Grundlage für die Prüfung „per Augenschein“ der *Arbeitshypothesen 1a* bis *1d*. Die Anordnung der 16 Items ist dabei analog zur Anordnung bei der Vorstellung des Erhebungsinstrumentes (siehe Abschnitt 2.4.1.2).



Legende. P_C…geschätzte Klassengröße, T_C…Treffsicherheit, MP…Modus Ponens, NA…Negation des Antezedens, BK…Bestätigung der Konsequenz, MT…Modus Tollens, 1…keine Negation in der Hauptprämisse, 2…Negation in der Konsequenz, 3…Negation im Antezedens, 4…Negation in Antezedens und Konsequenz.

Abbildung 3: Klassenbedingte Lösungswahrscheinlichkeiten der Fünf-Klassen-Lösung für die 16 Items des KKS

Es zeigt sich eine Klasse (P_C = .236, T_C = .964) mit hohen Lösungswahrscheinlichkeiten für MP- sowie MT-Aufgaben bei gleichzeitig geringen Lösungswahrscheinlichkeiten für NA- und BK-Aufgaben. Dieses Muster der Lösungswahrscheinlichkeiten entspricht der postulierten Stufe „Unsophistizierte". Damit kann *Arbeitshypothese 1a* beibehalten werden. In einer zweiten Klasse (P_C = .326, T_C = .923) zeigen sich hohe Lösungswahrscheinlichkeiten für MP-Aufgaben bei gleichzeitig geringen Lösungswahrscheinlichkeiten für BK-Aufgaben. Bei NA-Aufgaben ist die Lösungswahrscheinlichkeit für Item NA1, also die NA-Aufgabe ohne jegliche Negationen in der Hauptprämisse, deutlich geringer als die Lösungswahrscheinlichkeiten der drei anderen NA-Aufgaben. Gleichzeitig ist bei MT-Aufgaben die Lösungswahrscheinlichkeit für Item MT1, also die MT-Aufgabe ohne jegliche Negationen in der Hauptprämisse, deutlich höher als die Lösungswahrscheinlichkeiten der drei anderen MT-Aufgaben. Dieses Muster der Lösungswahrscheinlichkeiten entspricht der postulierten Stufe „Unsophistizierte II" des erweiterten Stufen-Modells. *Arbeitshypothese 1b* kann folglich ebenfalls beibehalten werden. Eine dritte Klasse (P_C = .191, T_C = .936) zeichnet sich durch hohe Lösungswahrscheinlichkeiten für MP-, NA- und BK-Aufgaben bei gleichzeitig niedrigen Lösungswahrscheinlichkeiten für MT-Aufgaben aus. Dieses Muster der Lösungswahrscheinlichkeiten entspricht der postulierten Stufe „Fortgeschrittene", wodurch *Arbeitshypothese 1c* beibehalten werden kann. In einer vierten Klasse (P_C = .155, T_C = .966) zeigen sich für sämtliche Aufgaben hohe Lösungswahrscheinlichkeiten. Dieses Muster der Lösungswahrscheinlichkeiten entspricht der postulierten Stufe „Sophistizierte". Entsprechend kann auch *Arbeitshypothese 1d* beibehalten werden. Jede der vier Stufen des erweiterten Stufen-Modells wird durch eine Klasse repräsentiert, in der sich das postulierte, stufen-spezifische Muster der Lösungswahrscheinlichkeiten zeigt. Damit kann auch *Arbeitshypothese 1* beibehalten werden, die postuliert, dass alle vier Stufen des erweiterten Stufen-Modells identifiziert werden können.

Dabei steht die identifizierte „Restklasse" dem Beibehalten von *Arbeitshypothese 1* nicht entgegen. Sie kann eine methodische Notwendigkeit sein, um bspw. stereotypes Antwortverhalten in Form bestimmter Antworttendenzen zu berücksichtigen (vgl. entsprechende Ausführungen in Abschnitt 2.4.1.5). Ebenso ist es jedoch möglich, dass das Muster der Lösungswahrscheinlichkeiten in dieser Klasse eine weitere Sophistiziertheits-Stufe Konditionalen Schlussfolgerns indiziert. Diese Frage soll jedoch erst an späterer Stelle diskutiert werden (siehe Abschnitt 2.4.3.1). Festzuhalten bleibt, dass die „Restklasse" die geringste Klassengröße (P_C = .091) und die niedrigste Treffsicherheit

(T_C = .888) aufweist. Es zeigen sich in dieser Klasse hohe Lösungswahrscheinlichkeiten für MP-Aufgaben sowie niedrige Lösungswahrscheinlichkeiten für MT-Aufgaben. Für NA- und BK-Aufgaben differieren die Lösungswahrscheinlichkeiten zwischen den Hauptprämissen. Sie sind jedoch für Aufgaben mit Negation in der Konsequenz jeweils höher.

Sämtliche Hypothesen des *Arbeitshypothesenkomplexes 1* können beibehalten werden. Es zeigen sich vier Klassen, deren Muster der klassenbedingten Lösungswahrscheinlichkeiten den postulierten vier Stufen des erweiterten Stufen-Modells entspricht, sowie eine „Restklasse".

Zusammenhang zwischen SKS und Reasoning

Bevor die multinomiale logistische Regression für latente Variablen berechnet wird, sind zunächst die Messmodelle zu überprüfen. Das Messmodell für SKS entspricht der (Fünf-)Klassen-Lösung der 16 Items, durch deren gute Modellpassung das erste Messmodell bereits als überprüft gelten kann. Obwohl mit der „Restklasse" eine zusätzliche (nicht postulierte) Klasse in das Messmodell aufgenommen werden muss, ändert das nichts an der grundlegenden Aussage von *Arbeitshypothese 2*. Selbst wenn für die „Restklasse" keine Vorhersage getroffen werden kann, so sollen doch nach wie vor die vier postulierten Klassen, die die vier Sophistiziertheits-Stufen Konditionalen Schlussfolgerns repräsentieren, durch die kontinuierliche latente Variable Reasoning vorhergesagt werden können. In einer multinomialen logistischen Regression werden die Ausprägungen der nominalen Variable jeweils binär logistisch in Beziehung zur Referenzklasse (Referenzkategorie) gesetzt (vgl. Abschnitt 2.4.1.5), weshalb Unterschiede zwischen den postulierten vier Klassen unabhängig von der Restklasse zu finden sein sollten. *Arbeitshypothese 2* kann also trotzdem geprüft werden. Vorher ist jedoch noch das Messmodell für die 10 Matrizenitems zu überprüfen. Die Stichprobe dafür umfasst 742 Personen[42]. Für die Prüfung des Modellfits werden wiederum 200 parametrische Bootstrap-Stichproben generiert. Dabei ergeben sich folgende *p*-Werte für Chi-Quadrat und Cressie Read:

Chi-Quadrat: $p = .28$
Cressie Read: $p = .06$

[42] Dabei handelt es sich um diejenigen Personen der (Modus-Ponens-)reduzierten Untersuchungsstichprobe, die alle 10 Matrizenitems bearbeitet haben.

Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) ergibt kein signifikantes Ergebnis. Die Nullhypothese braucht folglich nicht verworfen zu werden und die Gültigkeit des Messmodells kann angenommen werden. Andrichs Reliabilität liegt bei $Rel = .59$ und ist damit vergleichsweise niedrig. Allerdings gelten für Modelle der IRT die bereits beschriebenen Einschränkungen bezüglich der Interpretierbarkeit von Reliabilitätsschätzungen (vgl. Abschnitt 2.4.1.5). Zudem wird die beabsichtigte multinomiale logistische Regression für latente und damit messfehlerbereinigte Variablen berechnet. Die geringe Reliabilität ist daher – zumindest für diese Analyse – nicht von Bedeutung. Die *Z*-Werte der Q-Indizes liegen zwischen -1,05 und 1,19, sodass für keinen der Diskriminationsparameter β_i der 10 Matrizenitems die Nullhypothese ($\beta_i = 1$) verworfen zu werden braucht. Dies stützt zusätzlich zu den Ergebnissen des parametrischen Bootstrapverfahrens die Annahme eines Rasch-Modells für die 10 Matrizenitems. Ein Rasch-Modell als Messmodell für die (kontinuierliche) latente Variable Reasoning kann damit als passend angenommen werden.

Für die Berechnung der multinomialen logistischen Regression[43] wird für SKS ein LCM unter Vorgabe von fünf Klassen und für Reasoning ein Rasch-Modell der 10 Matrizenitems spezifiziert. Die Stichprobe für diese Analyse umfasst dieselben Personen ($N = 742$), auf die sich auch die Überprüfung des Messmodells der 10 Matrizenitems bezieht[44]. Die Log-Likelihood dieser multinomialen logistischen Regression beträgt $LogL_1 = -7768{,}202$, die der restriktiveren (alle vier Anstiegskoeffizienten auf null fixiert) $LogL_0 = -7807{,}128$. Das entspricht einer Differenz von $LogL_0 - LogL_1 = -38{,}926$, also einem Chi-Quadrat-Wert von $\chi^2 = 77{,}852$. Bei vier Freiheitsgraden[45] ($df = 4$) ergibt sich ein *p*-Wert von $p < .001$, der deutlich kleiner als das vorher festgelegte alpha-Niveau ($\alpha = .05$) ist. Folglich wird die Nullhypothese verworfen, der zufolge *keine* lineare regressive Abhängigkeit der logit-transformierten Klassenwahrscheinlichkeiten von der kontinuierlichen latenten Variable besteht, die durch die 10 Matrizenitems gemessen wird. Es kann daher von einer (multinomialen logistischen) regressiven Abhängigkeit der Sophistiziertheit Konditionalen Schlussfolgerns von Reasoning ausge-

[43] Das M*plus*-Inputfile der multinomialen logistischen Regression findet sich im Anhang D dieser Arbeit.
[44] Da es sich dabei nur um einen Teil der Personen handelt, für die das SKS-Messmodell überprüft wurde, erfolgte zumindest ein zusätzlicher „Augenschein-Vergleich" der geschätzten Parameter in dieser Teil-Stichprobe mit denen in der Gesamt-Stichprobe. Dieser Vergleich fiel positiv aus.
[45] Durch die zusätzliche Klasse („Restklasse") werden vier Klassen in Bezug zur Referenzklasse gesetzt und nicht wie ursprünglich beschrieben drei (siehe Abschnitt 2.4.1.5). Folglich ergeben sich vier Freiheitsgrade.

gangen werden. Nagelkerkes R-Quadrat beträgt dabei $R^2_{Nagelkerke} = .10$, was für einen eher geringen, tendenziell mittleren Effekt spricht.

In der Folge werden nun die Anstiegskoeffizienten analysiert, um die Ordnung der identifizierten Klassen hinsichtlich der kontinuierlichen latenten Variable Reasoning zu überprüfen. Die Referenzklasse (Referenzkategorie) wird durch die verwendete Software M*plus* (Muthén & Muthén, 1998-2007) automatisch festgelegt. Bei der vorliegenden Analyse handelt es sich dabei um die „Restklasse". Da für die „Restklasse" keine Vorhersagen bezüglich Reasoning existieren, sind die Ergebnisse dann konform mit *Arbeitshypothese 2*, wenn die Anstiegskoeffizienten folgende Ordnung aufweisen:

$\beta_{Sophistizierte\_0} > \beta_{Fortgeschrittene\_0} > \beta_{UnsophistizierteII\_0} > \beta_{Unsophistizierte\_0}$
(0 indiziert dabei die Referenzklasse, also in diesem Fall die „Restklasse".)

Es ergeben sich folgende Schätzungen für die Anstiegskoeffizienten:

$\beta_{Sophistizierte\_0} = 1{,}002$
$\beta_{Fortgeschrittene\_0} = 0{,}723$
$\beta_{UnsophistizierteII\_0} = 0{,}142$
$\beta_{Unsophistizierte\_0} = -0{,}129$

Damit zeigt sich die postulierte Ordnung der Anstiegskoeffizienten, sodass *Arbeitshypothese 2A*, die Präzisierung der inhaltlichen *Arbeitshypothese 2* hinsichtlich einer multinomialen logistischen Regression für latente Variablen, beibehalten werden kann. Auf eine zusätzliche Analyse, inwieweit die Anstiegskoeffizienten signifikant von null verschieden sind, wird verzichtet, da sich keinerlei Hypothese auf den Unterschied zwischen der „Restklasse" und den anderen Klassen bezieht.

Für die Rangkorrelation der manifesten Variablen SKS und Reasoning wird zunächst jedem Probanden ein Testwert zugeordnet. Für Reasoning ist das der Summenwert der 10 Matrizenitems, für SKS die Zuordnung zur wahrscheinlichsten Klasse. Dabei werden jedoch lediglich die Personen der Stichprobe berücksichtigt, die den vier, durch das erweiterte Stufen-Modell postulierten Klassen zugeordnet werden, da sich lediglich diese Klassen theoriegeleitet ordnen lassen (1…Unsophistizierte, 2…Unsophistizierte II, 3…Fortgeschrittene, 4…Sophistizierte; siehe auch Abschnitt 2.4.1.5). Personen, die der „Restklasse" zugeordnet werden (ca. 9%), werden von der Analyse ausgeschlossen. Der Stichprobenumfang beträgt daher für diese Analyse $N = 679$. Das Ergebnis der Rangkorrelation ist in Tabelle 11 dargestellt.

Tabelle 11: Rangkorrelation der manifesten Variablen Sophistiziertheit Konditionalen Schlussfolgerns (SKS_{man}) und Reasoning ($Reas_{man}$)

Variablen	Spearmans Rho (p-Wert)	Kendalls Tau (p-Wert)
SKS_{man} und $Reas_{man}$	.326 ($p < .001$)	.267 ($p < .001$)

Es ergeben sich signifikante Rangkorrelationskoeffizienten. Die Stärke des Effektes liegt dabei jeweils im mittleren Bereich. Folglich kann auch *Arbeitshypothese 2B*, die Präzisierung der inhaltlichen *Arbeitshypothese 2* hinsichtlich einer Rangkorrelation für manifeste Variablen, beibehalten werden. Da diese Analyse für manifeste Variablen berechnet wird, ist die vergleichsweise geringe Reliabilität (*Rel* = .59) des Summenwertes der 10 Matrizenitems durchaus von Bedeutung. Der „wahre" Zusammenhang der beiden manifesten Variablen kann demnach deutlich größer sein als es die Rangkorrelationskoeffizienten in Tabelle 11 indizieren.

Die Ergebnisse beider Verfahren zur Überprüfung der inhaltlichen *Arbeitshypothese 2* sprechen dafür, dass die Sophistiziertheits-Stufen Konditionalen Schlussfolgerns mit Reasoning dahingehend zusammenhängen, dass höhere Sophistiziertheits-Stufen Konditionalen Schlussfolgerns mit höheren Reasoning-Ausprägungen einhergehen. Sämtliche Ergebnisse sind signifikant. Die Stärke der Effekte liegt im geringen bis mittleren Bereich (multinomiale logistische Regression) bzw. mittleren Bereich (Rangkorrelation).

Überprüfung von Reihenfolgeeffekten

Weder der Chi-Quadrat-Test zur Prüfung eines Reihenfolgeeffektes bezüglich des KKS ($\chi^2 = 2{,}70$, $df = 4$, $p = .61$) noch der zweiseitige T-Test zur Prüfung eines Reihenfolgeeffektes bezüglich der 10 Matrizenitems ($t = 0{,}80$, $p = .42$) liefern signifikante Ergebnisse. Reihenfolgeeffekte scheinen daher unwahrscheinlich – zumindest bezogen auf die Reihenfolge von KKS und Matrizenaufgaben im Online-Test.

2.4.3 Diskussion der ersten empirischen Erprobung und Ausblick

Zum Abschluss der ersten empirischen Erprobung werden die Ergebnisse der Überprüfung der sechs Arbeitshypothesen vor dem theoretischen Hintergrund diskutiert. Der Fokus liegt dabei jedoch (noch) nicht auf einer ausführlichen theoretischen Diskussion. Hierzu ist die Betrachtung lediglich einer empirischen Studie aus Sicht des Autors zu wenig. Eine umfangreiche Diskussion erfolgt daher erst am Ende der vorliegenden Arbeit (siehe Kapitel 4) nach Ausführungen zu weiteren empirischen Studien. Ziel des folgenden Abschnitts ist die kritische Würdigung der Ergebnisse dieser ersten empirischen Erprobung und der verwendeten Methoden sowie das Ableiten von Implikationen für die Weiterentwicklung des KKS, seine weitere empirische Erprobung und insbesondere die Bestimmung seiner Testgütekriterien.

2.4.3.1 Diskussion der empirischen Überprüfung der Arbeitshypothesen

Nacheinander sollen nun die Ergebnisse der empirischen Überprüfung der beiden Arbeitshypothesenkomplexe diskutiert werden. Zunächst werden die Ergebnisse zum erweiterten Stufen-Modell (*Arbeitshypothesenkomplex 1*) betrachtet, anschließend die Ergebnisse zum Zusammenhang von SKS und Reasoning (*Arbeitshypothese 2*).

Empirische Überprüfung des erweiterten Stufen-Modells

Zunächst kann festgehalten werden, dass ca. 91% der untersuchten Personen aufgrund ihres Antwortverhaltens den vier Klassen zugeordnet werden können, die die vier Stufen des erweiterten Stufen-Modells repräsentieren. Es können also ca. 91% der Personen klassifiziert werden. Betrachtet man das sensu Clogg (1995) als Gütekriterium der durchgeführten LCA, so ist dies ein vielversprechendes Ergebnis, das dem Vergleich mit ähnlichen Untersuchungen durchaus standhält. So liegt bspw. bei Evans et al. (2007) der Anteil der auf Basis der theoretischen Konzeption klassifizierbaren Personen bei 86% und damit etwas niedriger als bei der ersten empirischen Erprobung des KKS. Des Weiteren kann festgehalten werden, dass sich die drei Klassen, die die drei Stufen des ursprünglichen Stufen-Modells repräsentieren (Unsophistizierte, Fortgeschrittene und Sophistizierte), sehr deutlich zeigen (vgl. Abbildung 3 in Abschnitt 2.4.2). Die durch das Stufen-Modell postulierten klassenbedingten Lösungswahrscheinlichkeiten zeigen sich bei jeder einzelnen der 16 Aufgaben mit einer Wahrscheinlichkeit größergleich .80. Angesichts der Vorhersage der Lösungswahrscheinlichkeiten für drei Klas-

sen bei jeweils 16 Items kann dies als deutlich hypothesenkonformes Ergebnis betrachtet werden. Empirische Evidenz für das ursprüngliche Stufen-Modell von Rijmen und De Boeck (2003) liegt damit vor. Gleichzeitig wird dadurch auch die Annahme gestützt, dass ein Großteil von Personen das in den inhaltlichen Arbeitshypothesen postulierte, stufen-spezifische Antwortverhalten unabhängig von der Verwendung von Negationen in der Hauptprämisse zeigt. Dies widerspricht der allgemeinen Auffassung der Kognitionspsychologie, dass es sich bei Effekten von Negationen wie bspw. dem Negative Conclusion Bias um allgemeinpsychologische Phänomene handelt. Das (differenzialpsychologische) erweiterte Stufen-Modell postuliert für die Effekte von Negationen eine zusätzliche Stufe im Konditionalen Schlussfolgern: Unsophistizierte II. Auch diese Stufe kann in der durchgeführten Studie identifiziert werden. Personen auf dieser Stufe (bzw. in der entsprechenden latenten Klasse) weisen bei Items ohne jegliche Negation in der Hauptprämisse bei Negation des Antezedens eine deutlich geringere und beim Modus Tollens eine deutlich höhere Lösungswahrscheinlichkeit auf als bei Items mit Negationen in der Hauptprämisse zu diesen Schlussfiguren (vgl. Abbildung 3 in Abschnitt 2.4.2). Dies ist konform zu dem in der inhaltlichen *Arbeitshypothese 1b* postulierten Antwortverhalten. Allerdings ist anzumerken, dass sich das (theoretisch vermutete) geringere Ausmaß dieses Effektes bei Item NA3 bzw. MT2 nicht zeigt. Bei diesen Items resultiert die als Ursache vermutete Schwierigkeit beim Umgang mit einer doppelten Negation lediglich aus deren Enkodierung, nicht aus deren aktivem Auflösen (vgl. Abschnitt 2.3.2.1). Offenbar scheint dies jedoch keinen Unterschied zu machen. Insgesamt kann dennoch festgehalten werden, dass sich auch die durch das postulierte Antwortverhalten der „Unsophistizierten-Stufe II“ indizierten klassenbedingten Lösungswahrscheinlichkeiten in der empirischen Erprobung zeigen[46]. Es findet sich also empirische Evidenz für alle vier postulierten Stufen des erweiterten Stufen-Modells und damit für *Arbeitshypothese 1*. Genauerer Betrachtung bedarf allerdings die bislang als „Restklasse“ bezeichnete zusätzliche Klasse, die sich in der bestpassenden Klassen-Lösung (Fünf-Klassen-Lösung; siehe Abschnitt 2.4.2) der 16 Items des KKS zeigt.

Zunächst wird davon ausgegangen, dass es sich bei der fünften Klasse tatsächlich um eine Restklasse handelt, die notwendig ist, um Personen zu berücksichtigen, die aus verschiedenen konstruktirrelevanten Gründen (siehe dazu Abschnitt 2.4.1.5) keiner

[46] Die ansonsten für diese Stufe postulierten hohen Lösewahrscheinlichkeiten für Modus-Ponens-Aufgaben und niedrigen Lösewahrscheinlichkeiten für Bestätigung-der-Konsequenz-Aufgaben zeigen sich ebenfalls konform zu *Arbeitshypothese 1b* (vgl. Abbildung 3 in Abschnitt 2.4.2).

der postulierten Klassen zugeordnet werden können. Dabei ist festzuhalten, dass es sich um die Klasse mit der geringsten Klassengröße (P_C = .091) und der geringsten Treffsicherheit (T_C = .888) handelt. Dennoch soll zumindest versucht werden, diese Klasse auch inhaltlich zu interpretieren. Während Modus-Ponens-Aufgaben in dieser Klasse insgesamt eher gelöst und Modus-Tollens-Aufgaben insgesamt eher nicht gelöst werden (vgl. Abbildung 3 in Abschnitt 2.4.2), zeigt sich für NA und BK gleichermaßen folgendes Phänomen: Bei den beiden Hauptprämissen mit Negation in der Konsequenz zeigen sich jeweils höhere Lösungswahrscheinlichkeiten als bei den beiden Hauptprämissen ohne Negation in der Konsequenz. Dies könnte ein konstruktrelevantes Phänomen sein. Allerdings scheint dies unwahrscheinlich, da es bei den Aufgaben zum Modus Tollens nicht auftritt. Eine mögliche Erklärung ist, dass dieses Phänomen ein Resultat der Vorgabe-Reihenfolge der Items ist. Da aufgrund der systematischen Itemvorgabe des KKS die Hauptprämisse für jeweils vier Items (eines zu jeder Schlussfigur) gleich bleibt (vgl. Abschnitt 2.4.1.2), werden also auch − gegeben der beiden Hauptprämissen mit einer Negation in der Konsequenz − die vier Schlussfiguren nacheinander vorgegeben. Ist durch die Hauptprämisse einmal ein Schema aktiviert (z.B. „Bei dieser Hauptprämisse ist keine Aussage über das Funktionieren der Schaltungen möglich.“), könnte dieses Schema erst bei Vorgabe einer neuen Hauptprämisse wieder „überschrieben“ werden. Das würde erklären, warum sich für beide Schlussfiguren (NA und BK) das gleiche Phänomen zeigt. Für weitere Anwendungen des KKS empfiehlt sich daher eine unsystematische Vorgabe der Items. Tritt das beschriebene Phänomen auch dann noch auf, wäre eine konstruktrelevante Ursache wahrscheinlich und weitere theoretische wie empirische Analysen würden notwendig.

Zusammenfassend kann festgehalten werden, dass sich empirische Evidenz für das erweiterte Stufen-Modell finden lässt. Dieser Befund ist in weiteren empirischen Studien zu überprüfen. Ferner ist in weiteren Studien zu überprüfen, ob es sich bei der „Restklasse“ um einen stabilen empirischen Befund handelt. Hierzu soll der KKS dahingehend modifiziert werden, dass die Items in unsystematischer Reihenfolge vorgegeben werden.

Empirische Überprüfung des Zusammenhangs von SKS und Reasoning

Der postulierte positive Zusammenhang zwischen SKS und Reasoning zeigt sich sowohl in Form einer (multinomialen logistischen) regressiven Abhängigkeit als auch bei Betrachtung einer Rangkorrelation. Die im mittleren Bereich vermuteten Effekte zeigen

sich dabei im geringen, tendenziell mittleren Bereich. Die theoretisch angenommene Ordnung der vier Stufen des erweiterten Stufen-Modells wird dadurch empirisch gestützt, dass die Anstiegskoeffizienten der multinomialen logistischen Regression die postulierte Ordnung aufweisen. Doch wenngleich sich der postulierte positive Zusammenhang über verschiedene Analysemethoden hinweg zeigt, so sind dennoch zwei Punkte kritisch anzumerken. Erstens wird bei der multinomialen logistischen Regression die „Restklasse" in die Analyse einbezogen. Dies ist bei der Betrachtung der Signifikanz des Ergebnisses unerheblich (vgl. Ausführungen zur Nullhypothese der multinomialen logistischen Regression in Abschnitt 2.4.1.5). Für das angegebene Effektstärkemaß (Nagelkerkes R-Quadrat) kann dies jedoch durchaus von Bedeutung sein, da unklar ist, inwieweit systematische Varianz an der „Restklasse" aufgeklärt wird. Bei der Rangkorrelation wird die „Restklasse" nicht berücksichtigt. Dadurch erklärt sich möglicherweise, dass die Effektstärken bei der Rangkorrelation tendenziell etwas höher ausfallen. Allerdings könnte dies auch darin begründet liegen, dass Effektstärkemaße der Pseudo-R-Quadrat-Familie (wie in diesem Falle Nagelkerkes R-Quadrat) generell eher niedrig ausfallen (Hosmer & Lemeshow, 2000). Ein zweiter kritischer Punkt betrifft die Operationalisierung von Reasoning. Dass Matrizenitems den dafür besten Marker darstellen, gilt als weitgehend unumstritten (Carpenter et al., 1990; Carroll, 1993). Dass in der vorliegenden Studie bislang noch nicht erprobte Items verwendet werden, ist jedoch durchaus kritisch zu bewerten. Zwar liegt aufgrund der regelgeleiteten Konstruktion (vgl. Ihme, 2007) nahe, dass die verwendeten 10 Matrizenitems das beabsichtigte Konstrukt (Reasoning) erfassen, dennoch handelt es sich nicht um einen etablierten Matrizentest wie bspw. bei den *Standard Progressive Matrices* (*SPM*; Raven, Raven & Court, 2004) oder beim *Wiener Matrizen-Test* (*WMT*; Formann, 1979). Um den gefundenen Zusammenhang zu replizieren, sollte für künftige Studien daher ein etablierter Matrizentest verwendet werden.

2.4.3.2 Kritische Würdigung der verwendeten Methoden

Der zweite Abschnitt der Diskussion dieser ersten empirischen Erprobung widmet sich ihren Methoden. Sie sollen in gleicher Reihenfolge betrachtet werden wie bei ihrer Vorstellung (siehe Abschnitt 2.4.1). Dabei werden jeweils Implikationen für weitere empirische Studien abgeleitet.

Allgemeines Untersuchungsdesign

Die Umsetzung eines quantitativen Untersuchungsdesigns scheint die einzige Möglichkeit, die theoretisch intendierten Analysen umzusetzen, eine Online-Erhebung die beste Möglichkeit, den notwendigen Stichprobenumfang ökonomisch realisieren zu können. Adäquate Methoden zur Sicherung der Datenqualität vorausgesetzt, sollte dieses Untersuchungsdesign daher für weitere empirische Erprobungen des KKS beibehalten werden. Ist dies in künftigen Studien der Fall, wird dies im weiteren Verlauf der Arbeit nicht mehr gesondert berichtet.

Erhebungsinstrumente

Die erste empirische Erprobung des KKS als Erhebungsinstrument für Sophistiziertheit Konditionalen Schlussfolgerns kann insgesamt als recht erfolgreich bewertet werden. So lassen sich die vier postulierten Stufen des erweiterten Stufen-Modells mit dem KKS identifizieren. Allerdings resultieren aus theoretischen Überlegungen wie auch aus den empirischen Ergebnissen einige Veränderungen bzw. Weiterentwicklungen, die bei weiteren Anwendungen des KKS berücksichtigt werden sollen:

1. Nicht-sequentielle Vorgabe der Items
   Möglicherweise ist die „Restklasse“ ein Resultat der bislang systematischen Vorgabe der Items (vgl. Abschnitt 2.4.3.1). Die Reihenfolge der Items wird daher entsprechend folgenden Richtlinien verändert:
   a. Es sollen keine zwei Items zu einer Schlussfigur aufeinanderfolgen.
   b. Es sollen keine zwei Items mit der gleichen Hauptprämisse aufeinanderfolgen.

   Ansonsten kann die Reihenfolge der Items in der überarbeiteten Version des KKS zufällig gewählt werden (für einen Überblick über die neue Reihenfolge der Items siehe Anhang A.2).
2. Einführung einer zusätzlichen vierten Antwortoption: „*Ich weiß es nicht.*“
   Dadurch soll verdeutlicht werden, dass „*Es ist keine eindeutige Aussage über das Funktionieren der Schaltung ableitbar.*“ eine genuine Antwortoption darstellt (siehe Abschnitt 2.4.1.2 für eine Erläuterung sowie die bisherigen Antwortenalternativen).
3. Geringfügige Modifikationen des Layouts

Diese Änderungen stellen die (erste) *Testrevision* des KKS dar. Die auf diese Weise revidierte Form wird in der Folge als *überarbeitete Version des KKS* bezeichnet. Eine

zweite Implikation bezüglich der verwendeten Erhebungsinstrumente ergibt sich für die Prüfung weiterer Zusammenhangshypothesen von SKS und Reasoning. Letzteres soll künftig durch einen etablierten Matrizentest operationalisiert werden (vgl. Abschnitt 2.4.3.1).

Durchführung

Wie bereits bei der Diskussion des allgemeinen Untersuchungsdesigns ausgeführt, scheint die Durchführung als Online-Test die beste Möglichkeit, ausreichend große Stichproben ökonomisch zu rekrutieren und zu erfassen. Dem Problem der Datenqualitätssicherung in Online-Erhebungen wird durch die Technik der Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5) begegnet. Effekte der Reihenfolge (KKS – Matrizentest vs. Matrizentest – KKS) zeigen sich nicht. Der Einsatz eines Online-Tests in der vorliegenden Form scheint also auch für weitere empirische Erprobungen des KKS adäquat. Nach Auswertung der Kommentare (standardmäßig vorgegebenes offenes Eingabefeld) einiger Probanden wird allerdings eine Reduktion der Gesamtbearbeitungszeit bei künftigen Untersuchungen angestrebt.

Untersuchungsstichprobe

Zunächst kann festgehalten werden, dass es sich bei den betrachteten 905 Personen um eine recht umfangreiche Untersuchungsstichprobe handelt. Evans et al. (2007) bspw. sprechen bei allgemeinpsychologischen Experimenten zum Konditionalen Schlussfolgern bereits ab ca. 100 Probanden von einer großen Stichprobe. Auffällig ist das vergleichsweise hohe Bildungsniveau der Untersuchungsstichprobe (vgl. Abschnitt 2.4.1.4). Dies ist zwar angesichts des Rekrutierungssystems (Rekrutierung durch Studierende in einer Art Schneeballsystem; siehe Abschnitt 2.4.1.4) nicht überraschend, stellt aber dennoch eine Besonderheit dieser Untersuchungsstichprobe dar. So kann das hohe Bildungsniveau bspw. eine Ursache dafür sein, dass ein so hoher Prozentsatz von Personen (15,5%) der Sophistizierten-Stufe und damit der höchsten Sophistiziertheits-Stufe Konditionalen Schlussfolgerns zugeordnet wird. Schließlich sprechen die angestellten theoretischen Überlegungen (siehe Abschnitt 2.2.3) zu den bei Kleinbeck (2005) berichteten Daten für ca. 3% Sophistizierte und Rijmen und De Boeck (2003) können gar überhaupt keine Personen auf dieser Stufe identifizieren[47]. Da es sich bei der Zielgruppe

[47] Allerdings werden hierfür noch andere Gründe als die Zusammensetzung der Untersuchungsstichprobe vermutet (vgl. Abschnitt 2.3.1).

Chipdesigner, für die der KKS (zunächst) entwickelt wird, allerdings auch um Personen mit vergleichsweise hohem Bildungsniveau handelt (ausnahmslos Abitur, größtenteils Hochschulabschluss), ist diese Besonderheit der Untersuchungsstichprobe unproblematisch und wahrscheinlich sogar förderlich für deren Repräsentativität. Einschränkend bezüglich der Generalisierbarkeit bleibt jedoch, dass der überwiegende Teil der Probanden aus Thüringen stammt.

Auswertungsmethoden

In der Folge sollen die Analyseverfahren der ersten empirischen Erprobung des KKS kritisch gewürdigt werden. Doch bevor sich den eigentlichen Analyseverfahren zugewendet wird, soll die zur Sicherung der Datenqualität entwickelte Modus-Ponens-Reduktion kurz betrachtet werden.

Das aus den theoretischen Überlegungen resultierende Zwei-(latente-)Klassen-Modell zeichnet sich durch eine gute Modellpassung, hohe Treffsicherheiten und gut interpretierbare (klassenbedingte) Lösungswahrscheinlichkeiten aus. Die Interpretation der beiden Klassen als „Modus-Ponens-Löser“ und „Durchklicker“ wird jedoch nicht nur durch die klassenbedingten Lösungswahrscheinlichkeiten, sondern auch durch die hypothesenkonforme Beziehung zum Außenkriterium „Frage nach ernsthafter Testbearbeitung“ gestützt. Die Modus-Ponens-Reduktion scheint damit in der vorliegenden Online-Erhebung eine angemessene Methode zur Sicherung der Datenqualität zu sein. Da die Sicherung der Datenqualität ansonsten in Online-Erhebungen nur schwer gewährleistet werden kann, soll die Modus-Ponens-Reduktion auch in sämtlichen weiteren empirischen Erprobungen des KKS eingesetzt werden.

Deutlich ausführlicher soll die LCA als Analysemethode für die 16 Items des KKS diskutiert werden. Zunächst stellt ein LCM eine Präzisierung des erweiterten Stufen-Modells in der Sprache der Wahrscheinlichkeitstheorie dar (vgl. Abschnitt 2.4.1.5). Allein das rechtfertigt die LCA als Analysemethode. Darüberhinaus sprechen auch die empirischen Ergebnisse für eine gute Anwendbarkeit dieses Verfahrens zur Beantwortung der relevanten Fragestellungen. Das zeigt sich zunächst in der guten Modellpassung (hohe *p*-Werte der beiden Prüfgrößen bei Anwendung eines parametrischen Bootstrapverfahrens). Dies ist keineswegs selbstverständlich. So argumentieren bspw. Bonnefon et al. (2008), dass für den Fall großer Stichproben das alpha-Niveau für derartige Modelltests durchaus auch auf $\alpha = .01$ festgelegt werden kann. Schließlich sei für die Modellierung latenter Variablen bekannt, dass bei großen Stichproben die hohe

Testpower dazu führen kann, dass auch gute Modelle zurückgewiesen werden (Bonnefon et al., 2008). Die Untersuchungsstichprobe der ersten empirischen Erprobung des KKS ($N = 867$) ist sogar noch umfangreicher als die von Bonnefon et al. (2008) und dennoch zeigen sich für die beiden Prüfgrößen (Chi-Quadrat, Cressie Read) *p*-Werte, die deutlich über dem für diese Untersuchung festgesetzten alpha-Niveau von $\alpha = .05$ liegen (vgl. Abschnitt 2.4.2). Auch für die weiteren empirischen Erprobungen des KKS soll diese Festlegung des alpha-Niveaus ($\alpha = .05$) beibehalten werden, da bei dieser Art der Prüfung der Modellgüte die Nullhypothese Wunschhypothese ist. Aus diesem Grund scheint dem Autor der vorliegenden Arbeit eine Festlegung des alpha-Niveaus auf $\alpha = .01$ wie bei Bonnefon et al. (2008) unangebracht. Das parametrische Bootstrapverfahren zur Prüfung der Modellgüte bietet aber über diese Bewährung hinaus noch einen weiteren Vorteil. Wenngleich die Untersuchungsstichprobe der vorliegenden Studie vergleichsweise umfangreich ist, wäre für eine LCA sogar eine noch größere Stichprobe ideal (Rost, 2004). Mithilfe des parametrischen Bootstrapverfahrens ist nun bereits für deutlich geringere Stichprobenumfänge (als die optimalen) eine Prüfung der Modellgüte möglich.

Über die gute Modellpassung hinaus zeichnet sich das berechnete Fünf-Klassen-Modell durch hohe Treffsicherheiten aus. Die getesteten Personen können also im Mittel mit einer hohen Wahrscheinlichkeit (.939) den entsprechenden Klassen zugeordnet werden. Da das Muster der klassenbedingten Lösungswahrscheinlichkeiten bereits per Augenschein den vier, durch das erweiterte Stufen-Modell postulierten Klassen zugeordnet werden kann, ist letztlich auch die Interpretierbarkeit der Klassen und damit ein weiteres, laut Clogg (1995) entscheidendes Gütekriterium einer LCA gegeben. Die LCA erweist sich also als adäquates Analyseverfahren für die 16 Items des KKS und bietet sich damit für weitere empirische Erprobungen des KKS an. Auch das spezifische Vorgehen bei der LCA in der vorliegenden Untersuchung soll in der folgenden Form beibehalten werden:

1. Bestimmung der bestpassenden Klassen-Lösung durch Vergleich der BIC-Werte
2. Bestimmung der Modellpassung dieser bestpassenden Klassen-Lösung mittels parametrischen Bootstrapverfahrens
3. Analyse der Treffsicherheiten des LCMs
4. Prüfung der Interpretierbarkeit der klassenbedingten Lösungswahrscheinlichkeiten

Abschließend sollen die Methoden zur Überprüfung der Zusammenhangshypothese von SKS und Reasoning kritisch betrachtet werden. Um der zweiten Forderung an die vorliegende Arbeit (empirische Hypothesenprüfung unter Verwendung von Latente-Variablen-Modellen) zu entsprechen, wird zur Überprüfung dieser Zusammenhangshypothese eine multinomiale logistische Regression für latente Variablen berechnet. Beide latenten Variablen lassen sich adäquat modellieren (Fünf-Klassen-Modell für die 16 Items des KKS, dichotomes Rasch-Modell für die 10 Matrizenitems). Die multinomiale logistische Regression lässt sich dann hinsichtlich Signifikanz, Effektstärke und Ordnung der Anstiegskoeffizienten interpretieren. Es sind also drei verschiedene Informationen zur Überprüfung von *Arbeitshypothese 2* (bzw. *Arbeitshypothese 2a*, der Präzisierung hinsichtlich einer multinomialen logistischen Regression) gegeben. Die multinomiale logistische Regression für latente Variablen scheint damit zur Analyse derartiger Zusammenhangshypothesen (latenter Variablen) gut geeignet und soll daher ebenfalls bei der weiteren empirischen Erprobung des KKS eingesetzt werden. Einschränkungen ergeben sich lediglich aus der automatischen Wahl der Referenzkategorie durch die verwendete Software M*plus*. Im vorliegenden Fall wird durch das Programm die „Restklasse“ als Referenzkategorie festgelegt. Sollte in künftigen Analysen eine andere Klasse als Referenzkategorie festgelegt werden, kann zusätzlich überprüft werden, welche der Anstiegskoeffizienten signifikant von null verschieden sind. Eine solche Klasse unterscheidet sich dann signifikant von der Referenzkategorie hinsichtlich des jeweiligen Regressors.

Um die Ergebnisse mit denen anderer Untersuchungen vergleichen zu können, wird zusätzlich eine Rangkorrelation betrachtet. Die Klassenzuordnungen indizieren dabei (manifest) SKS, der Summenwert der 10 Matrizenitems Reasoning. Beides lässt sich theoretisch begründen. Die Ordnung der mit dem KKS identifizierten Klassen ergibt sich schlüssig aus der Theorie zum erweiterten Stufen-Modell. Die hohen Treffsicherheiten der Klassen legitimieren dabei zusätzlich die Bildung dieser manifesten Variable. Bezüglich der 10 Matrizenitems sprechen das Ergebnis des parametrischen Bootstrapverfahrens wie auch die Analyse der Q-Indizes für ein Rasch-Modell zur Modellierung der (kontinuierlichen) latenten Variable Reasoning. Der Summenwert kann daher als suffiziente Statistik angenommen werden (Steyer & Eid, 2001). Demnach kann auch die Rangkorrelation zwischen der manifesten SKS-Variable und dem Summenscore der 10 Matrizenitems als adäquates Verfahren zur Prüfung dieser Zusammenhangshypothese angesehen werden. Die resultierenden (Rang-)Korrelations-

koeffizienten lassen sich hinsichtlich ihrer Höhe leicht mit den Ergebnissen anderer Untersuchungen zu ähnlichen Fragestellungen vergleichen, da dort meist auch Korrelationskoeffizienten berichtet werden.

Sämtliche verwendete Analysemethoden scheinen also für die Überprüfung der aufgestellten Arbeitshypothesen angemessen. Sie sollen daher auch bei der weiteren empirischen Erprobung des KKS (bei vergleichbaren Fragestellungen) angewendet werden.

2.4.3.3 Implikationen für die weitere empirische Erprobung des KKS

Nachdem die kritische Würdigung der ersten empirischen Erprobung des KKS sowohl hinsichtlich der Ergebnisse als auch hinsichtlich der verwendeten Methoden bislang recht positiv ausfällt, stellt sich nun die Frage, welche weiteren empirischen Erprobungen des KKS aus wissenschaftlicher Sicht indiziert sind. In erster Linie zählt dazu die Replizierbarkeit der Ergebnisse – sowohl zum erweiterten Stufen-Modell als auch zum Zusammenhang von SKS und Reasoning. Zusätzlich werden weitere relevante Forschungsfragen und insbesondere die Bestimmung der Testgütekriterien des KKS thematisiert.

Replizierbarkeit der Ergebnisse

Aufgrund der hypothesenkonformen Ergebnisse bei der Überprüfung des erweiterten Stufen-Modells, also des Messmodells des KKS, werden im Rahmen der Testrevision lediglich geringfügige Modifikationen am KKS vorgenommen (unsystematische Itemreihenfolge, vierte Antwortoption, geringfügige Layoutänderungen; vgl. Abschnitt 2.4.3.1). In weiteren empirischen Studien ist zu überprüfen, ob auch dort die vier Stufen des erweiterten Stufen-Modells identifiziert werden können. Ebenso ist zu prüfen, ob wiederum die Spezifikation einer zusätzlichen „Restklasse“ notwendig ist und ggf. ob sich deren klassenbedingte Lösungswahrscheinlichkeiten von denen in dieser Untersuchung unterscheiden. Erweist sich das Messmodell des KKS über verschiedene Untersuchungen hinweg als passend, ist der Zusammenhang mit einem etablierten Matrizentests zu überprüfen (vgl. Abschnitt 2.4.3.1).

Weitere Forschungsfragen

Um den KKS perspektivisch in der Psychodiagnostik einsetzen zu können, stellt sich nicht nur die Frage nach seinen Testgütekriterien, sondern auch die Frage nach der

Stabilität des damit gemessenen Konstruktes Sophistiziertheit Konditionalen Schlussfolgerns. Diese Frage wird zunächst natürlich durch die theoretische Konzeption des betrachteten Konstruktes beantwortet, sollte aber zusätzlich empirisch überprüft werden. Dafür bietet sich eine Längsschnittuntersuchung mit dem KKS an.

Bestimmung von Testgütekriterien

Der wesentliche Unterschied zwischen einem unwissenschaftlichen Test (bspw. einer einfachen Fragen- oder Aufgabensammlung) und einem wissenschaftlich fundierten, psychologischen Test besteht darin, dass letzterer hinsichtlich der Erfüllung seiner Testgütekriterien empirisch überprüft wird (Moosbrugger & Kelava, 2007). Nachdem die bisherige empirische Prüfung des KKS insgesamt positiv ausfällt, stellt die Bestimmung seiner Testgütekriterien die anstehende Herausforderung dar, um ihn als psychodiagnostisches Testverfahren zu etablieren. Die betrachteten Testgütekriterien werden an späterer Stelle (siehe dazu Kapitel 3) ausführlich beschrieben. Hier sei lediglich erwähnt, dass Beckmann und Guthke (1999) für die Mehrzahl der kognitionspsychologisch fundierten Tests zweierlei beklagen: Erstens fehle meist eine umfassende Konstruktvalidierung, zweitens fehle häufig die Bewährung bei der Vorhersage von Außenkriterien. Darauf Bezug nehmend wird auf beides im Rahmen der vorliegenden Arbeit besonderer Wert gelegt. Speziell zur Konstruktvalidierung bieten sich neben Reasoning weitere Konstrukte an, die theoriegeleitet einen Zusammenhang mit SKS aufweisen sollten. Erwähnt sei an dieser Stelle zunächst lediglich die Arbeitsgedächtniskapazität, der in beiden betrachteten kognitionspsychologischen Theorien zum Konditionalen Schlussfolgern eine entscheidende Rolle zugeschrieben wird (vgl. Abschnitte 2.1.2.1 und 2.1.2.2). Damit kann abschließend festgehalten werden, dass der Fokus der weiteren empirischen Erprobung des KKS auf der Replikation der bisherigen Ergebnisse, Betrachtungen zur Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns sowie der Bestimmung seiner Testgütekriterien liegt. Doch bevor sich diesen Fragen zugewandt wird, soll eine kurze Zusammenfassung der bisherigen Testentwicklungsarbeiten gegeben werden.

2.5 Zusammenfassung der bisherigen Testentwicklungsarbeiten

An die beschriebenen (bisherigen) Testentwicklungsarbeiten wurden zwei zentrale Forderungen gestellt: Zum einen sollte die Testkonstruktion kognitionspsychologisch fundiert sein, zum anderen unter Verwendung von Latente-Variablen-Modellen empirisch überprüft werden. Beide Forderungen können als umgesetzt angesehen werden.

Nach der Betrachtung Konditionalen Schlussfolgerns aus verschiedenen Perspektiven (Aussagenlogik, Kognitionspsychologie, Differenzielle Psychologie, Intelligenzstrukturforschung) wird entschieden, dass die Logiktheorie zum Konditionalen Schlussfolgern die kognitionspsychologische Grundlage der Testkonstruktion bilden soll. Es wird ein darauf aufbauendes Stufen-Modell vorgestellt, das differenzialpsychologische Aussagen zum Konditionalen Schlussfolgern erlaubt. Dieses Stufen-Modell wird auf Konditionalaussagen mit Negationen in der Hauptprämisse erweitert, woraus 16 Items (das sog. Negationsparadigma) resultieren, für die konkrete Arbeitshypothesen abgeleitet werden. Die Arbeitshypothesen beziehen sich jedoch nicht nur auf das erweiterte Stufen-Modell, sondern auch auf den Zusammenhang der so konstruierten Variable *Sophistiziertheit Konditionalen Schlussfolgerns* mit Reasoning. Zur Überprüfung der Arbeitshypothesen und damit der Vorform des *Kurztests zum Konditionalen Schlussfolgern* (*KKS*) werden Daten an einer umfangreichen Stichprobe ($N = 905$) erhoben, die unter Verwendung von Latente-Variablen-Modellen (Analyse latenter Klassen, multinomiale logistische Regression für latente Variablen) analysiert werden. Die Ergebnisse liefern empirische Evidenz für das erweiterte Stufen-Modell und sprechen gleichsam für eine adäquate Wahl der verwendeten Analysemethoden. Dies soll zunächst in weiteren empirischen Studien repliziert werden. Gleichzeitig stellen sich weitere Forschungsfragen und -aufgaben bezüglich des KKS, insbesondere die Bestimmung seiner Testgütekriterien. Vor allem dies soll Gegenstand des folgenden Kapitels sein.

3 Weitere Erprobung des Tests – Bestimmung seiner Testgütekriterien

Wenngleich die erste empirische Erprobung des KKS insgesamt hypothesenkonforme Ergebnisse liefert, ist es dennoch verfrüht, von einem „Psychologischen Test“ zu sprechen. Hierzu bedarf es – wie bereits ausgeführt – der Replikation der empirischen Befunde, der Beantwortung der Frage nach der Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns (SKS) sowie der Bestimmung von Testgütekriterien des KKS. Der Fokus der weiteren empirischen Erprobung des KKS (und damit des folgenden Kapitels) soll auf der Bestimmung seiner Testgütekriterien liegen, da diese ein wichtiges Instrument zur Beurteilung der Qualität eines Tests sind (z.B. Moosbrugger & Kelava, 2007). Bereits an dieser Stelle sei dabei auf eine Besonderheit des KKS hingewiesen. Sehr häufig werden die mit Tests gemessenen Konstrukte als kontinuierliche latente Variablen angenommen und entsprechend modelliert. Dies gilt in besonderem Maße für Leistungstests. Es existieren meist vielfältig erprobte Methoden und Analyseverfahren für die Bestimmung von Testgütekriterien im Falle kontinuierlicher latenter Variablen. Das Konstrukt SKS ist hingegen eine nominale latente Variable, deren Werte die vier Sophistiziertheits-Stufen Konditionalen Schlussfolgerns (Unsophistizierte, Unsophistizierte II, Fortgeschrittene, Sophistizierte) sind. Man könnte nun argumentieren, dass diese Variable durch das theoriegeleitete Ordnen der Stufen (vgl. Abschnitte 2.2.2 und 2.3.2.2) ordinales Datenniveau aufweist. Dies ändert jedoch nichts daran, dass viele der für kontinuierliche Variablen angemessenen Verfahren zur Testgütekriterienbestimmung in diesem Fall nicht anwendbar sind. Es ist also zunächst notwendig, sich zumindest für einige der betrachteten Testgütekriterien zu überlegen, welche inhaltlichen Konsequenzen sich ergeben, wenn das betrachtete Merkmal nominales Skalenniveau aufweist. Darauf aufbauend sind dann ggf. konkrete Verfahren zur Bestimmung der Testgütekriterien abzuleiten. Da jede Ordinalskala gleichzeitig eine Nominalskala ist, scheinen Überlegungen für nominale latente Variablen ausreichend. Wo nötig, werden sie für ordinale Variablen präzisiert.

Entsprechend sind die Abschnitte zu den einzelnen Testgütekriterien aufgebaut. Zunächst wird das jeweilige Testgütekriterium kurz vorgestellt und ggf. seine Bedeutung für ein nominales Konstrukt thematisiert. Anschließend wird für den KKS überlegt, ob das jeweilige Testgütekriterium als erfüllt betrachtet werden kann. Hierzu ist es mitunter nötig, entsprechende Methoden bzw. Analyseverfahren zur Testgütekriterienbestimmung abzuleiten und in zusätzlichen empirischen Studien auf den KKS anzuwenden.

Aus ökonomischen Gründen sollen für die anderen Forschungsfragen (siehe Abschnitt 2.4.3.3) keine zusätzlichen Studien durchgeführt werden. Vielmehr sollen diese in den Studien zur Bestimmung der Testgütekriterien mit beantwortet werden. So wird bspw. die Replizierbarkeit der empirischen Befunde zum erweiterten Stufen-Modell im Rahmen der Messmodellprüfung in jeder der in der Folge berichteten Studien überprüft. Ebenso werden die Replikation des Zusammenhangs von SKS und Reasoning sowie die Prüfung der Stabilität von SKS an passenden Stellen integriert.

Nun stellt sich noch die Frage, welche Testgütekriterien betrachtet werden sollen. Dazu findet sich in den meisten Lehrbüchern zur Psychologischen Diagnostik, zur Testtheorie und/oder zur Testkonstruktion (z.B. Guthke, Böttcher & Sprung, 1990; Kubinger & Jäger, 2003; Lienert & Ratz, 1998; Petermann & Eid, 2006; Rost, 2004) mindestens ein Kapitel zu Testgütekriterien. Konsens herrscht dabei bezüglich der sog. *Hauptgütekriterien* von Tests: *Objektivität*, *Reliabilität* und *Validität*. Sie werden in der Testkonstruktionspraxis bei jedem Test betrachtet und folglich auch für den KKS als erstes thematisiert. Anschließend folgt eine Reihe von Nebengütekriterien, deren Auswahl wie auch einleitende Kurzbeschreibungen sich überwiegend an Kubinger (2003) sowie Moosbrugger und Kelava (2007) orientieren. Darüber hinaus wird zusätzlich das Gütekriterium Attraktivität (siehe Abschnitt 3.11) thematisiert, da es für das Forschungsprojekt, im Rahmen dessen der KKS entstanden ist (siehe dazu Kapitel 1), von Bedeutung ist (siehe dazu Abschnitt 3.11).

3.1 Objektivität

Objektivität ist eines der drei Hauptgütekriterien eines Tests und soll sicherstellen, dass Testleistungen verschiedener getesteter Personen vergleichbar sind. Klassischerweise werden dabei drei Aspekte betrachtet: Objektivität bei der Durchführung und bei der

Auswertung des Tests sowie bei der Interpretation der Testergebnisse (vgl. z.B. Lienert & Raatz, 1998). Beim KKS handelt es sich gegenwärtig[48] um einen Online-Test (vgl. Abschnitt 2.4.1.3) und damit um einen computergestützten Test. Aussagen zur Objektivität von computergestützten Tests im Allgemeinen wie von Online-Tests im Speziellen gelten folglich auch für den KKS. Dies ist Ausgangspunkt der Betrachtungen, insbesondere zur Durchführungs- und zur Auswertungsobjektivität.

Durchführungsobjektivität

Generell liegt Durchführungsobjektivität dann vor, wenn das Testergebnis unabhängig vom Testleiter ist. Einigkeit herrscht darüber, dass das bei computergestützten Tests prinzipiell der Fall ist (Batinic & Bosnjak, 2000; Kleinmuntz & McLean, 1968; Schuler & Höft, 2006). Online-Tests (und damit der KKS) zählen zur Gruppe der computergestützten Tests, sodass sich die Testleiterunabhängigkeit unproblematisch auf Online-Tests übertragen lässt (vgl. Batinic & Bosnjak, 2000). Gleichzeitig ist jedoch anzumerken, dass sich durch den fehlenden Testleiter die Wahrscheinlichkeit von Missverständnissen bei der Testbearbeitung erhöht, da keine Nachfragen seitens der Teilnehmer möglich sind (Nosek, Banaji & Greenwald, 2002; Reips, 2000). Dies wiederum kann zu Testabbrüchen oder falschen Angaben führen (Nosek et al., 2002). Streng genommen ist das jedoch eher eine Frage der Qualität der Instruktion bzw. der Navigation durch den (Online-)Test. Im Falle des KKS besteht die Möglichkeit, zu jedem Zeitpunkt der Testung die Instruktion nochmals aufzurufen. Da es keine Zeitbeschränkung gibt, scheint dies unproblematisch. Das beschriebene Problem kann dadurch zwar nicht gelöst, aber zumindest „gemildert“ werden. Unstrittig ist allerdings, dass es andere Störquellen gibt, die die Durchführungsobjektivität für Online-Tests einschränken, seien es nun verschieden schnelle Internetverbindungen, unterschiedliche Browser oder die Umgebungsbedingungen (Geräusche, Lichtverhältnisse usw.). Häufig unterscheidet sich die Testdurchführung zwischen verschiedenen Probanden hinsichtlich dieser Punkte. Während die technischen Rahmenbedingungen (Browser, Internetverbindung, Bildschirmauflösung u.v.m.) in Online-Tests zumindest mit erhoben und bspw. als Kovariaten in statistischen Analysen berücksichtigt werden können, ist eine detaillierte Erhebung der Umgebungsbedingungen, geschweige denn deren Kontrolle kaum möglich. Hierzu sei angemerkt, dass die Kontrolle sämtlicher Bedingungen zur Steigerung der Durchfüh-

[48] Damit soll angedeutet werden, dass auch eine Papier-Bleistift-Version möglich wäre, ebenso wie eine computergestützte Variante, die nicht online bearbeitet wird.

rungsobjektivität zunehmend in Frage gestellt wird (z.B. Martin, 1996; Reips, 2000), da die Übertragbarkeit der Ergebnisse auf reale Alltagssituationen mit zunehmender Kontrolle sämtlicher Bedingungen sogar abnehmen kann. Dem Problem der Vergleichbarkeit technischer Rahmenbedingungen trägt der KKS Rechnung, indem nach seiner Programmierung Testläufe auf den drei am häufigsten verwendeten Browsern (Internet Explorer, Mozilla Firefox, Opera) bei einer Vielzahl gängiger Bildschirmauflösungen (ab 800 x 600 Pixel aufwärts) durchgeführt wurden. Erst nachdem der Test unter jeder der resultierenden Bedingungen (Browser x Bildschirmauflösung) in vergleichbarer Weise ablief, wurde der Test online geschaltet. Es bleibt insgesamt festzuhalten, dass die prinzipiell hohe Durchführungsobjektivität von Online-Tests auch auf den KKS übertragen werden kann. Zusätzlich wird beim KKS sogar noch einigen Problemen bezüglich der Durchführungsobjektivität von Online-Tests Rechnung getragen. Man kann also beim KKS durchaus von Durchführungsobjektivität ausgehen.

Auswertungsobjektivität

Auswertungsobjektivität liegt vor, wenn (bei gegebenem Testprotokoll) das Testergebnis unabhängig vom Testauswerter ist. Für Aufgaben mit gebundenem Antwortformat − wie die des KKS − ist dies im Allgemeinen gegeben (Moosbrugger & Kelava, 2007). Hinzu kommt der Vorteil der hohen Standardisierung der Auswertungsprozesse bei computergestützten Tests (Schuler & Höft, 2006), bspw. weil durch automatisierte Speicherungs- und Auswertungsprozeduren die Gefahr von Eingabefehlern nicht besteht (Kiesler & Sproull, 1986). Bereits deshalb kann für den KKS von Auswertungsobjektivität ausgegangen werden. Dies gilt umso mehr, da die Zuordnung eines Testwertes (Klassenzugehörigkeit) zu einem bestimmten Antwortmuster durch das statistische Analyseverfahren (LCA) erfolgt und daher a priori vom Testanwender unabhängig ist.

Interpretationsobjektivität

Interpretationsobjektivität liegt vor, wenn verschiedene Testanwender für Testpersonen mit demselben Testwert zu denselben Schlussfolgerungen kommen. Für den KKS ist der Testwert einer Person die Klasse, der das Antwortmuster dieser Person (und damit die Person) mit der höchsten Wahrscheinlichkeit zugeordnet wird. Die Interpretation dieser Testwerte (also der Klassenzuordnungen: Unsophistizierte, Unsophistizierte II, Fortgeschrittene und Sophistizierte) ist aufgrund der kognitionspsychologischen Fundie-

rung des Tests (Stufen-Modell siehe Abschnitt 2.2.2; erweitertes Stufen-Modell siehe Abschnitt 2.3.2.2) recht leicht möglich:

Unsophistizierte akzeptieren die einladende Inferenz einer Konditionalaussage.

Unsophistizierte II haben infolge von Schwierigkeiten beim Enkodieren bzw. aktiven Auflösen einer doppelten Negation eine höhere Wahrscheinlichkeit der einladenden Inferenz zu widerstehen, jedoch lediglich bei den Schlussfiguren NA und MT und auch dort nur, wenn in der Hauptprämisse mindestens eine Negation enthalten ist.

Fortgeschrittene widerstehen der einladenden Inferenz, sind jedoch nicht in der Lage, einen sophistizierten Modus-Tollens-Schluss zu ziehen. Hierfür fehlt es möglicherweise an Kenntnis (oder Verfügbarkeit) der reductio ad absurdum.

Sophistizierte schließlich sind in der Lage sowohl der einladenden Inferenz zu widerstehen als auch einen sophistizierten Modus-Tollens-Schluss zu ziehen.

Auf Basis dieser Interpretationen können entsprechende Feedbacks formuliert werden, die dann ebenfalls automatisiert erstellt und dargeboten werden können. Durch die Angabe einer konkreten Zuordnungswahrscheinlichkeit zur jeweiligen wahrscheinlichsten Klasse (und zwar für jedes Antwortmuster und damit für jede Person) kann sogar auf die Angemessenheit der entsprechenden Interpretation geschlossen werden. Diese ist umso höher, je höher die Zuordnungswahrscheinlichkeit zu der entsprechenden Klasse ist. Auch wenn die Umsetzung des KKS als Einzeltestung gegenwärtig noch nicht erfolgt ist, birgt er aufgrund der Automatisierungsmöglichkeiten Potenzial für eine hohe Interpretationsobjektivität. Allerdings lässt sich erst nach der Umsetzung des KKS für die Einzelfalldiagnostik (siehe dazu Abschnitt 4.3.2) seine Interpretationsobjektivität abschließend bewerten. Dennoch kann aufgrund der kognitionspsychologischen Fundierung der vier Klassen vorerst von Interpretationsobjektivität ausgegangen werden.

Zusammenfassende Bewertung der Objektivität des KKS

Da sowohl Durchführungs- als auch Auswertungsobjektivität sowie prinzipiell auch Interpretationsobjektivität für den KKS gegeben sind, kann das Gütekriterium Objektivität für den KKS insgesamt als erfüllt angesehen werden.

3.2 Reliabilität (Messgenauigkeit)

In der Testkonstruktionspraxis gilt ein Test dann als reliabel, messgenau oder auch zuverlässig, wenn er das Merkmal, das er misst, möglichst exakt (also möglichst messfehlerfrei) misst. Im Rahmen von Testtheorien ist Reliabilität ein *wohldefinierter* Begriff (für die KTT vgl. bspw. Steyer & Eid, 2001; für die IRT vgl. bspw. Andrich, 1988). So ist Reliabilität in der KTT als Determinationskoeffizient definiert, der angibt, wie viel der Varianz der *Testwertvariable* durch die Varianz der *true score-Variable* aufgeklärt wird (vgl. z.B. Steyer & Eid, 2001). Bei nominalen Variablen (wie bspw. SKS) ist eine sinnvolle Interpretation der Varianz jedoch nur schwer möglich (Collins, 2001), sodass diese klassische Definition von Reliabilität nicht angebracht ist. Daher wird im Folgenden von „Messgenauigkeit" gesprochen, wenn das entsprechende Testgütekriterium des KKS thematisiert wird. So soll Missverständnissen vorgebeugt werden, die bei Verwendung des Begriffes „Reliabilität" entstehen könnten. Zunächst wird überlegt, wie die Messgenauigkeit bei nominalen (latenten) Variablen ganz generell bestimmt werden kann. Des Weiteren wird mit der Bestimmung der Messgenauigkeit mittels wiederholter Testvorgabe eine weitere gängige Methode zur Messgenauigkeitsbestimmung vorgestellt (vgl. z.B. Schermelleh-Engel & Werner, 2007). Nach Überlegungen zur Anwendbarkeit dieser Methode auf nominale Variablen werden jeweils konkrete Ergebnisse für den KKS angegeben, die durch eine abschließende Bewertung seiner Messgenauigkeit zusammengefasst werden. Ergänzend sei erwähnt, dass in diese Betrachtungen auch die Überprüfung der Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns integriert wird.

3.2.1 Messgenauigkeit im Falle nominaler latenter Variablen

Was bedeutet nun Messgenauigkeit, wenn Ziel der Messung die Identifikation nominaler latenter Variablen ist, im Falle des KKS also die Identifikation latenter Klassen? Anders gefragt: Wie bestimmt man die Messgenauigkeit eines LCMs? Dieser Frage wird in der psychometrischen Forschung insgesamt eher wenig Aufmerksamkeit geschenkt (Clogg & Manning, 1996). In der Folge soll ein diesbezüglicher Vorschlag von Rost (2004) kurz beschrieben werden.

Bei der Personenparameterschätzung in einem LCM wird für jedes Antwortmuster (und damit für jede Person, die dieses Antwortmuster zeigt) die Zuordnungswahr-

scheinlichkeit zu jeder identifizierten Klasse bestimmt. Die Klasse, für die diese Zuordnungswahrscheinlichkeit am größten ist, stellt schließlich die Personenparameterschätzung für eine Person dar, die dieses Antwortmuster zeigt. Konkret heißt das, man erhält für jede Person (bzw. für deren Antwortmuster) zwei Kennwerte: zum einen die Klasse, zu der die Person am wahrscheinlichsten gehört, und zum anderen die Wahrscheinlichkeit, mit der die Person zu dieser Klasse gehört. Mittelt man diese Wahrscheinlichkeiten aller Personen, die einer bestimmten (nämlich der für sie wahrscheinlichsten) Klasse *C* zugeordnet werden, erhält man die Treffsicherheit T_C dieser Klasse (vgl. auch Abschnitt 2.4.1.5). Diese stellt nach Rost (2004) ein Messgenauigkeitsmaß für die Klassenzuordnungen zu dieser Klasse dar, für das die bereits berichteten Konventionen (siehe Abschnitt 2.4.1.5) gelten: $T_C > .85$ − relativ hohe Treffsicherheit, $T_C > .90$ − hohe Treffsicherheit. Wie bei anderen Modellen der IRT ist auch bei einem LCM die Messgenauigkeit abhängig von der Ausprägung des Personenparameters. Das heißt, für jede Klasse kann die Treffsicherheit und damit die Genauigkeit der Messung eine andere sein. Mittelt man die Zuordnungswahrscheinlichkeiten aller Personen über die Klassen hinweg, denen sie zugeordnet werden, ist dies die Treffsicherheit *T* der gesamten Klassen-Lösung, im Falle des KKS also die Treffsicherheit seines Messmodells.

Dabei stellt sich die prinzipielle Frage, ob die Treffsicherheit einen geeigneten Schätzer für Messgenauigkeit darstellt. Hierzu soll eine Definition für Messgenauigkeit im Rahmen eines LCMs versucht werden: Wenn das gemessene Konstrukt als nominale (latente) Klassenvariable konzipiert ist, dann ist der *wahre Wert* einer Person ihre „wahre“ Klasse. Der Testwert einer Person ist ihre (manifeste) Klassenzuordnung aufgrund ihres Antwortmusters. Demnach kann die Reliabilität auf Ebene der Person definiert werden als:

Wahrscheinlichkeit, einer Klasse zugeordnet zu werden, gegeben das Antwortmuster und gegeben tatsächlich zu dieser Klasse zu gehören.

Die Treffsicherheit ist die (gemittelte) Wahrscheinlichkeit, mit der Personen aufgrund ihres Antwortmusters der für sie wahrscheinlichsten Klasse zugeordnet werden. Sie ist demnach lediglich dann ein Schätzer dieser theoretischen Größe, wenn folgende zusätzliche Annahme getroffen wird:

Die für eine Person wahrscheinlichste Klasse ist ihre „wahre“ Klasse.

Auch im Rahmen der KTT sind für die Bestimmung der Reliabilität zusätzliche Annahmen zu treffen. Fragen nach der Angemessenheit dieser Annahme oder Alternativen für Reliabilitäts-Schätzer in einem LCM werden an späterer Stelle (siehe Abschnitt 4.2.4) diskutiert. Vorerst werden die Treffsicherheiten der einzelnen Klassenzuordnungen wie auch die Treffsicherheit der gesamten Klassen-Lösung (unter der Annahme, die für eine Person wahrscheinlichste Klasse sei ihre „wahre" Klasse) als Maße für die Messgenauigkeit verwendet. Dafür spricht auch, dass es sich um ein vergleichsweise intuitives Maß handelt, das zudem von der verwendeten Software (Winmira 2001; vgl. Abschnitt 2.4.1.5) automatisch ausgegeben wird. Es folgt die Angabe konkreter Werte der Treffsicherheiten für den KKS.

Messgenauigkeit des KKS

Die Messgenauigkeit der bei der ersten empirischen Erprobung verwendeten Version des KKS wird also durch die Treffsicherheit T der Fünf-Klassen-Lösung von $T = .939$ (vgl. Tabelle 10 in Abschnitt 2.4.2) indiziert. Die Personen werden mit nach Rost (2004) hohen Treffsicherheiten den postulierten Klassen zugeordnet ($T_C = .923$ bis $T_C = .966$) sowie mit einer relativ hohen Treffsicherheit ($T_C = .888$) der „Restklasse". Die Bestimmung der Klassenzugehörigkeit und damit des Personenparameters kann für den KKS daher als messgenau gelten. Da es sich beim KKS um eine Testneuentwicklung sowie bei der LCA um eine bislang eher wenig gebräuchliche Methode zur Bestimmung von Testwerten handelt, soll in jeder der im weiteren Verlauf der Arbeit erwähnten Studien zunächst immer die Messgenauigkeit des KKS in Form einer Analyse der Treffsicherheiten überprüft werden. Nicht zuletzt, da die Angemessenheit der Treffsicherheit als Messgenauigkeits-Schätzer noch nicht ausreichend diskutiert ist (siehe dazu Abschnitt 4.2.4), soll die Messgenauigkeit des KKS noch auf eine zweite Weise bestimmt werden: mittels wiederholter Testvorgabe.

3.2.2 Bestimmung der Messgenauigkeit des KKS mittels wiederholter Testvorgabe

Eine generelle Möglichkeit zur Bestimmung der Messgenauigkeit eines Tests ist, ihn denselben Personen mehrfach vorzugeben. Bei zweimaliger Vorgabe spricht man von der *Retest-Methode* zur Bestimmung der Reliabilität (vgl. z.B. Schermelleh-Engel & Werner, 2007), wobei unter Gültigkeit bestimmter Annahmen der KTT, z.B. der eines

Paralleltest-Modells (siehe z.B. Steyer, 2001), die *Test-Retest-Korrelation* einen Schätzer für die Reliabilität darstellt. Problematisch ist dabei jedoch, dass für eine hohe Test-Retest-Korrelation nicht nur eine hohe Messgenauigkeit des Tests notwendig ist, sondern auch, dass das gemessene Merkmal über die Zeit stabil ist. Die Stabilität des Merkmals ist sogar die Voraussetzung dafür, dass die Annahme gleicher wahrer Testwerte zu beiden Zeitpunkten gerechtfertigt ist. Die Annahme gleicher wahrer Testwerte ist wiederum eine Voraussetzung dafür, die Test-Retest-Korrelation als Reliabilitäts-Schätzer interpretieren zu können. Findet bezüglich des zu messenden Merkmals zwischen den Messzeitpunkten Entwicklung statt, kann die Test-Retest-Korrelation niedrig sein, obwohl der Test selber zu beiden Zeitpunkten genau misst. Es ist daher zur Bestimmung der Messgenauigkeit mittels wiederholter Testvorgabe notwendig, zunächst die Stabilität des gemessenen Merkmals zu betrachten, welche ohnehin Gegenstand der noch offenen Forschungsfragen zum KKS ist (vgl. Abschnitt 2.4.3.3). Anschließend können dann Aussagen über den Zusammenhang der Testwerte zu zwei Messzeitpunkten getroffen werden.

3.2.2.1 Stabilität der latenten Variable SKS

In der Psychologie werden bspw. Persönlichkeitseigenschaften gemeinhin als stabil postuliert (z.B. Asendorpf, 2007; Conley, 1984; Kruse & Schmitt, 2004) ebenso wie verschiedene Fähigkeiten, z.B. Intelligenz (Asendorpf, 2007; Deary et al., 2004; Wilson, 1983). Einige Arbeiten gehen auf Basis von Piagets (1971) Theorie kognitiver Entwicklung für das Schlussfolgern bei Aufgaben mit Konditionalaussagen zumindest im Kindes- und Jugendalter von einer stufenweisen Entwicklung aus (vgl. z.B. Spiel et al., 2001). Demnach müsste nach Erreichen der letzten Entwicklungsstufe (im Jugendalter) jeder die gleiche Fähigkeit und damit Sophistiziertheits-Stufe im Konditionalen Schlussfolgern zeigen. Dem widersprechen jedoch Befunde von Rumain et al. (1983), die sämtliche Sophistiziertheits-Stufen Konditionalen Schlussfolgerns auch bei Probanden im Erwachsenenalter beobachten können (vgl. Abschnitt 2.2.1), selbst die Unsophistizierten-Stufe, deren Antwortverhalten nach Spiel et al. (2001) eher dem Kindesalter zuzuordnen wäre. Auch die Ergebnisse der ersten empirischen Erprobung des KKS sprechen für einen nicht unerheblichen Anteil Erwachsener (ca. 56 %) auf den beiden Unsophistizierten-Stufen (vgl. Abschnitt 2.4.2). Diese Befunde sprechen eher gegen ein Erreichen der höchsten Entwicklungsstufe im Erwachsenenalter und eine daraus resultierende Stabilität. Auch aus der theoretischen Betrachtung der Sophistiziertheits-Stufen

als Kompetenz-Stufen (vgl. Abschnitt 2.2.2) ergeben sich bezüglich der Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns keine klaren Konsequenzen. SKS kann demnach durchaus stabil sein, wenn hinsichtlich der Komptenz-Stufen keine Entwicklung passiert. Ebenso ist es jedoch möglich, dass infolge entsprechender Trainingsmaßnahmen (siehe dazu auch Abschnitt 4.3.2) höhere Kompetenz-Stufen erreicht werden.

Da die Frage der Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns über die Zeit bislang weder Gegenstand theoretischer Arbeiten noch empirischer Studien ist, soll dieser Frage empirisch nachgegangen werden. Dies führt wiederum zu der Frage, mit welcher Analysemethode die Stabilität einer nominalen latenten Variable bestimmt werden kann. Generell bieten sich zur Bestimmung der Stabilität latenter Variablen Modelle der *Latent-State-Trait-Theorie* (*LST-Theorie*; Steyer, Ferring & Schmitt, 1992; Steyer, Schmitt & Eid, 1999 sowie speziell für nominale latente Variablen Eid & Langeheine, 1999). Bezieht man die Analyse latenter Klassen (als das Verfahren zur Modellierung der nominalen latenten Variable SKS) in die Überlegungen ein, dann ist die Stabilität der nominalen latenten Variable nichts anderes als die Stabilität der Klassen-Lösung über die Zeit. Ein Verfahren, mit dem die Stabilität einer Klassen-Lösung über die Zeit analysiert werden kann, ist die *Analyse latenter Transitionen* (*LTA*[49]; Collins & Wugalter, 1992; Graham et al., 1991), bei der Wahrscheinlichkeiten für latente Klassenübergänge zwischen zwei (oder auch mehr) Zeitpunkten geschätzt werden können. Auf Basis dieser latenten Klassenübergänge sollte dann die Bestimmung geeigneter Koeffizienten zur Beschreibung der Stabilität der nominalen latenten Variable SKS möglich sein.

3.2.2.2 Analyse latenter Transitionen

Die LTA wird häufig als eine Art Längsschnitt-LCA bezeichnet (Nylund, 2007), wobei dies streng genommen nur zutrifft, wenn zu sämtlichen Zeitpunkten jeweils ein LCM das Messmodell der Wahl ist. Für eine Analyse des KKS zu zwei Messzeitpunkten wäre dies der Fall. Theoriegeleitet würde zu beiden Zeitpunkten die gleiche Klassen-Lösung (vier Klassen im Sinne des erweiterten Stufen-Modells) postuliert. Hauptgegenstand einer LTA ist die Betrachtung latenter Transitionen, also latenter Klassenübergänge. Eines der Ergebnisse sind dabei Schätzungen für die Wahrscheinlichkeiten solcher latenter Klassenübergänge. Da in der vorliegenden Arbeit mit der LTA letztlich die Stabilität der latenten Variable SKS überprüft werden soll, ist es theoretisch plausibel,

[49] Abk. für die englische Bezeichnung *Latent Transition Analysis*

zu beiden Zeitpunkten das gleiche LCM[50] als Messmodell anzunehmen. Dies ist zunächst empirisch zu prüfen. Zeigt sich zu beiden Zeitpunkten die gleiche Anzahl an latenten Klassen als bestpassende Lösung und ähneln sich die Muster der klassenbedingten Lösungswahrscheinlichkeiten über die Zeitpunkte hinweg, dann ist für die LTA die Annahme der *Invarianz des Messmodells* über die Zeitpunkte angebracht (Nylund, 2007). Im Falle eines LCMs für dichotome Items heißt das, die bedingten Lösungswahrscheinlichkeiten innerhalb jeder Klasse werden über die (in diesem Fall zwei) Zeitpunkte hinweg gleichgesetzt (Nylund, 2007). Unter der Voraussetzung, dass die Bedeutung der Klassen über die zwei Zeitpunkte hinweg gleich ist, ermöglicht das eine einfache Interpretation der Wahrscheinlichkeiten für die Klassenübergänge. Die geschätzten Wahrscheinlichkeiten für die latenten Klassenübergänge können nun zur Bestimmung der Stabilität der nominalen latenten Variable SKS durch Berechnung geeigneter Koeffizienten verwendet werden. Um der Frage nach der Stabilität von SKS nachzugehen, soll eine eigenständige Studie durchgeführt werden, deren Erhebungsinstrumente, Stichprobe, Durchführung, Auswertungsmethoden sowie Ergebnisse im Folgenden kurz beschrieben werden. Diese Studie wird in der Folge als *Stabilitätsstudie* bezeichnet.

Erhebungsinstrumente

Zur Erfassung von SKS wurde die überarbeitete Version des KKS vorgegeben, die sich von der ersten Version (siehe Abschnitt 2.4.1.2) vor allem dahingehend unterscheidet, dass eine zusätzliche vierte Antwortoption („*Ich weiß es nicht.*“) eingeführt wird und die Items in nicht-sequentieller Reihenfolge vorgegeben werden (vgl. Abschnitt 2.4.3.2 für eine ausführliche Beschreibung wie auch Begründung der Änderungen).

Durchführung

Die Untersuchung fand in Form einer Online-Erhebung zu zwei Zeitpunkten[51] statt. Hierzu wurde die überarbeitete Version des KKS, ergänzt um Angaben zur Demographie, einzelne Übergangsseiten sowie weitere Aufgaben, die jedoch im Rahmen dieser Arbeit nicht ausgewertet werden, als Online-Test programmiert. Die Zuordnung der Daten beider Zeitpunkte zu dem jeweiligen Probanden erfolgte über einen individuellen

[50] „Gleiches LCM“ meint hier, dass die Anzahl der latenten Klassen sowie die klassenbedingten Lösungswahrscheinlichkeiten zu beiden Zeitpunkten als gleich angenommen werden.

[51] Wenn in der Folge von „Zeitpunkt 1“ oder „Zeitpunkt 2“ gesprochen wird, sind damit die beiden Erhebungszeitpunkte der Stabilitätsstudie gemeint.

Code[52], den die Personen zu Beginn jeder Erhebung eingaben. Der erste Erhebungszeitraum erstreckte sich von Anfang November 2007 bis Anfang Dezember 2007, der zweite Erhebungszeitraum von Mitte Januar 2008 und bis Mitte Februar 2008. Beide Erhebungszeiträume waren so gewählt, dass für jeden Testteilnehmer mindestens sechs Wochen zwischen beiden Messzeitpunkten lagen. Dieser Zeitraum scheint in Anlehnung an Eid (1997) angemessen, um davon ausgehen zu können, dass Erinnerungseffekte an die Aufgaben bzw. einzelne Items weitestgehend ausgeschlossen werden können. Zu beiden Zeitpunkten gab es keine Zeitbeschränkungen. Lediglich die voraussichtliche Testdauer (ca. 20 Minuten) wurde in der Instruktion angekündigt. Tabelle 12 illustriert den Aufbau dieses Online-Tests zu beiden Zeitpunkten.

Tabelle 12: Aufbau des Online-Tests zu beiden Zeitpunkten der Stabilitätsstudie

Zeitpunkt 1 (Nov.-Dez. 2007)	Zeitpunkt 2 (Jan.-Feb. 2008)
Einleitung	Einleitung
Codeabfrage	Codeabfrage
Demographie	16 Items zum Konditionalen Schlussfolgern *
16 Items des KKS	Demographie
Pause	Wason Selection Task *
Wason Selection Task *	Pause
Selbsteinschätzung Intelligenz *	16 Items des KKS
Verabschiedung	Selbsteinschätzung Intelligenz *
	Verabschiedung

Anmerkung. Die mit * gekennzeichneten Aufgaben bzw. Items werden im Rahmen der vorliegenden Arbeit nicht ausgewertet.

Untersuchungsstichprobe

Wie schon bei der ersten empirischen Erprobung erfolgte die Rekrutierung der Untersuchungsstichprobe über eine Art Schneeballsystem, bei dem Psychologiestudierende eines Seminars gebeten wurden, den Link zur Online-Studie an ca. 20 Bekannte zu versenden, mit der Bitte um Teilnahme an zwei Erhebungszeitpunkten (durchschnittlicher zeitlicher Abstand: 52 Tage). Auf diese Weise konnten 195 Personen rekrutiert

[52] Persönlicher Code: erster Buchstabe des Geburtsortes (Bsp.: **J**ena), zweiter Buchstabe des Vornamens (Bsp.: M**i**chael), dritter Buchstabe des Nachnamens (Bsp.: Ko**c**h), jeweils die letzte Ziffer des Geburtstags und des Geburtsmonats (Bsp.: 1**3**.0**8**.): **JIC38**

werden, die den KKS zu beiden Zeitpunkten vollständig bearbeiteten. Sämtliche Personen machten soziodemographische Angaben zur eigenen Person, die im Folgenden kurz zusammengefasst werden.

63,1% der Teilnehmer waren Frauen, entsprechend 36,9% Männer. Frauen sind damit in der Untersuchungsstichprobe überrepräsentiert. Das Durchschnittsalter lag bei 26 Jahren (*Sd* = 7,8 Jahre) und ähnelt dem der ersten empirischen Erprobung des KKS. Dies ist angesichts der vergleichbaren Rekrutierungsstrategie nicht überraschend. Als höchsten Bildungsabschluss gaben 2,6% Realschule oder niedriger an, 5,6% eine abgeschlossene Berufsausbildung, 66,2% Abitur und 20,5% einen Hochschulabschluss[53]. 73,8% befanden sich zum Untersuchungszeitpunkt in Ausbildung (auch Studium), 19,5% waren Angestellte[54]. Da der Altersdurchschnitt für eine Stichprobe junger Erwachsener spricht, kann aufgrund der Angaben zur aktuellen Tätigkeit und wegen des vergleichsweise hohen Bildungsniveaus davon ausgegangen werden, dass Studierende den überwiegenden Teil der Stichprobe darstellen. Wie auch schon bei der ersten empirischen Erprobung des KKS ist Thüringen als Herkunfts-Bundesland stark überrepräsentiert (44,6%).

Auswertungsmethoden

Zunächst wird für die Daten beider Zeitpunkte eine Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5) zur Sicherung der Datenqualität durchgeführt. Anschließend ist für jeden der beiden Zeitpunkte das Messmodell für die latente Variable SKS zu überprüfen. Dies erfolgt jeweils durch Beurteilung der Passung eines LCMs unter Vorgabe von vier Klassen bei gleichzeitig guter Interpretierbarkeit dieser vier Klassen im Sinne des erweiterten Stufen-Modells. Aufgrund der Ergebnisse der ersten empirischen Erprobung des KKS (siehe Abschnitt 2.4.2) sowie der durchgeführten Modifikationen des KKS (siehe Abschnitt 2.4.3.2) wird postuliert, dass eine Vier-Klassen-Lösung (zu beiden Zeitpunkten) den niedrigsten BIC aller getesteten Klassen-Lösungen aufweist und gleichzeitig die Kriterien für einen guten Modellfit bei der Durchführung eines parametrischen Bootstrapverfahrens erfüllt. Die Interpretation der Klassen erfolgt aufgrund der Zuordenbarkeit des Musters der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Klassen des erweiterten Stufen-Modells. Es wird wiederum davon

[53] Außerdem gaben 0,5% „Promotion“ an sowie 4,6% „Sonstiges“.
[54] Außerdem gaben 0,5% „Selbstständigkeit“, 0,5% „Leitende(r) Angestellte(r)“ sowie 5,6% „Sonstiges“ an.

ausgegangen, dass diese Zuordnung bereits per Augenschein möglich ist. Zudem soll noch zu jedem Messzeitpunkt die Messgenauigkeit der Vier-Klassen-Lösung durch Beurteilung der Treffsicherheiten bewertet werden (vgl. Abschnitt 3.2.1). Die LCAs werden mithilfe der Software Winmira 2001 (Davier, 2001) berechnet. Da es sich um die gleichen Auswertungsmethoden wie bei der ersten empirischen Erprobung des KKS handelt, wird auf ausführlichere Beschreibungen verzichtet (siehe dafür Abschnitt 2.4.1.5).

Erweist sich zu beiden Messzeitpunkten die gleiche Vier-Klassen-Lösung als passendes Messmodell, ist es gerechtfertigt, eine LTA zur Schätzung latenter Klassenübergänge durchzuführen. Zu beiden Messzeitpunkten wird dazu ein LCM unter Vorgabe von vier Klassen spezifiziert, dessen Parameter (Itemschwierigkeiten in jeder Klasse) als *invariant über die Messzeitpunkte* angenommen und entsprechend fixiert werden (vgl. Nylund, 2007). Als Ergebnis werden für dieses Modell 16 Wahrscheinlichkeiten für die latenten Klassenübergänge (Übergang jeder der vier Klassen zum Zeitpunkt 1 in jede der vier Klassen zum Zeitpunkt 2) geschätzt. Die LTA wird mit der Software M*plus* (Muthén & Muthén, 1998-2007) berechnet. Die Modellgüte des Latente-Transitionen-Modells kann allerdings nicht bestimmt werden, da in M*plus* keine Prozedur implementiert ist, die das sparse-data-Problem (siehe auch Abschnitt 2.4.1.5) des zu spezifizierenden Latente-Transitionen-Modells berücksichtigt (B. Muthén, 2009, persönliche Kommunikation). Folglich kann lediglich die Passung der LCMs zu beiden Zeitpunkten als Indikator für die Modellgüte dienen. Des Weiteren ist durch Analyse der Entropie *E* der LTA eine Aussage darüber möglich, wie gut Personen durch die LTA klassifiziert werden können (Nylund, 2007). Die Entropie ähnelt konzeptuell der Treffsicherheit *T* eines LCMs (zu Entropie siehe z.B. Ramaswamy, DeSarbo, Reibstein & Robinson, 1993). Allerdings bezieht sich die Entropie im Falle einer LTA nicht auf die Zuordnungswahrscheinlichkeiten zu den identifizierten latenten Klassen, sondern zu den latenten Klassenübergängen. Entropie-Werte größer als .80 indizieren dabei eine gute Klassifizierbarkeit (Clark & Muthén, 2009).

Ergebnisse

Gegenüber der ersten empirischen Erprobung des KKS soll die Darstellung der Ergebnisse deutlich verkürzt erfolgen. Zwar werden sämtliche relevanten Werte berichtet, jedoch nicht so ausführlich beschrieben (siehe dazu Abschnitt 2.4.2). Zunächst wird zu

jedem der beiden Zeitpunkte das Messmodell (also die Klassen-Lösung) für SKS überprüft.

Überprüfung des Messmodells für SKS zum Zeitpunkt 1

Für die Modus-Ponens-Reduktion werden Klassen-Lösungen für 1, 2 und 3 Klassen der vier Modus-Ponens-Aufgaben berechnet. Dabei weist die Ein-Klassen-Lösung (BIC = 165,82) den niedrigsten BIC-Wert auf (Zwei-Klassen-Lösung: BIC = 191,80; Drei-Klassen-Lösung: BIC = 218,05) und stellt damit die bestpassende Klassen-Lösung dar. Für die Prüfung des Modellfits der Ein-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein *p*-Wert für Chi-Quadrat von $p = .68$ bzw. für Cressie Read von ebenfalls $p = .68$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Ein-Klassen-Lösung. Für die eine identifizierte Klasse ergeben sich für die vier Modus-Ponens-Aufgaben Lösungswahrscheinlichkeiten zwischen .95 und .99. Da diese Lösungswahrscheinlichkeiten allesamt nahe eins liegen, kann die eine identifizierte Klasse als „Modus-Ponens-Löser“ interpretiert werden (vgl. Abschnitt 2.4.1.5). Eine (Modus-Ponens-)Reduktion der Stichprobe zum Zeitpunkt 1 ist demnach nicht nötig[55]. Die beabsichtigten Analysen werden daher an den Daten der kompletten Stichprobe ($N = 195$) durchgeführt. Für diese 195 Personen werden die Antworten auf die 16 Items des KKS mithilfe verschiedener LCMs analysiert, wobei wie schon bei der ersten empirischen Erprobung zwischen 2 und 7 Klassen vorgegeben werden (vgl. Abschnitt 2.4.1.5)[56]. Als bestpassendes Modell wird dasjenige mit dem niedrigsten BIC ausgewählt. Die BIC-Werte der berechneten Klassen-Lösungen sind in Tabelle 13 angegeben.

[55] Vermutlich wurde der Test lediglich von sehr motivierten Personen zweimal bearbeitet, die ein ernsthaftes Interesse an den Testergebnissen hatten. Dafür spricht auch, dass 193 der 195 getesteten Personen die Frage nach ernsthafter Testbearbeitung (zum Erhebungszeitpunkt 2) mit „ja“ beantworteten.

[56] Auf die Vorgabe einer Ein-Klassen-Lösung und damit eine Prüfung des Unabhängigkeitsmodells wird verzichtet, da es aufgrund der Ergebnisse der ersten empirischen Erprobung des KKS äußerst unwahrscheinlich ist, dass die 16 Items des KKS „nicht messen“ (vgl. Abschnitt 2.4.2).

Tabelle 13: BIC-Werte verschiedener Klassen-Lösungen der 16 Items des KKS zum Zeitpunkt 1 der Stabilitätsstudie

	Vorgegebene Klassenzahl					
	2	3	4	5	6	7
BIC:	2779,05	2652,15	2586,59	2620,50	2682,91	2730,75

Für die 16 Items des überarbeiteten KKS erweist sich demnach eine Vier-Klassen-Lösung als die bestpassende. Für die Prüfung des Modellfits der Vier-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein *p*-Wert für Chi-Quadrat von $p = .12$ bzw. für Cressie Read von $p = .09$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Vier-Klassen-Lösung. Die Treffsicherheiten liegen für die einzelnen Klassen zwischen .961 und .997 sowie für die gesamte Vier-Klassen-Lösung bei .975 und damit im hohen Bereich. Sie sind in Tabelle 14 zusammengefasst.

Tabelle 14: Geschätzte Klassengrößen (P_C) und Treffsicherheiten (T_C) der Vier-Klassen-Lösung der 16 Items des KKS zum Zeitpunkt 1 der Stabilitätsstudie

Klasse	geschätzte Klassengröße P_C	Treffsicherheit T_C
Unsophistizierte	$P_{C=1} = .20$	$T_1 = .961$
Unsophistizierte II	$P_{C=2} = .44$	$T_2 = .963$
Fortgeschrittene	$P_{C=3} = .20$	$T_3 = .993$
Sophistizierte	$P_{C=4} = .16$	$T_4 = .997$
Treffsicherheit der Vier-Klassen-Lösung:		$T = .975$

Anmerkung. Die Benennung der Klassen erfolgt auf Basis der Interpretation der klassenbedingten Lösungswahrscheinlichkeiten, welche später in Abbildung 4 (siehe dort) dargestellt sind.

Für den KKS kann zum Zeitpunkt 1 folglich von einer guten Passung der Vier-Klassen-Lösung bei hoher Treffsicherheit ausgegangen werden. Die Frage, ob die vier identifizierten Klassen im Sinne des erweiterten Stufen-Modells interpretiert werden können, soll erst nach Prüfung des Messmodells zum Zeitpunkt 2 beantwortet werden.

Überprüfung des Messmodells für SKS zum Zeitpunkt 2

Für die Modus-Ponens-Reduktion werden Klassen-Lösungen für 1, 2 und 3 Klassen der vier Modus-Ponens-Aufgaben berechnet. Dabei weist die Ein-Klassen-Lösung (BIC = 149,22) den niedrigsten BIC-Wert auf (Zwei-Klassen-Lösung: BIC = 171,94; Drei-Klassen-Lösung: BIC = 198,30) und stellt damit die bestpassende Klassen-Lösung dar. Für die Prüfung des Modellfits der Ein-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein *p*-Wert für Chi-Quadrat von $p = .22$ bzw. für Cressie Read von ebenfalls $p = .22$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Ein-Klassen-Lösung. Für die eine identifizierte Klasse ergeben sich für die vier Modus-Ponens-Aufgaben Lösungswahrscheinlichkeiten zwischen .97 und .99. Da diese Lösungswahrscheinlichkeiten allesamt nahe eins liegen, kann die eine identifizierte Klasse als „Modus-Ponens-Löser“ interpretiert werden (vgl. Abschnitt 2.4.1.5). Auch zum Zeitpunkt 2 der Stabilitätsstudie ist demnach keine Modus-Ponens-Reduktion der Stichprobe nötig. Die beabsichtigten Analysen werden also auch für den zweiten Erhebungszeitpunkt an den Daten der kompletten Stichprobe ($N = 195$) durchgeführt.

Auch für Erhebungszeitpunkt 2 werden die Antworten auf die 16 Items des KKS mithilfe verschiedener LCMs analysiert, wobei wiederum zwischen 2 und 7 Klassen vorgegeben werden. Als bestpassendes Modell wird dasjenige mit dem niedrigsten BIC ausgewählt. Die BIC-Werte der berechneten Klassen-Lösungen sind in Tabelle 15 angegeben.

Tabelle 15: BIC-Werte verschiedener Klassen-Lösungen der 16 Items des KKS zum Zeitpunkt 2 der Stabilitätsstudie

	Vorgegebene Klassenzahl					
	2	3	4	5	6	7
BIC:	2625,42	2378,57	2264,26	2289,51	2332,47	2399,37

Für die 16 Items des überarbeiteten KKS erweist sich demnach auch zum Zeitpunkt 2 eine Vier-Klassen-Lösung als die bestpassende. Für die Prüfung des Modellfits der Vier-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein *p*-Wert für Chi-Quadrat von $p = .11$ bzw. für Cressie Read von $p = .09$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher

festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Vier-Klassen-Lösung. Die Treffsicherheiten liegen für die einzelnen Klassen zwischen .987 und .991 sowie für die gesamte Vier-Klassen-Lösung bei .989 und damit wiederum im hohen Bereich. Sie sind in Tabelle 16 zusammengefasst.

Tabelle 16: Geschätzte Klassengrößen (P_C) und Treffsicherheiten (T_C) der Vier-Klassen-Lösung der 16 Items des KKS zum Zeitpunkt 2 der Stabilitätsstudie

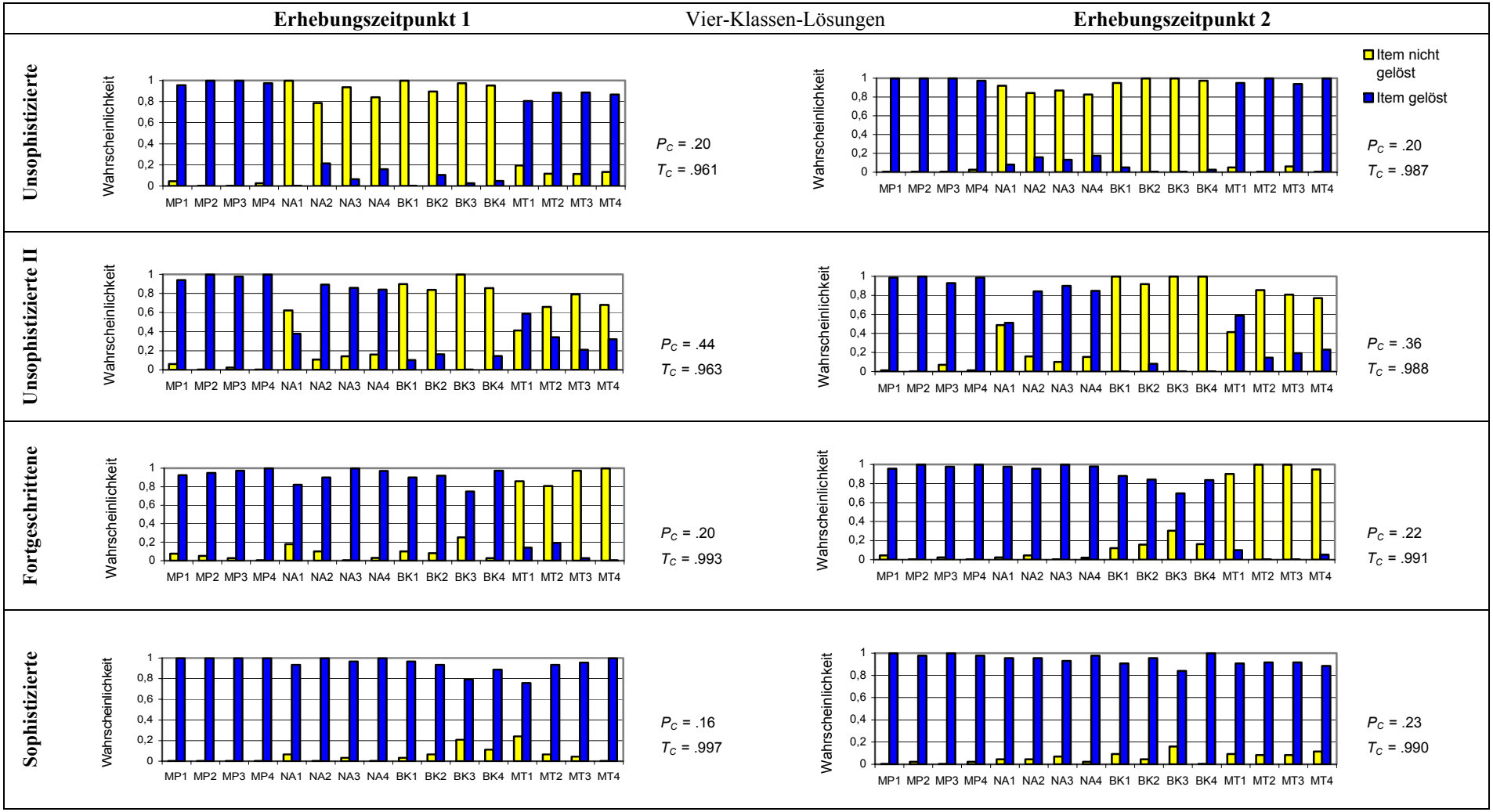
Klasse	geschätzte Klassengröße P_C	Treffsicherheit T_C
Unsophistizierte	$P_{C=1} = .19$	$T_1 = .987$
Unsophistizierte II	$P_{C=2} = .36$	$T_2 = .988$
Fortgeschrittene	$P_{C=3} = .22$	$T_3 = .991$
Sophistizierte	$P_{C=4} = .23$	$T_4 = .990$
Treffsicherheit der Vier-Klassen-Lösung:		$T = .989$

Anmerkung. Die Benennung der Klassen erfolgt auf Basis der Interpretation der klassenbedingten Lösungswahrscheinlichkeiten, welche später in Abbildung 4 (siehe dort) dargestellt sind.

Wie schon zum Zeitpunkt 1 kann auch zum Zeitpunkt 2 von einer guten Passung der Vier-Klassen-Lösung bei hoher Treffsicherheit ausgegangen werden. Es bleibt die Frage, ob die Klassen-Lösung zu beiden Erhebungszeitpunkten im Sinne des erweiterten Stufen-Modells interpretiert werden kann. Abbildung 4 zeigt die klassenbedingten Lösungswahrscheinlichkeiten der Vier-Klassen-Lösung zu beiden Erhebungszeitpunkten der Stabilitätsstudie. Dies illustriert zudem, welches Modell mit der LTA überprüft werden soll.

Da zu beiden Zeitpunkten die identifizierten vier Klassen das für die postulierten Sophistiziertheits-Stufen Konditionalen Schlussfolgerns charakteristische Muster der klassenbedingten Lösungswahrscheinlichkeiten aufweisen (vgl. Abbildung 4; für ausführliche Erläuterungen zur Augenscheinprüfung der Zuordnung des Musters der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Stufen des erweiterten Stufen-Modells siehe Abschnitt 2.4.2), kann die Vier-Klassen-Lösung (zu beiden Zeitpunkten) damit auch als gut interpretierbar gelten. Damit ist das letzte Kriterium für die Passung der Vier-Klassen-Lösung erfüllt. Somit spricht nichts gegen die Durchführung einer LTA in der zu Beginn dieses Abschnitts beschriebenen Weise[57].

[57] Das M*plus*-Inputfile der LTA findet sich im Anhang D dieser Arbeit.



Legende. P_C…geschätzte Klassengröße, T_C…Treffsicherheit, MP…Modus Ponens, NA…Negation des Antezedens, BK…Bestätigung der Konsequenz, MT…Modus Tollens, 1…keine Negation in der Hauptprämisse, 2…Negation in der Konsequenz, 3…Negation im Antezedens, 4…Negation in Antezedens und Konsequenz.

Abbildung 4: Klassenbedingte Lösungswahrscheinlichkeiten der Vier-Klassen-Lösung zu beiden Erhebungszeitpunkten der Stabilitätsstudie

Die Entropie *E* der berechneten LTA liegt bei $E = .95$ und damit nach Clark und Muthén (2009) im hohen Bereich. Personen lassen sich demnach durch diese LTA gut klassifizieren. Tabelle 17 zeigt als Ergebnis der LTA die Wahrscheinlichkeiten für die latenten Klassenübergänge[58].

Tabelle 17: Geschätzte Wahrscheinlichkeiten für die latenten Klassenübergänge zwischen den beiden Zeitpunkten der Stabilitätsstudie

		Zeitpunkt 2				
		Unsoph.	Unsoph. II	Fortg.	Soph.	Gesamt
Zeitpunkt 1	Unsophistizierte	**.120**	.059	.011	< .001	.190
	Unsophistizierte II	.058	**.321**	.050	.023	.452
	Fortgeschrittene	< .001	.027	**.136**	.045	.208
	Sophistizierte	.005	< .001	.013	**.132**	.150
Gesamt		.183	.407	.210	.201	1

Anmerkung. Stabilität-indizierende Wahrscheinlichkeiten sind fettgedruckt. Unsoph. … Unsophistizierte, Fortg. … Fortgeschrittene, Soph. … Sophistizierte.

Als Ergebnis der LTA zeigen sich die höchsten Wahrscheinlichkeiten (zwischen .120 und .321) in den Stabilität-indizierenden Zellen (gleiche Klassenzuordnung zum ersten wie zum zweiten Zeitpunkt). Die Wahrscheinlichkeiten in Zellen, die Veränderung indizieren (unterschiedliche Klassenzuordnungen zum ersten und zum zweiten Zeitpunkt), sind hingegen durchweg niedrig (allesamt < .06). Obwohl dies bereits auf den ersten Blick ersichtlich ist, scheint die Angabe konkreter Koeffizienten sinnvoll, um die Stabilität der Klassen-Lösung mit den Ergebnissen anderer Studien vergleichbar zu machen.

3.2.2.3 Koeffizientenbestimmung

Ein Koeffizient, der sich leicht aus den Wahrscheinlichkeiten für die latenten Klassenübergänge bilden lässt, ist die Wahrscheinlichkeit, zum zweiten Zeitpunkt der gleichen Klasse zugeordnet zu werden wie zum ersten Zeitpunkt. Hierzu brauchen einfach nur die Wahrscheinlichkeiten der Zellen der Hauptdiagonale (in Tabelle 17) aufsummiert werden, da diese Zellen Stabilität indizieren. In dieser Studie ergibt sich dabei für diese

[58] Die Unterschiede zwischen den geschätzten Klassengrößen (P_C) der mit Winmira 2001 berechneten LCAs und der mit M*plus* berechneten LTA resultieren aus den unterschiedlichen Schätzalgorithmen beider Programme.

Wahrscheinlichkeit ein Wert von .71, der durchaus als hoch zu werten ist. Dies gilt insbesondere, wenn man die Idee der „stayer“ and „mover“ (vgl. z.B. Eid & Langeheine, 2003, 2007) zugrunde legt, nach der bereits dann von einer Eigenschaft und damit einem vergleichsweise stabilen Merkmal gesprochen wird, wenn der Anteil der „stayer“ größer als der Anteil der „mover“ ist, also bei Werten größer als .50 für die Summe der Wahrscheinlichkeiten der Stabilität-indizierenden Zellen. Zusätzlich kann ein solcher Stabilitätskoeffizient nicht nur über die Klassen hinweg betrachtet werden, sondern auch für jede Klasse einzeln. Tabelle 18 zeigt die (geschätzten) bedingten Wahrscheinlichkeiten der Klassenzuordnung zum Zeitpunkt 2 gegeben der Klassenzuordnung zum Zeitpunkt 1.

Tabelle 18: Bedingte Wahrscheinlichkeiten der Klassenzuordnung zum Zeitpunkt 2 gegeben der Klassenzuordnung zum Zeitpunkt 1

		Zeitpunkt 2			
		Unsoph.	Unsoph. II	Fortg.	Soph.
Zeitpunkt 1	Unsophistizierte	**.632**	.310	.058	< .001
	Unsophistizierte II	.128	**.710**	.111	.051
	Fortgeschrittene	< .001	.130	**.654**	.216
	Sophistizierte	.033	< .001	.087	**.880**

Anmerkung. Stabilität-indizierende bedingte Wahrscheinlichkeiten sind fettgedruckt. Unsoph. … Unsophistizierte, Fortg. … Fortgeschrittene, Soph. … Sophistizierte.

Es zeigt sich, dass die bedingte Wahrscheinlichkeit, zum Zeitpunkt 2 einer Klasse zugeordnet zu werden, gegeben ihr bereits zum Zeitpunkt 1 zugeordnet worden zu sein, für alle vier identifizierten Klassen jeweils größer als .50 ist. Demnach kann nicht nur SKS insgesamt, sondern auch jede der vier Klassen als stabil angenommen werden. Die Ergebnisse der durchgeführten LTA sowie die daraus berechneten (Stabilitäts-)Koeffizienten sprechen also für die Stabilität der Klassen-Lösung (sowohl insgesamt als auch in Bezug auf jede einzelne Klasse) und damit für die Stabilität des gemessenen Merkmals. Es kann daher davon ausgegangen werden, dass diese Voraussetzung für die Betrachtung einer Test-Retest-Korrelation als Schätzer für Reliabilität erfüllt ist.

Bei kontinuierlichen Variablen wird die Test-Retest-Korrelation zur Reliabilitätsbestimmung üblicherweise als *Produkt-Moment-Korrelation* berechnet. Meist wird dabei ab einer Höhe dieses Reliabilitäts-Schätzers von $r = .70$ von ausreichend hoher

Messgenauigkeit ausgegangen (z.B. Moosbrugger & Kelava, 2007; Schermelleh-Engel & Werner, 2007). Die Bestimmung eines vergleichbaren (Korrelations-)Koeffizienten ist für nominale latente Variablen (wie SKS) zunächst nicht möglich. Ordnet man allerdings theoriegeleitet die identifizierten Klassen – wie bereits bei der ersten empirischen Erprobung – ergibt sich eine manifeste ordinale Variable SKS (vgl. Abschnitt 2.4.1.5). Für zwei so gebildete ordinale Variablen (SKS zum Zeitpunkt 1 und zum Zeitpunkt 2) kann nun zumindest eine Rangkorrelation berechnet werden. Aufgrund der hohen Treffsicherheit des LCMs zu beiden Zeitpunkten der Stabilitätsstudie (vgl. Abschnitt 3.2.2.2) scheint die Verwendung dieser manifesten Variablen unproblematisch, da das Messfehlerproblem als vergleichsweise gering angenommen werden kann. Die Ordnung der identifizierten Klassen wird theoriegeleitet wie folgt festgelegt:

1 (Rang 1) … Unsophistizierte
2 (Rang 2) … Unsophistizierte II
3 (Rang 3) … Fortgeschrittene
4 (Rang 4) … Sophistizierte

Für die Rangkorrelation der manifesten SKS-Variablen zum Zeitpunkt 1 und zum Zeitpunkt 2 ergeben sich die in Tabelle 19 dargestellten Ergebnisse.

Tabelle 19: Rangkorrelationen der manifesten Variable SKS zum Zeitpunkt 1 (SKS_{man_t1}) und zum Zeitpunkt 2 (SKS_{man_t2})

Variablen	Spearmans Rho (p-Wert)	Kendalls Tau (p-Wert)
SKS_{man_t1} und SKS_{man_t2}	.767 ($p < .001$)	.713 ($p < .001$)

Beide Rangkorrelationskoeffizienten liegen damit oberhalb der beschriebenen Grenze von .70, was für eine ausreichend hohe Messgenauigkeit des KKS bei wiederholter Testvorgabe spricht.

3.2.3 Zusammenfassende Bewertung der Messgenauigkeit des KKS

Insgesamt bleibt nach dieser zweiten empirischen Studie zum KKS dreierlei festzuhalten. Zum ersten scheinen die Modifikationen (siehe Abschnitt 2.4.3.2) gegenüber der ersten Version des KKS den gewünschten Effekt zu haben, dass tatsächlich keine „Restklasse“ mehr modelliert werden muss, um eine adäquate Modellpassung zu errei-

chen. Zu beiden Messzeitpunkten ist eine Vier-Klassen-Lösung das bestpassende Messmodell. Zum zweiten sprechen die Ergebnisse dafür, dass es sich bei SKS um ein über die Zeit stabiles Merkmal handelt. Zum dritten schließlich bleibt festzuhalten, dass beide vorgestellte Methoden zur Bestimmung der Messgenauigkeit des KKS zu ähnlichen Ergebnissen führen. Jeweils kann davon ausgegangen werden, dass es sich beim KKS um einen messgenauen und damit zuverlässigen Test handelt (für eine Diskussion der empirischen Ergebnisse zur Messgenauigkeit siehe Abschnitt 4.1.2). Offen bleibt jedoch, ob mit dem KKS tatsächlich Sophistiziertheit Konditionalen Schlussfolgerns im theoretisch postulierten Sinne gemessen wird. Dies wird auf Basis theoretischer Überlegungen sowie empirischer Studien im folgenden Abschnitt unter der Überschrift „Validität“ untersucht.

3.3 Validität

Die Frage nach der Validität eines Tests hat sich aus der grundlegenden Frage entwickelt, ob man beim Messen wirklich das misst, was man messen will (z.B. Cattell, 1946; Kelley, 1927). Bereits 1927 bezeichnete Kelley einen Test dann als valide, wenn er das misst, was er vorgibt zu messen. Ausgehend von dieser pragmatischen Beschreibung wird Validität – zumindest in der Testkonstruktionspraxis – als das wichtigste Gütekriterium eines Test bezeichnet (z.B. Rost, 2004; Schermelleh-Engel, Kelava & Moosbrugger, 2006). Allerdings handelt es sich bei Validität nicht um einen wohldefinierten Begriff (wie bspw. bei Reliabilität; vgl. Abschnitt 3.2). Die häufig verwendete KTT-basierte Definition als Korrelation $Kor(Y_i, K)$ der betrachteten Testwertvariable Y_i mit einem Kriterium K führt zwangsläufig zu dem Problem, dass das Kriterium K definiert werden muss und dabei keineswegs sicher ist, dass das Kriterium K seinerseits als valide bezeichnet werden kann. Es stellt sich die Frage, wie mit dieser Diskrepanz zwischen Wichtigkeit für die Testkonstruktionspraxis einerseits und fehlender Definiertheit andererseits umgegangen werden kann. Dazu sollen zunächst zwei aktuelle Konzeptionen von Validität skizziert werden. Als „state of the art“ gilt derzeit die Validitätskonzeption von Messick (1989), im Rahmen derer Validität als integriertes evaluatives Urteil über das Ausmaß bezeichnet wird, in dem empirische Evidenz und theoretische Überlegungen für die Angemessenheit von Maßnahmen sprechen, die auf den Testwerten basieren. Im Mittelpunkt der (Validitäts-)Betrachtungen stehen bei Messick

(1989) nicht die Testwerte selbst, sondern deren Interpretation sowie daraus abgeleitete Maßnahmen. Die drei klassischen Validitätstypen (*Inhalts*-, *Konstrukt*- und *Kriteriums-validität*) subsumiert Messick (1994) ebenso unter dem Begriff „Konstruktvalidität" wie auch soziale Konsequenzen des Testens. An Messicks (1989, 1994, 1995) „vereinheitlichter Validität" kritisieren Borsboom und Mellenbergh (2007; siehe auch Borsboom, van Heerden & Mellenbergh, 2003 sowie Borsboom, Mellenbergh & van Heerden, 2004) unter anderem, dass es schwer – wenn nicht unmöglich – ist, alle diese Aspekte zu einem integrativen Urteil über die Validität eines Tests zu aggregieren. Sie schlagen stattdessen eine einfachere Definition von Validität vor, nach der ein Test dann valide für die Messung eines theoretischen Konstruktes ist, wenn aus der Variation des Konstruktes eine Variation der Testwerte folgt, und zwar durch die vom Test hervorgerufenen Antwort- bzw. Reaktionsprozesse (Borsboom & Mellenbergh, 2007). Während nach dieser Definition Validität an sich kein komplexes Gütekriterium ist (wie in der Testkonstruktionspraxis häufig behauptet wird), können die Designs und Methoden, mit denen Validität (empirisch) überprüft wird, hingegen durchaus komplex sein.

Eines haben beide Konzeptionen von Validität gemeinsam: Unabhängig davon, ob der Testwert selbst (Borsboom) oder die daraus abgeleiteten Interpretationen und Maßnahmen (Messick) betrachtet werden, die Bewertung der Validität erfolgt auf Basis theoretischer Argumentation sowie empirischer Befunde. Dieser Prozess des theoretischen Argumentierens und empirischen Überprüfens wird als *Validierung* bezeichnet. In Anlehnung an Borsboom und Mellenbergh (2007) werden in der vorliegenden Arbeit die Aktivitäten zur Bestimmung der Validität des KKS unter dem Oberbegriff *Validierung* zusammengefasst. Diese Validierung erfolgt in Anlehnung an die Testkonstruktionspraxis bezüglich des (Item-)Inhaltes, des Konstruktes selbst sowie hinsichtlich verschiedener Kriterien. Dabei werden jeweils theoretische wie empirische Indizien dafür (oder ggf. dagegen) gesammelt, dass der KKS das misst, was er messen soll, nämlich die Sophistiziertheit Konditionalen Schlussfolgerns, also die Kompetenz (vgl. Abschnitt 2.2.2), logisch korrekte Schlussfolgerungen bei Konditionalaussagen zu ziehen. Zum Abschluss dieses Abschnitts wird ein integratives Gesamturteil versucht, indem die theoretischen Überlegungen und empirischen Befunde zur Validierung zusammengefasst werden.

3.3.1 Inhaltsvalidierung

Eine Möglichkeit der Inhaltsvalidierung eines Tests ist, Items sowie zugehörige Antwortoptionen von verschiedenen Experten hinsichtlich der Erfassung des theoretischen Konstruktes, das der Test messen soll, bewerten zu lassen. Bei Tests, deren Items und Antwortoptionen schlüssig aus kognitionspsychologischen Theorien hergeleitet sind, wird darauf jedoch häufig verzichtet, da bereits aufgrund dieser Herleitung (aus kognitionspsychologischen Theorien) von Inhaltsvalidität sensu Klauer (1978, 1984) ausgegangen werden kann (Beckmann & Guthke, 1999). Dies gilt auch für den KKS. Die 16 Items entsprechen strukturell dem in der Kognitionspsychologie etablierten Negationsparadigma (siehe Abschnitt 2.3.2.1), dessen Verwendung zur Untersuchung von kognitiven Prozessen beim Konditionalen Schlussfolgerns gegenwärtig unstrittig ist (z.B. Evans et al., 2007; Oberauer et al., 2007). Ebenso lässt sich die erschöpfende Menge an Antwortalternativen pro Item (Multiple-Choice-Generierungsparadigma) schlüssig theoretisch herleiten (vgl. Abschnitt 2.4.1.2). Einzig der Item-Inhalt „Funktionieren elektrischer Schaltungen“ stellt kein Ergebnis theoretischer Herleitungen dar, sondern ist Resultat pragmatischer Anforderungen an den KKS (vgl. Abschnitt 2.3.3). Die Frage nach der Übertragbarkeit der Ergebnisse des KKS auf andere Item-Inhalte bleibt vorerst unbeantwortet, wird aber im Diskussionsteil dieser Arbeit wieder aufgegriffen (siehe dazu Abschnitte 4.1.1 und 4.3.1). Eine zusätzliche Inhaltsvalidierung scheint für den KKS insgesamt unnötig. Die Items und zugehörigen Antwortoptionen können aufgrund der Herleitung aus kognitionspsychologischen Theorien als valide angenommen werden.

3.3.2 Konstruktvalidierung

Beckmann und Guthke (1999) kritisieren, dass die Konstruktvalidierung von Tests, die auf kognitionspsychologischen Theorien basieren, häufig unzureichend ist. Möglicherweise wird aufgrund der als valide angenommenen Items solcher Tests (vgl. Abschnitt 3.3.1) oft auf eine umfangreiche Konstruktvalidierung verzichtet. Um diesem Umstand Rechnung zu tragen, wird für den KKS insbesondere auf die Konstruktvalidierung sehr großer Wert gelegt. Dazu soll in der Folge theoretisch argumentiert und empirisch überprüft werden, ob aus dem Testergebnis des KKS auf das Konstrukt Sophistiziertheit Konditionalen Schlussfolgerns im theoretisch postulierten Sinne geschlossen werden

kann. Eine derartige Konstruktvalidierung kann sowohl *strukturprüfend* als auch *strukturtursuchend* erfolgen. Beide Vorgehensweisen werden im Falle des KKS umgesetzt.

3.3.2.1 Strukturprüfende Konstruktvalidierung

Bei der strukturprüfenden Konstruktvalidierung werden strukturelle Annahmen zur Beziehung zwischen den Items und der mit ihnen gemessenen latenten Variable inferenzstatistisch überprüft. Klassischerweise erfolgt dies in Form der Überprüfung von Messmodellen, wie sie bei der klassischen Strukturgleichungsmodellierung (z.B. Rose, Pohl, Böhme & Steyer, im Druck) üblich sind oder auch durch Überprüfung von Modellen der IRT (siehe z.B. Hambleton & Swaminathan, 1985). Im Falle des KKS erfolgt dies durch Überprüfung der zentralen Annahmen des erweiterten Stufen-Modells (siehe Abschnitt 2.3.2.2) für die 16 Items des KKS. Konkret wird also die Passung eines LCMs unter Vorgabe von vier latenten Klassen überprüft, wobei die Klassen im Sinne des erweiterten Stufen-Modells interpretierbar sein sollen. So empfehlen bspw. auch Borsboom und Mellenbergh (2007) die LCA prinzipiell als Verfahren zur Validierung von Stufen-Modellen. Die strukturprüfende Konstruktvalidierung des KKS besteht also in dem Vorgehen, dass bereits bei der ersten empirischen Erprobung des KKS ausführlich beschrieben wird (vgl. Abschnitt 2.4.1.5). Dort führt diese Prüfung zu weitgehend hypothesenkonformen Ergebnissen (vgl. Abschnitt 2.4.2) mit der Einschränkung, dass ein Fünf-Klassen-Modell mit einer zusätzlichen „Restklasse“ am besten auf die Daten passt. Die vier Klassen mit den höchsten geschätzten Klassengrößen sowie höchsten Treffsicherheiten können allerdings im Sinne des erweiterten Stufen-Modells interpretiert werden. Die überarbeitete Version des KKS (siehe Abschnitt 2.4.3.2 für einen Überblick über die Änderungen) wird in der Stabilitätsstudie (siehe Abschnitt 3.2.2.2) einer ersten empirischen Erprobung unterzogen. Die Ergebnisse stützen die strukturellen Annahmen für die überarbeitete Version des KKS: Ein Vier-Klassen-Modell passt am besten auf die Daten, weist dabei eine adäquate Modellpassung auf und die vier Klassen lassen sich − bei hoher Treffsicherheit − entsprechend den Annahmen des erweiterten Stufen-Modells interpretieren.

Der eingangs dieses Abschnitts erwähnten besonderen Bedeutung der Konstruktvalidierung soll auch dadurch Rechnung getragen werden, dass eigens dafür eine weitere Studie durchgeführt wird. Diese Studie wird im Folgenden als *Konstruktvalidierungsstudie* bezeichnet. Für die überarbeitete Version des KKS wird also die aufgrund der Annahmen des erweiterten Stufen-Modells postulierte Struktur überprüft. Es folgen

einige kurze Ausführungen zu verwendeten Erhebungsinstrumenten, Untersuchungsstichprobe, Durchführung, Auswertungsmethoden sowie Ergebnissen der Konstruktvalidierungsstudie.

Erhebungsinstrumente

Für diesen Abschnitt der Arbeit ist zunächst nur relevant, dass die überarbeitete Version des KKS (siehe Anhang A) eines der verwendeten Erhebungsinstrumente war. Ausführungen zu anderen verwendeten Testverfahren erfolgen zu einem späteren Zeitpunkt (Abschnitt *Struktursuchende Konstruktvalidierung*), zu dem ihre theoretische Einbettung besser möglich ist.

Durchführung

Der KKS wurde auch im Rahmen der Konstruktvalidierungsstudie in Form eines Online-Tests dargeboten. Allerdings erfolgte diese Online-Testung insoweit kontrolliert, dass jeweils Gruppen zu ca. 25 Personen den Test in einem Computer-Pool online bearbeiteten. Die Probanden folgten den Anweisungen auf dem Bildschirm. Die anwesende Testleiterin stand für Rückfragen zur Verfügung, wurde jedoch in keiner der insgesamt sieben Sitzungen in Anspruch genommen, die zwischen November 2007 und Juni 2008 stattfanden.

Untersuchungsstichprobe

Die Stichprobe der Konstruktvalidierungsstudie bestand aus 154 Studierenden, von denen 72,7% im Hauptfach sowie 24,7% im Nebenfach Psychologie an der Friedrich-Schiller-Universität Jena studierten[59], die meisten von ihnen im 1. Semester. Anreiz zur Teilnahme bildete die Verlosung von insgesamt sechs Gutscheinen im Wert von jeweils 10 Euro, die Vergabe von Versuchspersonenstunden sowie die Möglichkeit, nach Auswertung der Daten ein anonymes, rechnergestütztes Feedbacks zu den Ergebnissen in einzelnen Leistungstests zu kognitiven Fähigkeiten zu erhalten. Sämtliche Personen machten soziodemographische Angaben zur eigenen Person, die im Folgenden kurz zusammengefasst werden.

85,1% der Teilnehmer waren Frauen, entsprechend 14,9% Männer. Diese Überrepräsentation von Frauen entspricht dem typischen Geschlechterverhältnis des Studiengangs Psychologie. Das Durchschnittsalter lag bei 21 Jahren (*Sd* = 2,9 Jahre) und ist

[59] 2,6% gaben „Sonstiges“ als Studiengang an.

ebenfalls charakteristisch für Studierende der Psychologie früher Semester. Die meisten der Studierenden stammten ursprünglich aus Thüringen (33,1%) und Sachsen (14,9%). Der prozentuale Anteil Studierender aus anderen Bundesländern lag jeweils im einstelligen Bereich.

Auswertungsmethoden

Zur Sicherung der Datenqualität wird zunächst die Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5) durchgeführt. Die Strukturprüfung erfolgt durch Beurteilung der Modellpassung eines LCMs unter Vorgabe von vier Klassen bei gleichzeitig guter Interpretierbarkeit dieser vier Klassen im Sinne des erweiterten Stufen-Modells. Analog zum Vorgehen bei der Stabilitätsstudie (siehe Abschnitt 3.2.2.2) wird postuliert, dass die Vier-Klassen-Lösung den niedrigsten BIC-Wert aller getesteten Klassen-Lösungen aufweist und gleichzeitig die Kriterien für einen guten Modellfit bei der Durchführung eines parametrischen Bootstrapverfahrens erfüllt. Die Interpretation der Klassen erfolgt aufgrund der Zuordenbarkeit des Musters der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Klassen des erweiterten Stufen-Modells. Es wird wiederum davon ausgegangen, dass diese Zuordnung bereits per Augenschein möglich ist. Schließlich soll noch die Messgenauigkeit der Vier-Klassen-Lösung durch Beurteilung der Treffsicherheiten bewertet werden (vgl. Abschnitt 3.2.1). Die LCA wird mithilfe der Software Winmira 2001 (Davier, 2001) berechnet. Da es sich um die gleichen Auswertungsmethoden wie bei der ersten empirischen Erprobung des KKS handelt, wird auf ausführlichere Beschreibungen verzichtet (siehe dafür Abschnitt 2.4.1.5).

Ergebnisse

Für die Modus-Ponens-Reduktion werden Klassen-Lösungen für 1, 2 und 3 Klassen der vier Modus-Ponens-Aufgaben berechnet. Dabei weist die Ein-Klassen-Lösung (BIC = 111,28) den niedrigsten BIC-Wert auf (Zwei-Klassen-Lösung: BIC = 136,28; Drei-Klassen-Lösung: BIC = 161,44) und stellt damit die bestpassende Klassen-Lösung dar. Für die Prüfung des Modellfits der Ein-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein p-Wert für Chi-Quadrat von $p = .44$ bzw. für Cressie Read von ebenfalls $p = .44$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Ein-Klassen-Lösung. Für die eine identifizierte Klasse ergeben sich für die vier Modus-Ponens-Aufgaben Lösungswahrschein-

lichkeiten zwischen .95 und .99. Da diese Lösungswahrscheinlichkeiten allesamt nahe eins liegen, kann die eine identifizierte Klasse als „Modus-Ponens-Löser“ interpretiert werden (vgl. Abschnitt 2.4.1.5). Eine (Modus-Ponens-)Reduktion der Stichprobe der Konstruktvalidierungsstudie ist demnach nicht nötig[60]. Die beabsichtigten Analysen werden daher an den Daten der kompletten Stichprobe ($N = 154$) durchgeführt.

Für diese 154 Personen werden die Antworten auf die 16 Items des KKS mithilfe verschiedener LCMs analysiert, wobei wie schon bei der ersten empirischen Erprobung zwischen 2 und 7 Klassen vorgegeben werden (vgl. Abschnitt 2.4.1.5). Als bestpassendes Modell wird dasjenige mit dem niedrigsten BIC ausgewählt. Die BIC-Werte der berechneten Klassen-Lösungen sind in Tabelle 20 angegeben.

Tabelle 20: BIC-Werte verschiedener Klassen-Lösungen der 16 Items des KKS in der Konstruktvalidierungsstudie

	Vorgegebene Klassenzahl					
	2	3	4	5	6	7
BIC:	2176,04	2089,37	2079,09	2133,96	2195,45	2270,40

Für die 16 Items des überarbeiteten KKS erweist sich demnach eine Vier-Klassen-Lösung als die bestpassende. Für die Prüfung des Modellfits der Vier-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein p-Wert für Chi-Quadrat von $p = .20$ bzw. für Cressie Read von $p = .12$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Vier-Klassen-Lösung. Die Treffsicherheiten liegen für die einzelnen Klassen zwischen .946 und .993 sowie für die gesamte Vier-Klassen-Lösung bei .957 und damit im hohen Bereich. Sie sind in Tabelle 21 zusammengefasst.

[60] Wahrscheinlich resultiert die durchweg ernsthafte Testbearbeitung aus der vergleichsweise kontrollierten Untersuchungssituation (vgl. Abschnitt zur Durchführung). Dafür spricht auch, dass keine der 154 getesteten Personen die Frage nach ernsthafter Testbearbeitung mit „nein“ beantwortete.

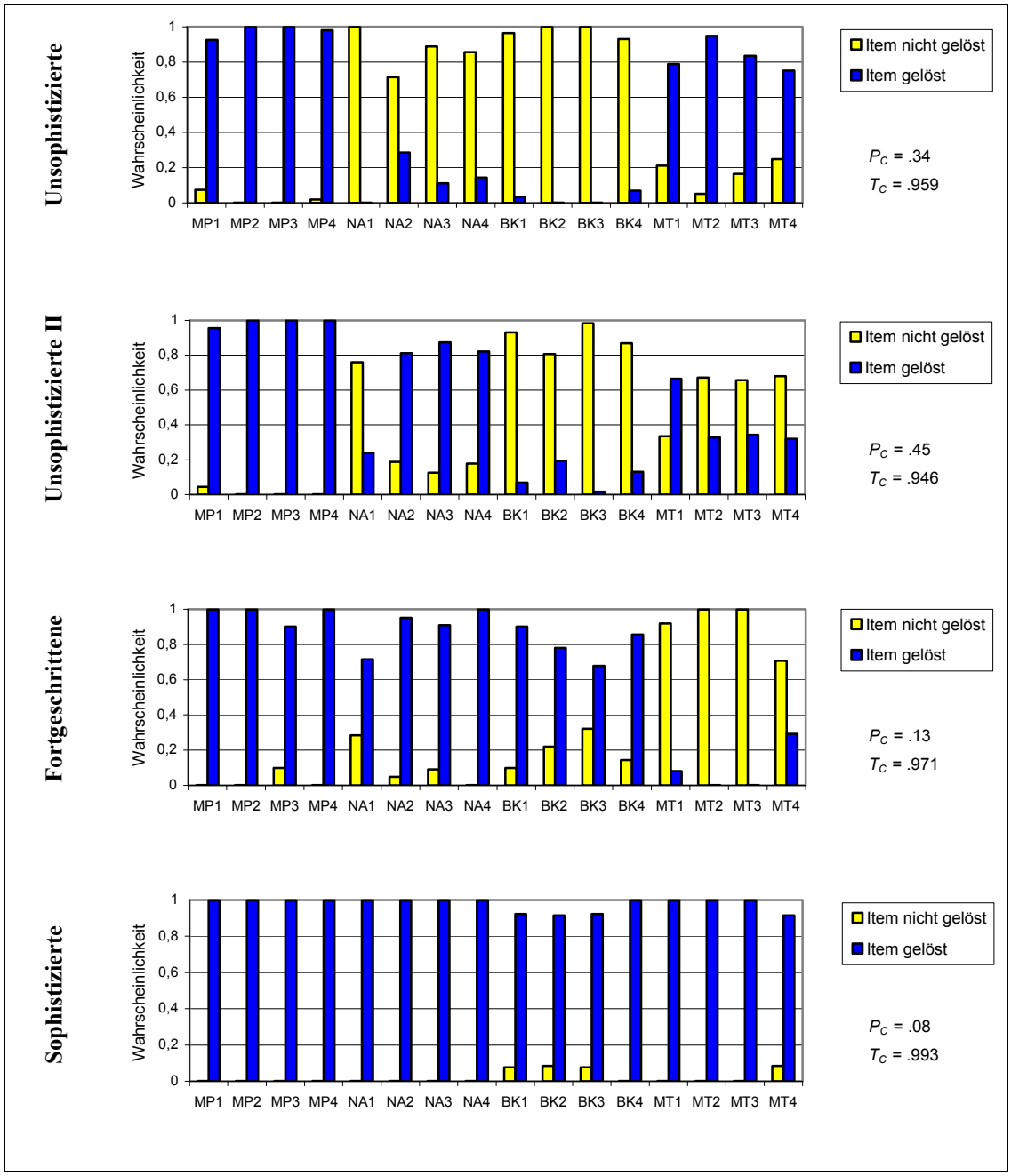
Tabelle 21: Geschätzte Klassengrößen (P_C) und Treffsicherheiten (T_C) der Vier-Klassen-Lösung der 16 Items des KKS in der Konstruktvalidierungsstudie

Klasse	geschätzte Klassengröße P_C	Treffsicherheit T_C
Unsophistizierte	$P_{C=1} = .34$	$T_1 = .959$
Unsophistizierte II	$P_{C=2} = .45$	$T_2 = .946$
Fortgeschrittene	$P_{C=3} = .13$	$T_3 = .971$
Sophistizierte	$P_{C=4} = .08$	$T_4 = .993$
Treffsicherheit der Vier-Klassen-Lösung:		$T = .957$

Anmerkung. Die Benennung der Klassen erfolgt auf Basis der Interpretation der klassenbedingten Lösungswahrscheinlichkeiten, welche später in Abbildung 5 (siehe dort) dargestellt sind.

Für den KKS kann in der Konstruktvalidierungsstudie folglich von einer guten Passung der Vier-Klassen-Lösung bei hoher Treffsicherheit ausgegangen werden. Es bleibt die Frage, ob die vier identifizierten Klassen im Sinne des erweiterten Stufen-Modells interpretiert werden können. Abbildung 5 zeigt die Antwortwahrscheinlichkeiten innerhalb der vier identifizierten Klassen. Diese weisen das für die postulierten Sophistiziertheits-Stufen Konditionalen Schlussfolgerns charakteristische Muster der klassenbedingten Lösungswahrscheinlichkeiten auf (vgl. Abbildung 5; für ausführliche Erläuterungen zur Augenscheinprüfung der Zuordnung des Musters der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Stufen des erweiterten Stufen-Modells siehe Abschnitt 2.4.2). Damit kann die Vier-Klassen-Lösung auch als gut interpretierbar gelten. Es braucht daher keine der Annahmen der strukturprüfenden Konstruktvalidierung aufgrund der Daten verworfen zu werden.

Eine Strukturprüfung in dieser Form ist für die überarbeitete Version des KKS auch im Rahmen der Stabilitätsstudie erfolgt (vgl. Abschnitt 3.2.2.2). Auch die Stabilitätsstudie liefert demnach Ergebnisse, die zur (strukturprüfenden) Konstruktvalidierung dienen können und die ebenfalls in postulierter Weise ausfallen.



Legende. P_C…geschätzte Klassengröße, T_C…Treffsicherheit, MP…Modus Ponens, NA…Negation des Antezedens, BK…Bestätigung der Konsequenz, MT…Modus Tollens, 1…keine Negation in der Hauptprämisse, 2…Negation in der Konsequenz, 3…Negation im Antezedens, 4…Negation in Antezedens und Konsequenz.

Abbildung 5: Klassenbedingte Lösungswahrscheinlichkeiten der Vier-Klassen-Lösung für die Stichprobe der Konstruktvalidierungsstudie

3.3.2.2 Struktursuchende Konstruktvalidierung

Bei der struktursuchenden Konstruktvalidierung kann bspw. *explorativ* nach Strukturen innerhalb des betrachteten Tests gesucht werden (z.B. Moosbrugger & Kelava, 2007). Aufgrund der klaren Annahmen über die Struktur des KKS und deren bereits erfolgter Überprüfung (vgl. Abschnitt 3.3.2.1) ist dies jedoch unnötig. Viel angebrachter scheint die ebenfalls der struktursuchenden Konstruktvalidierung zugeordnete Einordnung des gemessenen Merkmals in ein Gefüge weiterer theoretischer Konstrukte, ein sog. *nomologisches Netzwerk* (Cronbach & Meehl, 1955). Hierbei wird davon ausgegangen, dass das gemessene Merkmal mit den Ergebnissen anderer Tests, die das gleiche oder ein ähnliches Merkmal messen (sollen), hoch zusammenhängt. Ist dies der Fall, spricht man von *konvergenter Validität* des gemessenen Merkmals. Gleichzeitig sollte das gemessene Merkmal mit den Ergebnissen von Tests, die ein diskriminierbares eigenständiges Konstrukt messen, kaum oder gar nicht zusammenhängen. Bestätigt sich dies, wird von *diskriminanter* oder *divergenter Validität* gesprochen. Beide Konzepte sind bspw. auch im Rahmen sog. *Multitrait-Multimethod-Analysen* (Campbell & Fiske, 1959; siehe auch Eid, 2000 sowie Pohl & Steyer, in press) von zentraler Bedeutung. Die Einordnung in ein nomologisches Netzwerk auf Basis betrachteter Zusammenhänge mit den Ergebnissen anderer Tests ist seit einigen Jahren in der Testkonstruktionspraxis häufig *die* Methode zur Konstruktvalidierung (Hartig, Frey & Jude, 2007). Sie wird daher auch im Rahmen dieser Arbeit verfolgt, wenngleich sensu Borsboom und Mellenbergh (2007) den Ergebnissen der strukturprüfenden Konstruktvalidierung (siehe Abschnitt 3.3.2.1) deutlich größeres Gewicht beigemessen werden sollte.

Zunächst stellt sich also die Frage, mit welchen Konstrukten die mit dem KKS gemessene Sophistiziertheit Konditionalen Schlussfolgerns theoretisch zusammenhängen sollte und mit welchen nicht. Natürlich sollte SKS mit den Ergebnissen anderer Tests zum Konditionalen Schlussfolgern zusammenhängen. Doch wie bereits ausgeführt, existieren keine etablierten, kognitionspsychologisch fundierten Testverfahren zur Erfassung dieses Konstruktes im Erwachsenenbereich (vgl. Abschnitt 2.1.4.2). So bleibt lediglich die Möglichkeit, Zusammenhänge zwischen SKS und anderen Tests zu betrachten, die Merkmale erfassen, mit denen theoretisch ein Zusammenhang bestehen sollte. Ein erster solcher Zusammenhang ist der zwischen SKS und Reasoning (zur theoretischen Herleitung siehe Abschnitte 2.2.2 und 2.3.2.2; zu den Ergebnissen der empirischen Überprüfung siehe Abschnitt 2.4.2). Die Replizierbarkeit dieses Zusammenhangs stellt eine der Forschungsfragen dar, die sich in der Diskussion der ersten

empirischen Erprobung des KKS ergeben haben (vgl. Abschnitt 2.4.3.3). Reasoning ist damit das erste Konstrukt, das in das nomologische Netzwerk aufgenommen wird. Wiederum soll es durch Matrizenitems erhoben werden, da diese als bester Indikator für Reasoning gelten (Carpenter et al., 1990; Carroll, 1993; vgl. auch Abschnitt 2.4.1.2). Wie jedoch in der Diskussion der ersten empirischen Erprobung bereits ausgeführt, soll anstelle der zunächst verwendeten 10 Matrizenitems, die mit dem Programm ITEMGENERATOR (Ihme, 2007) regelgeleitet generiert wurden, ein etablierter Matrizentest verwendet werden (vgl. Abschnitt 2.4.3.1).

Für die Einordnung des gemessenen Merkmals in ein nomologisches Netzwerk ist es üblich, nicht nur Voraussagen darüber zu treffen, ob ein Zusammenhang mit anderen Konstrukten des nomologischen Netzwerkes vorliegt, sondern auch in welcher Höhe Effekte zu erwarten sind (z.B. Hartig et al., 2007). Sowohl die bisherigen empirischen Ergebnisse für die erste Version des KKS (siehe Abschnitt 2.4.2) als auch die Befunde von Stanovich zum Zusammenhang allgemeinpsychologischer Aufgaben mit klassischen Intelligenztests (zusammenfassend Stanovich, 1999; Stanovich & West, 2000) legen nahe, dass die Stärke des Zusammenhangs zwischen SKS und dem etablierten Matrizentest im mittleren Bereich liegt. Dies ist Hypothese 1 der strukturensuchenden Konstruktvalidierung.

Hypothese Konstruktvalidierung_1: SKS hängt positiv mit Reasoning (indiziert durch einen etablierten Matrizentest) zusammen. Der Effekt liegt im mittleren Bereich.

Es stellt sich nun die Frage nach weiteren Konstrukten (und damit nach entsprechenden Tests), die als konvergent bzw. diskriminant angenommen werden können. Durch die theoretische Ordnung der Sophistiziertheits-Stufen, die Betrachtung der geordneten Stufen als kognitive Leistung (vgl. Abschnitte 2.2.2 und 2.3.2.2) und den postulierten Zusammenhang mit Reasoning (siehe oben) würde sich prinzipiell ein intelligenzbezogenes nomologisches Netzwerk für die strukturensuchende Konstruktvalidierung anbieten. Ausgangspunkt soll dabei ein zentraler Punkt der kognitionspsychologischen Fundierung der Testentwicklung sein. Demnach wird unabhängig von der zugrundegelegten kognitionspsychologischen Theorie (Modelltheorie oder Logiktheorie; siehe Abschnitt 2.1.2) Arbeitsgedächtniskapazität als limitierend für (korrektes) Konditionales Schlussfolgern angenommen (zur theoretischen Begründung siehe Abschnitte 2.1.2.1 und 2.1.2.2). Im Kontext der Intelligenzstrukturforschung weist das Konstrukt Verarbeitungskapazität konzeptuell hochgradige Überlappungen sowie empirisch hohe Zusam-

menhänge mit Arbeitsgedächtniskapazität auf (Sander, 2005; Wilhelm, 2000; Wittmann, Süß, Oberauer, Schulze & Wilhelm, 1995). Damit stellt Verarbeitungskapazität das zweite Konstrukt dar, mit dem ein theoretischer Zusammenhang bestehen sollte. Da Reasoning (*Hypothese Konstruktvalidierung_1*) und Verarbeitungskapazität auf Konstruktebene praktisch identisch sind (Wilhelm, 2000), soll eine Abgrenzung dadurch erfolgen, dass Verarbeitungskapazität explizit nicht durch einen Matrizentest operationalisiert wird, sondern idealerweise durch mehrere Tests mit unterschiedlichem Aufgabenmaterial. Durch die Vergleichbarkeit mit Reasoning (auf Konstruktebene) wird die theoretische Argumentation zur Stärke des Effektes übernommen und ebenfalls ein Effekt im mittleren Bereich postuliert. Die zweite Hypothese der (konvergenten) struktursuchenden Konstruktvalidierung lautet daher:

Hypothese Konstruktvalidierung_2: SKS hängt positiv mit Verarbeitungskapazität zusammen. Der Effekt liegt im mittleren Bereich.

Ein Zusammenhang von SKS und Verarbeitungskapazität wird auch dadurch gestützt, dass Wilhelm (1995) für sämtliche dem Deduktiven Denken zugeordneten Formen des Schlussfolgerns (z.B. Propositionales Schlussfolgern, Syllogistisches Schlussfolgern, Relationales Schlussfolgern) in seinem Modell zur „Fähigkeit zum Lösen logischer Denkprobleme“ die Operationsklasse Verarbeitungskapazität (K) des Berliner Intelligenzstrukturmodells (BIS; Jäger, 1982; siehe Abbildung 1 in Abschnitt 2.1.4.1 sowie Abschnitt 2.1.4.1 für eine kurze Beschreibung) als limitierend annimmt. Nun wird zwar für SKS keine dimensionale Fähigkeitskonzeption postuliert, sondern die beschriebenen (Kompetenz-)Stufen (vgl. Abschnitt 2.2.2), dennoch kann Wilhelms (1995) Argumentation aufgrund der Zuordnung Konditionaler Schlüsse zu den Propositionalen Schlüssen im Rahmen der Aussagenlogik (vgl. Abschnitt 2.1.1) durchaus als theoretischer Indikator für den postulierten Zusammenhang zwischen SKS und Verarbeitungskapazität dienen.

Nun sollen jedoch nicht nur einzelne Konstrukte bezüglich ihres Zusammenhangs mit SKS betrachtet werden, sondern eben ein (intelligenzbezogenes) nomologisches Netzwerk, das heißt verschiedene Intelligenz-Konstrukte, die in einem theoretischen Zusammenhang miteinander stehen. Als nomologisches Netzwerk sollte demnach ein komplettes Intelligenzstrukturmodell betrachtet werden, insbesondere da Intelligenz mehr ist als nur Reasoning bzw. Verarbeitungskapazität. Hierfür bietet sich das Berliner Intelligenzstrukturmodell an, da Verarbeitungskapazität dort bereits eine eigene Opera-

tionsklasse bildet (siehe auch Abschnitt 2.1.4.1). Das Berliner Intelligenzstrukturmodell zählt ohnehin zu den populärsten Intelligenzstrukturmodellen (Beckmann & Guthke, 1999). Zudem hat es sich für die Validierung von Tests zum Logischen bzw. Deduktiven Denken bewährt (Wilhelm & Conrad, 1998; Wilhelm, 2000).

Für die anderen drei Operationsklassen des Berliner Intelligenzstrukturmodells (*Bearbeitungsgeschwindigkeit BIS-B*, *Merkfähigkeit BIS-M, Einfallsreichtum BIS-E*; vgl. Abbildung 1 in Abschnitt 2.1.4.1) ergeben sich keine plausiblen Gründe, die einen Zusammenhang mit SKS nahelegen. SKS wird durch den KKS erhoben. Da dieser ohne Zeitbegrenzung vorgegeben wird, sollte Bearbeitungsgeschwindigkeit keine Rolle spielen. Des Weiteren sollte auch Merkfähigkeit unerheblich sein, da jedes Item stets alle relevanten Informationen enthält (vgl. Abschnitt 2.4.1.2), also keine Informationen im Gedächtnis behalten zu werden brauchen. Schließlich scheint auch Einfallsreichtum zum Lösen der Items des KKS unnötig, da es weder um das Generieren einer Menge von Lösungen geht wie bspw. in einem klassischen Generierungsparadigma (siehe Abschnitt 2.4.1.2) noch um das Erkennen von Regeln wie bei Tests zum Induktiven Denken. Das bloße Auswählen der korrekt abgeleiteten Lösung wie im Falle des Multiple-Choice-Generierungsparadigmas des KKS (vgl. Abschnitt 2.4.1.2) sollte demnach unabhängig vom Einfallsreichtum einer Person sein. Entsprechend lauten die *Hypothesen Konstruktvalidierung_3*, *Konstruktvalidierung_4* und *Konstruktvalidierung_5* zur (divergenten) struktursuchenden Konstruktvalidierung.

Hypothese Konstruktvalidierung_3: SKS hängt nicht mit der Operationsklasse Bearbeitungsgeschwindigkeit (BIS-B) des Berliner Intelligenzstrukturmodells zusammen.

Hypothese Konstruktvalidierung_4: SKS hängt nicht mit der Operationsklasse Merkfähigkeit (BIS-M) des Berliner Intelligenzstrukturmodells zusammen.

Hypothese Konstruktvalidierung_5: SKS hängt nicht mit der Operationsklasse Einfallsreichtum (BIS-E) des Berliner Intelligenzstrukturmodells zusammen.

Zudem lässt sich *Hypothese Konstruktvalidierung_2* nach Wahl des Berliner Intelligenzstrukturmodells als Basis für das nomologische Netzwerk wie folgt konkretisieren:

Hypothese Konstruktvalidierung_2: SKS hängt positiv mit der Operationsklasse Verarbeitungskapazität (BIS-K) des Berliner Intelligenzstrukturmodells zusammen. Der Effekt liegt im mittleren Bereich.

Aus den *Hypothesen Konstruktvalidierung_1* bis *Konstruktvalidierung_5* resultiert ein auf dem Berliner Intelligenzstrukturmodell aufbauendes nomologisches Netzwerk als Basis der struktursuchenden Konstruktvalidierung von SKS. Abbildung 6 illustriert dieses nomologische Netzwerk.

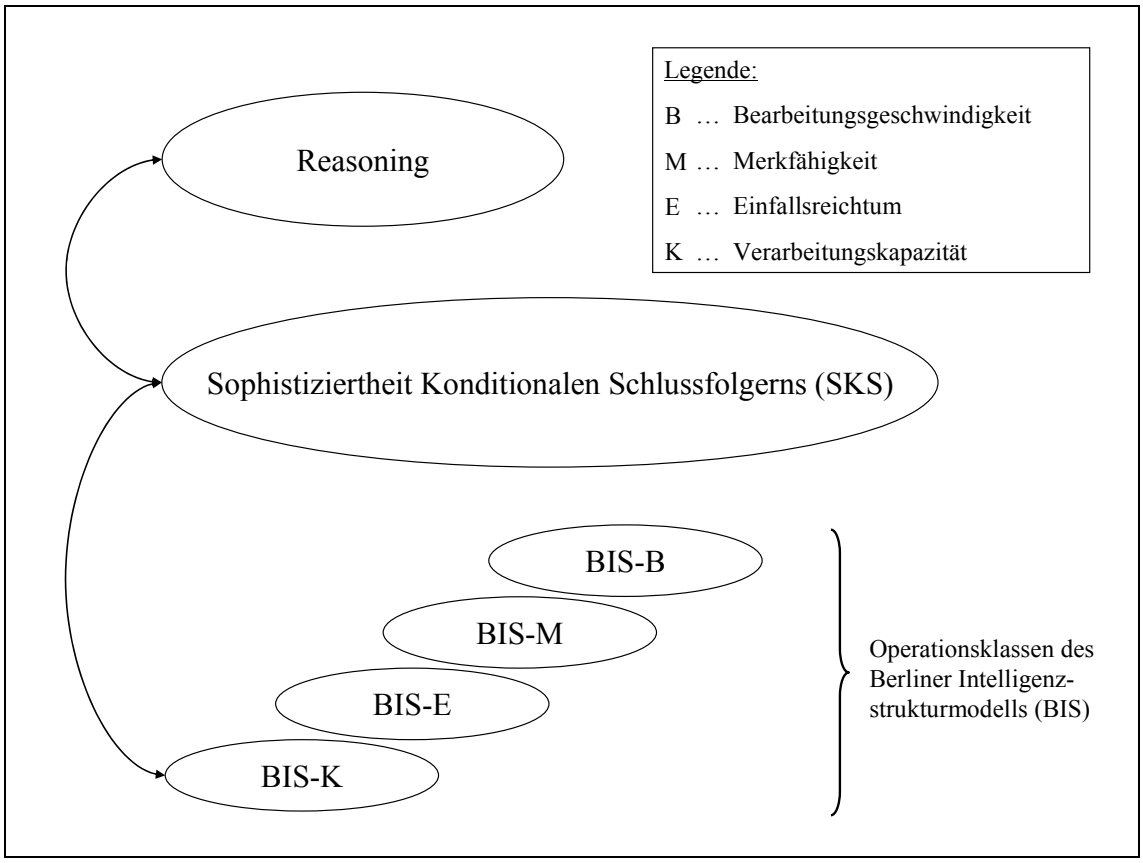


Abbildung 6: Einordnung von SKS in ein auf dem Berliner Intelligenzstrukturmodell basierendes nomologisches Netzwerk

Zur Überprüfung der *Hypothesen Konstruktvalidierung_1* bis *Konstruktvalidierung_5* und damit dieses nomologischen Netzwerks wird die teilweise bereits im Rahmen der strukturprüfenden Konstruktvalidierung (siehe Abschnitt 3.3.2.1) vorgestellte Konstruktvalidierungsstudie durchgeführt. Sie soll im Folgenden etwas ausführlicher beschrieben werden, indem wiederum auf Erhebungsinstrumente, Untersuchungsstichprobe, Durchführung, Auswertungsmethoden und Ergebnisse eingegangen wird.

Erhebungsinstrumente

Zur Erhebung von SKS wurde die überarbeitete Version des KKS verwendet, als (etablierter) Matrizentest der Wiener Matrizentest (WMT, Formann, 1979). Der WMT ist ein

klassischer Matrizentest, dessen Rasch-Skalierung die Testwertbildung durch einfaches Aufsummieren legitimiert. Er besteht aus 24 Aufgaben, zu denen jeweils acht Antwortalternativen vorgegeben werden, unter denen die richtige auszuwählen ist. Die Zeitbegrenzung beträgt 25 Minuten. Da keine Individualdiagnostik erfolgen sollte, waren die häufig kritisierten „veralteten" Normen von 1979 unproblematisch. Die vier Operationsklassen des Berliner Intelligenzstrukturmodells wurden mit der Kurzform (Testheft 2) des *Berliner Intelligenzstruktur Tests* (*BIS-4*, Jäger, Süß & Beauducel, 1997) erhoben. Auf eine Vorgabe der Langform des BIS-4 wurde verzichtet, da deren Bearbeitungszeit (laut Testheft mehr als 2 Stunden) die Probanden aus Sicht des Autors zu stark belastet hätte, insbesondere da neben dem BIS-4 auch noch der KKS und der WMT in dieser Studie vorgegeben wurden. Die hohe zeitliche Belastung hätte zu verfälschten Ergebnissen aufgrund von Ermüdung oder nachlassender Konzentration führen können. In der BIS-4-Kurzform wird Verarbeitungskapazität (BIS-K) durch sechs Aufgaben erhoben, Bearbeitungsgeschwindigkeit (BIS-B), Merkfähigkeit (BIS-M) und Einfallsreichtum (BIS-E) jeweils durch drei Aufgaben. Tabelle 22 zeigt die Zuordnung der 15 Aufgaben der BIS-4-Kurzform zu den vier Operationsklassen.

Tabelle 22: Zuordnung der Aufgaben der BIS-4-Kurzform zu den vier Operationsklassen des Berliner Intelligenzstrukturmodells

Operationsklasse des Berliner Intelligenzstrukturmodells			
BIS-K	BIS-B	BIS-M	BIS-E
Analogien, Charkow, Wort-Analogien, Tatsache-Meinung, Zahlenreihen, Schätzen	Teil-Ganzes, X-Größer, Buchstaben-Durchstreichen	Sinnvoller Test, Zahlen-Paare, Orientierungs-Gedächtnis	Layout, Eigenschaften-Fähigkeiten, Divergentes Rechnen

Jede Aufgabe wiederum besteht aus einer unterschiedlichen Anzahl von Items. Der Punktwert jeder Aufgabe ergibt sich beim BIS-4 durch Umwandeln der Summe richtig gelöster Items in einen konkreten Punktwert zwischen 70 und 130 (für die Aufgabe). Diese Zuweisung erfolgt über konkrete Auswertevorschriften auf Basis der Verteilung der Summenwerte jeder Aufgabe in der Normierungsstichprobe des BIS-4 (Jäger et al., 1997). So ergeben sich für sämtliche Aufgaben des BIS-4 vergleichbare Punktwerte,

obwohl sich die Zahl der Items zwischen den einzelnen Aufgaben mitunter stark unterscheidet. Der Gesamtpunktwert für eine Operationsklasse wiederum ergibt sich durch Aufsummieren der Punktwerte aller Aufgaben, die dieser Operationsklasse zugeordnet werden. Für die Kurzform existieren lediglich Normwerte für die Operationsklasse Verarbeitungskapazität sowie für allgemeine Intelligenz (Summe der Punktwerte sämtlicher Aufgaben aller vier Operationsklassen). Dies war jedoch nicht von Bedeutung, da keine Individualdiagnostik beabsichtigt war. Die Zeitbegrenzung und damit die Bearbeitungszeit der Kurzform liegt bei 47 Minuten.

Durchführung

Die Datenerhebung zur Konstruktvalidierungsstudie bestand aus einem schriftlichen und einem computergestützten Teil. Der schriftliche Teil wurde durch eine Testleiterin angeleitet, welche für die Instruktionen, die Ausgabe der Materialien, die Zeitnahme usw. verantwortlich war. Er fand in mehreren Gruppentestungen zu je etwa 25 Probanden statt. Nach Angabe eines persönlichen Codes und Vorgabe eines Befindlichkeits-Fragebogens (MDBF; Steyer, Schwenkmezger, Notz & Eid, 1997) wurde zunächst der WMT und nach einer kurzen Pause der BIS-4 vorgegeben. Es folgten weitere Testverfahren, die jedoch im Rahmen der vorliegenden Arbeit nicht ausgewertet werden. Insgesamt nahm der schriftliche Teil etwa 90-120 Minuten in Anspruch. Der computergestützte Teil fand als Online-Test zu einem späteren Zeitpunkt statt, wiederum als Gruppentestung zu je etwa 25 Personen in einem eigens dafür reservierten Computer-Pool. Die Rechner dieses Computer-Pools waren hinsichtlich Ausstattung und Internetverbindung vergleichbar. Auf diese Weise sollten für den computergestützten Teil vergleichbare Bedingungen geschaffen werden, die ansonsten bei Online-Tests kaum kontrollierbar sind. Wiederum war eine Testleiterin anwesend, die die Testdurchführung jedoch lediglich beaufsichtigte, nicht anleitete. Auch der computergestützte Teil begann mit der Eingabe des persönlichen Codes[61], gefolgt vom MDBF. Anschließend wurden die 16 Items der überarbeiteten Version des KKS vorgegeben sowie in der Folge weitere Items und Testverfahren, die jedoch im Rahmen der vorliegenden Arbeit nicht ausgewertet werden. Etwa nach der Hälfte der Testung wurden die Probanden gebeten, einige soziodemographische Angaben zur eigenen Person zu machen. Der computergestützte Teil nahm ca. 90 Minuten in Anspruch. Die gesamte Datenerhebung der Kon-

[61] Über den persönlichen Code ist die Zuordnung der Testergebnisse beider Teile zu den jeweiligen Personen möglich.

struktvalidierungsstudie erstreckte sich über einen Zeitraum von November 2007 bis Juni 2008. Anschließend wurden die Tests des schriftlichen Teils ausgewertet und die Ergebnisse in eine Datenbank eingegeben. Zugeordnet über den persönlichen Code wurden dann die Daten der Tests des computergestützten Teils eingelesen, woraus der vollständige Datensatz der Konstruktvalidierungsstudie resultierte.

Untersuchungsstichprobe

Auf eine Beschreibung der Untersuchungsstichprobe soll an dieser Stelle verzichtet werden, da die Stichprobe der Konstruktvalidierungsstudie bereits ausführlich unter dem entsprechenden Punkt der strukturprüfenden Konstruktvalidierung (siehe Abschnitt 3.3.2.1) beschrieben wird.

Auswertungsmethoden

Die Überprüfung der in den *Hypothesen Konstruktvalidierung_1* bis *Konstruktvalidierung_5* postulierten Zusammenhänge soll analog der Überprüfung von *Arbeitshypothese 2* (Zusammenhang SKS mit Reasoning) bei der ersten empirischen Erprobung des KKS erfolgen (siehe Abschnitt 2.4.1.5). Die Präzisierung der inhaltlichen Hypothesen hinsichtlich der gewählten Analyseverfahren wird aus Platzgründen nicht gesondert dargestellt. Sie erfolgt analog zu *Arbeitshypothese 2* bei der ersten empirischen Erprobung des KKS (siehe Abschnitt 2.4.1.5). Wiederum wird eine multinomiale logistische Regression für latente Variablen als statistisches Analyseverfahren gewählt, da so der Zusammenhang zwischen latenten nominalen Variablen (hier SKS) und latenten kontinuierlichen Variablen (hier die mit WMT bzw. den BIS-4-Skalen gemessenen latenten Variablen) analysiert werden kann. Gerade bei der Konstruktvalidierung werden häufig Zusammenhänge zwischen latenten Variablen betrachtet (vgl. z.B. Borsboom & Mellenbergh, 2007; Hartig et al., 2007), um so das Messfehlerproblem zu berücksichtigen. Zur besseren Vergleichbarkeit mit Untersuchungsergebnissen anderer Studien (zumeist Korrelationen) sowie aufgrund der leichteren Interpretierbarkeit werden zudem paarweise Rangkorrelationen berechnet. Dabei wird je nach Hypothese der entsprechende Testwert (nach Auswertevorschrift des jeweiligen Testmanuals der Summenwert im WMT bzw. jeweils der Summenwert der Aufgaben-Punktwerte der vier BIS-4-Skalen zu den Operationsklassen) mit der manifesten, theoriegeleitet ordinalen Variable SKS korreliert. Wenngleich dadurch dem Messfehlerproblem nicht mehr Rechnung getragen werden kann, scheint die Verwendung der Summenwerts der Punktwerte der BIS-

Operationsklassen bzw. des Summenwerts des WMT durchaus gerechtfertigt. Schließlich handelt es sich um etablierte und bereits vielfach eingesetzte Messinstrumente mit vergleichsweise hoher Reliabilität (vgl. Formann, 1979; Jäger et al., 1997), deren Testwerte per Auswertevorschrift des jeweiligen Testmanuals durch Aufsummieren zu bilden sind. Bei SKS spricht die hohe Treffsicherheit der überarbeiteten Version des KKS (Treffsicherheit $T = .957$; vgl. Abschnitt 3.3.2.1) für eine hohe Messgenauigkeit und damit nach Clogg und Manning (1996) für eher geringe Messfehler, sodass eine Verwendung in manifester Form gerechtfertigt scheint.

Vor der Berechnung der multinomialen logistischen Regressionen sind jedoch zunächst Messmodelle für die latenten Variablen zu überprüfen. Eine Überprüfung des Messmodells für SKS ist bereits erfolgt (strukturprüfende Konstruktvalidierung; vgl. Abschnitt 3.3.2.1), sodass lediglich für die latenten Variablen, die mit dem WMT bzw. den vier BIS-4-Skalen gemessen werden, noch Messmodelle zu überprüfen sind. Nach Festlegung der Messmodelle können dann die latenten Variablen entsprechend diesen Messmodellen in der multinomialen logistischen Regression modelliert werden. Zur besseren Unterscheidung werden die latenten Variablen, die mit den beschriebenen Testverfahren (WMT, BIS-4) gemessen werden, in der Folge mit dem Index „*lat*“ versehen. So soll deutlich gemacht werden, dass es sich um latente Variablen handelt.

Messmodell für WMT_{lat}

Die 24 Items des WMT werden als Rasch-skaliert angenommen (Formann, 1979), was ein Rasch-Modell zur Modellierung der latenten Variable WMT_{lat} (als Indikator für Reasoning) nahelegt. Dies soll analog zu den 10 Matrizenitems aus der ersten empirischen Erprobung des KKS überprüft werden (vgl. Abschnitt 2.4.1.5). Die Modellpassung eines Rasch-Modells wird mittels parametrischen Bootstrapverfahrens bestimmt. Brauchen die Annahmen eines Rasch-Modells nicht verworfen zu werden, werden zusätzlich Andrichs Reliabilität (Andrich, 1988) als Maß für die Messgenauigkeit des WMT sowie die *Z*-Werte der Q-Indices als Indikatoren für von eins verschiedene Diskriminationsparameter überprüft. Die Modellprüfung erfolgt unter Verwendung der Software Winmira 2001 (Davier, 2001), da – wie bereits andernorts erwähnt – das parametrische Bootstrapverfahren zur Modellprüfung in die ansonsten verwendete Software M*plus* (Muthén & Muthén, 1998-2007) nicht implementiert ist.

Messmodelle für die Operationsklassen des Berliner Intelligenzstrukturmodells

Die Indikatoren für die latenten Variablen Verarbeitungskapazität (*BIS-*K_{lat}), Einfallsreichtum (*BIS-*E_{lat}), Bearbeitungsgeschwindigkeit (*BIS-*B_{lat}) und Merkfähigkeit (*BIS-*M_{lat}) sind die Punktwerte der ihnen zugeordneten Aufgaben (siehe Tabelle 22). Zur Modellierung bieten sich Modelle der KTT an (für einen Überblick siehe z.B. Steyer, 2001). Da im Testmanual des BIS-4 keine konkreten Vorschriften zur Modellierung latenter Variablen aus den Punktwerten zu finden sind, soll ein passendes Messmodell wie folgt bestimmt werden. Für jede der vier latenten Variablen wird zunächst ein (restriktives) Paralleltestmodell (definiert durch τ-Äquivalenz, unkorrelierte Messfehler und homogene Fehlervarianzen; vgl. Steyer, 2001) angenommen. Bei Nichtpassung werden nach und nach Restriktionen gelockert und weniger restriktive Modelle (essentiell τ-äquivalentes Modell, τ-kongenerisches Modell; vgl. Steyer, 2001 oder Steyer & Eid, 2001) geprüft. Dies erfolgt solange, bis ein passendes Modell gefunden wird. Auf diese Weise soll das am sparsamsten parametrisierte und dennoch passende Messmodell gefunden werden (*Parsimonitätsprinzip*). Dem Autor ist bewusst, dass sich das Messmodell eigentlich aus theoretischen Annahmen ableiten soll. Es sei aber darauf hingewiesen, dass die Überprüfung eines Messmodells für die BIS-Operationsklassen nicht Ziel dieser Arbeit ist. Vielmehr soll ein Messmodell gefunden werden, dass aufgrund der Daten nicht verworfen zu werden braucht, um in der multinomialen logistischen Regression davon ausgehen zu können, dass auf eine latente Variable regrediert wird, die adäquat durch ihre Indikatoren gemessen wird. Die Überprüfung der Messmodelle erfolgt unter Verwendung der Software M*plus* (Muthén & Muthén, 1998-2007). Die Passung des Modells wird mithilfe verschiedener Modellfit-Indizes beurteilt. Entsprechend der Empfehlung von Bollen und Long (1993) sowie von Schermelleh-Engel, Moosbrugger und Müller (2003) werden dabei mehrere Fit-Indizes betrachtet. Die Chi-Quadrat-Statistik wird zur inferenzstatistischen Überprüfung des Modellfits verwendet, wobei ein nicht-signifikanter *p*-Wert ($p > .05$ bei konventioneller Festlegung des alpha-Niveaus) einen guten Modellfit indiziert. Weiterhin werden der *Root Mean Square Error of Approximation* (*RMSEA*; Steiger, 1990) und das *Standardized Root Mean Square Residual* (*SRMR*; Jöreskog & Sörbom, 1986) als deskriptive Indizes in die Modellbewertung einbezogen. Bei beiden sprechen Werte kleiner .05 für einen guten Modellfit, Werte kleiner .08 für einen akzeptablen und Werte kleiner .10 für einen zumindest noch ausreichenden Modellfit (Browne & Cudeck, 1993). Schließlich werden noch zwei vergleichende Fit-Indizes betrachtet: der *Comparative Fit Index* (*CFI*; Bentler,

1990) und der *Tucker-Lewis Index* (*TLI*; Tucker & Lewis, 1973; Bentler & Bonnett, 1980), bei denen Werte größer .97 einen guten Modellfit und Werte zwischen .95 und .97 einen akzeptablen Modellfit indizieren (Hu & Bentler, 1995, 1998, 1999).

Multinomiale logistische Regressionen

Sind die Messmodelle überprüft, wird jeweils einzeln die multinomiale logistische Regression von SKS_{lat} auf die jeweilige Variable (WMT_{lat}, $BIS\text{-}B_{lat}$, $BIS\text{-}M_{lat}$, $BIS\text{-}E_{lat}$, $BIS\text{-}K_{lat}$) berechnet. Eine simultane Prüfung des nomologischen Netzwerks wäre zur Vermeidung einer Kumulierung des alpha-Fehlers zwar wünschenswert, scheint jedoch aufgrund der geringen Stichprobengröße ($N = 154$) unangebracht. Die Beurteilung, ob ein paarweiser Zusammenhang zwischen den latenten Variablen besteht, erfolgt wie auch bei der ersten empirischen Erprobung (siehe Abschnitt 2.4.1.5) durch Bestimmung des aus dem LR-Test resultierenden Chi-Quadrat-Wertes bei drei Freiheitsgraden[62] ($df = 3$), dessen p-Wert mit dem konventionellen alpha-Niveau ($\alpha = .05$) verglichen wird. Nagelkerkes R-Quadrat dient zur Beschreibung der Stärke des Effektes (vgl. Ausführungen in Abschnitt 2.4.1.5).

Rangkorrelationen

Wie bereits bei der ersten empirischen Erprobung des KKS ausführlich beschrieben, kann die nominale latente Klassenvariable SKS_{lat} theoriegeleitet auf ein ordinales Datenniveau transformiert werden (vgl. Abschnitt 2.4.1.5). Hierzu wird den Klassenzuordnungen (theoriegeleitet) folgende Ordnung zugewiesen:

1 (Rang 1) … Unsophistizierte
2 (Rang 2) … Unsophistizierte II
3 (Rang 3) … Fortgeschrittene
4 (Rang 4) … Sophistizierte

Als Ergebnis der durchgeführten LCA ergibt sich für jede Person eine konkrete (manifeste) Klassenzuordnung, der entsprechend eben beschriebener Ordnung konkrete Zahlen zugewiesen werden. Die so entstandene manifeste ordinale Variable SKS_{man}[63] kann nun durch Rangkorrelationen mit den nach Auswertevorschrift bestimmten Testwerten

[62] Da die Vier-Klassen-Lösung als Messmodell für SKS an dieser Stelle bereits überprüft ist, ergeben sich für die Chi-Quadrat-Statistik de facto drei Freiheitsgrade. In Abschnitt 2.4.1.5 wird die Bestimmung der Freiheitsgrade im Falle einer Vier-Klassen-Lösung ausführlich beschrieben.
[63] Der Index „*man*“ soll verdeutlichen, dass es sich nun um eine manifeste Variable handelt, die Ergebnis der beschriebenen Auswertevorschrift ist.

WMT_{man}, $BIS\text{-}K_{man}$, $BIS\text{-}B_{man}$, $BIS\text{-}M_{man}$ und $BIS\text{-}E_{man}$ korreliert werden. Zur Bestimmung dieser Rangkorrelationen werden jeweils Spearmans Rho und Kendalls Tau als Koeffizienten angegeben. Da für WMT_{man} und $BIS\text{-}K_{man}$ ein positiver Zusammenhang mit SKS_{man} vermutet wird, erfolgt dort die Signifikanzprüfung einseitig. Für $BIS\text{-}B_{man}$, $BIS\text{-}M_{man}$ und $BIS\text{-}E_{man}$ wird kein Zusammenhang mit SKS_{man} postuliert. Die Nullhypothese (Spearmans Rho = 0 bzw. Kendalls Tau = 0) ist also Wunschhypothese, weshalb der zugehörige *p*-Wert mit Werten $p > .05$ diese Hypothesen stützen würde. Die Stärke der Effekte wird gemäß den Konventionen von Cohen (1990, 1992) für Korrelationskoeffizienten beurteilt (.10 – geringer Effekt, .30 – mittlerer Effekt, .50 – hoher Effekt).

Ergebnisse

Zunächst sind die Messmodelle für die betrachteten Variablen zu überprüfen. Die LCA zur Bestimmung des Messmodells für SKS_{lat} ist bereits beim strukturprüfenden Vorgehen beschrieben (vgl. Abschnitt 3.3.2.1). SKS_{lat} kann demnach durch ein LCM unter Vorgabe von vier Klassen gemessen werden. Im Folgenden werden die Messmodelle für WMT_{lat} sowie für die vier latenten BIS-Variablen vorgestellt.

Prüfung des Messmodells für WMT_{lat}

Für die Prüfung des Modellfits der 24 Items des WMT werden unter Annahme eines Rasch-Modells 200 parametrische Bootstrap-Stichproben generiert. Dabei ergibt sich ein *p*-Wert für Chi-Quadrat von $p = .34$ bzw. für Cressie Read von ebenfalls $p = .34$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung des Rasch-Modells. Andrichs Reliabilität liegt bei $Rel = .73$ und ist damit zufriedenstellend hoch, insbesondere da für Modelle der IRT die bereits beschriebenen Einschränkungen bezüglich der Interpretierbarkeit von Reliabilitätsschätzungen gelten (vgl. Abschnitt 2.4.1.5). Die *Z*-Werte der Q-Indizes liegen zwischen -0,80 und 0,78, sodass für keinen der Diskriminationsparameter der 24 Items des WMT die Nullhypothese ($\beta_i = 1$) verworfen zu werden braucht. Dies stützt zusätzlich zu den Ergebnissen des Bootstrapverfahrens die Annahme eines Rasch-Modells für die 24 Items des WMT. Das postulierte Rasch-Modell als Messmodell für den WMT und die damit gemessene latente Variable WMT_{lat}, die als Indikator für Reasoning verwendet wird, kann also als passend angenommen werden.

Prüfung der Messmodelle für die BIS-Operationsklassen

Wie bereits ausgeführt, erfolgt die Auswahl des Messmodells nach dem Parsimonitätsprinzip, das heißt, das Messmodell soll so sparsam wie möglich parametrisiert sein und dennoch einen guten Modellfit aufweisen. Getestet werden die klassischen Messmodelle der KTT: Paralleltestmodell, essentiell τ-äquivalentes Modell und τ-kongenerisches Modell. Tabelle 23 fasst die Ergebnisse für die vier Skalen des BIS zusammen.

Es lässt sich für jede der vier Operationsklassen des Berliner Intelligenzstrukturmodells ein Messmodell finden, dass über alle betrachteten Modellfit-Indizes hinweg einen zumindest akzeptablen Modellfit aufweist. Für die Modellierung von $BIS\text{-}K_{lat}$ ist dies ein τ-kongenerisches Messmodell, bei dem eine Fehlerkorrelation zwischen den Aufgaben Zahlenreihen und Schätzen zugelassen wird. Da beide Aufgaben im Gegensatz zu den restlichen Aufgaben der Inhaltsklasse „numerisch“ (Ausführungen zu den Inhaltsklassen des Berliner Intelligenzstrukturmodells siehe Abschnitt 2.1.4.1) zugeordnet werden, ist diese zusätzlich notwendige Korrelation durchaus erklärbar. Für die Modellierung von $BIS\text{-}B_{lat}$ weist ein prinzipiell τ-kongenerisches Messmodell mit Einser-Ladungen auf den Aufgaben X-Größer und Buchstabendurchstreichen einen adäquaten Modellfit auf, für die Modellierung von $BIS\text{-}M_{lat}$ und $BIS\text{-}E_{lat}$ jeweils essentiell τ-äquivalente Messmodelle. Somit ist für jede der BIS-Operationsklassen ein Messmodell identifiziert und getestet, das für die Modellierung der entsprechenden latenten Variable in der multinomialen logistischen Regression verwendet werden kann.

Tabelle 23: Prüfung verschiedener Messmodelle der KTT für die vier BIS-Operationsklassen (BIS-K, BIS-B, BIS-M, BIS-E) durch Bestimmung des Modellfits

		Modellfit-Indikatoren						
	Messmodell	χ^2	*df*	*p*-Wert	RMSEA	SRMR	CFI	TLI
BIS-K	parallel	178.18	24	< .001	.204	.343	.17	.48
	ess. *τ*-äquiv.	40.09	14	< .001	.110	.274	.86	.85
	τ-kongen.*	**13.13**	**8**	**.108**	**.065**	**.035**	**.97**	**.95**
BIS-B	parallel	51.24	6	< .001	.221	.554	.00	.40
	ess. *τ*-äquiv.	9.32	2	.010	.154	.186	.81	.71
	τ-kongen.**	**1.66**	**1**	**.197**	**.066**	**.100**	**.98**	**.95**
BIS-M	parallel	38.55	6	< .001	.375	.841	.00	.00
	ess. τ-äquiv.	**1.05**	**2**	**.592**	**.000**	**.072**	**1**	**1**
	τ-kongen.***	–	–	–	–	–	–	–
BIS-E	parallel	50.65	6	< .001	.220	.441	.00	.00
	ess. τ-äquiv.	**1.28**	**2**	**.528**	**.000**	**.066**	**1**	**1**
	τ-kongen.***	–	–	–	–	–	–	–

Anmerkung. Die passenden Messmodelle sind jeweils fettgedruckt. ess. *τ*-äquiv. …essentiell *τ*-äquivalent, *τ*-kongen. … *τ*-kongenerisch

* Für dieses *τ*-kongenerische Messmodell wird eine Korrelation der Fehler der Aufgaben Zahlenreihen (ZN) und Schätzen (SC) zugelassen.

** Für dieses prinzipiell *τ*-kongenerische Messmodell werden die Ladungen der Aufgaben X-Größer (XG) und Buchstabendurchstreichen (BD) auf eins fixiert und lediglich die Ladung für die Aufgabe Teil-Ganzes (TG) frei geschätzt. So ergibt sich ein Freiheitsgrad und das Modell ist testbar.

*** Diese Modelle werden gemäß dem Parsimonitätsprinzip nicht (mehr) berechnet, da bereits ein sparsamer parametrisiertes Modell auf die Daten passt.

Ergebnisse der multinomialen logistischen Regressionen

Für die Berechnung der multinomialen logistischen Regression[64] werden für Sophistiziertheit Konditionalen Schlussfolgerns ein LCM unter Vorgabe von vier Klassen, für die anderen Variablen die eben beschriebenen Messmodelle spezifiziert. Da von allen Personen vollständige Datensätze vorliegen, umfasst die Stichprobe alle $N = 154$ Personen. Die Signifikanzprüfung erfolgt via LR-Test (vgl. Abschnitt 2.4.1.5). Die Differenz $LogL_0$ - $LogL_1$ aus der Log-Likelihood $LogL_0$ des restriktiven Modells (alle Anstiegskoeffizienten auf null fixiert) und der Log-Likelihood $LogL_1$ des weniger restriktiven

[64] Die M*plus*-Inputfiles der multinomialen logistischen Regressionen finden sich im Anhang D dieser Arbeit.

Modells (alle Anstiegskoeffizienten frei geschätzt) wird dazu in eine χ^2-verteilte Prüfgröße umgerechnet (Multiplikation mit -2), deren *p*-Wert bei drei Freiheitsgraden bestimmt wird. Tabelle 24 fasst die LR-Tests der fünf durchgeführten multinomialen logistischen Regressionen zusammen und gibt zudem Nagelkerkes R-Quadrat als Effektstärkemaß an.

Tabelle 24: Multinomiale logistische regressive Abhängigkeiten der latenten SKS-Variable (SKS_{lat}) von den latenten Variablen des nomologischen Netzwerks zur Konstruktvalidierung

Regression	$LogL_0$	$LogL_1$	χ^2	*df*	*p*	R^2_{Nag}
SKS_{lat} auf WMT_{lat}	-2571,67	-2560,52	22,30	3	< .001	.13
SKS_{lat} auf $BIS\text{-}K_{lat}$	-4257,44	-4244,60	25,72	3	< .001	.15
SKS_{lat} auf $BIS\text{-}B_{lat}$	-2629,90	-2627,02	5,76	3	.124	(.04)
SKS_{lat} auf $BIS\text{-}M_{lat}$	-2584,69	-2583,59	2,20	3	.532	(.01)
SKS_{lat} auf $BIS\text{-}E_{lat}$	-2564,15	-2557,95	12,40	3	.006	(.08)

Anmerkung. $LogL_0$ … Log-Likelihood des restriktiven Modells (Anstiegskoeffizienten auf null fixiert), $LogL_1$ … Log-Likelihood des weniger restriktiven Modells (Anstiegskoeffizienten frei geschätzt), $\chi^2 = -2(LogL_0 - LogL_1)$, R^2_{Nag} … Nagelkerkes R-Quadrat (für die Berechnungsformel siehe Anhang C).

Hypothesenkonform ergeben sich (multinomiale logistische) regressive Abhängigkeiten der Sophistiziertheit Konditionalen Schlussfolgerns von WMT_{lat} (*Hypothese Konstruktvalidierung_1*) sowie von $BIS\text{-}K_{lat}$ (*Hypothese Konstruktvalidierung_2*). Die Effekte (indiziert durch Nagelkerkes R-Quadrat) liegen im mittleren Bereich. Ebenfalls hypothesenkonform zeigt sich keine (multinomiale logistische) regressive Abhängigkeit von $BIS\text{-}B_{lat}$ (*Hypothese Konstruktvalidierung_3*) und $BIS\text{-}M_{lat}$ (*Hypothese Konstruktvalidierung_4*). Jedoch ergibt sich entgegen *Hypothese Konstruktvalidierung_5* eine (multinomiale logistische) regressive Abhängigkeit von $BIS\text{-}E_{lat}$, allerdings bei einem lediglich geringen Effekt.

Zusätzlich werden die Anstiegskoeffizienten analysiert, um die Ordnung der identifizierten Klassen hinsichtlich der jeweiligen kontinuierlichen latenten Variable zu überprüfen. Allerdings soll dies nur für die drei Variablen (WMT_{lat}, $BIS\text{-}K_{lat}$, $BIS\text{-}E_{lat}$) durchgeführt werden, für die sich auch eine (multinomiale logistische) regressive Abhängigkeit ergibt. Da die Referenzkategorie durch die verwendete Software M*plus*

(Muthén & Muthén, 1998-2007) automatisch festgelegt wird, sind die Ergebnisse dann hypothesenkonform, wenn die Anstiegskoeffizienten folgende Ordnung aufweisen:

$\beta_{\text{Sophistizierte\_0}} > \beta_{\text{Fortgeschrittene\_0}} > \beta_{\text{UnsophistizierteII\_0}} > \beta_{\text{Unsophistizierte\_0}}$
(0 indiziert dabei die Referenzklasse bzw. -kategorie)

Der Wert des Anstiegskoeffizienten der Referenzkategorie liegt bei null, die anderen drei Anstiegskoeffizienten können entsprechend eingeordnet werden. Zusätzlich kann geprüft werden, ob ein Anstiegskoeffizient signifikant von null verschieden ist. Die Ausprägung der kontinuierlichen latenten Variable (also des Regressors) in dieser Klasse unterscheidet sich dann signifikant von der in der Referenzkategorie. Tabelle 25 zeigt die Ordnung der Anstiegskoeffizienten für die drei betrachteten regressiven Abhängigkeiten.

Tabelle 25: Ordnung der Anstiegskoeffizienten der multinomialen logistischen regressiven Abhängigkeiten der latenten SKS-Variable (SKS_{lat}) von den latenten Variablen des nomologischen Netzwerks zur Konstruktvalidierung

Regression	Referenzkategorie	$\beta_{\text{Uns\_0}}$	$\beta_{\text{Uns.II\_0}}$	$\beta_{\text{Fort\_0}}$	$\beta_{\text{Soph\_0}}$
SKS_{lat} auf WMT_{lat}	Unsophistizierte II	-0,291	**0,000**	0,121	1,717*
SKS_{lat} auf $BIS\text{-}K_{lat}$	Unsophistizierte	**0,000**	0,146	0,162	0,503*
SKS_{lat} auf $BIS\text{-}E_{lat}$	Unsophistizierte II	-0,038	**0,000**	0,044	0,388

Anmerkung. Der Anstiegskoeffizient (0,000) der durch M*plus* jeweils festgelegten Referenzkategorie ist fettgedruckt. Uns … Unsophistizierte, Fort … Fortgeschrittene, Soph … Sophistizierte.
* signifikant von null verschieden bei konventionellem alpha-Niveau ($\alpha = .05$)

Die postulierte Ordnung der Anstiegskoeffizienten zeigt sich für alle drei betrachteten regressiven Abhängigkeiten. Die *Hypothesen Konstruktvalidierung_1* und *Konstruktvalidierung_2* werden dadurch zusätzlich gestützt. Signifikant von null verschieden sind jedoch lediglich die Koeffizienten der Sophistizierten-Stufe gegenüber der Referenzkategorie bei WMT_{lat} und $BIS\text{-}K_{lat}$.

Ergebnisse der Rangkorrelationen

Zur besseren Vergleichbarkeit mit den Ergebnissen anderer Untersuchungen werden zusätzlich Rangkorrelationen der manifesten Variable Sophistiziertheit Konditionalen Schlussfolgerns (SKS_{man}) mit den Summenwerten des WMT bzw. der vier BIS-4-Skalen

(zu den vier Operationsklassen) bestimmt, also mit ebenfalls manifesten Variablen. Die Ergebnisse sind in Tabelle 26 dargestellt.

Tabelle 26: Rangkorrelationen der Variable SKS_{man} mit dem Summenwert des WMT sowie Summenwerten der BIS-4-Skalen BIS-4-K, BIS-4-B, BIS-4-M und BIS-4-E

Variablen	Spearmans Rho (p-Wert)	Kendalls Tau (p-Wert)
SKS_{man} und WMT_{man}	.262 ($p < .001$)	.212 ($p < .001$)
SKS_{man} und $BIS\text{-}K_{man}$	.310 ($p < .001$)	.242 ($p < .001$)
SKS_{man} und $BIS\text{-}B_{man}$	-.049 ($p = .272$)	-.039 ($p = .266$)
SKS_{man} und $BIS\text{-}M_{man}$	.045 ($p = .289$)	.036 ($p = .282$)
SKS_{man} und $BIS\text{-}E_{man}$	.169 ($p = .018$)	.132 ($p = .017$)

Für SKS_{man} zeigen sich (rang-)korrelative Zusammenhänge bei mittleren Effekten mit WMT_{man} und $BIS\text{-}K_{man}$ sowie bei geringem Effekt mit $BIS\text{-}E_{man}$. Keine (rang-)korrelativen Zusammenhänge zeigen sich hingegen zwischen SKS_{man} und $BIS\text{-}B_{man}$ bzw. $BIS\text{-}M_{man}$.

Die Ergebnisse sowohl der multinomialen logistischen Regressionen als auch der Rangkorrelationen lassen sich wie folgt zusammenfassen: Es zeigt sich eine (multinomiale logistische) regressive Abhängigkeit der latenten Variable SKS sowohl von der latenten Variable, die der WMT misst als auch von der latenten Variable, die mit den sechs Aufgaben zur Verarbeitungskapazität des BIS-4 gemessen wird. Der Effekt liegt jeweils im mittleren Bereich. Ebenfalls signifikant bei mittleren Effekten fallen die Ergebnisse der berechneten Rangkorrelationen der manifesten Variable SKS mit den Summenwerten des WMT und der Skala BIS-K aus. Diese Ergebnisse sprechen also *für* die Hypothesen der (konvergenten) struktursuchenden Konstruktvalidierung (*Konstruktvalidierung_1* und *Konstruktvalidierung_2*).

Ferner zeigt sich weder eine (multinomiale logistische) regressive Abhängigkeit der latenten Variable SKS von der latenten Variable, die mit den drei Aufgaben des BIS-4 zur Bearbeitungsgeschwindigkeit gemessen wird, noch von der latenten Variable, die mit den drei Aufgaben des BIS-4 zur Merkfähigkeit gemessen wird. Auf manifester Ebene ergeben sich sehr geringe Rangkorrelationen, bei nicht-signifikanten p-Werten. Diese Ergebnisse sprechen also für zwei der Hypothesen zur (divergenten) struktursuchenden Konstruktvalidierung (*Konstruktvalidierung_3* und *Konstruktvalidierung_4*).

Was sich jedoch zeigt, ist eine (multinomiale logistische) regressive Abhängigkeit der latenten Variable SKS von der latenten Variable, die mit den drei Aufgaben des BIS-4 zum Einfallsreichtum gemessen wird. Der Effekt liegt im geringen Bereich. Ebenfalls signifikant bei eher geringen Effekten fallen die Ergebnisse der berechneten Rangkorrelation der manifesten Variable SKS mit dem Summenwert der Skala BIS-E aus. *Hypothese Konstruktvalidierung_5* zur (divergenten) struktursuchenden Konstruktvalidierung sollte aufgrund dieser empirischen Befunde nicht beibehalten werden, gleichwohl der Effekt jeweils nur im geringen Bereich liegt.

Insgesamt fallen die empirischen Befunde zur (konvergenten wie divergenten) struktursuchenden Konstruktvalidierung hypothesenkonform aus, abgesehen von geringen Effekten für den Zusammenhang von SKS und BIS-Einfallsreichtum (für eine Diskussion der empirischen Ergebnisse zur struktursuchenden Konstruktvalidierung siehe Abschnitt 4.1.2).

3.3.3 Kriteriumsvalidierung

Bei der Kriteriumsvalidierung wird überprüft, ob aus dem Ergebnis eines Tests auf ein für diagnostische Entscheidungen relevantes Kriterium geschlossen werden kann (z.B. Hartig et al., 2007). In der Eignungsdiagnostik werden dazu üblicherweise Zusammenhänge bzw. Vorhersagekraft eines Testergebnisses in Bezug auf Schul-, Studiums- oder Berufserfolg betrachtet. Wird der Zusammenhang mit dem aktuellen Erfolg, also einem zeitgleichen Kriterium, analysiert, betrachtet man die sog. *konkurrente Validität* des Tests, wird zukünftiger Erfolg vorhergesagt, die sog. *prognostische Validität* des Tests. Eine dritte, weit weniger verbreitete Art der Kriteriumsvalidierung ist die Betrachtung von Zusammenhängen des Testergebnisses mit zeitlich bereits zurückliegenden (Erfolgs-)Kriterien, die bspw. Bühner (2006) als *retrospektive Validität* bezeichnet.

Borsboom und Mellenbergh (2007) rechnen die Zusammenhänge mit solchen Außenkriterien nicht zur Validität, sondern eher zu anderen Testgütekriterien wie bspw. Nützlichkeit. In Messicks (1989) „vereinheitlichter“ Validität spielen solche Zusammenhänge hingegen durchaus eine Rolle, sind jedoch nicht unter dem Begriff „Kriteriumsvalidität“ zusammengefasst, sondern zählen zur Evaluation der Maßnahmen, die aus der Interpretation der Testergebnisse abgeleitet werden. Dennoch stimmen Borsboom und Mellenbergh (2007) sowie Messick (1989) darin überein, dass es sich bei der Vorhersagekraft des Testergebnisses für derartige Außenkriterien um ein wichtiges Gütekri-

terium eines Tests handelt. Daher sollen diesbezügliche empirische Befunde auch im Rahmen der vorliegenden Arbeit betrachtet werden, in Anlehnung an die gängige Testkonstruktionspraxis unter der Überschrift „Kriteriumsvalidierung“ (gleichwohl eine Thematisierung unter einer anderen Überschrift ebenso möglich wäre).

Bei der Auswahl geeigneter Kriterien für die Kriteriumsvalidierung des KKS sind einige der Rahmenbedingungen des Projektes „Konstruktion psychometrischer Fähigkeitstests für Chipdesigner“ (siehe auch Kapitel 1) von Bedeutung, im Rahmen dessen der KKS entstanden ist. Im Folgenden sollen daher einige Überlegungen bezüglich Kriteriumsvalidierung aus der Historie dieses Projektes vorgestellt werden.

Zunächst lag es nahe, dass der zukünftige Berufserfolg von Chipdesignern ein hervorragendes Kriterium für die Bestimmung der prognostischen Validität des KKS wäre. Schließlich identifizierte eine im Rahmen des Projektes durchgeführte Anforderungsanalyse das Ziehen logisch korrekter Konditionaler Schlüsse als tätigkeitsrelevant für Chipdesigner (Böhme & Steyer, 2008). Die Umsetzung einer entsprechenden Studie war im Rahmen dieses Projektes aus einer Vielzahl von Gründen jedoch nicht möglich (vgl. Böhme & Steyer, 2008), sodass sich die Frage stellte, welches zumindest ähnliche Kriterium verwendet werden könnte. Der Berufserfolg einer arbeitstätigen Stichprobe schien ungeeignet, da eine solche Stichprobe vermutlich zu heterogen wäre, um Rückschlüsse auf eine spezielle Tätigkeit wie Chipdesigner ziehen zu können. Zudem schien die Rekrutierung einer solchen Stichprobe generell schwierig. Hinzu kommt die umstrittene Operationalisierbarkeit von Berufserfolg (Stern & Hardy, 2004). Realistisch schien hingegen die Rekrutierbarkeit einer studentischen Stichprobe, die zudem den Vorteil hätte, dass das entsprechende Erfolgskriterium „Studiumserfolg“ durch Studiumsnoten recht einfach zu operationalisieren ist (für einen Überblick über Operationalisierungsmöglichkeiten von Studiumserfolg siehe z.B. Trapmann, 2007). Um nun jedoch die interessierende Tätigkeit Chipdesigner zu berücksichtigen, sollte der Zusammenhang des KKS mit dem Studiumserfolg speziell für solche Studiengänge betrachtet werden, deren Absolventen zumindest potenziell Chipdesigner werden können. Das sind laut der 2005 durchgeführten VDE-Ingenieurstudie[65] vor allem die Studiengänge Informatik, Physik und Elektrotechnik. Gleichwohl Sophistiziertheit Konditionalen Schlussfolgerns möglicherweise auch im Rahmen anderer Studiengänge relevant ist, sollen darüber keine Aussagen getroffen werden. Aufgrund der Ergebnisse der erwähnten Anforde-

[65] erhältlich unter www.vde.com/reports (Zugriffsmöglichkeit letztmalig überprüft am 21.02.2010)

rungsanalyse wird dies speziell für die Studiengänge Informatik, Physik und Elektrotechnik postuliert und überprüft. Innerhalb einer recht umfangreichen Studierenden-Stichprobe (vgl. Grohmann, 2008), die den KKS bearbeitet hatte, konnte eine Teilstichprobe von Studierenden der Informatik, Physik oder Elektrotechnik identifiziert werden. Allerdings existierten von dieser Stichprobe lediglich Daten zu einem Erhebungszeitpunkt, sodass die Vorhersage künftigen Studiumserfolges im Sinne prognostischer Validität nicht möglich war. Jedoch war es möglich, zumindest folgende Hypothese zur Kriteriumsvalidierung zu prüfen:

Hypothese Kriteriumsvalidierung_1: Es besteht ein Zusammenhang zwischen Sophistiziertheit Konditionalen Schlussfolgerns und dem aktuellen Studiumserfolg von Studierenden der Informatik, Elektrotechnik und Physik. Der Effekt liegt dabei im mittleren Bereich.

Der Effekt wird im mittleren Bereich vermutet, da selbst Zusammenhänge zwischen Intelligenz und Studiumserfolg lediglich mittlere Effekte aufweisen (z.B. Giesen, Gold, Hummer & Jansen, 1986; Kuncel, Hezlett & Ones, 2004). Angesichts einer Vielzahl weiterer relevanter Prädiktoren für Studiumserfolg wie bspw. Leistungsmotivation (Robbins et al., 2004) oder auch Persönlichkeitseigenschaften und dabei insbesondere Gewissenhaftigkeit (O'Connor & Paunonen, 2007; Trapmann, Hell, Hirn & Schuler, 2007) überrascht die lediglich moderate Prädiktion durch Intelligenz nicht. Da SKS lediglich durch ein eng umrissenes Spektrum verschiedener Aufgabeninhalte zum Konditionalen Schlussfolgern erfasst wird, während Intelligenz durch ein deutlich vielfältigeres Aufgabenspektrum indiziert wird (vgl. z.B. Intelligenzstrukturtheorien und -modelle in Abschnitt 2.1.4.1), ist es unplausibel, anzunehmen, dass die Effektstärke des Zusammenhangs mit SKS höher ausfallen könnte als mit Intelligenz. Viel wahrscheinlicher ist sogar, dass der Effekt geringer ausfällt.

Im Rahmen der Kriteriumsvalidierung lassen sich so viele verschiedene kriteriumsbezogene Zusammenhänge analysieren (und damit validitätsbezogene Aussagen treffen) wie es Kriterien gibt. Daher soll der Zusammenhang von SKS mit mindestens noch einem weiteren (Erfolgs-)Kriterium betrachtet werden. Hierfür wird ein Kriterium gesucht, das bei studentischen Stichproben generell verfügbar ist. Im Sinne der retrospektiven Validität bietet sich dazu die Abiturnote als Indikator für den (zurückliegenden) Schulerfolg an. Es ist davon auszugehen, dass alle Studierenden eine entsprechende Note erhalten haben und diese bei einer Befragung auch erinnern. Gleichwohl sich

die Abitur-Gesamtnote verschieden zusammensetzen kann, ist sie ein generelles Schulerfolgskriterium, das auch bei Studierenden unterschiedlicher Fachrichtungen eine zumindest ähnliche Bedeutung hat. Eine Beschränkung auf Studierende der Informatik, Elektrotechnik und Physik scheint daher nicht nötig. Für die studentische Stichprobe von Grohmann (2008) liegt die Abiturnote vor, ebenso wie für die Stichprobe der Konstruktvalidierungsstudie (siehe Abschnitt 3.3.2.1). Die Betrachtung soll für beide Studien getrennt erfolgen, um eine Art *Kreuzvalidierung* durchführen zu können. Die Betrachtung des Zusammenhangs eines Testergebnisses während des Studiums mit der zurückliegenden Abiturnote ist ein häufig gewählter Indikator für retrospektive Validität (Bühner, 2006). Es leitet sich also folgende weitere Hypothese zur Kriteriumsvalidierung ab:

Hypothese Kriteriumsvalidierung_2: Es besteht ein Zusammenhang zwischen Sophistiziertheit Konditionalen Schlussfolgerns und (retrospektiv) der Abiturnote von Studierenden. Der Effekt liegt dabei im geringen Bereich.

Der lediglich geringe Effekt wird aufgrund bereits angeführter Argumentation zu *Hypothese Kriteriumsvalidierung_1* sowie des zum Teil recht großen Zeitabstandes zwischen der Testung und dem Zeitpunkt des Abiturs vermutet. Es folgen kurze Ausführungen zu verwendeten Erhebungsinstrumenten, Untersuchungsstichprobe, Durchführung, Auswertungsmethoden sowie Ergebnissen der Studie, im Rahmen derer die *Hypothesen Kriteriumsvalidierung_1* und *Kriteriumsvalidierung_2* überprüft werden sollen. Die Studie wird im Folgenden als *Kriteriumsvalidierungsstudie* bezeichnet. Kurze Ausführungen zur Überprüfung von *Hypothese Kriteriumsvalidierung_2* an der Stichprobe der Konstruktvalidierungsstudie werden an passender Stelle integriert.

Erhebungsinstrumente

Zur Erhebung von SKS wurde die überarbeitete Version des KKS verwendet. Als Kriterium für Studiumserfolg boten sich wie bereits erwähnt Studiumsnoten an. Um mögliche Erinnerungsfehler zu minimieren, wurde eine konkrete Note gewählt, die Studierende besser erinnern sollten als Einzelnoten: die Durchschnittsnote des Vordiploms bzw. einer vergleichbaren Prüfung. Nach dieser Note wurde direkt gefragt. Es handelt sich also um einen Selbstbericht. Die Abitur-Gesamtnote wurde ebenfalls direkt abgefragt, also auch in Form eines Selbstberichtes erhoben.

Durchführung

Eine ausführliche Beschreibung der Durchführung dieser Studie findet sich bei Grohmann (2008). Es sei an dieser Stelle lediglich erwähnt, dass die Erhebung von September bis Oktober 2007 ebenfalls in Form eines Online-Tests stattfand und die Probanden nach der Angabe einiger demographischer Daten den KKS bearbeiteten, gefolgt von einigen Items zu ihrem Abitur sowie Studium (unter anderem Abiturnote und Durchschnittsnote des Vordiploms). In der Konstruktvalidierungsstudie (Durchführung siehe Abschnitt 3.3.2.2) war die Abiturnote Teil der erfragten demographischen Angaben.

Untersuchungsstichprobe

Aus der von Grohmann (2008) per E-Mail-Aufruf deutschlandweit rekrutierten Studierendenstichprobe bearbeiteten $N = 305$ Personen den KKS vollständig. Sie stellen die „Ausgangsstichprobe" der Kriteriumsvalidierungsstudie dar. Anreiz zur Teilnahme bildete die Verlosung von insgesamt sechs Gutscheinen im Wert von jeweils 10 bis 20 Euro sowie die Möglichkeit, nach Auswertung der Daten ein anonymes, rechnergestütztes Feedback zu den Ergebnissen zu erhalten. Der überwiegende Teil der Probanden machte soziodemographische Angaben zur eigenen Person. Das Durchschnittsalter lag bei 24 Jahren ($Sd = 4{,}2$ Jahre). 31% der Teilnehmer waren Frauen, entsprechend 69% Männer. Im Gegensatz zu den anderen durchgeführten Studien sind hier Männer überrepräsentiert. Dies ist jedoch nicht überraschend, da der Großteil der von Grohmann (2008) rekrutierten Studierenden Informatik (42,8%), Elektrotechnik (13,6%) oder Physik (8,4%) studierte. In diesen Studiengängen ist eine Überrepräsentation von Männern üblich. Von 78 Studierenden dieser drei Fächer existiert eine selbst berichtete Vordiplomnote ($M = 2{,}48$; $Sd = 0{,}76$). Sie stellen die Stichprobe der „potenziellen Chipdesigner" zur Überprüfung von *Hypothese Kriteriumsvalidierung_1* dar. 88,5% dieser „potenziellen Chipdesigner" waren Männer, 10,3% Frauen[66]. Das Durchschnittsalter dieser Substichprobe lag bei 24,5 Jahren ($Sd = 2{,}4$ Jahre).

Die Überprüfung von *Hypothese Kriteriumsvalidierung_2* erfolgt anhand der Daten derjenigen 250 Studierenden der Ausgangsstichprobe, die Angaben zu ihrer Abiturnote ($M = 2{,}16$; $Sd = 0{,}62$) machten. 30,8% von ihnen waren Frauen, 68,4% Männer[67]. Das Durchschnittsalter lag bei 23,5 Jahren ($Sd = 3{,}2$ Jahre). Quasi zur Kreuzvalidierung der Ergebnisse wird *Hypothese Kriteriumsvalidierung_2* zusätzlich für die

[66] 1,2% machte keine Angaben.
[67] 0,8% machte keine Angaben.

154 Studierenden der Konstruktvalidierungsstudie (Stichprobenbeschreibung siehe Abschnitt 3.3.2.1) überprüft, von denen ebenfalls Angaben zu ihrer Abiturnote existieren. Abschließend sei erwähnt, dass aus dem deutschlandweiten E-Mail-Aufruf eine recht ausgeglichene Verteilung der Herkunftsbundesländer resultierte. So handelte es sich bei den beiden am häufigsten vertretenen Bundesländern Nordrhein-Westfalen (16,8%) und Baden-Württemberg (16,8%) tatsächlich um bevölkerungsreiche deutsche Bundesländer.

Auswertungsmethoden

Bei den Angaben zur Abiturnote sowie zur Durchschnittsnote des Vordiploms handelt es sich um (manifeste) Werte, über deren Messgenauigkeit keine Aussagen getroffen werden können. Für SKS hingegen ist zunächst wieder das postulierte Messmodell zu überprüfen. Dies soll für die gesamte Stichprobe ($N = 305$) geschehen, da die Schätzung bei diesem Stichprobenumfang deutlich robuster ist, als bei bloßer Betrachtung der 78 Studierenden der Informatik, Elektrotechnik und Physik. Zudem bietet sich einmal mehr die Möglichkeit, die Struktur der überarbeiteten Version des KKS an einer weiteren Stichprobe zu überprüfen. Zur Sicherung der Datenqualität wird wieder die Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5) durchgeführt. Die Prüfung des Messmodells für SKS erfolgt dann durch Beurteilung der Passung eines LCMs unter Vorgabe von vier Klassen bei gleichzeitig guter Interpretierbarkeit dieser vier Klassen im Sinne des erweiterten Stufen-Modells. Analog zum Vorgehen in Abschnitt 3.2.2.2 wird postuliert, dass die Vier-Klassen-Lösung den niedrigsten BIC-Wert aller getesteten Klassen-Lösungen aufweist und gleichzeitig die Kriterien für eine gute Modellpassung bei der Durchführung eines parametrischen Bootstraps erfüllt. Die Interpretation der Klassen erfolgt aufgrund der Zuordenbarkeit der Muster der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Klassen des erweiterten Stufen-Modells. Es wird erneut davon ausgegangen, dass diese Zuordnung bereits per Augenschein möglich ist. Schließlich wird noch die Messgenauigkeit der Vier-Klassen-Lösung durch Beurteilung der Treffsicherheiten bewertet. Die LCA wird mithilfe der Software Winmira 2001 (Davier, 2001) berechnet. Da es sich um die gleichen Auswertungsmethoden wie bei der ersten empirischen Erprobung des KKS handelt, wird auf ausführlichere Beschreibungen verzichtet (siehe dafür Abschnitt 2.4.1.5). Fällt die Treffsicherheit des KKS und damit seine Messgenauigkeit ausreichend hoch aus, soll die Vorhersagbarkeit der Durchschnittsnote des Vordiploms sowie der Abiturnote durch eine Rangkorrelation

manifester Variablen bestimmt werden. Dies ist auch insofern angebracht, da sich Angaben zur Kriteriumsvalidierung eines Tests üblicherweise auf Zusammenhänge des manifesten Testergebnisses mit den gewählten Außenkriterien beziehen. Es wird ein negativer Zusammenhang angenommen, da für klassische (Schul-)Noten (1, …, 6) gilt „je geringer, desto besser" und für die manifeste SKS-Variable (SKS_{man}) „je höher, desto besser", da diese bei passendem Messmodell erneut wie folgt gebildet werden soll:

1 (Rang 1) … Unsophistizierte
2 (Rang 2) … Unsophistizierte II
3 (Rang 3) … Fortgeschrittene
4 (Rang 4) … Sophistizierte

Da es sich um gerichtete Hypothesen handelt, erfolgt die Signifikanzprüfung einseitig. Die Stärke des Effektes wird dann gemäß den Konventionen von Cohen (1990, 1992) für Korrelationskoeffizienten beurteilt (.10 – geringer Effekt, .30 – mittlerer Effekt, .50 – hoher Effekt).

Ergebnisse

In der Folge werden die Ergebnisse der Kriteriumsvalidierungsstudie dargestellt, zunächst in Bezug auf das Messmodell für SKS, anschließend in Bezug auf die Rangkorrelationen.

Messmodell für SKS

Für die Modus-Ponens-Reduktion werden Klassen-Lösungen für 1, 2 und 3 Klassen der vier Modus-Ponens-Aufgaben berechnet. Dabei weist die Ein-Klassen-Lösung (BIC = 123,61) den niedrigsten BIC-Wert auf (Zwei-Klassen-Lösung: BIC = 152,21; Drei-Klassen-Lösung: BIC = 180,81) und stellt damit die bestpassende Klassen-Lösung dar. Für die Prüfung des Modellfits der Ein-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein p-Wert für Chi-Quadrat von $p = .16$ bzw. für Cressie Read von ebenfalls $p = .16$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Ein-Klassen-Lösung. Für die eine identifizierte Klasse ergeben sich für die vier Modus-Ponens-Aufgaben Lösungswahrscheinlichkeiten zwischen .98 und .99. Da diese Lösungswahrscheinlichkeiten allesamt nahe

eins liegen, kann die eine identifizierte Klasse als „Modus-Ponens-Löser“ interpretiert werden (vgl. Abschnitt 2.4.1.5). Eine (Modus-Ponens-)Reduktion der Stichprobe der Konstruktvalidierungsstudie ist demnach nicht nötig[68]. Die beabsichtigten Analysen werden daher an den Daten der kompletten Stichprobe ($N = 305$) durchgeführt.

Für diese 305 Personen werden die Antworten auf die 16 Items des KKS mithilfe verschiedener LCMs analysiert, wobei wie schon bei der ersten empirischen Erprobung zwischen 2 und 7 Klassen vorgegeben werden (vgl. Abschnitt 2.4.1.5). Als bestpassendes Modell wird dasjenige mit dem niedrigsten BIC ausgewählt. Die BIC-Werte der berechneten Klassen-Lösungen sind in Tabelle 27 angegeben.

Tabelle 27: BIC-Werte verschiedener Klassen-Lösungen der 16 Items des KKS in der Kriteriumsvalidierungsstudie

	Vorgegebene Klassenzahl					
	2	3	4	5	6	7
BIC:	3718,41	3529,59	3317,32	3380,09	3442,40	3526,72

Für die 16 Items des überarbeiteten KKS erweist sich demnach eine Vier-Klassen-Lösung als die bestpassende. Für die Prüfung des Modellfits der Vier-Klassen-Lösung werden 200 parametrische Bootstrap-Stichproben gezogen. Dabei ergibt sich ein p-Wert für Chi-Quadrat von $p = .24$ bzw. für Cressie Read von $p = .18$ in der jeweiligen, via Bootstrap generierten Verteilung. Der Vergleich mit dem vorher festgelegten alpha-Niveau ($\alpha = .05$) indiziert eine gute Modellpassung der Vier-Klassen-Lösung.

Die Treffsicherheiten liegen für die einzelnen Klassen zwischen .972 und .987 sowie für die gesamte Vier-Klassen-Lösung bei .980 und damit im hohen Bereich. Sie sind in Tabelle 28 zusammengefasst.

[68] Wahrscheinlich resultiert die durchweg ernsthafte Testbearbeitung aus der Freiwilligkeit der Untersuchung. Es wird davon ausgegangen, dass lediglich solche Personen den kompletten Test bearbeiteten, die auch ein Interesse an einer realistischen Rückmeldung hatten. Dafür spricht auch, dass alle der 305 getesteten Personen die Frage nach ernsthafter Testbearbeitung mit „ja“ beantworteten.

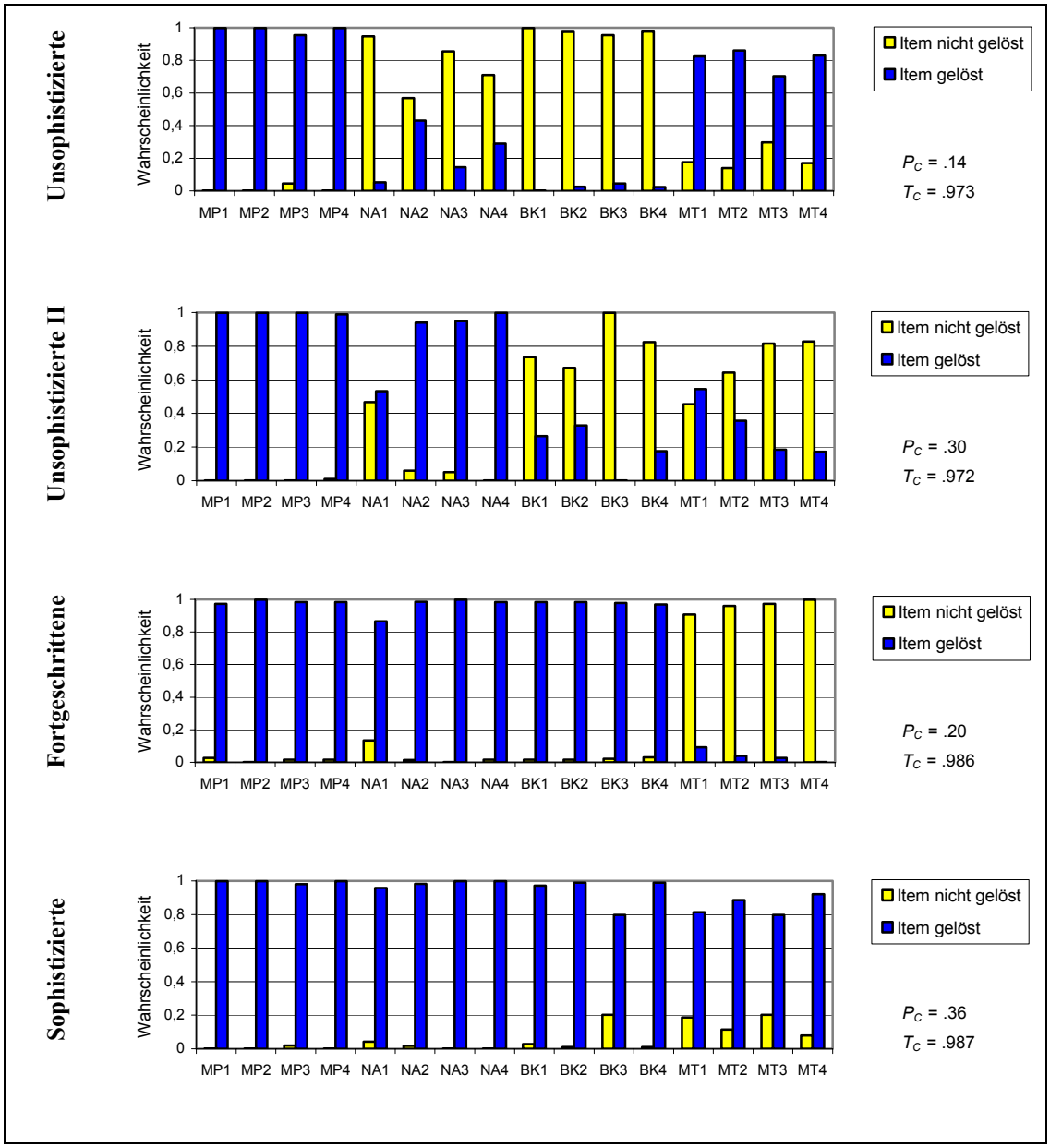
Tabelle 28: Geschätzte Klassengrößen (P_C) und Treffsicherheiten (T_C) der Vier-Klassen-Lösung der 16 Items des KKS in der Kriteriumsvalidierungsstudie

Klasse	geschätzte Klassengröße P_C	Treffsicherheit T_C
Unsophistizierte	$P_{C=1} = .14$	$T_1 = .973$
Unsophistizierte II	$P_{C=2} = .30$	$T_2 = .972$
Fortgeschrittene	$P_{C=3} = .20$	$T_3 = .986$
Sophistizierte	$P_{C=4} = .36$	$T_4 = .987$
Treffsicherheit der Vier-Klassen-Lösung:		$T = .980$

Anmerkung. Die Benennung der Klassen erfolgt auf Basis der Interpretation der klassenbedingten Lösungswahrscheinlichkeiten, welche später in Abbildung 7 (siehe dort) dargestellt sind.

Für den KKS kann in der Kriteriumsvalidierungsstudie folglich von einer guten Passung der Vier-Klassen-Lösung bei hoher Treffsicherheit ausgegangen werden. Bleibt die Frage, ob die vier identifizierten Klassen im Sinne des erweiterten Stufen-Modells interpretiert werden können. Abbildung 7 zeigt die klassenbedingten Lösungswahrscheinlichkeiten der vier identifizierten Klassen.

Da die vier identifizierten Klassen das für die postulierten Sophistiziertheits-Stufen Konditionalen Schlussfolgerns charakteristische Muster der klassenbedingten Lösungswahrscheinlichkeiten aufweisen (vgl. Abbildung 7; für ausführliche Erläuterungen zur Augenscheinprüfung der Zuordnung des Musters der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Stufen des erweiterten Stufen-Modells siehe Abschnitt 2.4.2), kann die Vier-Klassen-Lösung damit auch als gut interpretierbar gelten. Daher braucht keine der Annahmen zur Strukturprüfung aufgrund der Daten verworfen zu werden. Einmal mehr erweist sich damit das postulierte Messmodell für SKS als passend.



Legende. P_C…geschätzte Klassengröße, T_C…Treffsicherheit, MP…Modus Ponens, NA…Negation des Antezedens, BK…Bestätigung der Konsequenz, MT…Modus Tollens, 1…keine Negation in der Hauptprämisse, 2…Negation in der Konsequenz, 3…Negation im Antezedens, 4…Negation in Antezedens und Konsequenz.

Abbildung 7: Klassenbedingte Lösungswahrscheinlichkeiten der Vier-Klassen-Lösung für die Stichprobe der Kriteriumsvalidierungsstudie

Rangkorrelationen

Aufgrund der hohen Treffsicherheit (für die gesamte Vier-Klassen-Lösung: $T = .986$) scheint es vertretbar, SKS als manifeste Variable (SKS_{man}) in den beabsichtigten Rangkorrelationen zu verwenden. Die Ergebnisse für die Koeffizienten Spearmans Rho und Kendalls Tau der Rangkorrelationen von SKS_{man} und der Durchschnittsnote des Vordiploms sowie der Abiturnote sind in Tabelle 29 dargestellt.

Tabelle 29: Rangkorrelationen der manifesten Variable SKS_{man} und der Durchschnittsnote des Vordiploms von Studierenden der Informatik, Elektrotechnik und Physik ($N = 78$) sowie mit der Abiturnote von Studierenden verschiedener Fachrichtungen ($N = 250$)

Variablen	Spearmans Rho (p-Wert)	Kendalls Tau (p-Wert)
SKS_{man} und Durchschnittsnote des Vordiploms ($N = 78$)	-.268 ($p = .009$)	-.212 ($p = .009$)
SKS_{man} und Abiturnote ($N = 250$*)	-.180 ($p = .002$)	-.139 ($p = .002$)

Anmerkung. Der mit * gekennzeichnete reduzierte Stichprobenumfang ergibt sich, weil lediglich 250 der getesteten 305 Studierenden Angaben zu ihrer Abiturnote machten.

Außerdem zeigt Tabelle 30 die Ergebnisse der Rangkorrelation von SKS_{man} und der Abiturnote für die Stichprobe der Konstruktvalidierungsstudie.

Tabelle 30: Rangkorrelationen der manifesten Variable SKS_{man} und der Abiturnote für Psychologie-Studierende der Konstruktvalidierungsstudie ($N = 153$)

Variablen	Spearmans Rho (p-Wert)	Kendalls Tau (p-Wert)
SKS_{man} und Abiturnote ($N = 153$*)	-.180 ($p = .013$)	-.143 ($p = .013$)

Anmerkung. Der mit * gekennzeichnete reduzierte Stichprobenumfang ergibt sich, weil lediglich 153 der getesteten 154 Studierenden der Konstruktvalidierungsstudie Angaben zu ihrer Abiturnote machten.

Es zeigt sich der postulierte negative Zusammenhang zwischen SKS_{man} und der Durchschnittsnote des Vordiploms. Die Stärke des Effektes liegt dabei im geringen bis mittleren Bereich. Der postulierte negative Zusammenhang zeigt sich ebenfalls zwischen SKS_{man} und der Abiturnote, und zwar sowohl für die Stichprobe der Kriteriumsvalidierungsstudie als auch für die Stichprobe der Konstruktvalidierungsstudie, jeweils bei geringen Effekten. Aufgrund dieser Ergebnisse können die *Hypothesen Kriteriumsvali-*

dierung_1 und *Kriteriumsvalidierung_2* beibehalten werden (für eine Diskussion der empirischen Ergebnisse zur Kriteriumsvalidierung siehe Abschnitt 4.1.2).

3.3.4 Zusammenfassende Bewertung der Validierung des KKS

Zusammenfassend lässt sich sagen, dass sich im Rahmen von Inhalts-, Konstrukt- und Kriteriumsvalidierung theoretische Argumente sowie empirische Befunde finden lassen, die eine Erfüllung des Testgütekriteriums Validität für den KKS nahelegen. Unabhängig davon, ob man diese Befunde unter der Überschrift „Validität" zusammenfasst oder entsprechend der eingangs dargestellten kritischen Überlegungen (kein wohldefinierter Begriff etc.) eine andere Überschrift wählt, so kann doch festgehalten werden:

1. Items und Antwortalternativen des KKS lassen sich schlüssig aus kognitionspsychologischen Theorien zum Konditionalen Schlussfolgern ableiten.
2. Die postulierte Struktur des mit dem KKS gemessenen Konstruktes SKS zeigt sich auch empirisch. Außerdem sprechen weitere empirische Befunde für Zusammenhänge zwischen SKS und Konstrukten, mit denen theoretisch ein Zusammenhang bestehen sollte sowie gegen Zusammenhänge zwischen SKS und Konstrukten, mit denen theoretisch kein Zusammenhang bestehen sollte.
3. Zumindest für ein zeitgleich erhobenes, selbst berichtetes Studiumserfolgskriterium (Durchschnittsnote des Vordiploms) wie auch für ein zeitlich zurückliegendes Kriterium (Abiturnote) zeigen sich Zusammenhänge mit dem Testergebnis des KKS.

Es kann also davon ausgegangen werden, dass mit dem KKS tatsächlich Sophistiziertheit Konditionalen Schlussfolgerns gemessen wird, und dass zudem Zusammenhänge mit relevanten Außenkriterien bestehen. Auf Basis theoretischer Überlegungen und empirischer Befunde kann der KKS in diesem Sinne als valide gelten.

Abschließend sei zu Validität angemerkt, dass einige der von Messick (1989) mit zu Validität gerechneten Aspekte in der Testkonstruktionspraxis oft als eigenständige Testgütekriterien behandelt werden. So werden bspw. aus dem Testergebnis resultierende Maßnahmen und deren soziale Konsequenzen häufig unter der Überschrift *Nützlichkeit* (in dieser Arbeit Abschnitt 3.7) oder ethische Fairness nicht selten unter der Überschrift *Fairness* (in dieser Arbeit Abschnitt 3.10) als eigenständige Testgütekriterien betrachtet. Festzuhalten bleibt, dass es sich stets um testrelevante Fragestellungen handelt, die auch im Rahmen der vorliegenden Arbeit betrachtet werden, wenngleich sie

hier (in Anlehnung an die Testkonstruktionspraxis) unter anderen Überschriften behandelt werden.

3.4 Normierung

Normierung oder auch Eichung eines Tests liegt vor, wenn bei der Testung einer Person Vergleichswerte von Personen mit ähnlichen Merkmalsausprägungen auf soziodemographischen Variablen (z.B. Geschlecht, Alter, Schulabschluss) existieren, im Vergleich zu denen die Merkmalsausprägung der Testperson interpretiert werden kann. Dies setzt die Rekrutierung einer umfangreichen Norm- oder Eichstichprobe voraus. Als eine erste solche Normstichprobe könnte die Untersuchungsstichprobe der ersten empirischen Erprobung des KKS (siehe Abschnitt 2.4.1.4) betrachtet werden. Mit einem Umfang von $N = 867$ Personen[69] liegt der Umfang deutlich über der Grenze von $N = 300$ Personen, ab welcher der Umfang einer Normstichprobe als hoch gilt (European Federation of Psychologists' Associations, EFPA, 2008; Fisseni, 1997). Auch der laut COTAN-System (Evers, 2001) erforderliche (Mindest-)Umfang für Normstichproben bei Tests für wichtige Entscheidungen auf individueller Ebene von $N = 400$ wäre in diesem Falle erfüllt.

Neben der Verteilung der Stufen über alle 867 Personen als globale Normwerte, wären die Verteilungen der Stufen innerhalb bestimmter Substichproben (z.B. Frauen, Männer, bestimmte Altersstufen, Hauptschüler, Realschüler, Abiturienten) konkrete Normwerte für vergleichbare Testpersonen. Nun ist jedoch zu beachten, dass nach der ersten empirischen Erprobung leichte Modifikationen des KKS vorgenommen wurden (vgl. Abschnitt 2.4.3.2). Diese überarbeitete Version hat sich in der Stabilitätsstudie (siehe Abschnitt 3.2.2.2), in der Konstruktvalidierungsstudie (siehe Abschnitt 3.3.2) und in der Kriteriumsvalidierungsstudie (siehe Abschnitt 3.3.3) empirisch bewährt und soll daher auch künftig eingesetzt werden. Folglich sollten auch Eichstichproben auf Daten basieren, die mit dieser überarbeiteten Version des KKS gewonnen wurden. Nach Fisseni (1997) sowie den Empfehlungen der EFPA (2008) sind Stichproben ab einem Umfang von $N = 150$ als Normstichproben geeignet. Dies trifft für alle drei mit der überarbeiteten Version des KKS durchgeführten Studien zu. Da die Zusammensetzung der Stichprobe der Kriteriumsvalidierungsstudie (siehe Abschnitt 3.3.3) sehr speziell ist

[69] diejenigen der insgesamt 905 Personen, die nach der Modus-Ponens-Reduktion tatsächlich bei den Analysen berücksichtigt worden sind (vgl. Abschnitt 2.4.2)

(vgl. Grohmann, 2008), ist sie als Normstichprobe jedoch ungeeignet. Aufgrund der Daten der anderen beiden Studien sowie der Zusammensetzung der jeweiligen Untersuchungsstichproben (siehe Abschnitte 3.2.2.2, 3.3.2.1) können für den KKS folgende zwei Normstichproben angegeben werden:

1. die Stichprobe der Stabilitätsstudie[70] für junge Erwachsene mit vergleichsweise hohem Bildungsabschluss (vgl. Abschnitt 3.2.2.2),
2. die Stichprobe der Konstruktvalidierungsstudie für Psychologiestudierende, vor allem Studienanfänger (vgl. Abschnitt 3.3.2.1).

Als Normwerte für Sophistiziertheit Konditionalen Schlussfolgerns können nun die geschätzten (prozentualen) Klassengrößen gelten. Die vorläufigen Normen des KKS sind in Tabelle 31 dargestellt.

Tabelle 31: Vorläufige Normen des KKS

	Normgruppe	
Sophistiziertheits-Stufe	junge Erwachsene mit vergleichsweise hohem Bildungsabschluss (N = 195)	Psychologiestudierende, vor allem Studienanfänger (N = 154)
Unsophistizierte	20%	34%
Unsophistizierte II	45%	45%
Fortgeschrittene	20%	13%
Sophistizierte	15%	8%

Aufgrund des vergleichsweise geringen Umfangs der beiden Normstichproben ist die Angabe differenzierterer Normen (z.B. nach Geschlecht oder Alter) nicht empfehlenswert. Je nach Fragestellung sind weitere spezifische Normstichproben zu erheben. Als Normen für die Normalbevölkerung können vorerst lediglich die Daten der jungen Erwachsenen mit vergleichsweise hohem Bildungsabschluss dienen. Insbesondere von einer Stichprobe der Normalbevölkerung sollten daher in der Folge Normdaten gewon-

[70] Dabei werden die Ergebnisse zum Erhebungszeitpunkt 1 verwendet. Wenngleich Sophistiziertheit Konditionalen Schlussfolgerns aufgrund der Ergebnisse zur Stabilität durchaus als stabiles Merkmal angenommen werden kann (vgl. Abschnitt 3.2.2), scheinen dennoch die Daten zum Zeitpunkt 1 angebrachter. Die Personen werden wahrscheinlich erstmalig mit Aufgaben dieses Typs konfrontiert und gleiches gilt für Personen, die künftig mit dem KKS getestet werden sollen.

nen werden. Da für den KKS jedoch prinzipiell Normen vorliegen, kann das Kriterium der Normierung als erfüllt betrachtet werden.

Ergänzend soll erwähnt werden, dass sich neben der beschriebenen normorientierten Auswertung im Falle des KKS auch eine *kriteriumsorientierte Auswertung* anbieten würde. Da die Testentwicklung auf kognitionspsychologischen Grundlagen aufbaut, ist für jeden Testwert, also jede Sophistiziertheits-Stufe eine spezifische Interpretation möglich (vgl. Ausführungen zur Interpretationsobjektivität in Abschnitt 3.1), aus der dann bspw. Trainingsmaßnahmen abgeleitet werden können (siehe dazu Abschnitt 4.3.2) – ganz im Sinne von Messicks Validitätskonzeption (vgl. Abschnitt 3.3).

3.5 Skalierung

Das Gütekriterium Skalierung wird für einen Leistungstest im Allgemeinen dann als erfüllt betrachtet, wenn eine leistungsfähigere Person auch einen besseren Testwert in dem entsprechenden Test erreicht. Meist sind diese Testwerte Summen- oder Durchschnittswerte einzelner oder aller Aufgaben des Tests. Für den KKS ist diese „übliche" Testwertbestimmung nicht möglich, da anstelle von Summenwerten hier spezielle Muster von Lösungswahrscheinlichkeiten die Grundlage für die Testwertbestimmung (Klassenzuordnung als Ergebnis des LCMs) bilden. Betrachtet man die Testwerte des KKS, also die resultierenden Klassen ihrem Skalenniveau entsprechend als nominal, ist das Gütekriterium der Skalierung nicht umsetzbar (z.B. Moosbrugger & Kelava, 2007). Nach dem theoriegeleiteten Ordnen der Klassen (theoretisch siehe Abschnitt 2.3.2.2, praktisch siehe Abschnitt 2.4.1.5) ist ein Vergleich zwischen Testpersonen jedoch durchaus möglich. Die empirische Ordnung bildet dann auch die konzeptualisierte Ordnung der Sophistiziertheits-Stufen Konditionalen Schlussfolgerns ab. Inwieweit diese Ordnung auch tatsächlich die empirischen Relationen zwischen Merkmalsträgern abbildet, kann nur durch empirische Studien überprüft werden, wie sie bspw. im Rahmen der struktursuchenden Konstruktvalidierung (siehe Abschnitt 3.3.2.2) durchgeführt werden. Es bleibt festzuhalten, dass zumindest nach dem theoriegeleiteten Ordnen der Klassen auch das Gütekriterium der Skalierung für den KKS als erfüllt betrachtet werden kann.

3.6 Testökonomie

Das Gütekriterium der Testökonomie gilt dann als erfüllt, wenn ein Test in Relation zu seinem diagnostischen Erkenntnisgewinn wenige Ressourcen benötigt. Neben zeitlichen und finanziellen Ressourcen zählen dazu insbesondere auch personelle Ressourcen für die Testdurchführung oder -auswertung.

Für computergestützte Tests (wie auch den KKS) kann dieses Gütekriterium bereits aufgrund der Einsparung zeitlicher und personeller Ressourcen (Batinic & Bosniak, 2000; Kleinmuntz & McLean, 1968) wie auch der damit einhergehenden Rationalisierung des gesamten Erhebungsprozesses (Schuler & Höft, 2006) als erfüllt gelten. Die Bearbeitungszeit des KKS liegt lediglich bei 5 bis 10 Minuten, so ist der Zeitaufwand für Probanden äußerst gering. Für Testanwender ist der Zeitaufwand sogar noch geringer, da sie lediglich die computergestützte Auswertung abwarten müssen, welche im Idealfall innerhalb weniger Sekunden erfolgen sollte (für Ausführungen zum konkreten Einsatz des KKS in der Individualdiagnostik siehe Abschnitt 4.3.2). Sind der Test und seine Auswerteroutinen einmal programmiert, entstehen keinerlei zusätzliche Kosten für neue Materialien. Lediglich Pflege und Unterhalt des Testprogramms sollten gewährleistet sein.

In der Beschreibung dieses Testgütekriteriums wird außerdem die Relation zum diagnostischen Erkenntnisgewinn betont, bei der die Beanspruchung von Ressourcen mit der von Tests verglichen wird, die etwas Ähnliches bzw. das Gleiche messen (vgl. z.B. Moosbrugger & Kelava, 2007). Da der KKS der erste Test zur Identifikation der Sophistiziertheits-Stufen Konditionalen Schlussfolgerns ist, ist ein solcher Vergleich jedoch nicht möglich.

Wenngleich das Gütekriterium Testökonomie für den KKS insgesamt also als erfüllt betrachtet werden kann, könnte dennoch ein Kritikpunkt sein, dass vier Items des KKS, nämlich die vier zum Modus Ponens, keinerlei Informationen zur Diskrimination der Stufen liefern, da sie von nahezu allen Personen (unabhängig von deren Sophistiziertheits-Stufe) gelöst werden. Die Diskriminationsfähigkeit der Items zwischen den identifizierten Klassen ist jedoch ein Gütekriterium eines LCMs (Rost, 2004). Im Gegensatz zu den vier Modus-Ponens-Items ist das Antwortverhalten bei den anderen 12 Items indikativ für die Stufen-Zuordnungen der Personen. Diese 12 Items erfüllen also das Gütekriterium der Diskriminationsfähigkeit. Der Test könnte folglich durch Eliminierung der vier Modus-Ponens-Items noch ökonomischer werden bei gleichzeitiger

Verbesserung der (durchschnittlichen) Diskriminationsfähigkeit der Testitems. Allerdings hat die Verwendung der vier Modus-Ponens-Items auch zwei klare Vorteile. Zum einen komplettieren sie das in allgemeinpsychologischen Experimenten vielfach erprobte Negationsparadigma (siehe Abschnitt 2.3.2.1) und erhöhen dadurch die Vergleichbarkeit empirischer Befunde. Zum anderen ermöglichen sie die zur Sicherung der Datenqualität verwendete Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5), deren zentrale Bedeutung in Abschnitt 4.3.1 nochmals thematisiert wird. Die vier Modus-Ponens-Items sollen daher auch künftig Bestandteil des KKS sein, der auch dann noch als sehr ökonomisch bezeichnet werden kann.

3.7 Nützlichkeit

Ein Test gilt als nützlich, wenn das mit ihm gemessene Merkmal praktische Relevanz besitzt. Praktische Relevanz ist dabei durch zweierlei gekennzeichnet. Zum einen besteht praktische Relevanz, wenn ein Test nützliche Anwendungsmöglichkeiten bietet, das heißt, wenn er sich im Rahmen der Kriteriumsvalidierung bei der Vorhersage relevanter Außenkriterien bewährt hat. Dafür sprechen die bereits angeführten empirischen Ergebnisse der Kriteriumsvalidierung des KKS (vgl. Abschnitt 3.3.3). Zum anderen liegt praktische Relevanz vor, wenn kein anderer Test zur Messung des entsprechenden Merkmals existiert, der gleichgute oder bessere Testgütekriterien aufweist. Eine solche Überprüfung ist zunächst nicht möglich, da keine anderen psychodiagnostischen Testverfahren zur Erfassung der Sophistiziertheit Konditionalen Schlussfolgerns existieren. So ist lediglich der Vergleich mit einem Testverfahren möglich, bei dem ebenfalls Konditionale Syllogismen (logisch korrekt) zu lösen sind. Dabei handelt es sich um den „Leistungsprofiltest Schlussfolgerndes Denken – Verbal“ (SDV; Spiel et al., 2004; für einen Überblick siehe Abschnitt 2.1.4.2). Dieser ist damit das einzige (deutschsprachige) Testverfahren, mit dem eine Erfassung einer Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern − zumindest theoretisch − möglich wäre. Dahingehend soll in der Folge ein Vergleich mit dem KKS erfolgen, wenngleich des SDV nicht explizit zur Erfassung einer solchen Fähigkeit oder Kompetenz konzipiert wurde. Es spräche dennoch gegen die Nützlichkeit des KKS, wenn er einem diesbezüglichen Vergleich mit dem SDV nicht standhalten würde.

Zur Vereinfachung werden die in beiden Tests zu lösenden Konditionalen Syllogismen (unabhängig vom betrachteten Test) in der Folge als „Items zum Konditionalen Schlussfolgern“ bezeichnet. Der Vergleich von KKS und SDV soll zuerst auf konzeptueller, dann auf methodischer Ebene erfolgen und schließlich auf Ebene der Testgütekriterien. Zunächst zeigt Tabelle 32 einen groben Vergleich beider Testkonzeptionen.

Tabelle 32: Vergleich von KKS und SDV hinsichtlich theoretischer Grundlage, Zielgruppe und Indikation der Items zum Konditionalen Schlussfolgern

Testkonzeption	KKS	SDV
Theoretische Grundlage der Testkonstruktion	kognitionspsychologisch	entwicklungspsychologisch
Zielgruppe	Erwachsene	Jugendliche
Indikation der Items zum Konditionalen Schlussfolgern	als Indikator für SKS	als Indikator für Schlussfolgerndes Denken

Anmerkung. KKS … Kurztest zum Konditionalen Schlussfolgern, SDV … Leistungsprofiltest Schlussfolgerndes Denken – Verbal (Spiel et al., 2004), SKS … Sophistiziertheit Konditionalen Schlussfolgerns.

Bereits dieser erste Überblick zeigt, dass sich beide Tests in sehr zentralen Punkten unterscheiden. Für die Zielgruppe Erwachsene scheint der SDV aufgrund der (auf Jugendliche bezogenen) entwicklungspsychologischen Grundlagen ungeeignet. Insgesamt spricht für den KKS, dass er explizit auf die Messung einer Kompetenz zum Konditionalen Schlussfolgern (SKS) abzielt, während der SDV die Items zum Konditionalen Schlussfolgern lediglich als Indikatoren für Schlussfolgerndes Denken verwendet.

Eine weitere Frage ist, welche Variationen der Items zum Konditionalen Schlussfolgern in den beiden Tests umgesetzt werden. Dies können Variationen der aussagenlogischen Form oder Variationen des Aufgabeninhaltes sein (vgl. Abschnitt 2.3.1). Einen Überblick dazu gibt Tabelle 33.

Tabelle 33: Vergleichende Bewertung von KKS und SDV hinsichtlich der Variation der Items zum Konditionalen Schlussfolgern

Variation der Items zum Konditionalen Schlussfolgern	KKS	SDV
getrennte Betrachtung aller vier Schlussfiguren	+	–
Variation durch zusätzliche Negation im Antezedenz	+	+
Variation durch zusätzliche Negation in der Konsequenz	+	–
Variation des Aufgabeninhaltes	–	+

Anmerkung. KKS … Kurztest zum Konditionalen Schlussfolgern, SDV … Leistungsprofiltest Schlussfolgerndes Denken – Verbal (Spiel et al., 2004), + … umgesetzt, − … nicht umgesetzt.

Ein Nachteil des SDV (hinsichtlich der Erfassung einer Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern) besteht darin, dass bei der Testauswertung Aufgaben zu verschiedenen Schlussfiguren zusammengefasst werden. Unterschiede in den Lösungswahrscheinlichkeiten zwischen MP- und MT-Aufgaben bspw. werden ignoriert. Diese Unterschiede stellen aber einen zentralen Punkt beider kognitionspsychologischer Theorien zum Konditionalen Schlussfolgern dar (vgl. Abschnitt 2.1.2). Stattdessen fassen Spiel et al. (2004) beide Schlussfiguren zusammen und vernachlässigen diesen kognitionspsychologischen Befund damit völlig. Ebenso werden Aufgaben zu den beiden invaliden Schlussfiguren (NA, BK) von Spiel et al. (2004) zusammengefasst. Dies ist zwar auch auf Basis des ursprünglichen Stufen-Modells theoretisch möglich (vgl. Abschnitt 2.2.2), allerdings nur, wenn keine zusätzlichen Negationen in den Hauptprämissen verwendet werden. Spiel et al. (2004) variieren aber die Hauptprämisse durch eine zusätzliche Negation im Antezedens (vgl. Tabelle 33). Dann sollte sich das Antwortverhalten auf Aufgaben zu diesen beiden Schlussfiguren durchaus unterscheiden (vgl. Abschnitt 2.3.2.1), sodass auch die Zusammenfassung von NA- und BK-Aufgaben aus kognitionspsychologischer Sicht fraglich ist. Da der SDV nicht explizit zur Differenzierung hinsichtlich einer Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern konzipiert ist, wiegt die fehlende Berücksichtigung kognitionspsychologischer Befunde bei der Testauswertung weniger schwer. Dennoch wäre der SDV aufgrund dessen für die Erfassung einer solchen Fähigkeit oder Kompetenz ungeeignet.

Ein Vorteil des SDV liegt hingegen in der Berücksichtigung verschiedener Aufgabeninhalte. Hierbei ist jedoch einschränkend anzumerken, dass Spiel et al. (2004) die Aufgaben nicht getrennt nach den drei Inhaltsbereichen analysieren, sodass unklar ist,

inwieweit hierdurch tatsächlich Vorteile entstehen. Dennoch bleibt festzuhalten, dass die Analyse verschiedener Aufgabeninhalte mit dem SDV prinzipiell möglich wäre.

Hinsichtlich der verwendeten Analysemethoden unterscheiden sich KKS und SDV dahingehend, dass zwar beide Tests mit Modellen der IRT analysiert werden, der KKS jedoch mittels LCA und der SDV unter Verwendung von Mixed-Rasch-Modellen (MRM; Rost, 1989, 1990; siehe auch Rost, 2004). Letztere sind im Grunde eine Kombination aus LCM und klassischem Rasch-Modell. Dabei wird innerhalb der (latenten) Klassen für die Items jeweils die Gültigkeit eines Rasch-Modells angenommen, welches sich hinsichtlich seiner Parameter (z.B. Itemschwierigkeiten) zwischen den Klassen unterscheiden kann. Das klassische Rasch-Modell stellt einen Spezialfall eines MRMs dar, bei dem lediglich eine latente Klasse angenommen wird. LCM und MRM unterscheiden sich dahingehend, dass bei einem LCM innerhalb der Klassen von konstantem Antwortverhalten ausgegangen wird (siehe Abschnitt 2.4.1.5), während man bei einem MRM innerhalb der Klassen noch eine kontinuierliche latente Variable (mit eigener Metrik) annimmt. Dass ein LCM ein passendes Messmodell für SKS und damit den KKS darstellt, zeigt sich in den durchgeführten empirischen Erprobungen (vgl. Abschnitte 2.4.2, 3.2.2.2, 3.3.2.1, 3.3.3). Für den SDV wird die Passung des verwendeten MRMs lediglich vergleichend evaluiert. Es passt besser als andere MRMs, allerdings besteht keinerlei Indikation, dass das Modell überhaupt passt. Dies ist insofern fragwürdig, als bspw. mit parametrischen Bootstrapverfahren (z.B. Efron & Tibshirani, 1993; siehe auch Abschnitt 2.4.1.5) durchaus Methoden existieren, um die Modellpassung eines MRMs zu überprüfen. Ob das postulierte Modell die Daten also tatsächlich gut beschreibt (und nicht nur besser als andere Modelle), bleibt von Spiel et al. (2004) letztlich unbeantwortet. Zudem bietet sich zur Präzisierung eines entwicklungspsychologischen Stufen-Modells prinzipiell ein LCM an (z.B. Borsboom & Mellenbergh, 2007), weshalb die Konzeptualisierung bzw. Modellierung als MRM von Spiel et al. (2004) auch inhaltlich fragwürdig scheint. Dennoch ergibt sich dadurch eine Gelegenheit, das Messmodell für SKS mit einem weiteren, empirisch bereits angewendeten Messmodell zu vergleichen. Aus den theoretischen Annahmen des erweiterten Stufen-Modells (siehe Abschnitt 2.3.2.2) und deren Präzisierung als LCM (vgl. Abschnitt 2.4.1.5) leitet sich für die 16 Items des KKS ein Vier-Klassen-LCM schlüssig als Messmodell für SKS ab (vgl. Abschnitt 2.4.1.5). Demnach sollte ein solches Modell die

Daten der drei (mit der überarbeiteten Version des KKS) durchgeführten Studien[71] jeweils besser beschreiben als ein Vier-Klassen-MRM sensu der Konzeptualisierung von Spiel et al. (2004). Der Vergleich von Vier-Klassen-LCM und Vier-Klassen-MRM erfolgt durch einen Vergleich der BIC-Werte beider Modelle für jede der drei Studien und ist in Tabelle 34 zusammengefasst.

Tabelle 34: Vergleich der BIC-Werte eines Vier-Klassen-LCMs und eines Vier-Klassen-MRMs für die drei durchgeführten Studien

	BIC	
Studie	Vier-Klassen-LCM	Vier-Klassen-MRM
Stabilitätsstudie (Zeitpunkt 1)	2586,59	2594,66
Stabilitätsstudie (Zeitpunkt 2)	2264,26	2267,27
Konstruktvalidierungsstudie	2079,09	2084,62
Kriteriumsvalidierungsstudie	3317,32	3364,33

Anmerkung. LCM … Latente-Klassen-Modell, MRM … Mixed-Rasch-Modell.

Die BIC-Werte sind für das Vier-Klassen-LCM jeweils geringer, was heißt, dass ein Vier-Klassen-LCM die Daten in jeder der drei durchgeführten Studien besser beschreibt als ein Vier-Klassen-MRM. Die Modellierung einer kontinuierlichen (latenten) Variablen innerhalb der latenten Klassen scheint demnach unnötig. Dies stützt die durch das erweiterte Stufen-Modell (siehe Abschnitt 2.3.2.2) implizierte Annahme konstanten Antwortverhaltens auf jeder Stufe und damit innerhalb jeder der vier Klassen. Das sparsamer parametrisierte Vier-Klassen-LCM[72] ist zudem aufgrund des Parsimonitätsprinzips (siehe auch Abschnitt 3.3.2.2) als Messmodell zu bevorzugen. Dieser Vergleich von Vier-Klassen-LCM und Vier-Klassen-MRM könnte auch im Rahmen der strukturprüfenden Konstruktvalidierung (siehe Abschnitt 3.3.2.1) betrachtet werden. Neben der guten Passung des Vier-Klassen-LCMs (sowohl absolut als auch relativ zu anderen Klassen-Lösungen; vgl. Abschnitte 3.2.2.2, 3.3.2.1, 3.3.3), liegen nun auch Befunde zum Vergleich mit einem alternativen und für Items zum Konditionalen

[71] Stabilitätsstudie (siehe Abschnitt 3.2.2.2), Konstruktvalidierungsstudie (siehe Abschnitt 3.3.2.1), Kriteriumsvalidierungsstudie (siehe Abschnitt 3.3.3)

[72] Im Vier-Klassen-MRM sind (bei 16 Items) pro Klasse zusätzlich 17 Personenparameter zu schätzen. Demnach sind im Vier-Klassen-LCM insgesamt 68 Parameter weniger zu schätzen. Es stellt folglich das sparsamer parametrisierte Modell dar.

Schlussfolgern bereits andernorts (Spiel et al., 2004) verwendeten Messmodell vor, die ebenfalls für das Vier-Klassen-LCM als Messmodell für SKS sprechen.

Schließlich fällt noch ein weiterer Vergleich empirischer Befunde zugunsten des KKS aus. Mit dem KKS können alle vier postulierten Sophistiziertheits-Stufen Konditionalen Schlussfolgerns identifiziert werden. Spiel et al. (2004) finden hingegen für eine der vier von ihnen postulierten Stufen, die Stufe der formalen Operationen, keine empirische Evidenz. So bleibt offen, ob diese Stufe mit dem SDV tatsächlich identifiziert werden kann. Abschließend werden beide Tests noch hinsichtlich der Erfüllung der (Haupt-)Testgütekriterien betrachtet. In Tabelle 35 ist dies zusammengefasst.

Tabelle 35: Vergleichende Bewertung von KKS und SDV hinsichtlich der Erfüllung der (Haupt-)Testgütekriterien

(Haupt-)Testgütekriterien	KKS	SDV
Objektivität	+	+
Messgenauigkeit (Reliabilität)	+	+
Validierung in Form von:		
strukturprüfender Konstruktvalidierung	+	–
struktursuchender Konstruktvalidierung	+	–
Kriteriumsvalidierung	+	+

Anmerkung. KKS … Kurztest zum Konditionalen Schlussfolgern, SDV … Leistungsprofiltest Schlussfolgerndes Denken – Verbal (Spiel et al., 2004), + … kann als erfüllt gelten, − … kann nicht als erfüllt gelten (bspw. weil nicht überprüft).

Wenngleich bei beiden Tests alle drei (Haupt-)Testgütekriterien thematisiert werden, so wurden für den KKS auch noch empirische Studien zur strukturprüfenden wie struktursuchenden Konstruktvalidierung durchgeführt. Die empirische Basis für ein integratives Gesamturteil ist demnach umfangreicher als im Falle des SDV.

Gegenüber dem SDV weist der KKS also – zumindest hinsichtlich der Erfassung einer Fähigkeit oder Kompetenz zum Konditionalen Schlussfolgern – Vorteile in der Konzeptualisierung, den verwendeten Methoden, den empirischen Ergebnissen und hinsichtlich der betrachteten Testgütekriterien auf. Beide Aspekte von praktischer Relevanz scheinen damit für den KKS gegeben, sodass auch das Testgütekriterium Nützlichkeit als erfüllt gelten kann.

3.8 Zumutbarkeit

Das Gütekriterium Zumutbarkeit eines Tests bezieht sich auf die mit ihm getestete Person. Diese sollte weder zeitlich noch körperlich oder psychisch über Gebühr belastet werden, und zwar sowohl absolut als auch in Relation zum Nutzen, der aus der Anwendung des Tests resultiert. Mit einer Länge von lediglich 5 bis 10 Minuten kann der KKS zeitlich als vergleichsweise kurz eingeschätzt werden. Er ist körperlich nicht anstrengend und für „Augenfreundlichkeit“ wird durch eine vergleichsweise große und kontrastreiche Schrift gesorgt (siehe Screenshot eines KKS-Items in Anhang A.3). Von psychischen Belastungen ist ebenfalls nicht auszugehen. Lediglich negative Ergebnisrückmeldungen wie bspw. die Zugehörigkeit zur Klasse mit der niedrigsten Kompetenzausprägung könnten Probanden psychisch belasten. Allerdings sollte sich aufgrund der substanziellen Besetzung auch dieser Klasse in den Normstichproben (20% bzw. 34%; vgl. Abschnitt 3.4) diese psychische Belastung in Grenzen halten. Die Zumutbarkeit des KKS scheint damit ebenfalls gewährleistet.

3.9 Unverfälschbarkeit

Das Gütekriterium Unverfälschbarkeit liegt bei einem Test vor, wenn eine Testperson ihr Testergebnis nicht durch gezieltes Verhalten steuern oder verzerren kann. Das gilt insbesondere für Persönlichkeitstests, bei denen bspw. im Sinne sozialer Erwünschtheit geantwortet werden kann, insbesondere dann, wenn die Testperson das Messprinzip durchschaut. Für Leistungstests ist dies im Allgemeinen unproblematisch (Moosbrugger & Kelava, 2007), selbst wenn die Testpersonen das Messprinzip durchschauen. Zudem ist zu beobachten, dass Personen auf niedrigeren Sophistiziertheits-Stufen (Unsophistizierte, Unsophistizierte II, Fortgeschrittene) ihr Antwortverhalten als durchaus richtig annehmen. Bewusste Verfälschungen wären also nur denkbar, wenn eine Testperson ein schlechteres Testergebnis beabsichtigt, was allerdings auch für andere Leistungstests gilt. So kann auch das Gütekriterium der Unverfälschbarkeit für den KKS als erfüllt betrachtet werden.

3.10 Fairness

Das Gütekriterium Fairness gewinnt durch die zunehmende Verbreitung sog. *Culture-Fair-Tests* immer mehr an Bedeutung. Dabei handelt es sich um Tests, die kulturunabhängig eingesetzt werden können, da sie bspw. sprachfrei sind. Fairness liegt vor, wenn bei einem Test bestimmte ethnische, soziokulturelle oder geschlechtsspezifische Gruppen nicht benachteiligt werden. Sprachfrei ist der KKS nicht, allerdings existiert eine englische Version, die kulturübergreifend verständlich sein sollte, vorausgesetzt, gewisse Basiskenntnisse der englischen Sprache sind vorhanden. Die Geschäftssprache der (vorläufigen) Zielgruppe Chipdesigner ist ebenfalls Englisch und der sprachliche Anspruch der Items vergleichsweise niedrig, sodass Fairness zumindest für die vorläufige Zielgruppe gegeben scheint. Bezüglich des Geschlechts ist unklar, ob Frauen durch den Inhalt (Funktionieren elektrischer Schaltungen) möglicherweise benachteiligt werden. In der Instruktion wird zwar explizit darauf hingewiesen, dass zur Bearbeitung der Aufgaben keinerlei Wissen über Elektrotechnik notwendig ist (vgl. Anhang A.1), dennoch kann nicht ausgeschlossen werden, dass Frauen durch solche Inhalte bspw. „abgeschreckt“ werden. Für die Stichprobe zum Erhebungszeitpunkt 1 der Stabilitätsstudie (siehe Abschnitt 3.2.2.2) zeigt sich jedenfalls, dass mehr Männer den Klassen „Fortgeschrittene“ und „Sophistizierte“ zugeordnet werden als erwartet und umgekehrt mehr Frauen den Klassen „Unsophistizierte“ und „Unsophistizierte II“. Die inferenzstatistische Prüfung erfolgt via Chi-Quadrat-Test und liefert verglichen mit dem konventionellen alpha-Niveau ($\alpha = .05$) ein signifikantes Ergebnis ($\chi^2 = 11{,}51$, $df = 3$, $p = .009$). Tabelle 36 zeigt die entsprechende Kreuztabelle.

Dies muss jedoch nicht Folge des Inhalts der Items sein, da bspw. auch für Logisches Denken im Allgemeinen davon ausgegangen wird, dass Männer bessere Ergebnisse erzielen als Frauen (z.B. Halpern, 2000). Die Frage nach der Geschlechter-Fairness kann also nicht endgültig beantwortet werden und bedarf weiterer Forschung. Ebenso sind Studien zur ethnischen Fairness notwendig. Dabei sollte es jedoch leicht sein, entsprechende Items in verschiedenen Sprachen zu generieren, da die Konstruktionsstrategie für die Items und für die Antwortalternativen klar vorgegeben ist (vgl. Abschnitte 2.3.4 und 2.4.1.2). Das Gütekriterium der Fairness kann also nicht a priori als vollständig erfüllt betrachtet werden und bedarf weiterer Forschung.

Tabelle 36: Kreuztabelle „Klasse Sophistiziertheit Konditionalen Schlussfolgerns (SKS) gekreuzt mit Geschlecht“

			Geschlecht		
			weiblich	männlich	gesamt
Klasse SKS	Unsophistizierte	Anzahl	28	11	39
		erwartete Anzahl	24,6	14,4	
	Unsophistizierte II	Anzahl	63	25	88
		erwartete Anzahl	55,5	32,5	
	Fortgeschrittene	Anzahl	18	20	38
		erwartete Anzahl	24	14	
	Sophistizierte	Anzahl	14	16	30
		erwartete Anzahl	18,9	11,1	
gesamt			123	72	195

3.11 Attraktivität

Dieses eher selten betrachtete Testgütekriterium soll ergänzend angeführt werden, da es im Entstehungskontext des KKS von besonderer Bedeutung ist. Attraktivität bezieht sich dabei auf das Testmaterial, und zwar für die Zielgruppe des Tests. Wie in Kapitel 1 dieser Arbeit beschrieben, ist der Entstehungshintergrund des KKS das Projekt „Konstruktion Psychometrischer Fähigkeitstests für Chipdesigner“. Ziel dieses Projektes war es, einen Test für die Zielgruppe Chipdesigner zu entwickeln und im Sinne der Attraktivität auch das Aufgabenmaterial entsprechend zu gestalten. Wenngleich für Konditionalaussagen vielfältige Inhalte denkbar sind, so kann der tatsächlich verwendete Inhalt „Funktionieren elektrischer Schaltungen“ als für Chipdesigner besonders attraktiv gewertet werden. Dies jedenfalls war das Ergebnis eines entsprechenden Workshops mit Experten aus der Chipdesign-Branche (siehe dazu Böhme & Steyer, 2008). Ebenfalls attraktiv für die Zielgruppe Chipdesigner war die Darbietung als Online-Test, die sich seit einigen Jahren zunehmender Beliebtheit und wachsender Verbreitung erfreut (Batinic & Bosniak, 2000; Wilhelm & McKnight, 2002). Das Gütekriterium der Attraktivität kann für die Zielgruppe Chipdesigner in jedem Falle als erfüllt betrachtet werden. Bezüglich anderer Zielgruppen scheint Attraktivität aufgrund der zeitgemäßen Darbietung als Online-Test ebenfalls gegeben.

3.12 Zusammenfassung der Testgütekriterien

Tabelle 37 fasst die für den KKS betrachteten Testgütekriterien zusammen. Dabei wird sowohl dargestellt, ob sie als erfüllt betrachtet werden können als auch wie diese Bewertung begründet wird, ob allein durch theoretische Herleitung oder zusätzlich durch eine empirische Überprüfung.

Tabelle 37: Zusammenfassende Bewertung der Testgütekriterien des KKS

Testgütekriterium	Bewertung	Basis der Bewertung
Objektivität	erfüllt	theoretisch
Messgenauigkeit (Reliabilität)	erfüllt	theoretisch und empirisch
Validität	erfüllt	theoretisch und empirisch
Normierung	erfüllt	theoretisch
Skalierung	erfüllt	theoretisch
Testökonomie	erfüllt	theoretisch
Nützlichkeit	erfüllt	theoretisch
Zumutbarkeit	erfüllt	theoretisch
Unverfälschbarkeit	erfüllt	theoretisch
Fairness	zum Teil erfüllt	theoretisch
Attraktivität	erfüllt	theoretisch

Zusammenfassend bleibt festzuhalten, dass ein Großteil der betrachteten Gütekriterien für den KKS als erfüllt angesehen werden kann. Insbesondere die Hauptgütekriterien Objektivität, Messgenauigkeit (Reliabilität) und Validität können als theoretisch und zum Teil auch als empirisch geprüft gelten. Für die betrachteten Nebengütekriterien gilt dies ebenfalls, mit Ausnahme der Testfairness, welche nur bedingt als erfüllt betrachtet werden kann (vgl. Abschnitt 3.10). Zumindest wird für Testfairness jedoch aufgezeigt, wie deren Prüfung erfolgen kann bzw. welche Entwicklungen noch notwendig sind, um auch Testfairness als erfüllt betrachten zu können. Infolge der angestellten Betrachtungen zu Testgütekriterien scheint es durchaus legitim, den KKS als „Psychologischen Test“ im eingangs dieses Kapitels beschriebenen Sinne zu bezeichnen. Dennoch existieren auch eine Reihe kritischer Punkte sowie eine Vielzahl resultierender Forschungsfragen. Unter anderem diese sollen Gegenstand der nun folgenden Diskussion sein.

4 Diskussion

Zum Abschluss dieser Arbeit erfolgt eine ausführliche Diskussion der bisherigen Teile. Zwar wird die erste empirische Erprobung des KKS bereits in Abschnitt 2.4.3 diskutiert, allerdings wird dort darauf verwiesen, dass eine ausführliche theoretische Diskussion erst am Ende der Arbeit erfolgt, da dann aufgrund weiterer durchgeführter Studien eine sehr viel umfangreichere empirische Grundlage gegeben ist. Daher werden zunächst die empirischen Befunde vor dem Hintergrund der theoretischen Ausführungen zu Beginn dieser Arbeit (siehe Abschnitte 2.1 bis 2.3) sowie die empirisch überprüften Testgütekriterien kritisch gewürdigt. Anschließend werden die verwendeten Methoden diskutiert und zum Abschluss Implikationen für die Forschung und die psychologische Praxis abgeleitet.

4.1 Diskussion der Ergebnisse vor dem theoretischen Hintergrund

Gegenstand dieses Abschnitts ist zunächst die kritische Bewertung der Annahmen, die im Rahmen der Testentwicklung (siehe Abschnitte 2.1 bis 2.3) getroffen werden. Ein Schwerpunkt liegt dabei auf der theoretischen Herleitung des erweiterten Stufen-Modells, da dieses die konzeptionelle Basis der Testentwicklung bildet. Nach der nunmehr mehrfachen empirischen Überprüfung ist es möglich, auch zentrale theoretische Annahmen auf Basis einer umfangreichen empirischen Grundlage zu bewerten. Anschließend werden diejenigen der betrachteten Testgütekriterien nochmals thematisiert, die neben der theoretischen Herleitung auch empirisch überprüft werden (Messgenauigkeit und Validität; siehe Tabelle 37 in Abschnitt 3.12).

4.1.1 Diskussion der Testentwicklung

Die theoretischen Grundlagen der Testentwicklung werden in der Folge vor dem Hintergrund der durchgeführten empirischen Studien betrachtet. Doch zunächst sollen zwei der wichtigsten Testentwicklungsschritte kurz diskutiert werden.

Im ersten Schritt der Testkonstruktion wird die Aussagenlogik als Bezugssystem zur Bewertung der Korrektheit der Konklusion eines Konditionalen Schlusses festgelegt (vgl. Abschnitt 2.1.1). Sensu Gigerenzer und Gaissmaier (2006) muss ein (psychologisch) korrekter Schluss jedoch nicht zwingend formal-logisch korrekt sein. So kann im Alltag auch eine bikonditionale Interpretation einer Konditionalaussage durchaus (psychologisch) sinnvoll sein, vorausgesetzt, sie erfasst die Absichten des Senders korrekt. Man betrachte hierzu das Beispiel aus Abschnitt 2.1.2.2 „*Wenn Du den Rasen mähst, gebe ich Dir 5 Dollar*“. Das Akzeptieren der einladenden Inferenz „*Wenn Du den Rasen nicht mähst, gebe ich Dir keine 5 Dollar*“ und die daraus resultierende bikonditionale Interpretation der ursprünglichen Konditionalaussage können für den Erhalt der „*5 Dollar*“ durchaus förderlich sein. Dennoch spricht im Falle des KKS − neben den bereits in Abschnitt 2.1.1 angeführten Argumenten − ein letztlich entscheidender Grund für die Wahl der Aussagenlogik, und zwar die Zielgruppe des KKS: Chipdesigner (vgl. Kapitel 1). Korrektes Konditionales Schlussfolgern ist für Chipdesignern bspw. bei der Durchführung von Simulationen relevant (siehe dazu das Beispiel in der Einleitung dieser Arbeit). Die bei diesen Simulationen ablaufenden informationstechnischen Prozesse sind streng logisch aufgebaut. Demnach kommen im Falle des KKS nur formal-logisch korrekte Schlüsse als korrekte Schlüsse in Frage, andernfalls wäre der Test in Bezug auf diese Zielgruppe und deren Tätigkeit ungeeignet.

In weiteren Vorüberlegungen zur Testentwicklung wird eine *Fähigkeit zum Konditionalen Schlussfolgern* als eine „Teilfähigkeit“ des Deduktiven Denkens hergeleitet (vgl. Abschnitt 2.1.4.1), bei der es sich folglich − wie beim Deduktiven Denken − um eine dimensionale Fähigkeit handeln sollte. Bereits in den weiteren (theoretischen) Testentwicklungsschritten (siehe insbesondere Abschnitt 2.2) zeigt sich jedoch, dass sich aus den betrachteten kognitionspsychologischen Theorien keine solche dimensionale Fähigkeit ableitet, sondern ein (Kompetenz-)Stufen-Modell Konditionalen Schlussfolgerns. Dies hat Konsequenzen, bspw. für die Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns (siehe dazu auch Abschnitt 4.1.2) oder auch für deren Trainierbarkeit (siehe dazu Abschnitt 4.3.2). Auf Grundlage der empirischen Ergebnisse folgt nun die Diskussion der weiteren (theoretischen) Testentwicklungsschritte, wobei der Fokus insbesondere auf der Herleitung dieses Kompetenz-Stufen-Modells, also des erweiterten Stufen-Modells liegt. Dazu wird zunächst die Argumentation bei der Testentwicklung (siehe Abschnitte 2.1 bis 2.3) jeweils kurz nachvollzogen und dann auf Grundlage der empirischen Ergebnisse bewertet. Abschließend wird zudem überprüft,

inwieweit die in Abschnitt 2.4.3.3 als Ergebnis der Testentwicklung resultierenden Forschungsfragen beantwortet werden können.

Diskussion zum Stufen-Modell Konditionalen Schlussfolgerns

Entsprechend der Forderung einer kognitionspsychologischen Fundierung der Testentwicklung wird das aus der Logiktheorie (Braine & O'Brien, 1991; siehe auch Abschnitt 2.1.2.2) abgeleitete Stufen-Modell zum Konditionalen Schlussfolgern (Rijmen & De Boeck, 2003) als Grundlage für die Testentwicklung gewählt (vgl. Abschnitt 2.2). Die darin postulierten drei Stufen (Unsophistizierte, Fortgeschrittene, Sophistizierte) lassen sich in jeder der vier durchgeführten Studien in einem passenden, messgenauen und gut interpretierbaren Messmodell identifizieren (vgl. Abschnitte 2.4.2, 3.2.2.2, 3.3.2.1, 3.3.3). Damit findet sich erstmals auch empirische Evidenz für die Sophistizierten-Stufe, welche Rijmen und De Boeck (2003) nicht identifizieren können. Eine Ursache könnte das vergleichsweise hohe Bildungsniveau aller vier Untersuchungsstichproben der vorliegenden Arbeit sein (vgl. Abschnitte 2.4.1.4, 3.2.2.2, 3.3.2.1, 3.3.3) oder aber die Verwendung von Disjunktionen und Konjunktionen in den Items von Rijmen und De Boeck (2003; vgl. Abschnitt 2.3.1). Neben der Identifizierbarkeit der Stufen ist eine zweite entscheidende Annahme zum Stufen-Modell die Ordnung der Stufen hinsichtlich Reasoning (vgl. z.B. Abbildung 2 in Abschnitt 2.2.3), die sich empirisch ebenfalls bestätigt, und zwar konsistent in beiden dazu durchgeführten Studien (vgl. Abschnitte 2.4.2, 3.3.2.2). Die vorliegende Arbeit liefert damit empirische Evidenz für das Stufen-Modell von Rijmen und De Boeck (2003) und stützt so die Entscheidung, die Testkonstruktion auf diesem Modell und damit auf der Logiktheorie (Braine & O'Brien, 1991) aufzubauen. Die immense Bedeutung der Modelltheorie (Johnson-Laird & Byrne, 2002) für das Konditionale Schlussfolgern bleibt davon unberührt. So lassen sich bspw. auch die Befunde zum Stufen-Modell – zumindest teilweise – auf Basis der Modelltheorie erklären (vgl. Rijmen & De Boeck, 2003).

Als weiteres Argument für die Wahl des Stufen-Modells als Grundlage für die Testentwicklung wird in Abschnitt 2.2.3 seine Erklärungskraft für zum Teil kontraintuitive empirische Befunde angeführt, wie bspw. die negative Korrelation zwischen der Anzahl gelöster Modus-Tollens-Probleme und Reasoning. In der vorliegenden Arbeit zeigt sich – wie postuliert – ein positiver Zusammenhang von SKS und Reasoning (vgl. Abschnitte 2.4.2, 3.3.2.2). Korreliert man jedoch bspw. die Anzahl der gelösten Modus-Tollens-Aufgaben mit Reasoning (indiziert durch den Summenwert der 10 Matrizeni-

tems; vgl. Abschnitt 2.4.1.2), so ergibt sich für die Daten der ersten empirischen Erprobung des KKS ($N = 742$)[73] eine Nullkorrelation ($r = -.001$, $p = .484$). Dieses Ergebnis passt zu den Befunden von Evans et al. (2007), die ebenfalls keinen (signifikanten) Zusammenhang zwischen dem Lösen von Modus-Tollens-Aufgaben und den Ergebnissen klassischer Reasoning-Tests finden. Dass diese Korrelation bei der ersten empirischen Erprobung des KKS nicht negativ ausfällt, liegt möglicherweise an dem hohen Anteil Sophistizierter (15,5%; vgl. Abschnitt 2.4.2) in dieser Stichprobe. Nach dem Stufen-Modell lösen schließlich auch Sophistizierte Modus-Tollens-Aufgaben korrekt, sodass mit steigendem Anteil Sophistizierter auch die (lineare) Korrelation ansteigen sollte.

Als weiterer durch das Stufen-Modell erklärbarer empirischer Befund wird in Abschnitt 2.3.1 der positive Zusammenhang zwischen dem Lösen von NA- sowie BK-Aufgaben mit Reasoning angeführt. Betrachtet man für diese Schlussfiguren die einfache lineare Korrelation zwischen der Anzahl gelöster Aufgaben und dem Summenwert der 10 Matrizenitems in der ersten empirischen Erprobung des KKS, so zeigt sich der beschriebene Effekt (NA: $r = .278$, $p < .001$; BK: $r = .287$, $p < .001$). Damit lassen sich bisherige Befunde (De Neys et al., 2005; Evans et al., 2007; teilweise Newstead et al., 2004) anhand der Daten der ersten empirischen Erprobung des KKS durchaus replizieren, wenngleich eine Zusammenhangsanalyse mittels linearer Korrelationen unangebracht ist, zumindest wenn man Theorie und Empirie der vorliegenden Arbeit berücksichtigt. Dennoch lässt sich so die (bislang lediglich theoretisch argumentierte) Erklärungskraft des Stufen-Modells für diese zum Teil kontraintuitiven empirischen Befunde (vgl. Abschnitt 2.3.1) an den Daten der ersten empirischen Erprobung des KKS illustrieren.

Diskussion zum erweiterten Stufen-Modell Konditionalen Schlussfolgerns

Den nächsten wichtigen (theoretischen) Testentwicklungsschritt stellt die Erweiterung des Stufen-Modells auf das Negationsparadigma (also die Verwendung zusätzlicher Negationen in der Hauptprämisse; vgl. Abschnitt 2.3.2.2) dar. Zunächst wird davon ausgegangen, dass sich die drei Stufen des ursprünglichen Stufen-Modells für einen Teil der Personen unabhängig von der Verwendung zusätzlicher Negationen in der Haupt-

[73] Es handelt sich um die gleiche Stichprobe, für die auch die multinomiale logistische regressive Abhängigkeit der Sophistiziertheit Konditionalen Schlussfolgerns von Reasoning berechnet wurde (vgl. Abschnitt 2.4.2).

prämisse zeigen. Dies wird durch die in diesem Abschnitt bereits angeführten Befunde zum ursprünglichen Stufen-Modell gestützt. Darüber hinaus werden (hauptsächlich basierend auf dem Phänomen des Negative Conclusion Bias, NCB) vier weitere Annahmen getroffen, aus denen dann eine Negationen-sensitive Stufe (Unsophistizierte II) hergeleitet wird (vgl. Abschnitt 2.3.2.1):

1. Effekte zusätzlicher Negationen in der Hauptprämisse zeigen sich lediglich bei den Schlussfiguren NA und MT (vgl. Abschnitt 2.3.2.1).
2. Ist zur Lösung einer Aufgabe das Auflösen einer doppelten Negation notwendig, werden weniger eindeutige Inferenzen gezogen (und entsprechend häufiger bspw. die „keine Aussage möglich"-Reaktion gewählt). Dies trifft auf die Items NA2 und NA4 sowie MT3 und MT4 zu (Nomenklatur siehe Abschnitt 2.3.4; Items siehe Anhang A.2). Entsprechend steigt für die Items NA2 und NA4 die Lösungswahrscheinlichkeit, während sie für MT3 und MT4 sinkt (vgl. Abschnitt 2.3.2.2).
3. Beim Lösen der Items NA3 und MT2 tritt ebenfalls eine doppelte Negation auf, allerdings ist diese nicht aktiv aufzulösen, sondern lediglich zu enkodieren. Dennoch sollte auch für Item NA3 die Lösungswahrscheinlichkeit ansteigen sowie für Item MT2 sinken, jedoch nicht so stark wie bei den jeweils anderen beiden in Annahme 2 angeführten Items zu diesen Schlussfiguren (vgl. Abschnitt 2.3.2.2).
4. Die Unsophistizierten-Stufe II ordnet sich hinsichtlich Reasoning zwischen der Unsophistizierten- und der Fortgeschrittenen-Stufe ein (vgl. Abschnitt 2.3.2.2).

Für die erste Annahme sprechen die empirischen Befunde aller vier durchgeführten Studien. Die Lösungswahrscheinlichkeiten bei den vier MP-Items wie auch bei den vier BK-Items sind innerhalb jeder Klasse (also auf jeder Sophistiziertheits-Stufe) vergleichsweise konstant und damit unabhängig von der Verwendung zusätzlicher Negationen in der Hauptprämisse (vgl. Abschnitte 2.4.2, 3.2.2.2, 3.3.2.1, 3.3.3). Einzige Ausnahme bildet die „Restklasse" der ersten empirischen Erprobung des KKS, in der die vier BK-Items unterschiedliche Lösungswahrscheinlichkeiten aufweisen (vgl. Abbildung 3 in Abschnitt 2.4.2). Da diese „Restklasse" jedoch als Resultat der sequentiellen Itemvorgabe in der ersten empirischen Erprobung des KKS angenommen wird (vgl. Abschnitt 2.4.3.1) und sich (folgerichtig) nicht stabil über die vier durchgeführten Studien hinweg zeigt, soll dieser Befund vernachlässigt werden.

Auch die zweite und die dritte Annahme werden durch das Antwortverhalten jeweils einer der identifizierten Klassen[74] in allen vier Studien gestützt (vgl. Abschnitte 2.4.2, 3.2.2.2, 3.3.2.1, 3.3.3). Es zeigt sich das für die Unsophistizierten-Stufe II postulierte Muster der Lösungswahrscheinlichkeiten (siehe präzisierte *Arbeitshypothese 1b* in Abschnitt 2.4.1.5). Allerdings fällt die Lösungswahrscheinlichkeit des Items NA3 (in dieser Klasse) nicht geringer aus als die der Items NA2 und NA4. Stattdessen sind die Lösungswahrscheinlichkeiten dieser drei Items in etwa gleich hoch. Der gleiche Effekt, jedoch in umgekehrter Richtung, zeigt sich für Item MT2. Auch dessen Lösungswahrscheinlichkeit ist (in dieser Klasse) in etwa so gering wie die der Items MT3 und MT4 und nicht – wie postuliert – etwas höher (vgl. Abschnitte 2.4.2, 3.2.2.2, 3.3.2.1, 3.3.3). Der Argumentation aus Abschnitt 2.3.2.2 folgend wäre demnach das Enkodieren einer doppelten Negation ebenso schwierig wie deren aktives Auflösen. Dies widerspricht jedoch der bisherigen Interpretation empirischer Befunde zum NCB (z.B. Kleinbeck, 2005). Allerdings beziehen sich diese stets auf prozentuale (Lösungs-)Häufigkeiten jedes einzelnen Items und nie auf das Antwortmuster aller 16 Items des Negationsparadigmas. Zunächst wäre daher in Reanalysen der Daten klassischer Negationsparadigma-Experimente zu prüfen, wie sich das Antwortverhalten bei Betrachtung der Antwortmuster darstellt, bspw. mittels Analyse latenter Klassen. Damit könnte gleichzeitig der Frage nachgegangen werden, ob möglicherweise der Item-Inhalt des KKS „Funktionieren elektrischer Schaltungen“ (siehe Abschnitt 2.3.3) Ursache dieser Befunde ist. Zeigen sich die − durch das erweiterte Stufen-Modell intendierten − charakteristischen Muster der klassenbedingten Lösungswahrscheinlichkeiten auch bei Items anderen (bspw. abstrakten) Inhaltes, dann kann der Item-Inhalt „Funktionieren elektrischer Schaltungen“ als Ursache ausgeschlossen werden.

Die vierte Annahme wird durch die Ergebnisse ebenfalls gestützt. Sowohl in der ersten empirischen Erprobung des KKS (siehe Abschnitt 2.4.2) als auch in der Konstruktvalidierungsstudie (siehe Abschnitt 3.3.2.2) kann aufgrund der Ordnung der Anstiegskoeffizienten in der multinomialen logistischen Regression davon ausgegangen werden, dass sich die Unsophistizierten-Stufe II hinsichtlich Reasoning zwischen der Unsophistizierten- und der Fortgeschrittenen-Stufe einordnet.

Dass die Ergebnisse aller vier durchgeführten Studien das erweiterte Stufen-Modell prinzipiell stützen, spricht gleichzeitig dagegen, dass es sich beim NCB um ein

[74] Dabei handelt es sich jeweils um die Klasse, die als „Unsophistizierte II“ bezeichnet wird.

allgemeinpsychologisches Phänomen handelt. Ansonsten wäre dieser Effekt zusätzlicher Negationen in der Hauptprämisse auf allen drei Stufen des ursprünglichen Stufen-Modells (Rijmen & De Boeck, 2003; siehe Abschnitt 2.2.2) zu beobachten. Gegen ein allgemeinpsychologisches Phänomen spricht zudem, dass sich die Negationen-sensitive Stufe (Unsophistizierte II) offenbar auch zeitlich stabil zeigt (vgl. Ergebnisse der Stabilitätsstudie in Abschnitt 3.2.2.2, insbesondere Tabelle 18). Anfällig für die Effekte von zusätzlichen Negationen in der Hauptprämisse scheinen demnach zum zweiten Erhebungszeitpunkt vorrangig wieder dieselben Personen zu sein wie zum ersten Erhebungszeitpunkt. So ergeben sich aus der differenzialpsychologischen Betrachtung des Negationsparadigmas zum Konditionalen Schlussfolgern (theoretisch begründete) empirische Befunde, die der kognitionspsychologischen Forschung zum Negationsparadigma neue Impulse geben können.

Diskussion zu Inhaltseffekten

Ein weiterer (theoretischer) Diskussionspunkt zur Testentwicklung adressiert die mangelnde Berücksichtigung von Inhaltseffekten, deren Bedeutung beim Konditionalen Schlussfolgern unumstritten ist (vgl. Abschnitt 2.3.3). Zwar ist der gewählte Item-Inhalt „Funktionieren elektrischer Schaltungen“ das Resultat pragmatischer Vorgaben der Testentwicklung (vgl. Abschnitt 2.3.3) und sollte aufgrund dessen zumindest zu hoher Akzeptanz bei der Zielgruppe (Chipdesigner) führen, dennoch bleibt unklar, inwieweit das erweiterte Stufen-Modell gilt, wenn die aussagenlogischen Veränderungen in andere Inhalte eingebettet werden. Da der Inhalt der Items jedoch in keiner der durchgeführten Studien experimentell manipuliert wird, liegen hierzu in dieser Arbeit keine empirischen Ergebnisse vor. Stattdessen ergibt sich eine Vielzahl von Implikationen für weitere Studien, weshalb eine ausführlichere Diskussion von Inhaltseffekten erst erfolgen soll, wenn aus den Ergebnissen dieser Arbeit weitere Forschungsfragen abgeleitet werden (siehe dazu Abschnitt 4.3.1).

Diskussion der resultierenden Forschungsfragen der Testentwicklung

Neben der Bestimmung von Testgütekriterien ergeben sich zwei wesentliche Forschungsfragen aus der Diskussion der ersten empirischen Erprobung des KKS (siehe Abschnitt 2.4.3.3), zum einen die Frage nach der Replizierbarkeit der Ergebnisse und zum anderen die Frage nach der Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns. Nach drei weiteren durchgeführten Studien lässt sich festhalten, dass die vier

Stufen des erweiterten Stufen-Modells jeweils repliziert werden können – stets bei adäquater Modellpassung, hoher Treffsicherheit und guter Interpretierbarkeit (vgl. Abschnitte 3.2.2.2, 3.3.2.1, 3.3.3). Auch der postulierte Zusammenhang zwischen SKS und Reasoning kann repliziert werden (vgl. Abschnitt 3.3.2.2), wobei letzteres – wie in Abschnitt 2.4.3.1 gefordert – mit einem etablierten Matrizentest, dem WMT, erhoben wird (vgl. Abschnitt 3.3.2.2). Die zweite aus der Testentwicklung resultierende Forschungsfrage adressiert die Stabilität von SKS. Gemäß der zweiten Forderung an die vorliegende Arbeit (Verwendung von Latente-Variablen-Modellen) wird diese Stabilität mittels einer LTA analysiert (vgl. Abschnitt 3.2.2.2). Die dabei betrachteten Wahrscheinlichkeiten der latenten Klassenübergänge (siehe Tabelle 17 in Abschnitt 3.2.2.2) sowie die daraus resultierenden Stabilitätskoeffizienten sprechen für die Stabilität der Sophistiziertheit Konditionalen Schlussfolgerns über die Zeit (vgl. Abschnitt 3.2.2.3). Allerdings ist der zeitliche Abstand zwischen den beiden Erhebungszeitpunkten der Stabilitätsstudie vergleichsweise kurz (im Durchschnitt ca. 7,5 Wochen). In weiteren Studien sollte daher die Langzeitstabilität (zeitlicher Abstand z.B. ein Jahr) von SKS überprüft werden.

4.1.2 Diskussion der Testgütekriterien

Gegenstand von Abschnitt 4.1 ist die Diskussion der (empirischen) Ergebnisse vor dem theoretischen Hintergrund. Empirische Befunde liegen jedoch nur für die Gütekriterien Messgenauigkeit und Validität vor (vgl. Tabelle 37 in Abschnitt 3.12). Daher werden auch nur diese Gütekriterien im folgenden Abschnitt diskutiert. Die Diskussion der anderen Testgütekriterien ist ausschließlich theoretischer Natur und erfolgt bereits bei deren Vorstellung (siehe entsprechende Abschnitte in Kapitel 3). Da sämtliche betrachteten Testgütekriterien für die überarbeitete Version des KKS bestimmt werden, beziehen sich auch die folgenden Ausführungen stets auf diese überarbeitete Version.

Diskussion der Messgenauigkeit des KKS

In Abschnitt 3.2 werden zwei Indikatoren zur Bestimmung der Messgenauigkeit des KKS vorgestellt, zum einen die Treffsicherheit *T* der Vier-Klassen-Lösung, zum anderen die Rangkorrelation der (manifesten) SKS-Variable zu zwei Erhebungszeitpunkten. Letztere stellt eine konkrete Umsetzung der Messgenauigkeitsbestimmung via Retest-Methode dar. Besonders die Analyse der Treffsicherheiten spricht in allen drei (mit der

überarbeiteten Version des KKS) durchgeführten Studien für eine sehr hohe Messgenauigkeit[75], und zwar nicht nur für die gesamte Vier-Klassen-Lösung, sondern auch für jede Ausprägung (also jede Klasse) der latenten Variable (vgl. Abschnitte 3.2.2.2., 3.3.2.1, 3.3.3). Die Messgenauigkeitsbestimmung via Retest-Methode in Form der berechneten Rangkorrelationen (siehe Tabelle 19 in Abschnitt 3.2.2.3) spricht ebenfalls für eine ausreichend hohe Messgenauigkeit des KKS. Voraussetzung für eine sinnvolle Interpretierbarkeit einer Test-Retest-Korrelation ist zum einen, dass zwischen den Erhebungszeitpunkten keine Entwicklung (hinsichtlich der gemessenen Variable) passiert und zum anderen, dass ausreichend zeitlicher Abstand gegeben ist, um Erinnerungseffekte an einzelne Items auszuschließen. Von ersterem kann aufgrund der Ergebnisse zur Stabilität der latenten SKS-Variable ausgegangen werden (vgl. Abschnitt 3.2.2.3). Zweiteres scheint nach Eid (1997) ebenfalls gewährleistet. Zeigt sich die latente Variable SKS auch über einen längeren Zeitraum stabil, stellt die Test-Retest-(Rang-)Korrelation auch für Erhebungszeitpunkte in größerem zeitlichen Abstand einen Indikator für die Messgenauigkeit dar. Da auf Basis des erweiterten Stufen-Modells bei SKS von einer Kompetenz ausgegangen wird, ist ihre Langzeit-Stabilität − zumindest theoretisch − jedoch fraglich. Dennoch ist es denkbar, dass sich die Sophistiziertheits-Stufen auch über einen längeren Zeitraum hinweg stabil zeigen können, vorausgesetzt, diese Kompetenz wird in der Zwischenzeit nicht trainiert (zum Training von SKS siehe Abschnitt 4.3.2).

Diskussion der Validierungsstudien des KKS

Bei den Ausführungen zur Validierung (siehe Abschnitt 3.3) werden Inhalts-, Konstrukt- und Kriteriumsvalidierung des KKS vorgestellt. Es wird argumentiert, dass eine gesonderte Inhaltsvalidierung des KKS nicht notwendig ist, da Items und Antwortoptionen schlüssig aus kognitionspsychologischen Theorien hergeleitet werden. Inhaltsvalidierung ist daher nicht Gegenstand der folgenden Diskussion. Auch auf eine kritische Bewertung der strukturprüfenden Konstruktvalidierung wird verzichtet. Die diesbezüglichen Ergebnisse (siehe Abschnitt 3.3.2.1) werden gemeinsam mit den Ergebnissen der anderen drei durchgeführten Studien bereits mehrfach als empirische Evidenz für das erweiterte Stufen-Modell berichtet (vgl. z.B. Abschnitt 4.1.1). Stattdessen liegt der

[75] Zumindest, wenn zusätzlich die Annahme getroffen wird, dass die für eine Person wahrscheinlichste Klasse ihre „wahre“ Klasse darstellt (vgl. Abschnitt 3.2.1 sowie die entsprechende Diskussion in Abschnitt 4.2.4).

Fokus des folgenden Abschnitts auf den Ergebnissen der struktursuchenden Konstruktvalidierung sowie denen der Kriteriumsvalidierung.

Diskussion der struktursuchenden Konstruktvalidierung

In Abschnitt 3.3.2.2 wird die mit dem KKS gemessene Sophistiziertheit Konditionalen Schlussfolgerns in ein intelligenzbezogenes nomologisches Netzwerk eingeordnet, welches Reasoning sowie die vier Operationsklassen des Berliner Intelligenzstrukturmodells umfasst (vgl. Abbildung 6 in Abschnitt 3.3.2.2). Bei der empirischen Überprüfung ergibt sich zunächst ein Zusammenhang von SKS und Reasoning (indiziert durch die Matrizenitems des WMT), wie postuliert bei einem mittleren Effekt. Dieses Ergebnis ist eine Replikation des Befundes aus der ersten empirischen Erprobung des KKS (siehe Abschnitt 2.4.2) und stützt einmal mehr die Ordnung der Stufen des erweiterten Stufen-Modells hinsichtlich Reasoning. Des Weiteren zeigt sich − ebenfalls wie postuliert − ein Zusammenhang mittlerer Stärke zwischen SKS und der Operationsklasse Verarbeitungskapazität des Berliner Intelligenzstrukturmodells. Aufgrund der konzeptionell hochgradigen Überlappung von Arbeitsgedächtniskapazität und Verarbeitungskapazität spricht dieser Befund für die angenommene Abhängigkeit der Sophistiziertheit Konditionalen Schlussfolgerns von der Arbeitsgedächtniskapazität, die durch beide kognitionspsychologische Theorien (Logiktheorie und Modelltheorie) nahegelegt wird.

Da Verarbeitungskapazität und Reasoning auf Konstruktebene als praktisch identisch angenommen werden (Wilhelm, 2000), stützt der gefundene Zusammenhang von SKS und Verarbeitungskapazität zudem die Annahme eines Zusammenhangs von SKS und Reasoning. Dies gilt insbesondere, da Verarbeitungskapazität im BIS-4 durch insgesamt sechs verschiedene Aufgabentypen (vgl. Tabelle 22 in Abschnitt 3.3.2.2) erfasst wird, die jeweils auch adäquate Operationalisierungen von Reasoning darstellen. Die Stärke des Effektes erweist sich in ähnlicher Höhe wie bei Operationalisierung von Reasoning durch die Matrizenitems des WMT. Der postulierte Zusammenhang zwischen SKS und Reasoning zeigt sich damit auch für andere mögliche Operationalisierungen von Reasoning.

Allerdings fällt bei Betrachtung der Anstiegskoeffizienten der multinomialen logistischen Regressionen (SKS_{lat} auf WMT_{lat} sowie SKS_{lat} auf $BIS\text{-}K_{lat}$; siehe Abschnitt 3.3.2.2) auf, dass diese zwar die postulierte Ordnung aufweisen, jedoch jeweils nur der Unterschied der Sophistizierten-Stufe gegenüber der Referenzkategorie signifikant von null verschieden ist (vgl. Tabelle 25 in Abschnitt 3.3.2.2). Eine mögliche theoretische

Erklärung dafür liefert die Zwei-Prozess-Theorie von Evans (1984). Sie verbindet rationale und nicht-rationale Ansätze und findet ebenso wie die beiden klassischen Theorien (Logiktheorie und Modelltheorie) bei deutlich mehr Denkprozessen Anwendung als nur beim Konditionalen Schlussfolgern. Kern der Theorie ist, dass menschliches Schlussfolgern in zwei verschiedenen Systemen ablaufen kann (Evans & Over, 1996). System 1 (im Original *rationality*$_1$) werden Schlussfolgerungen zugeordnet, die ein zuverlässiges und effizientes Erreichen individueller Ziele ermöglichen, mitunter aber irrational im Sinne der Nichtvereinbarkeit mit den Gesetzen und Regeln der Logik sein können. Schlussfolgern in System 1 erfolgt eher assoziativ, heuristisch und schnell. System 2 (im Original *rationality*$_2$) repräsentiert hingegen die Schlussfolgerungen, die im Einklang mit einem rationalen System (z.B. der Logik) stehen. Schlussfolgern in System 2 erfolgt eher analytisch, rational und langsam. Mit Reasoning bzw. Verarbeitungskapazität korrelieren vor allem die Intelligenzleistungen, die von System 2 erbracht werden (Wilhelm, 2000). Nimmt man nun an, dass Inferenzen auf den niedrigen Sophistiziertheits-Stufen zum Großteil unter Nutzung von System 1 gezogen werden, dann ergäben sich zwischen den niedrigen Stufen deutlich geringere Unterschiede hinsichtlich Reasoning bzw. Verarbeitungskapazität als gegenüber den höheren Stufen und insbesondere gegenüber der Sophistizierten-Stufe. So wäre erklärbar, dass bezüglich Reasoning bzw. Verarbeitungskapazität lediglich die Unterschiede gegenüber der Sophistizierten-Stufe signifikant ausfallen. Für diese − auf der Zwei-Prozess-Theorie basierende − Erklärung sprechen auch Befunde, nach denen sich intelligentere Personen generell rationaler (im formal-logischen Sinne) verhalten als weniger intelligente (Stanovich, 1999; Wilhelm, 2000).

Im Rahmen der divergenten Konstruktvalidierung zeigen sich – wie postuliert – keine Zusammenhänge zwischen SKS und Bearbeitungsgeschwindigkeit sowie Merkfähigkeit. SKS lässt sich demnach klar von diesen Konstrukten abgrenzen. Da dieser Befund den Erwartungen entspricht, soll er nicht weiter diskutiert werden. Entgegen den Erwartungen zeigt sich jedoch ein – wenn auch geringer – Zusammenhang zwischen SKS und Einfallsreichtum. Inhaltlich ist dieser Zusammenhang schwer erklärbar. Möglicherweise fällt es einfallsreichen Personen leichter, die elaborierte, auf der reductio ad absurdum basierende Strategie zur korrekten Lösung eines Modus-Tollens-Problems zu finden. Dann wären unter den Einfallsreichen mehr Sophistizierte, was zu dem beschriebenen Ergebnis führen könnte. Es ist jedoch ebenso denkbar, dass es sich um ein Stichproben- oder Methodenartefakt handelt. So fällt bspw. die Korrelation zwischen

den Skalenwerten des BIS-4 für Verarbeitungskapazität (*BIS-K_{man}*) und Einfallsreichtum (*BIS-E_{man}*) in der Stichprobe der Konstruktvalidierungsstudie recht hoch aus ($r = .37$, $p < .001$). Der gefundene Zusammenhang zwischen SKS und Einfallsreichtum könnte daher nur eine „Scheinkorrelation" infolge des Zusammenhangs von SKS und Verarbeitungskapazität sein. In jedem Falle ist in weiteren Studien zu überprüfen, ob sich dieser Effekt stabil zeigt. Theoretische Überlegungen sprechen jedoch nicht dafür.

Die empirischen Befunde der Konstruktvalidierung des KKS sind damit ausreichend diskutiert worden und sprechen dafür, dass es sich beim KKS um ein valides Testverfahren handelt. Aus den möglichen Erklärungen für vereinzelte geringe Abweichungen von den postulierten Effekten ergibt sich jedoch Bedarf für weitere empirische Überprüfungen.

Diskussion der Kriteriumsvalidierung

Prinzipiell fallen die Ergebnisse der Kriteriumsvalidierung des KKS wie postuliert aus. Die Zusammenhänge von SKS und der Durchschnittsnote des Vordiploms bzw. der Abiturnote zeigen sich hypothesenkonform im geringen bis mittleren Bereich. Dass der (retrospektive) Zusammenhang mit der Abiturnote an einer zweiten Stichprobe kreuzvalidiert wird, erhöht zusätzlich die statistische Sicherheit dieses Befundes. Ein großes Defizit hinsichtlich der Kriteriumsvalidierung des KKS bleibt jedoch, dass für den KKS keine Befunde zur prognostischen Validität existieren. Insbesondere die Vorhersage zukünftiger (Erfolgs-)Kriterien sollte daher Gegenstand weiterer Untersuchungen sein. Im Falle der Studierendenstichprobe der Kriteriumsvalidierungsstudie (siehe Abschnitt 3.3.3) könnte bspw. die Diplomnote zum Ende des Studiums (oder eine vergleichbare Note) als Erfolgskriterium dienen. Die Teilnehmer könnten nach einiger Zeit eingeladen werden, an einer Folgebefragung teilzunehmen und entsprechende Angaben zu ihrem Studienabschluss zu machen. Mindestens ebenso großes Potenzial für weitere Studien resultiert daraus, dass es für einen Test so viele Möglichkeiten zur Kriteriumsvalidierung gibt wie Kriterien (z.B. Moosbrugger & Kelava, 2007). Um den KKS bspw. im Arbeitskontext einsetzen zu können, ist die Überprüfung von Zusammenhängen mit Berufserfolg sinnvoll, wenngleich dabei den bereits in Abschnitt 3.3 beschriebenen Problemen begegnet werden muss. Die Aufzählung weiterer relevanter Kriterien ließe sich sicher noch fortsetzen. Festzuhalten bleibt, dass hier weiterer Untersuchungsbedarf besteht, der KKS jedoch vorerst auch auf Basis der Kriteriumsvalidierung als valide gelten kann.

Abschließend sei noch bemerkt, dass in der vorliegenden Arbeit unter der Überschrift „Validität“ (sowie teilweise auch unter anderen Überschriften wie „Nützlichkeit“ oder „Fairness“) sowohl Fragestellungen im Sinne von Messicks (1989) Validitätskonzeption beantwortet wurden als auch Fragestellungen im Sinne der Validitätskonzeption von Borsboom und Mellenbergh (2007). Dennoch wäre auf Basis beider Validitätskonzeptionen eine Vielzahl weiterer Fragestellungen und empirischer Studien zur Validierung des KKS denkbar (siehe Abschnitt 4.3.1 für weitere mögliche Forschungsfragen zur Validierung des KKS).

4.2 Diskussion der verwendeten Methoden

Der zweite Abschnitt der Diskussion widmet sich dem methodischen Vorgehen in den durchgeführten Studien. Es werden Erhebungsinstrumente, Durchführung, Untersuchungsstichproben und Auswertungsmethoden thematisiert. Auf eine Diskussion des allgemeinen Untersuchungsdesigns wird hingegen verzichtet. Bereits in Abschnitt 2.4.3.2 wird ausgeführt, dass das verwendete Untersuchungsdesign den Forschungsfragen zum KKS angemessen ist. Gleiches gilt auch nach Durchführung der drei weiteren Studien.

4.2.1 Erhebungsinstrumente

Die Erhebungsinstrumente der ersten empirischen Erprobung des KKS sind bereits diskutiert (vgl. Abschnitt 2.4.3.2). In der Stabilitätsstudie (siehe Abschnitt 3.2.2.2) wird ausschließlich der KKS eingesetzt. Dieser soll an späterer Stelle thematisiert werden. In der Konstruktvalidierungsstudie (siehe Abschnitt 3.3.2) werden neben dem KKS der Wiener Matrizentest (WMT) sowie der Berliner Intelligenzstrukturtest (BIS-4) eingesetzt. Bei beiden Tests handelt es sich um etablierte diagnostische Instrumente, die sich empirisch vielfach bewährt haben. Eine kritische Würdigung scheint daher unnötig. Um die Operationsklassen des Berliner Intelligenzstrukturmodells noch genauer zu erfassen, könnte in weiteren Studien die Langform des BIS-4 eingesetzt werden. Bei dieser wird jede Operationsklasse mit einer größeren Zahl an Aufgaben erhoben als in der Kurzform, woraus eine Erhöhung der Messgenauigkeit resultiert (Jäger et al., 1997). In der Kriteriumsvalidierungsstudie (siehe Abschnitt 3.3.3) wird Studiums- bzw. Schulerfolg anhand berichteter Noten (Durchschnittsnote des Vordiploms, Abiturnote) erhoben. Da

Noten nach wie vor einen der besten Indikatoren für Studiums- bzw. Schulerfolg darstellen, könnte lediglich die Art der Erhebung in künftigen Studien geändert werden. Der tatsächliche Notenschnitt anstelle eines Selbstberichtes könnte Erinnerungsfehler und Verzerrungstendenzen (z.B. aufgrund sozialer Erwünschtheit) verhindern. Dennoch scheinen sämtliche verwendete Erhebungsinstrumente den Fragestellungen insgesamt angemessen.

Als zentrales Erhebungsinstrument dieser Arbeit soll nun der KKS erneut thematisiert werden, insbesondere die Resultate seiner Testrevision (siehe Abschnitt 2.4.3.2). Infolge der hypothesenkonformen Ergebnisse der ersten empirischen Erprobung des KKS (vgl. Abschnitt 2.4.2) fällt diese Testrevision vergleichsweise sparsam aus (vgl. Abschnitt 2.4.3.2). Eine der beiden wichtigsten resultierenden Veränderungen in der überarbeiteten Version stellt die nicht-sequentielle Vorgabe der Items dar. So soll einem sequentiellen Antwortverhalten entgegengewirkt werden, welches im Rahmen der ersten empirischen Erprobung des KKS als ursächlich für die „Restklasse“ angenommen wird (vgl. Abschnitt 2.4.3.1). Es bleibt festzuhalten, dass in keiner der drei Studien, die mit der überarbeiteten Version des KKS durchgeführt werden, die Modellierung einer „Restklasse“ notwendig ist. Jeweils erweist sich eine Vier-Klassen-Lösung als passend, messgenau und im Sinne des erweiterten Stufen-Modells interpretierbar (vgl. Abschnitte 3.2.2.2, 3.3.2.1, 3.3.3). Diese Befunde sprechen dafür, dass diese Änderung im Rahmen der Testrevision den gewünschten Effekt hat. Die zweite wichtige Änderung stellt die Vorgabe einer zusätzlichen vierten Antwortoption („*Ich weiß es nicht*“) dar. Diese soll eine bessere Abgrenzbarkeit gegenüber der genuinen Antwortoption „*Es ist keine eindeutige Aussage über das Funktionieren der Schaltung ableitbar*“ gewährleisten. Welche konkreten Auswirkungen diese Änderung hat, lässt sich anhand der empirischen Befunde nicht feststellen. Die nach wie vor gute Modellpassung und die sogar gestiegenen Treffsicherheiten (vgl. Abschnitte 3.2.2.2, 3.3.2.1, 3.3.3 mit 2.4.2) sprechen jedenfalls dafür, dass sich durch diese Änderung zumindest keine negativen Auswirkungen ergeben. Insgesamt kann davon ausgegangen werden, dass die Testrevision zu einer Verbesserung des KKS geführt hat. Zudem können die Testgütekriterien der überarbeiteten Version des KKS (fast) ausnahmslos als erfüllt betrachtet werden (vgl. Kapitel 3; zusammenfassend Tabelle 37 in Abschnitt 3.12). Diese Version kann daher auch in der Folge eingesetzt werden. Eine erneute Testrevision ist unnötig.

4.2.2 Durchführung

Über alle vier durchgeführten Studien hinweg wird der KKS computergestützt als Online-Test durchgeführt. Daraus resultieren einige Testbedingungen, die nur schwer kontrolliert werden können (vgl. Ausführungen zur Durchführungsobjektivität von Online-Tests in Abschnitt 3.1). Daher ist die Frage berechtigt, ob sich einige der Testergebnisse auf diese − kaum kontrollierbaren − Testbedingungen zurückführen lassen. Zwei Gründe sprechen jedoch dagegen. Zum einen zeigen sich die empirischen Befunde (insbesondere zum erweiterten Stufen-Modell) konsistent über die verschiedenen Online-Studien hinweg (vgl. Abschnitte 2.4.2, 3.2.2.2, 3.3.2.1, 3.3.3), zum anderen wird die Konstruktvalidierungsstudie (siehe Abschnitt 3.3.2) unter kontrollierten Bedingungen durchgeführt. Zwar wird online auf den KKS zugegriffen, jedoch für alle Testteilnehmer unter vergleichbaren Bedingungen (gleicher Raum, Rechner gleichen Typs, anwesende Testleiterin; vgl. Ausführungen zur Durchführung in Abschnitt 3.3.2.2). Da die Ergebnisse der Konstruktvalidierungsstudie (insbesondere zum erweiterten Stufen-Modell) mit denen der anderen drei Studien vergleichbar sind, scheint die Online-Testung als Varianzquelle für den KKS von untergeordneter Bedeutung. Zudem existiert mit der Modus-Ponens-Reduktion (siehe Abschnitt 2.4.1.5) eine empirisch überprüfte Möglichkeit zur Sicherung der Datenqualität (vgl. Abschnitt 2.4.2), wenngleich sie in keiner der drei weiteren Studien (Stabilitätsstudie, Konstruktvalidierungsstudie, Kriteriumsvalidierungsstudie) angewendet werden muss (vgl. Ausführungen zu Ergebnissen in den Abschnitten 3.2.2.2, 3.3.2.1, 3.3.3).

Es kann festgehalten werden, dass die Durchführung als Online-Test keine negativen Konsequenzen für die Testergebnisse des KKS zu haben scheint, jedoch als Vorteile aufweist, dass große Stichproben ökonomisch rekrutiert und erfasst werden können. Dennoch wären auch eine Papier-Bleistift-Version oder eine computergestützte Version, die nicht online bearbeitet wird, möglich.

4.2.3 Untersuchungsstichproben

Die Untersuchungsstichprobe der ersten empirischen Erprobung des KKS wird bereits in Abschnitt 2.4.3.2 diskutiert. Wird dort noch positiv bewertet, dass es sich um eine vergleichsweise umfangreiche Untersuchungsstichprobe ($N = 905$) handelt, ist festzuhalten, dass der Umfang in den drei weiteren durchgeführten Studien deutlich geringer ist ($N = 195$, $N = 154$, $N = 305$). Zwar bezeichnen Evans et al. (2007) bei allgemein-

psychologischen Experimenten zum Konditionalen Schlussfolgern bereits Stichproben ab ca. 100 Probanden als umfangreiche Stichproben, dennoch ist fraglich, ob die Parameterschätzungen in den LCAs bei diesen Stichprobenumfängen ausreichend robust sind (Rost, 2004). Es ist daher umso bemerkenswerter, dass die Ergebnisse der LCAs dieser drei Studien zu vergleichbaren Ergebnissen bezüglich des erweiterten Stufen-Modells führen. Dadurch erhöht sich die Plausibilität, dass es sich trotz der eher geringen Stichprobenumfänge um robuste Schätzungen handelt. Hinzu kommt, dass mit dem parametrischen Bootstrapverfahren eine Methode zur Modellprüfung angewendet wird, die in eben solchen „sparse data"-Fällen eine robuste Bewertung der Modellgüte erlaubt (Davier, 1997). Dennoch sollte die Replizierbarkeit der Ergebnisse dieser drei Studien in Untersuchungen mit umfangreicheren Stichproben überprüft werden.

Ein zweiter zu betrachtender Aspekt betrifft die Zusammensetzungen der Untersuchungsstichproben. Ausgangspunkt ist, dass sich eine Vier-Klassen-Lösung im Sinne des erweiterten Stufen-Modells stets als passendes, messgenaues und gut interpretierbares Messmodell für SKS erweist. Über alle vier Untersuchungsstichproben hinweg ist das Durchschnittsalter vergleichsweise niedrig (zwischen 21 und 26 Jahren) und der Bildungsabschluss vergleichsweise hoch (vgl. Abschnitte 2.4.1.4, 3.2.2.2, 3.3.2.1, 3.3.3). Bezüglich Alterseffekten sollte in weiteren Studien überprüft werden, ob sich auch für Stichproben älterer Personen die Vier-Klassen-Lösung als adäquates Messmodell für SKS erweist. Der vergleichsweise hohe Bildungsabschluss ist zunächst unkritisch, da die vorläufige Zielgruppe (Chipdesigner) ebenfalls hohe Bildungsabschlüsse aufweist (vgl. Abschnitt 2.4.3.2). Dennoch kann nicht ausgeschlossen werden, dass dieses Messmodell für SKS lediglich für junge Erwachsene mit vergleichsweise hohem Bildungsabschluss gilt. Weitestgehend ausgeschlossen werden kann jedoch eine Abhängigkeit vom Geschlecht. Zwar wird dies nicht explizit überprüft, jedoch variiert das Geschlechterverhältnis stark zwischen den einzelnen Untersuchungen (Gleichverteilung bei der ersten empirischen Erprobung des KKS, Überrepräsentation von Frauen in der Konstruktvalidierungsstudie, Überrepräsentation von Männern in der Kriteriumsvalidierungsstudie). Auch die Überrepräsentation Thüringens als Herkunftsbundesland (vgl. z.B. Abschnitt 2.4.1.4) gilt nicht für alle Untersuchungsstichproben. So entspricht die Verteilung der Herkunftsbundesländer unter den Teilnehmern der Kriteriumsvalidierungsstudie annähernd einer repräsentativen Verteilung (vgl. Grohmann, 2008). Da sich der KKS hinsichtlich der Passung des Messmodells auch in der Kriteriumsvalidie-

rungsstudie bewährt, scheint die Überrepräsentation Thüringens in den Stichproben der anderen durchgeführten Studien unproblematisch.

Es bleibt festzuhalten, dass nach der mehrfachen empirischen Erprobung des KKS Einflüsse einiger demographischer Besonderheiten der Untersuchungsstichproben weitgehend ausgeschlossen werden können, bezüglich anderer besteht jedoch weiterer Forschungsbedarf.

4.2.4 Auswertungsmethoden

Bereits in Abschnitt 2.4.3.2 werden die Analyseverfahren der ersten empirischen Erprobung des KKS diskutiert. Sowohl LCA als auch multinomiale logistische Regression für latente Variablen werden als geeignet für die Beantwortung der aufgestellten Forschungsfragen bewertet. Entsprechend wird die Anwendung dieser Verfahren für weitere Studien empfohlen (vgl. Abschnitt 2.4.3.2). Gleiches gilt für die Rangkorrelationen, deren Zweck vor allem die Vergleichbarkeit mit den Ergebnissen anderer Untersuchungen ist. Nach Anwendung dieser drei Analyseverfahren in den in der Folge durchgeführten Studien (Stabilitätsstudie, Konstruktvalidierungsstudie, Kriteriumsvalidierungsstudie) bleibt festzuhalten, dass sich die aufgestellten Hypothesen stets adäquat überprüfen lassen. Auf eine neuerliche Diskussion dieser drei Verfahren soll daher verzichtet werden. Die in Abschnitt 2.4.3.2 angeführten Argumente gelten auch nach Anwendung der Verfahren in drei weiteren empirischen Studien. Ergänzend sei lediglich erwähnt, dass sich für die multinomialen logistischen Regressionen sogar zusätzliche Auswertemöglichkeiten ergeben, da in keiner dieser drei Studien die Modellierung einer „Restklasse“ notwendig ist. Daher können die Anstiegskoeffizienten der einzelnen Kategorien − zumindest gegenüber der Referenzkategorie − hinsichtlich ihrer Verschiedenheit von null analysiert werden. Dies ermöglicht zusätzlich zur deskriptiven Analyse der Ordnung der Anstiegskoeffizienten eine inferenzstatistische Prüfung. Einschränkend bleibt jedoch, dass dieser Vergleich lediglich gegenüber der Referenzkategorie möglich ist. Da diese durch die verwendete Software (M*plus*) automatisch festgelegt wird, existiert bspw. kein inferenzstatistischer Vergleich der Fortgeschrittenen-Stufe gegenüber der Sophistizierten-Stufe (vgl. Abschnitt 3.3.2.2). Schließlich soll noch ein weiterer, die LCA betreffender Punkt festgehalten werden. Wie auch schon bei der ersten empirischen Erprobung des KKS ist auch in allen drei weiteren Untersuchungen die Zuordnung der klassenbedingten Lösungswahrscheinlichkeiten zu den postulierten Stufen des

erweiterten Stufen-Modells per Augenschein möglich. Das Setzen zusätzlicher Restriktionen erweist sich also auch bei weiteren Anwendungen des KKS als nicht notwendig. Insgesamt bleibt festzuhalten, dass sich alle drei Analyseverfahren auch bei der weiteren empirischen Erprobung des KKS bewährt haben.

Eine nach wie vor offene Frage zur LCA ist hingegen die Angemessenheit der Treffsicherheit *T* als Maß für die Messgenauigkeit eines LCMs (vgl. Abschnitt 3.2.1). Dabei sind bspw. folgende Kritikpunkte naheliegend:

1. In Abschnitt 3.2.1 muss eine zusätzliche Annahme (*die für eine Person wahrscheinlichste Klasse ist ihre „wahre" Klasse*) getroffen werden, damit die Treffsicherheit einen Schätzer für die Messgenauigkeit im dort definierten Sinne darstellt.
2. Die Metrik der Treffsicherheit entspricht nicht der von Korrelationskoeffizienten, die für Reliabilitäts-Schätzer häufig verwendet werden (vgl. Abschnitt 3.2.2.3).
3. Es wird lediglich die Zuordnungswahrscheinlichkeit zur wahrscheinlichsten Klasse betrachtet. Die Zuordnungswahrscheinlichkeiten zu den anderen Klassen werden nicht berücksichtigt.
4. Die Anzahl der Klassen wird ebenfalls nicht berücksichtigt.

Zum ersten Punkt sei angemerkt, dass es auch in der Klassischen Testtheorie keine Rolle spielt, ob der *true score* der „wirkliche" Wert der Person ist oder nicht. Dies gilt ebenso für ein LCM (Collins, 2001). Messgenauigkeit in einem LCM heißt demnach: Wie gut ist es möglich, mit dem Test (also mit den Items) Personen zu klassifizieren? Die Frage der Korrektheit dieser Klassifikation ist dabei zunächst nicht von Bedeutung (Collins, 2001). Daher wären für die Messgenauigkeit in einem LCM auch andere Bezeichnungen wie *Klassifikationsgüte* denkbar und möglicherweise sogar angemessener. Dennoch ist die getroffene Annahme dem Grunde nach unnötig und spricht nicht gegen die Verwendung von Treffsicherheiten als Indikatoren für die Messgenauigkeit eines LCMs.

Eine Überführung in eine Korrelationsmetrik, wie im zweiten Punkt kritisiert, ist bspw. unter Verwendung von *Yules Q* möglich (Clogg & Manning, 1996). Dabei wird zunächst *Odds Ratio* (*OR*) gebildet, indem die jeweilige Treffsicherheit (die eine Wahrscheinlichkeit ist; vgl. z.B. Abschnitt 2.4.1.5) durch ihre Gegenwahrscheinlichkeit dividiert wird. Yules *Q* berechnet sich dann:

$$Q = (OR - 1) / (OR + 1)$$

Durch diese Überführung in eine Korrelationsmetrik sind Vergleiche mit Messgenauigkeitsschätzungen möglich, die auf Korrelationskoeffizienten basieren und in der Testkonstruktionspraxis sehr viel gebräuchlicher sind als die Analyse von Treffsicherheiten. Tabelle 38 gibt einen Überblick über die Werte für Yules *Q* in sämtlichen mit der überarbeiteten Version des KKS durchgeführten Studien.

Tabelle 38: Yules *Q* (klassenspezifisch wie gesamt) in den mit der überarbeiteten Version des KKS durchgeführten Studien

	Studie			
	Stabilitäts-studie (Zeitpunkt 1)	Stabilitäts-studie (Zeitpunkt 2)	Konstrukt-validierungs-studie	Kriteriums-validierungs-studie
Unsophistizierte	.922	.974	.918	.946
Unsophistizierte II	.926	.976	.892	.944
Fortgeschrittene	.986	.982	.942	.972
Sophistizierte	.994	.980	.986	.974
gesamte Vier-Klassen-Lösung	.950	.978	.914	.960

Anmerkung. Die Werte für Yules *Q* berechnen sich direkt aus den allgemeinen wie klassenspezifischen Treffsicherheiten in den entsprechenden Studien (vgl. Tabelle 14, Tabelle 16, Tabelle 21, Tabelle 28).

Die Werte für Yules *Q* liegen ausnahmslos in einem Bereich, der für Korrelationskoeffizienten als sehr hoch zu bewerten wäre. Dies stützt zusätzlich die Ergebnisse der Analyse der Treffsicherheiten.

Bleiben die Kritikpunkte 3 und 4, dass bei der Analyse von Treffsicherheiten weder die Zuordnungswahrscheinlichkeiten zu den anderen Klassen berücksichtigt werden noch die Anzahl der Klassen. Letzteres scheint durchaus relevant. So bedeutet bspw. bei nur zwei Klassen eine Treffsicherheit von .50 eine Zuordnungswahrscheinlichkeit von .50 zu jeder der beiden Klassen. Das entspricht einer zufälligen Zuordnung und damit der theoretisch schlechtestmöglichen Klassifizierbarkeit bzw. Messgenauigkeit. Im Falle von vier Klassen ist dies bei Zuordnungswahrscheinlichkeiten von jeweils .25 der Fall. Demnach ist die theoretische untere Grenze der Treffsicherheit je nach Klassenzahl eine andere.

Auch die Zuordnungswahrscheinlichkeiten zu den anderen Klassen liefern Informationen darüber, ob ein Antwortmuster (und damit eine Person) bspw. allen ande-

ren Klassen mit gleich hoher Wahrscheinlichkeit zugeordnet wird oder einer der anderen Klassen mit besonders hoher Wahrscheinlichkeit. Dies kann für Aussagen über die Messgenauigkeit durchaus relevant sein. Beide Punkte sind jedoch lediglich von Relevanz, wenn die Treffsicherheiten eher niedrig ausfallen. Bei Werten in der Höhe, wie sie sich in den mit der überarbeiteten Version des KKS durchgeführten Studien zeigen, sind beide Punkte − zumindest praktisch − ohne Bedeutung, da die Differenz zur theoretischen unteren Grenze der Treffsicherheit (bei vier Klassen $T_C = .25$) sehr groß ist und gleichzeitig die Zuordnungswahrscheinlichkeiten zu den anderen Klassen so gering sind, dass ihre Verteilung unerheblich ist.

Festzuhalten bleibt, dass es sich bei der Treffsicherheit *T* um ein intuitives Maß zur Beurteilung der Klassifizierbarkeit von Personen mit den 16 Items des KKS handelt, welches den Vorteil hat, dass es zudem leicht bestimmbar ist und in der verwendeten Software (Winmira 2001; vgl. Abschnitt 2.4.1.5) bei Berechnung einer LCA automatisch ausgegeben wird. Zweifelsohne weist dieses Maß Nachteile auf, die jedoch bei den hohen Treffsicherheits-Werten in den durchgeführten Studien unerheblich sind. Wenngleich in anderen Fällen sicher bessere Schätzer für die Messgenauigkeit eines LCMs existieren, so kann dennoch konstatiert werden: Zeigen sich hohe Treffsicherheits-Werte für alle identifizierten Klassen, indiziert das eine gute Klassifizierbarkeit von Personen mithilfe der (Test-)Items und damit eine hohe Messgenauigkeit.

Eine adäquate Alternative zur Treffsicherheit stellt der von Collins (2001) vorgeschlagene Reliabilitäts-Schätzer *RC* für kategoriale latente Variablen dar. Er ist konzeptionell an die Reliabilitätsdefinition der Klassischen Testtheorie angelehnt und berücksichtigt sowohl die Zuordnungswahrscheinlichkeiten zu den anderen Klassen als auch deren Anzahl. Aus Platzgründen soll hier jedoch auf eine ausführlichere Beschreibung verzichtet werden (siehe dazu Collins, 2001).

Im Rahmen der Stabilitätsstudie wird noch ein weiteres Analyseverfahren vorgestellt und angewendet: die Analyse latenter Transitionen (LTA). Die LTA wird in der vorliegenden Arbeit genutzt, um die Stabilität nominaler latenter Variablen zu analysieren. Zunächst wird das Messmodell (Vier-Klassen-Lösung) zu beiden Erhebungszeitpunkten überprüft. Anschließend werden – unter der Annahme der Invarianz des Messmodells über die Zeit – Wahrscheinlichkeiten für die latenten Klassenübergänge geschätzt (siehe Tabelle 17 in Abschnitt 3.2.2.2). Für die Argumentation in dieser Arbeit sind lediglich die berechneten Stabilitätskoeffizienten (siehe Abschnitt 3.2.2.3) interessant, da sie Rückschlüsse auf die Stabilität von SKS zulassen. Allerdings soll an

dieser Stelle nicht unerwähnt bleiben, dass noch vielfältige andere Analysen zur Stabilität auf Basis der Wahrscheinlichkeiten der latenten Klassenübergänge möglich wären. Modelle zur Bestimmung von Interrater-Übereinstimmungen (siehe z.B. Agresti, 1992) bspw. basieren auf der Analyse ebensolcher Wahrscheinlichkeiten und könnten durchaus auf die interessierenden Inhalte übertragen werden. Gleiches gilt für den direkten Vergleich von Stabilitäts- und Veränderungsmodellen (siehe z.B. Rajulton & Ravanera, 2001) bspw. mittels Likelihood-Ratio-Test. So bleibt zweierlei festzuhalten. Zum einen erweist sich die LTA als adäquates Analyseverfahren zur Beantwortung der Frage nach der Stabilität einer nominalen latenten Variable über die Zeit. Zum anderen existiert eine Vielzahl weiterer Auswertemöglichkeiten der Wahrscheinlichkeiten latenter Klassenübergänge, die differenziertere Aussagen zur Stabilität nominaler latenter Variablen ermöglichen.

Ein abschließender, durchaus kritischer Punkt betrifft den notwendigen „Software-Wechsel“ zwischen den Programmen Winmira 2001 und M*plus* innerhalb der Überprüfung einzelner Hypothesen. So wird das Messmodell für SKS stets mit der Software Winmira 2001 (Davier, 2001) überprüft, die Zusammenhangshypothesen hingegen mit der Software M*plus* (Muthén & Muthén, 1998-2007). In M*plus* wird dabei jeweils das mit Winmira 2001 überprüfte Messmodell „nachgebaut“, wobei die geschätzten Parameter per Augenschein miteinander verglichen werden sollten, da sich die Schätzalgorithmen der beiden Programme zum Teil unterscheiden. Dies ist gegenwärtig noch etwas umständlich und eine Integration bspw. des parametrischen Bootstrapverfahrens zur Modellprüfung in *Mplus* wäre wünschenswert.

4.3 Implikationen für Forschung und Praxis

Der folgende Abschnitt der Diskussion widmet sich zunächst den Implikationen der vorliegenden Arbeit für die Forschung. Dabei werden jedoch nicht nur resultierende Forschungsfragen thematisiert, sondern auch methodische Implikationen. Abschließend wird auf die Bedeutung für die psychologische, und dabei insbesondere für die psychodiagnostische Praxis eingegangen.

4.3.1 Implikationen für die Forschung

Eine Vielzahl an weiteren Forschungsfragen ergibt sich bereits bei der Diskussion der Ergebnisse vor dem theoretischen Hintergrund (vgl. insbesondere Abschnitt 4.1.1). In der Folge sollen lediglich einige ausgewählte thematisiert werden. Anschließend werden methodische Implikationen der vorliegenden Arbeit abgeleitet, die auch über die Erfassung der Sophistiziertheit Konditionalen Schlussfolgerns hinaus Anwendung finden können.

Resultierende Forschungsfragen

Die naheliegendste Forschungsfrage ist zweifelsohne die Replizierbarkeit der empirischen Befunde. In diesem Falle ist jedoch eine Besonderheit, dass dafür nicht zwingend weitere Studien durchgeführt werden müssen. Die empirische Überprüfung des erweiterten Stufen-Modells ist ebenso in Form von Reanalysen der Daten klassischer Experimente zum Negationsparadigma denkbar. Wie bereits mehrfach erwähnt, werden bei klassischen Experimenten meist lediglich Häufigkeiten analysiert und nicht Antwortmuster der 16 Aufgaben des Negationsparadigmas (vgl. z.B. Abschnitt 2.2.3). Bei Anwendung der in dieser Arbeit vorgestellten Methoden (und im Falle einer entsprechend umfangreichen Datengrundlage) könnte die Replizierbarkeit der Ergebnisse damit sogar ohne weitere Datenerhebungen überprüft werden. Selbstverständlich sollten auch die in dieser Arbeit beschriebenen Studien wiederholt werden, idealerweise mit größeren Stichproben und unter zum Teil modifizierten Bedingungen. Ansätze dafür werden bereits in Abschnitt 4.1 vorgestellt.

Eine zweite resultierende Forschungsfrage adressiert Inhaltseffekte bei den verwendeten Items. Wie bereits in Abschnitt 4.1.1 ausgeführt, wird der Item-Inhalt „Funktionieren elektrischer Schaltungen“ im Falle des KKS konstant gehalten (vgl. Abschnitt 2.3.3). Nun gelten jedoch Inhaltseffekte bei Aufgaben zum Konditionalen Schlussfolgern als unumstritten (vgl. Abschnitt 2.3.3). Sei es nun, dass bestimmte Inhalte das korrekte Lösen der Aufgaben erleichtern, wie die Modelltheorie postuliert (vgl. Abschnitt 2.1.2.1), oder auch, dass inhaltliche Einkleidungen das korrekte Lösen eher inhibieren (z.B. Neth, Beller & Spada, 1999). Die Zusammenhänge der Sophistiziertheit Konditionalen Schlussfolgerns mit Reasoning werden jedenfalls in Anlehnung an Wilhelm (2000) umso größer vermutet, je abstrakter der Inhalt der Aufgaben ist. Bislang wird davon ausgegangen, dass der Item-Inhalt „Funktionieren elektrischer Schaltungen“

des KKS ausreichend abstrakt ist, um derartige Inhaltseffekte zu minimieren. Um diese Annahme zu überprüfen, könnte in einer weiteren Studie der KKS zu 16 strukturgleichen Aufgaben mit offensichtlich abstraktem Inhalt[76] in Beziehung gesetzt werden. Eine weitere Möglichkeit Inhaltseffekte des „Funktionierens elektrischer Schaltungen" zu überprüfen, wäre eine Befragung, die die Glaubhaftigkeit der vier Hauptprämissen des KKS in der beabsichtigten Zielgruppe erhebt. Ist bspw. „*Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2*" genauso glaubhaft wie „*Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2 nicht*"? Ist dies für alle vier Hauptprämissen der Fall, sind differenzielle Inhaltseffekte der vier Hauptprämissen des KKS eher unwahrscheinlich. Beides wären Möglichkeiten, der Frage nachzugehen, ob das Schlussfolgern bei den Items des KKS durch deren Inhalt beeinflusst wird. Vorerst bleibt jedoch festzuhalten, dass sich die 16 Items des Negationsparadigmas zur Prüfung des erweiterten Stufen-Modells zunächst lediglich mit dem Item-Inhalt „Funktionieren elektrischer Schaltungen" empirisch bewährt haben. Die bislang nicht erfolgte Manipulation des Item-Inhaltes bietet damit Potenzial für eine Vielzahl (vor allem) experimenteller Studien. So kann der Frage nachgegangen werden, inwieweit eine Übertragbarkeit auf andere Item-Inhalte gewährleistet ist. Konsequenz für die Forschung wäre die Erhöhung der Plausibilität des erweiterten Stufen-Modells, Konsequenz für die Praxis die Möglichkeit, den Inhalt der 16 Items des Negationsparadigmas an verschiedene andere Arbeitskontexte anzupassen (siehe dazu auch Abschnitt 4.3.2).

Ausgehend von der Modelltheorie Konditionalen Schlussfolgerns (Johnson-Laird & Byrne, 2002) sind Effekte von Vorwissen eine weitere Varianzquelle. Im Falle des KKS können sich diese auf Vorwissen zum Funktionieren elektrischer Schaltungen oder aber auf Vorwissen zur Aussagenlogik beziehen. Beides sollte in künftigen Studien mit erhoben werden, um diese Vermutung empirisch zu überprüfen. Gegebenenfalls sind diese Effekte von Vorwissen bei der Auswertung dann künftig zu berücksichtigen.

Wenngleich in der vorliegenden Arbeit das der Testkonstruktion zugrundeliegende Modell (erweitertes Stufen-Modell; siehe Abschnitt 2.3.2.2) aus der Logiktheorie (Braine & O'Brien, 1991) abgeleitet wird, können die beschriebenen Phänomene zum Teil auch auf Basis der Modelltheorie (Johnson-Laird & Byrne, 2002) erklärt werden (vgl. Rijmen & De Boeck, 2003). Auch wenn die empirischen Befunde also für das erweiterte Stufen-Modell sprechen, ist dennoch nicht klar entscheidbar, welche dieser

[76] z.B.: Hauptprämisse: „Wenn A, dann B", Nebenprämisse: „B", Antwortoptionen: „A" / „nicht A" / „keine Aussage bezüglich A möglich"

beiden Theorien den ablaufenden Denkprozessen zugrunde liegt. Eine Möglichkeit, dieser Frage nachzugehen, stellt der Einsatz *bildgebender Verfahren* wie bspw. der *Magnetresonanztomographie* dar. Dazu müssten die Gehirnaktivitäten von Probanden bei der Bearbeitung des KKS (oder auch anderer Aufgaben zum Konditionalen Schlussfolgern) analysiert werden. Nach Knauff (2006) würden Aktivierungen im temporalen Kortex der linken Hemisphäre während der Aufgabenbearbeitung für Denkprozesse im Sinne der Logiktheorie sprechen, während Gehirnaktivierungen im okzipito-parietalen Kortex der rechten Hemisphäre Denkprozesse im Sinne der Modelltheorie intendieren würden. Möglicherweise ist es also der Einsatz bildgebender Verfahren beim Lösen von Aufgaben zum Konditionalen Schlussfolgern, der dem langjährigen Disput zwischen beiden Theorien (vgl. auch Abschnitt 2.1.2.3) neue Impulse geben könnte.

Zum Abschluss der Diskussion resultierender Forschungsfragen soll die in Abschnitt 4.1.2 angekündigte weitere Möglichkeit zur Validierung des KKS vorgestellt werden. Diese resultiert daraus, dass die Testentwicklung auf kognitionspsychologischen Theorien aufbaut. Aus diesen kognitionspsychologischen Grundlagen kann für jede der Sophistiziertheits-Stufen eine eigene Ursache für das Zustandekommen der Leistung abgeleitet werden (siehe auch Ausführungen zur Interpretationsobjektivität in Abschnitt 3.1):

Unsophistizierte akzeptieren die einladende Inferenz einer Konditionalaussage.

Unsophistizierte II haben infolge von Schwierigkeiten beim Enkodieren bzw. aktiven Auflösen einer doppelten Negation eine höhere Wahrscheinlichkeit der einladenden Inferenz zu widerstehen, jedoch lediglich bei den Schlussfiguren NA und MT und auch dort nur, wenn in der Hauptprämisse mindestens eine Negation enthalten ist.

Fortgeschrittene widerstehen der einladenden Inferenz, sind jedoch nicht in der Lage, einen sophistizierten Modus-Tollens-Schluss zu ziehen. Hierfür fehlt es möglicherweise an Kenntnis (oder Verfügbarkeit) der reductio ad absurdum.

Sophistizierte schließlich sind in der Lage, sowohl der einladenden Inferenz zu widerstehen als auch einen sophistizierten Modus-Tollens-Schluss zu ziehen.

Diese vermuteten Ursachen für die Sophistiziertheits-Stufen könnten einzeln validiert werden, indem Beziehungen zu elementaren kognitionspsychologischen Aufgaben (z.B. zum Akzeptieren von einladenden Inferenzen, zum Auflösen doppelter Negationen oder

zum Verständnis der reductio ad absurdum) überprüft werden. Daraus resultiert ein ganzer Komplex weiterer Fragestellungen zur Konstruktvalidierung von SKS und damit eine weitere Möglichkeit, den KKS zu validieren.

Methodische Implikationen

In der Folge sollen noch einige methodische Implikationen der vorliegenden Arbeit abgeleitet werden. Die wichtigste ist aus Sicht des Autors, dass die Entwicklung und Erprobung eines Tests auch dann möglich ist, wenn das zugrundeliegende Konstrukt als nominal angenommen wird. Auch ein Test, der auf einer so konstruierten latenten Variable basiert, lässt sich hinsichtlich der Erfüllung klassischer Testgütekriterien bewerten. Dies kann sogar auf der Ebene latenter Variablen, also unter Berücksichtigung von Messfehlern geschehen und ist bspw. mit den in dieser Arbeit vorgestellten Analyseverfahren möglich.

Durch eine Analyse latenter Klassen (siehe Abschnitt 2.4.1.5) kann die Modellprüfung und damit die strukturprüfende Konstruktvalidierung (siehe Abschnitt 3.3.2.1) erfolgen. Durch eine Analyse der Treffsicherheiten kann die Messgenauigkeit des Tests überprüft werden (vgl. Abschnitte 3.2.1 und 4.2.4), wenngleich hierfür auch andere Reliabilitäts-Schätzer verwendet werden können (vgl. Abschnitt 4.2.4). Folgendes Vorgehen hat sich in der vorliegenden Arbeit für eine Analyse latenter Klassen bewährt:

1. Bestimmung der bestpassenden Klassen-Lösung durch Vergleich der BIC-Werte
2. Bestimmung der Modellpassung dieser bestpassenden Klassen-Lösung mittels parametrischen Bootstrapverfahrens
3. Analyse der Treffsicherheiten des LCMs
4. Prüfung der Interpretierbarkeit der klassenbedingten Lösungswahrscheinlichkeiten

Dieses Vorgehen hat neben der inhaltlichen Angemessenheit zudem den Vorteil, dass zur Umsetzung lediglich eine Software (Winmira 2001; vgl. Abschnitt 2.4.1.5) notwendig ist. Das als nominale latente Variable präzisierte Konstrukt kann in multinomialen logistischen Regressionen (siehe Abschnitt 2.4.1.5) zu manifesten oder latenten Regressoren jeglichen Skalenniveaus in Beziehung gesetzt werden. Dieses Verfahren empfiehlt sich bspw. bei der struktursuchenden Konstruktvalidierung oder bei der Kriteriumsvalidierung. Mit der Analyse latenter Transitionen (siehe Abschnitt 3.2.2.2) steht ein Verfahren zur Verfügung, mit dem die Stabilität nominaler latenter Variablen über

die Zeit überprüft werden kann. Dies ermöglicht die Voraussetzungsprüfung für eine Bestimmung der Messgenauigkeit mittels wiederholter Testvorgabe.

Da für die Bestimmung von Testgütekriterien also auch im Falle nominaler latenter Variablen entsprechende Analyseverfahren zur Verfügung stehen, stellt sich die Frage, warum diese dafür (bislang) so selten eingesetzt werden. Wahrscheinlich liegt es daran, dass die beschriebenen Analyseverfahren vergleichsweise unbekannt sind und in den Standardlehrbüchern statistischer Verfahren meist nicht einmal Erwähnung finden (Rost, 2006). Dennoch werden aktuell im Rahmen des sog. *Mixture Modeling* immer häufiger latente Klassen in Modellierungen einbezogen, wodurch auch die Analyse latenter Klassen zunehmend an Bedeutung gewinnt.

Eine weitere Implikation der Ergebnisse dieser Arbeit ergibt sich aus dem erfolgreichen Einsatz der Modus-Ponens-Reduktion (siehe Kapitel 2.4.1.5) zur Sicherung der Datenqualität in der in Abschnitt 2.4 vorgestellten Online-Erhebung. Diese Methode ist prinzipiell auf jegliche Art von Leistungstests übertragbar und bedarf lediglich einiger sehr leichter Items, bei denen davon ausgegangen werden kann, dass sie jeder Proband leicht lösen kann, vorausgesetzt, er bearbeitet den Test ernsthaft. Als verallgemeinerte Bezeichnung bietet sich bspw. *Leichte-Items-Reduktion* an. Nach Vorgabe des Tests werden lediglich die sehr leichten Items mittels eines Zwei-Klassen-Modells analysiert. Dadurch sollten zwei Klassen von Probanden identifiziert werden können: diejenigen, die den Test ernsthaft bearbeiten, und diejenigen, die den Test zufällig oder systematisch durchklicken (oder aber eine derart niedrige Ausprägung der betrachteten Fähigkeit aufweisen, dass sie für die Untersuchung höchstwahrscheinlich uninteressant sind). Charakteristisch für die erste Klasse („Leichte-Items-Löser“) wären Lösungswahrscheinlichkeiten für sämtliche dieser sehr leichten Items nahe eins. Im Falle eines gebundenen Antwortformates wären für die zweite Klasse („Durchklicker“) Lösungswahrscheinlichkeiten nahe $1/k$ zu erwarten, wobei k der Anzahl der Antwortalternativen entspricht. Bei offenem Antwortformat sollten die Lösungswahrscheinlichkeiten in der zweiten Klasse in jedem Falle deutlich von eins verschieden sein. In einem nächsten Schritt folgt eine Reduktion der Stichprobe auf diejenigen Personen, die mit hinreichend hoher Wahrscheinlichkeit der ersten Klasse („Leichte-Items-Löser“) zugeordnet werden können. Zuordnungswahrscheinlichkeiten ab .85 scheinen dafür ausreichend hoch. In Abhängigkeit von der Forschungsfrage kann diese Untergrenze natürlich variiert werden. Da zur Identifikation von zwei latenten Klassen lediglich vier Items notwendig sind (Clogg, 1995), ist dies eine vergleichsweise ökonomische Methode, die einen

Beitrag zur Sicherung der Datenqualität leisten kann. Da es sich zudem um sehr leichte Items handelt, ist der zusätzliche Zeitbedarf voraussichtlich sehr gering. Wenngleich die Ergebnisse zur Modus-Ponens-Reduktion im Rahmen dieser Arbeit (siehe Abschnitt 2.4.2) die Eignung einer solchen Leichte-Items-Reduktion zur Sicherung der Datenqualität in Online-Erhebungen stützen, besteht weiterer Forschungsbedarf hinsichtlich Zuverlässigkeit, Validierung und Übertragbarkeit auf andere Items als Modus-Ponens-Aufgaben. Die Replizierbarkeit der Ergebnisse dieser Arbeit sollte ebenso überprüft werden wie strukturelle Veränderungen von Klassen- oder Faktorlösungen nach einer solchen Leichte-Items-Reduktion.

Zu den Forschungs-Implikationen soll abschließend erwähnt werden, dass aufgrund der Ergebnisse der vorliegenden Arbeit bei jeglichen Aufgaben zum Konditionalen Schlussfolgern das Antwortverhalten nicht aufgabenweise, sondern stets über die Schlussfiguren (und damit über die Aufgaben) hinweg betrachtet werden sollte. Insbesondere das Löseverhalten bei Modus-Tollens-Aufgaben sollte stets im Zusammenhang mit dem Löseverhalten bei NA- und BK-Aufgaben analysiert werden. Andernfalls ist eine Differenzierung zwischen der Unsophistizierten- und der Sophistizierten-Stufe nicht möglich (vgl. auch Rijmen & De Boeck, 2003). Dies gilt ebenso für andere Aufgabentypen zum Konditionalen Schlussfolgern, wie bspw. die Wason Selection Task (Wason, 1966). Dabei handelt es sich um eine sehr bekannte Aufgabe zum Konditionalen Schlussfolgern. Es werden vier Karten vorgegeben und der Proband soll entscheiden, durch Umdrehen welcher der vier Karte eine gegebene Konditionalaussage überprüft werden kann. Jede dieser vier Karten repräsentiert dabei – gegeben der Konditionalaussage – eine der vier Schlussfiguren. Auch im Falle der Wason Selection Task sprechen neuere Untersuchungen dafür, nicht mehr die Entscheidung bezüglich einzelner Karten zu analysieren, sondern die Entscheidung und damit das Antwortmuster über alle vier Karten hinweg (vgl. z.B. Klauer, Stahl & Erdfelder, 2007).

4.3.2 Implikationen für die psychologische Praxis

In der Folge werden Implikationen der vorliegenden Arbeit für die psychologische Praxis abgeleitet. Zunächst wird kurz auf die Trainierbarkeit von SKS eingegangen, anschließend folgen − deutlich ausführlicher − Implikationen für die psychodiagnostische Praxis. Dabei wird zuerst thematisiert, wie der KKS zur Schließung einer bislang existierenden Lücke in der Intelligenzdiagnostik beitragen kann. Danach wird eine

Möglichkeit für die konkrete Umsetzung des KKS zur Individualdiagnostik vorgestellt, bevor abschließend mögliche Weiterentwicklungen des KKS überlegt werden.

Eine erste praktische Implikation der Ergebnisse der vorliegenden Arbeit ist die Möglichkeit, Sophistiziertheit Konditionalen Schlussfolgerns zu trainieren. Sie ergibt sich vor allem aus der Betrachtung der vier Sophistiziertheits-Stufen als Kompetenz-Stufen. Gegenüber einer Fähigkeit zum Konditionalen Schlussfolgern als „Teilfähigkeit" des Deduktiven Denkens (wie zunächst vermutet; vgl. Abschnitt 2.1.4.1) haben Kompetenzen den Vorteil, dass sie (leichter) trainier- und erlernbar sind. Inhalte von Trainingsmaßnahmen zum Erreichen höherer Sophistiziertheits-Stufen Konditionalen Schlussfolgerns könnten aufbauend auf den kognitionspsychologischen Grundlagen der Testentwicklung der Umgang mit einladenden Inferenzen, mit doppelten Negationen und mit der reductio ad absurdum sein. Weitere Ansätze für Trainings zum Konditionalen Schlussfolgern in Verbindung mit einer kurzen Bestandsaufnahme zu Trainings zum logischen Schlussfolgern allgemein finden sich bspw. bei Kleinbeck (2005).

In der Folge sollen nun mögliche Implikationen für die psychodiagnostische Praxis vorgestellt werden. Unter den ohnehin wenigen deutschsprachigen Tests zum Deduktiven Denken findet sich bislang im Erwachsenenbereich kein Test, der Aufgaben zum Konditionalen Schlussfolgern verwendet (vgl. Abschnitt 2.1.4.2). Selbst zu Aufgaben mit Propositionalen Schlüssen (denen Konditionale Schlüsse in der Aussagenlogik zugeordnet werden; vgl. Abschnitt 2.1.1) existiert kein entsprechendes Testverfahren. Wie bereits in Abschnitt 2.2 beschrieben, unternimmt Wilhelm zumindest den Versuch einer Testkonstruktion unter Verwendung von Items zum Konditionalen Schlussfolgern, wobei davon ausgegangen wird, dass das Lösen solcher Items auf der Erzeugung, Aufrechterhaltung und Bearbeitung mentaler Modelle basiert. Die zugrundeliegende Fähigkeit zum Lösen logischer Denkprobleme nimmt Wilhelm als dimensional an. Es ergeben sich jedoch Probleme bezüglich der psychometrischen Eigenschaften der Items zum Konditionalen Schlussfolgern (Wilhelm, 2000). Die Existenz verschiedener Sophistiziertheits-Stufen Konditionalen Schlussfolgerns (als eines der Hauptergebnisse der vorliegenden Arbeit) liefert eine mögliche Erklärung für diese Probleme. So könnten diese auf die unterschiedlichen Ausprägungen von SKS zurückzuführen sein. Für Personen auf den beiden Unsophistizierten- oder der Fortgeschrittenen-Stufe resultieren möglicherweise Folgefehler für alle weiteren mentalen Modelle, die zum Lösen einer Aufgabe notwendig sind. Lediglich für Personen auf der Sophistizierten-Stufe ist sichergestellt, dass sie prinzipiell in der Lage sind, Aufgaben zu allen vier Schlussfiguren

korrekt zu lösen. Nur dann können weitere Manipulationen der Aufgabenschwierigkeit − wie von Wilhelm auf Basis der Modelltheorie vorgenommen − zu sinnvoll interpretierbaren Ergebnissen führen.

Um diese Vermutung zu überprüfen, könnte man die Daten der Aufgaben mit Konditionalaussagen in dem Test zum Propositionalen Schlussfolgern (Wilhelm, 2000; Wilhelm, Witthöft & Größler, 1999) bspw. unter Verwendung von Mixed-Rasch-Modellen (Rost, 1990; siehe auch Rost 2004) reanalysieren. Dabei würden drei (im Falle zusätzlicher Negationen in der Hauptprämisse vier) latente Klassen angenommen und innerhalb dieser Klassen die aus der Modelltheorie abgeleiteten Annahmen über eine kontinuierliche latente Variable überprüft. Möglicherweise sind dadurch die psychometrischen Eigenschaften bislang ausgeschlossener Items (vgl. Wilhelm, 2000) theoriekonform erklärbar. Daraus ergäben sich für Tests zum Propositionalen Schlussfolgern, die auf Basis der Theorie mentaler Modelle konstruiert werden, zwei Möglichkeiten. Zum einen könnten prinzipiell keine Items verwendet werden, die Konditionalaussagen enthalten. Diese Strategie ist pragmatisch, scheint jedoch angesichts der enormen Bedeutung des Konditionalen Schlussfolgerns in der Kognitionspsychologie (vgl. Abschnitt 2.1.2) unangebracht. Angemessener wäre eine zweite Variante, in der mittels eines kurzen Screening-Tests (bspw. des KKS) erhoben wird, welcher Sophistiziertheits-Stufe Konditionalen Schlussfolgerns der Proband zuzuordnen ist. Gegeben der Stufe können die später vorgegebenen Schlussfolgerungs-Aufgaben auch bei einer Falschlösung prinzipiell korrekt bearbeitet worden sein. Lediglich die Unsophistizierten- oder Fortgeschrittenen-Stufe im Konditionalen Schlussfolgern führt zu dieser Falschlösung, während alle anderen Lösungsschritte korrekt durchgeführt werden. Bei Berücksichtigung der Ergebnisse der vorliegenden Arbeit scheint also auch die Konstruktion von Tests zum Propositionalen Schlussfolgern möglich, ohne dass bei den Items auf Konditionalaussagen verzichtet werden muss. Möglicherweise ist so die Konstruktion einer dimensionalen Fähigkeit auf Basis von Aufgaben zum Propositionalen oder auch Konditionalen Schlussfolgern möglich. Damit könnten dann, ergänzend zu den Testverfahren mit Aufgaben zum Syllogistischen und Relationalen Schlussfolgern von Wilhelm (1995; siehe auch Wilhelm & Conrad, 1998) die nach Knauff (2006) drei wichtigsten Formen Deduktiven Denkens erfasst werden. Eine bislang in der Intelligenzdiagnostik existierende Lücke kann so möglicherweise geschlossen werden.

In der vorliegenden Arbeit wird der KKS in empirischen Studien mit umfangreichen Stichproben zur Beantwortung von Forschungsfragen eingesetzt, vorrangig mit

dem Ziel, Beziehungen zwischen latenten Variablen zu analysieren. Dafür erweist er sich als prinzipiell geeignet. Um den Test jedoch in der psychodiagnostischen Praxis bspw. zur Einzelfalldiagnostik einzusetzen, bedarf es einer entsprechenden Konzeption. Eine Möglichkeit wird in der Folge skizziert.

Die zu testende Person bearbeitet den KKS am Rechner. Dies kann online geschehen, ist aber auch offline möglich. Die Daten der Testung werden an eine Datenbank geschickt, die mit diesen Daten und einem bestehenden Normierungsdatensatz ein LCM unter Vorgabe von vier Klassen berechnet. Die resultierende Klassenzuordnung (der Person) wird zusammen mit der entsprechenden Zuordnungswahrscheinlichkeit an den Testrechner zurückgemeldet. Basierend auf diesen beiden Informationen erhält die getestete Person ein Feedback zu ihrer Leistung im Test[77]. Aufgrund der rechnergestützten Testdurchführung und der gegebenen Interpretationsobjektivität (vgl. Abschnitt 3.1) ist eine automatisierte Generierung dieses Feedbacks durch Kombination verschiedener Textbausteine denkbar. Für das bei jeder Testung berechnete LCM werden (z.B. mittels parametrischen Bootstrapverfahrens) Gütekriterien der Modellpassung berechnet (Chi-Quadrat, Cressie Read). So ist zusätzlich eine kontinuierliche Modellprüfung möglich, die die Interpretierbarkeit der geschätzten Personenparameter (Klassenzuordnung und zugehörige Zuordnungswahrscheinlichkeit) legitimiert.

Abschließend sollen noch einige mögliche Weiterentwicklungen des KKS aufgezählt werden. Dazu zählt zunächst – je nach Ergebnissen weiterer Studien zu Inhaltseffekten – die berufsspezifische Anpassung der Iteminhalte. Beim Einsatz bspw. zur Personalauswahl in einer nicht-technischen Branche können anstelle des Funktionierens elektrischer Schaltungen branchenspezifische Inhalte verwendet werden. Dies wäre eine Möglichkeit, die Attraktivität des Tests für die jeweilige Zielgruppe zu erhöhen. Inwieweit für diese Item-Inhalte Inhaltseffekte und Effekte von Vorwissen ausgeschlossen werden können, ist jedoch im Vorhinein zu überprüfen. In Abschnitt 4.3.1 finden sich Vorschläge für diesbezügliche Studien. Des Weiteren könnten sowohl Haupt- als auch Nebenprämisse um weitere aussagenlogische Propositionen (Konjunktionen, Disjunktionen) erweitert werden. Eine sukzessive Erweiterung des KKS zu einem Test zum Propositionalen Schlussfolgern wäre auf diese Weise möglich.

[77] Ansatzpunkte für entsprechende Feedbacks finden sich bei den Ausführungen zur Interpretationsobjektivität in Abschnitt 3.1.

4.4 Resümee

Ein allgemeiner Trend der psychologischen Forschung ist die steigende Interdisziplinarität innerhalb dieser Wissenschaft (vgl. z.B. Borkenau et al., 2005; Schneider, 2005). Die verschiedenen Grundlagenfächer der Psychologie vernetzen sich zunehmend. Immer häufiger werden vielbetrachtete Phänomene eines Grundlagenfaches aus Perspektive eines anderen analysiert, bewertet und häufig um neue Blickwinkel ergänzt. In der vorliegenden Arbeit erfolgt die Entwicklung eines psychodiagnostischen Testverfahrens zur Erfassung einer Kompetenz zum Konditionalen Schlussfolgern auf Basis kognitionspsychologischer Theorien (gemäß der ersten Forderung an die vorliegende Arbeit; vgl. Abschnitt 2). Aus der Testentwicklung und -erprobung ergeben sich daher auch Implikationen für eben diese kognitionspsychologischen Theorien und damit für die Grundlagenpsychologie. Die vorliegende Arbeit folgt damit ebenfalls dem beschriebenen Trend.

Der entwickelte Test erfüllt die betrachteten Testgütekriterien. Eine Besonderheit bei deren Bewertung und empirischer Überprüfung besteht darin, dass *Sophistiziertheit Konditionalen Schlussfolgerns*, das zugrundeliegende Konstrukt, als nominale latente Variable definiert wird. Der in der zweiten Forderung an die vorliegende Arbeit formulierte Einsatz von Latente-Variablen-Modellen zur empirischen Hypothesenprüfung (vgl. Abschnitt 2.4.1) erfordert deshalb die Anwendung vergleichsweise unbekannter Analyseverfahren. Diese liefern adäquate Informationen und gut interpretierbare Ergebnisse und bieten sich daher generell für die Entwicklung und Erprobung psychodiagnostischer Testverfahren zur Erfassung nominaler Konstrukte an.

Mit dem *Kurztest zum Konditionalen Schlussfolgern* (*KKS*) existiert damit ein auf kognitionspsychologischen Grundlagen aufbauender Test, mit dem Personen verschiedenen Sophistiziertheits-Stufen im Konditionalen Schlussfolgern zugeordnet werden können. Durch den Einsatz des KKS in der psychologischen Forschung und in der psychodiagnostischen Praxis lässt sich möglicherweise eine bislang existierende Lücke in der Diagnostik Deduktiven Denkens schließen.

5 Literatur

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1*(2), 201-218.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioural and Brain Sciences, 14*, 471-517.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Asendorpf, J. B. (2007). *Psychologie der Persönlichkeit*. Berlin: Springer.

Batinic, B. (2006). *Internetbasierte Befragungen und Experimente*. Tutorial auf dem 45. Kongress der Deutschen Gesellschaft für Psychologie, Nürnberg.

Batinic, B. & Bosnjak, M. (2000). Fragebogenuntersuchungen im Internet. In B. Batinic (Hrsg.), *Internet für Psychologen, 2. Aufl.* (S. 287-317). Göttingen: Hogrefe.

Beck, C. (2007). *Kompetenz-Studie. Welche Kompetenzen fordern die Unternehmen von Bewerbern? -Ergebnisse-*. Koblenz 01/2007.

Beckmann, J. F. & Guthke, J. (1999). *Psychodiagnostik des schlußfolgernden Denkens*. Göttingen: Hogrefe.

Begg, I. & Denny, J. (1969). Empirical reconciliation of atmosphere and conversion interpretations of syllogistic reasoning. *Journal of Experimental Psychology, 81*, 351-354.

Beller, S. (1997). *Inhaltseffekte beim logischen Denken - Der Fall der Wason´schen Wahlaufgabe: Eine wissensbasierte Lösung für ein altes Problem*. Lengerich: Pabst.

Beller, S. & Spada, H. (2003). The logic of content effects in propositional reasoning: the case of conditional reasoning with a point of view. *Thinking and Reasoning, 9*(4), 335-378.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M. & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Berg, M. & Schaarschmidt, U. (1984). Überlegungen zu neuen Wegen in der Intelligenzdiagnostik. *Wissenschaftliche Zeitschrift der Humboldt-Universität zu Berlin. Mathematisch-naturwissenschaftliche Reihe, 6*, 565-573.

Böhme, H. F. & Steyer, R. (2008). *Konstruktion Psychometrischer Fähigkeitstests für Chip-Designer*. Friedrich-Schiller-Universität Jena: Unveröffentlichter Projektbericht.

Bollen, K. A. & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Bonatti, L. (1994). Propositional reasoning by model? *Psychological Review, 101*, 725-733.

Bonnefon, J. F., Eid, M., Vautier, S. & Jmel, S. (2008). A mixed Rasch model of dual-process conditional reasoning. *Quarterly Journal of Experimental Psychology, 61*, 809-824.

Borkenau, P., Egloff, B., Eid, M., Hennig, J., Kersting, M., Neubauer, A. C. & Spinath, F. M. (2005). Persönlichkeitspsychologie: Stand und Perspektiven. *Psychologische Rundschau*, *56*(4), 271-290.

Borsboom, D. & Mellenbergh, G. J. (2007). Test validity in cognitive assessment. In J. Leighton & M. Gierl (Hrsg.). *Cognitive diagnostic assessment for education: Theory and applications* (S. 85-115). Cambridge: Cambridge University Press.

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

Borsboom, D., van Heerden, J. & Mellenbergh, G. J. (2003). Validity and truth. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano & J. J. Meulman (Hrsg.), *New developments in psychometrics. Proceedings of the International Meeting of the Psychometric Society 2001* (S. 321-328). Tokio: Springer.

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

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation (2. Aufl.)*. Berlin: Springer.

Bortz, J., Lienert G. A. & Boehnke, K. (2000). *Verteilungsfreie Methoden in der Biostatistik (2. Aufl.)*. Heidelberg: Springer.

Braine, M. D. S. (1978). On the relation between the natural logic of reasoning and standard logic. *Psychological Review, 85*, 1-21.

Braine, M. D. S. & O'Brien, D. P. (1991). A theory of if. A lexical entry, reasoning program, and pragmatic principles. *Psychological Review, 98*, 182-203.

Braine, M. D. S. & O'Brien, D. P. (1998). *Mental logic*. Mahwah, NJ: Lawrence Erlbaum.

Braine, M. D. S., Reiser, B. J. & Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In G. H. Bower (Hrsg.), *The psychology of learning and motivation, 18* (S. 313–371). New York: Academic Press.

Braine, M. D. S. & Rumain, B. (1983). Logical reasoning. In J. H. Flavell & E. M. Markman (Hrsg.), *Handbook of child psychology, Vol. 3.: Cognitive development* (S. 263-339). New York: John Wiley & Sons.

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Hrsg.), *Testing structural equation models* (S. 136-161). Newbury Park, CA: Sage.

Bühner M. (2006). *Einführung in die Test- und Fragebogenkonstruktion (2. Aufl.)*. München: Pearson Studium.

Byrne, R. M. J. & Handley, S. J. (1992). Reasoning strategies. *Irish Journal of Psychology, 13*, 111-124.

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

Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404-431.

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

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues* (S. 122-130). New York: Guilford.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues, 2. Aufl.* (S. 69-76). New York: Guilford.

Cattell, R. B. (1946). *Description and measurement of personality*. New York: World Book Company.

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

Chater, N. & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38*, 191-258.

Clark, S. & Muthén, B. (2009). Relating latent class analysis results to variables not included in the analysis. Submitted for publication.

Clogg, C. C. (1995). Latent class models: Recent developements and prospects for the future. In G. Arminger, C. C. Clogg & M. E. Sobel (Hrsg.), *Handbook of statistical modeling in the social sciences* (S. 311-359). New York: Plenum.

Clogg, C. C. & Manning, W. D. (1996). Assessing reliability of categorical measurements using latent class models. In A. von Eye & C. C. Clogg (Hrsg.), *Categorical variables in developmental research: Methods of analysis* (S.169-182). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1034-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Collins, L. M. (2001). Measurement reliability for static and dynamic categorical latent variables. In L. M. Collins & A. G. Sayer (Hrsg.), *New methods for the analysis of change* (S. 273-288). Washington, DC: American Psychological Association.

Collins, L. M. & Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27*, 131-137.

Conley, J. J. (1984). The hierarchy of consistency: A review and model of longitudinal findings on adult individual differences in intelligence, personality and self-opinion. *Personality and Individual Differences, 5*, 11-25.

Cox, D. R. & Snell, E. J. (1989). *The analysis of binary data (2. Aufl.)*. London: Chapman and Hall.

Cressie, N. & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society, Series B, 46*, 440-464.

Cronbach, L. J. & Mehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Croon, M. (1990). Latent class analysis with ordered latent classes. *British Journal of Mathematical and Statistical Psychology, 43*, 171-192.

Davier, M. v. (1997). Bootstrapping Goodness-of-Fit Statistics for Sparse Categorical Data - Results of a Monte Carlo Study. *Methods of Psychological Research - MPR online, 2*(2), Lengerich: Pabst Science Publishers.

Davier, M. v. (2001). *WINMIRA user manual*. Kiel: Institut für die Pädagogik der Naturwissenschaften.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J. & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish Mental Surveys of 1932 and 1947. *Journal of Personality and Social Psychology, 86*, 130-147.

De Neys, W., Schaeken, W. & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking and Reasoning, 11*(4), 349-381.

Dörner. D. (1989). *Die Logik des Mißlingens.* Reinbek: Rowohlt Verlag.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eid, M. (1997). Happiness and satisfaction: An application of a latent state-trait model for ordinal variables. In J. Rost & R. Langeheine (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 145-151). Münster: Waxmann.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241-261.

Eid, M. & Langeheine, R. (1999). Measuring consistency and occasion specificity with latent class models: A new model and its application to the measurement of affect. *Psychological Methods, 4*, 100-116.

Eid, M. & Langeheine, R. (2003). Separating stable from variable individuals in longitudinal studies by mixture distribution models. *Measurement: Interdisciplinary Research and Perspectives, 1*, 179-206.

Eid, M. & Langeheine, R. (2007). Detecting population heterogeneity in stability and change of subjective well-being by mixture distribution models. In A. Ong & M. van Dulmen (Hrsg.), *Handbook of methods in positive psychology* (S. 609-632). Oxford: Oxford University Press.

European Federation of Psychologists' Associations (EFPA, 2008). EFPA review model for the description and evaluation of psychological tests. Test review form and notes for reviewers. Version 3.42. Verfügbar unter: www.efpa.eu/download/9044bd41c7953b956876e06c797f8c9f [11. März 2010].

Evans, J. St. B. T. (1977a). Linguistic factors in reasoning. *Quarterly Journal of Experimental Psychology, 29*, 297-306.

Evans, J. St. B. T. (1977b). Toward a statistical theory of reasoning. *Quarterly Journal of Experimental Psychology, 29*, 621-635.

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. London: Routledge & Kegan Paul.

Evans, J. St. B. T. (1984). Heuristic and analytical processes in reasoning. *British Journal of Psychology, 75*, 451-468.

Evans, J. St. B. T. (1993). The mental model theory of conditional reasoning: critical appraisal and revision. *Cognition, 48*, 1-20.

Evans, J. St. B. T. (1998). Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking and Reasoning, 4*, 45-82.

Evans, J. St. B. T., Clibbens, J. & Rood, B. (1995). Bias in conditional inference: Implications for mental models and mental logic. *The Quarterly Journal of Experimental Psychology, 48A*(3), 644-670.

Evans, J. St. B. T. & Handley, S. J. (1999). The role of negation in conditional inference. *The Quarterly Journal of Experimental Psychology, 52A*(3), 739-769.

Evans, J. St. B. T., Handley, S. J., Neilens, H. & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition, 35*(7), 1772-1784.

Evans, J. St. B. T., Handley, S. J. & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*, 321-355.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Erlbaum.

Evans, J. St. B. T. & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Lawrence Erlbaum.

Evans, J. St. B. T. & Over, D. E. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., Over, D. E. & Handley, S. J. (2005). Supposition, extensionality and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review, 112*(4), 1040-1052.

Evers, A. (2001). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing, 1*, 137-153.

Fahrmeir, L., Kneib, T. & Lang, S. (2007). *Regression*. Berlin: Springer.

Falmagne, R. J. (1993). On modes of explanation. *Behavioral and Brain Sciences, 16*(2), 346-347.

Falmagne, R. J. & Gonsalves, J. (1995). Deductive inference. *Annual Review of Psychology, 46*, 525-559.

Fillenbaum, S. (1993). Deductive reasoning: What are taken to be the premises and how are they interpreted? *Behavioral and Brain Sciences, 16*(2), 348-349.

Fisseni, H.-J. (1997). *Lehrbuch der psychologischen Diagnostik (2. Aufl.)*. Göttingen: Hogrefe.

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

Formann, A. K. (1984). *Die Latent-Class-Analyse*. Weinheim: Beltz.

Funke, F. & Reips, U.-D. (2007). Messinstrumente und Skalen. In M. Welker & O. Wenzel (Hrsg.), *Online-Forschung 2007: Grundlagen und Fallstudien* (S. 52-76). Köln: Herbert von Halem.

Geis, M. C. & Zwicky, A. M. (1971). On invited inferences. *Linguistic Inquiry, 2*, 561-566.

Giesen, H., Gold, A., Hummer, A. & Jansen, R. (1986). *Prognose des Studienerfolgs*. Frankfurt am Main: Institut für Pädagogische Psychologie.

Gigerenzer, G. & Gaissmaier, W. (2006). Denken und Urteilen unter Unsicherheit: Kognitive Heuristiken. In J. Funke (Hrsg.), *Band C/II/8 der Enzyklopädie der Psychologie „Denken und Problemlösen“* (S. 329-374). Göttingen: Hogrefe.

Gollwitzer, M. (2007). Latent-Class-Analysis. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 279-306). Heidelberg: Springer.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215-231.

Graham, J. W., Collins, L. M., Wugalter, S. E., Chung, N. K. & Hansen, W. B. (1991). Modeling transitions in latent stage-sequential processes: A substance use prevention example. *Journal of Consulting and Clinical Psychology, 59*, 48-57.

Green, D. M. & Swets J. A. (1966). *Signal Detection Theory and Psychophysics*. New York: John Wiley & Sons.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Hrsg.), *Syntax and semantics, III: Speech acts* (S. 41-58). New York: Academic Press.

Grice, H. P. (1978). Further notes on logic and conversation. In P. Cole (Hrsg.), *Syntax and semantics, IV: Pragmatics* (S. 113-127). New York: Academic Press.

Grohmann, A. (2008). *Erprobung eines internetbasierten, eignungsdiagnostischen Verfahrens aus Perspektive der Latent-State-Trait-Theorie*. Friedrich-Schiller-Universität Jena: Unveröffentlichte Diplomarbeit.

Guthke, J., Böttcher, H. R. & Sprung, L. (1990). *Psychodiagnostik. Ein Lehrbuch für Psychologen sowie empirisch arbeitende Humanwissenschaftler. Band 1*. Berlin: Deutscher Verlag der Wissenschaften.

Guttman, L. & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence, 15*, 79-104.

Halpern, D. F. (2000). *Sex differences and cognitive abilities (3. Aufl.)*. Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hartenstein, S. (2007). *flexSURVEY 0.9.4 (Software)*. www.flexsurvey.de.

Hartig, J., Frey, A. & Jude, N. (2007). Validität. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 135-163). Heidelberg: Springer.

Horn, J. L. & Blankson, N. (2005). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues, 2. Aufl.* (S. 41-68). New York: Guilford.

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

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression (2. Aufl.)*. New York: John Wiley & Sons.

Hu, L.-T. & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Hrsg.), *Structural equation modeling. Concepts, issues, and applications* (S. 76-99). London: Sage.

Hu, L.-T. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424-453.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Ihme, J. M. (2007). Ergänzung des Itempools des Matrizentests der adaptiven eignungsdiagnostischen Untersuchung. *Untersuchungen des Psychologischen Dienstes der Bundeswehr, 42*, 93-110.

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica, 28*, 195-226.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau, 35*, 21-35.

Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur Test. BIS-Test, Form 4*. Göttingen: Hogrefe.

Johnson-Laird, P. N. (1994). A model theory of induction. *International Studies in the Philosophy of Science, 8*, 5-29.

Johnson-Laird, P. N. (1997a). Rules and illusions: A critical study of Rips's The Psychology of Proof. *Mind and Machines, 7*(3), 387-407.

Johnson-Laird, P. N. (1997b). An end to the controversy? A reply to Rips. *Mind and Machines, 7*(3), 425-432.

Johnson-Laird, P. N. (1999) Reasoning: Formal rules vs. mental models. In R. J. Sternberg (Hrsg.), *Conceptual issues in psychology.* Cambridge, MA: MIT Press.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Science 5*(10), 434-442.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Lawrence Erlbaum.

Johnson-Laird, P. N. & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review, 109*(4), 646–678.

Johnson-Laird, P. N., Byrne, R. M. J. & Schaeken, W. (1994). Why models rather than rules give a better account of propositional reasoning: A reply to Bonatti and to O'Brien, Braine, and Yang. *Psychological Review 101*(4), 734-739.

Jöreskog, K. & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least square methods*. Mooresville, IN: Scientific Software.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York: Macmillian.

Kiesler, S. & Sproull, L. S. (1986). Response effects in the electronic survey. *Public Opinion Quarterly, 50*, 402-413.

Klaczynski, P. A. & Daniel, D. B. (2005). Individual differences in conditional reasoning: A dual-process account. *Thinking and Reasoning 11*(4), 305-325.

Klauer, K. C., Stahl, C. & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 680-703.

Klauer, K. J. (1978). Kontentvalidität. In K. J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik, Band 1* (S. 225-255). Düsseldorf: Schwann.

Klauer, K. J. (1984). Kontentvalidität. *Diagnostica, 30*, 1-23.

Kleinbeck, S. (2005). *Untersuchung zum Umgang mit abstrakten Konditionalaussagen und ihren negierten Komponenten*. Freiburg i. Br.: Dissertation.

Kleinmuntz, B. & McLean, R. S. (1968). Computers in behavioral science: Diagnostic interviewing by digital computer. *Behavioral Science, 13*, 75-80.

Knauff, M. (2006). Deduktion, logisches Denken. In J. Funke (Hrsg.), *Band C/II/8 der Enzyklopädie der Psychologie „Denken und Problemlösen“* (S. 167-264). Göttingen: Hogrefe.

Knauff, M., Rauh, R. & Schlieder, C. (1995). Preferred mental models in qualitative spatial reasoning: A cognitive assessment of Allen's calculus. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (S. 200-205). Mahwah, NJ: Lawrence Erlbaum.

Kruse, A. & Schmitt, E. (2004). Differentielle Psychologie des Alterns. In K. Pawlik (Hrsg.), *Enzyklopädie der Psychologie, Band VII: Differentielle Psychologie und Persönlichkeitspsychologie*. (S. 533-571). Göttingen: Hogrefe.

Kubinger K. D. (2003). Gütekriterien. In K. D. Kubinger & R. S. Jäger (Hrsg.), *Schlüsselbegriffe der Psychologischen Diagnostik* (S. 195-204). Weinheim: Beltz, PVU.

Kubinger, K. D. & Jäger, R. S. (2003). *Schlüsselbegriffe der Psychologischen Diagnostik*. Weinheim: Beltz, PVU.

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

Kutschera, F. v. & Breitkopf, A. (2007). *Einführung in die moderne Logik. (8. Aufl.)*. Freiburg: Alber.

Kyllonen, P. C. (1994). Aptitude testing inspired by information processing: A test of the four-sources model. *Journal of General Psychology, 120*, 375-405.

Kyllonen, P. C. & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*, 389-433.

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.

Lea, R. B. & Mulligan, E. J. (2002). The effect of negation on deductive inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(2), 303-317.

Lienert, G. & Raatz, U. (1998). *Testaufbau und Testanalyse (6. Aufl.)*. Weinheim: Beltz, PVU.

Lüer, G. & Spada, H. (1992). Denken und Problemlösen. In H. Spada (Hrsg.), *Lehrbuch Allgemeine Psychologie, 2. Aufl.* (S. 189-280). Bern: Huber.

Manktelow, K. I. (1999). *Reasoning and thinking*. Hove, UK: Psychology Press.

Martin, D. W. (1996). *Doing psychological experiments (4. Aufl.)*. Pacific Grove, CA: Brooks/Cole.

McCutcheon, A. L. (1987). *Latent class analysis*. Beverly Hills and London: Sage.

McGrew, K. S. (2005). The Cattell–Horn–Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues, 2. Aufl.* (S. 136-181). New York: Guilford.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1-10.

Messick, S. (1989). Validity. In R. L. Linn (Hrsg.), *Educational measurement, 3. Aufl.* (S. 13-103). New York: Macmillan.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessment, 10*(1), 1-9.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*(9), 741-749.

Moosbrugger, H. & Kelava, A. (2007). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 7-26). Heidelberg: Springer.

Muthén, L. K. and Muthén, B. O. (1998-2007). *Mplus User's Guide. (5. Aufl.)*. Los Angeles, CA: Muthén & Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Neth, H., Beller, S. & Spada, H. (1999). How knowledge interferes with reasoning - Suppression effects by content and context. In M. Hahn, & S. C. Stones (Hrsg.), *Proceedings of the twenty first annual conference of the Cognitive Science Society* (S. 468-473). Mahwah, NJ: Lawrence Erlbaum.

Newstead, S. E., Handley, S. J., Harley, C., Wright, H. & Farrelly, D. (2004). Individual differences in deductive reasoning. *The Quarterly Journal of Experimental Psychology, 57A*(1), 33-60.

Nosek, B. A., Banaji, M. R. & Greenwald, A. G. (2002). E-Research: Ethics, security, design, and control in psychological research on the internet. *Journal of Social Issues, 58*(1), 161-176.

Nylund, K. L. (2007). *Latent transition analysis: Modeling extensions and an application to peer victimization*. Doctoral dissertation: University of California, Los Angeles.

Nylund, K. L., Asparouhov, T. & Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling. A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535-569.

Oaksford, M. & Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(4), 835-854.

Oberauer, K., Geiger, S. M., Fischer, K. & Weidenfeld, A. (2007). Two meanings of "if"? Individual differences in the interpretation of conditionals. *The Quarterly Journal of Experimental Psychology, 60*(6), 790-819.

Oberauer, K. & Wilhelm, O. (2003). The meaning(s) of conditionals – Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 680-693.

O'Brien, D. P. (1987). The development of conditional reasoning: An iffy proposition. In H. Reese (Hrsg.), *Advances in child behaviour and development, Vol. 20* (S. 61-90). New York: Academic Press.

O'Brien, D. P. (2004). Mental-logic theory: What it proposes, and reasons to take this proposal seriously. In J. P. Leighton & R. J. Sternberg (Hrsg.), *The nature of reasoning* (S. 205-233). New York: Cambridge University Press.

O'Brien, D. P., Braine, M. D. S. & Yang, Y. (1994). Propositional reasoning by mental models? Simple to refute in principle and in practice. *Psychological Review, 101*, 725-733.

O'Brien, D. P. & Overton, W. F. (1982). Conditional reasoning and the competence-performance issue: A developmental analysis of a training task. *Journal of Experimental Child Psychology, 34*, 274-290.

O'Connor, M. & Paunonen, S. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43*, 971-990.

Petermann, F. & Eid, M. (2006). *Handbuch der Psychologischen Diagnostik*. Göttingen: Hogrefe.

Piaget, J. (1971). *Biology and Knowledge*. Chicago: University of Chicago Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y. & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903.

Pohl, S. & Steyer, R. (in press). Modelling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research.*

Pollard, P. & Evans, J. St. B. T. (1980). The influence of logic on conditional reasoning performance. *Quarterly Journal of Experimental Psychology, 32*, 605-624.

Preckel, F. (2003). *Diagnostik intellektueller Hochbegabung. Testentwicklung zur Erfassung der fluiden Intelligenz*. Göttingen: Hogrefe.

Rajulton, F. & Ravanera, Z. R. (2001). Stability and change: Illustrations with categorical and binary responses. *Canadian Studies in Population, Special Issue on Longitudinal Methodology, 28*(2), 491-512.

Ramaswamy, V., DeSarbo, W., Reibstein, D. & Robinson, W. (1993). An empirical pooling approach for estimating marketing mix elasticities with PIMS data. *Marketin Science, 12*, 103-124.

Rao, C. R. (1973). *Linear statistical inference and its applications (2. Aufl.)*. New York: John Wiley & Sons.

Raven, J., Raven, J. C. & Court, J. H. (2004). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3: The Standard Progressive Matrices*. San Antonio, TX: Harcourt Assessment.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Hrsg.), *Psychological experiments on the Internet* (S. 89-114). San Diego, CA: Academic Press.

Reips, U.-D. (2001). The Web Experimental Psychology Lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, and Computers, 33*(2), 201-211.

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49*(4), 243-256.

Rijmen, F. & De Boeck, P. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research, 67*, 219-231.

Rips, L. J. (1983). Cognitive processes in propositional reasoning. *Psychological Review, 90*, 38-71.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*(2), 261-288.

Roberge, J. J. (1971). Some effects of negation on adults' conditional reasoning abilities. *Psychological Reports, 29*(3), 839-844.

Roberge, J. J. (1976). The effect of negation in adults' disjunctive reasoning abilities. *Journal of General Psychology, 91*, 23-28.

Roberts, M. J. (1993). Human reasoning: Deduction rules or mental models, or both? *Quarterly Journal of Experimental Psychology, 46A*(4), 569-589.

Rose, N., Pohl, S., Böhme, H. F. & Steyer, R. (im Druck). Strukturgleichungsmodelle. In H. Holling & B. Schmitz (Hrsg.), *Handbuch der Psychologischen Methoden und Evaluation* (im Druck). Berlin: Springer.

Rost, D. H. (2009). *Intelligenz: Mythen und Fakten.* Weinheim: Beltz/Psychologie Verlagsunion.

Rost, J. (1989). Rasch models and latent class models for measuring change with ordinal variables. In R. Coppi & S. Bolasco (Hrsg.), *Multiway Data Analysis* (S. 473-483). Amsterdam: Elsevier.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion (2. Aufl.)*. Bern: Huber.

Rost, J. (2006). Latent-Class-Analyse. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 275-287). Göttingen: Hogrefe.

Rumain, B., Connell, J. & Braine, M. D. S. (1983). Conversational comprehension processes are responsible for reasoning fallacies in children as well as adults: If is not the biconditional. *Developmental Psychology, 19*, 471-481.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C. & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, *56*(3), 573-605.

Sander, N. (2005). *Inhibitory and executive functions in cognitive psychology: An individual differences approach examining structure and overlap with working memory capacity and intelligence.* Aachen: Shaker.

Schaeken, W. & Schroyens, W. (2000). The effect of explicit negatives and of different contrast classes on conditional syllogisms. *British Journal of Psychology, 91*, 533-550.

Schermelleh-Engel, K., Kelava, A. & Moosbrugger, H. (2006). Gütekriterien. In F. Petermann & M. Eid (Hrsg.), *Handbuch der Psychologischen Diagnostik* (S. 420-433). Göttingen: Hogrefe.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*(2), S. 23-74.

Schermelleh-Engel, K. & Werner, C. (2007). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 113-133). Heidelberg: Springer.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274.

Schneider, W. (2005). Zur Lage der Psychologie in Zeiten hinreichender, knapper und immer knapperer finanzieller Ressourcen: Entwicklungstrends der letzten 35 Jahre. *Psychologische Rundschau*, *56*(1), 2-19.

Schroyens, W., Schaeken, W. & d'Ydewalle, G. (2001). The processing of negations in conditional reasoning: A meta-analytic case study in mental model and/or mental logic theory. *Thinking and Reasoning, 7*(2), 121-172.

Schroyens, W., Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2000). Conditional reasoning with negations: Implicit and explicit affirmation or denial and the role of contrast classes. *Thinking and Reasoning, 6*(3), 221-251.

Schuler, H. & Höft, S. (2006). Konstruktorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie, 2. Aufl.* (S. 101-144). Göttingen: Hogrefe.

Skyrms, B. (1989). *Einführung in die induktive Logik*. Frankfurt am Main: Lang.

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

Spearman, C. (1938). Measurement of intelligence. *Scientia, 64*, 75-82.

Spearman, C. (1939). "Intelligence" tests. *Eugenics-Review, 30*, 249-254.

Spiel, C., Gittler, G., Sirsch, U. & Glück, J. (1997). Application of the Rasch model for testing Piaget's theory of cognitive development. In R. Langeheine & J. Rost (Hrsg.), *Applications of latent trait and latent class models in the social sciences* (S. 115-123). Münster: Waxmann.

Spiel, C., Glück, J. & Gößler, H. (2001). Stability and change of unidimensionality: The sample case of deductive reasoning. *Journal of Adolescent Research, 16*(2), 150-168.

Spiel, C., Glück, J. & Gössler, M. (2004). Messung von Leistungsprofil und Leistungshöhe im schlussfolgernden Denken im SDV - Die Integration von Piagets Entwicklungskonzept und Item-Response Modellen. *Diagnostica, 50*(3), 145-152.

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning.* Mahwah, NJ: Lawrence Erlbaum.

Stanovich, K. E. & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*(2), 161-188.

Stanovich, K. E. & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645-665.

Stegmüller, W. (1996). *Das Problem der Induktion: Humes Herausforderung und moderne Antworten*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research, 25*, 173-180.

Stern, E. & Hardy, I. (2004). Differentielle Psychologie des Lernens in Schule und Ausbildung. In K. Pawlik (Hrsg.), *Theorien und Anwendungen der Differentiellen Psychologie* (S. 573-618). Göttingen: Hogrefe.

Steyer, R. (2001). Classical (Psychometric) Test Theory. In C. Ragin & T. Cook (Hrsg.), *International Encyclopedia of the Social and Behavioural Sciences. Logic of Inquiry and Research Design.* (S. 481-520). Oxford: Pergamon.

Steyer, R. (2003). *Wahrscheinlichkeit und Regression*. Heidelberg: Springer.

Steyer, R. & Eid, M. (2001). *Messen und Testen (2. Aufl.)*. Heidelberg: Springer.

Steyer, R., Ferring, D. & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment, 8*(2), 79-98.

Steyer, R., Schmitt, M. J. & Eid, M. (1999). Latent State-Trait Theory and research in personality and individual differences. *European Journal of Personality, 13,* 389-408.

Steyer, R., Schwenkmezger, P., Notz, P. & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen (MDBF)*. Göttingen: Hogrefe.

Süß, H.-M. (2007). Eine Intelligenz - viele Intelligenzen? Neuere Intelligenztheorien im Widerstreit - 1. Teil. *news@science. Begabtenförderung und Begabtenforschung, 15*, 18-27.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O. & Schulze, R. (2002). Working memory capacity explains reasoning ability – and a little bit more. *Intelligence, 30*, 261-288.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Thurstone, L. L. & Thurstone, T. G. (1941). *Factorial studies of intelligence*. Chicago: University of Chicago Press.

Trapmann, S. (2007). *Mehrdimensionale Studienerfolgsprognose. Die Bedeutung kognitiver, temperamentsbedingter und motivationaler Prädiktoren für verschiedene Kriterien des Studienerfolgs*. Berlin: Logos Verlag.

Trapmann, S., Hell, B., Hirn, J.-O. W. & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift für Psychologie / Journal of Psychology, 215*(2), 132-151.

Tucker, L. R. & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Tutz, G. (2000). *Die Analyse kategorialer Daten*. München: Oldenbourg.

Velden, M. (1982). *Die Signalentdeckungstheorie in der Psychologie*. Stuttgart: Kohlhammer.

Waldmann, M. R. & Weinert, F. E. (1990). *Intelligenz und Denken*. Göttingen: Hogrefe.

Wason, P. C. (1966). Reasoning. In B. Foss (Hrsg.), *New horizons in psychology* (S. 135-151). Harmonsworth, UK: Penguin.

Wechsler, D. (1964). *Die Messung der Intelligenz Erwachsener*. Bern: Huber.

Wildman, T. M. & Fletcher, H. J. (1977). Developmental increases and decreases in solutions of conditional syllogism problems. *Developmental Psychology, 13*, 630-636.

Wilhelm, O. (1995). *Entwicklung und Erprobung logischer Denktests*. Mannheim: Unveröffentlichte Diplomarbeit.

Wilhelm, O. (2000). *Psychologie des schlussfolgernden Denkens: Differentialpsychologische Prüfung von Strukturüberlegungen*. Hamburg: Dr. Kovac.

Wilhelm, O. & Conrad, W. (1998). Entwicklung und Erprobung von Tests zur Erfassung des logischen Denkens. *Diagnostica, 44*, 71-83.

Wilhelm, O. & McKnight, P. E. (2002). Ability and achievement testing on the World Wide Web. In B. Batinic, U.-D. Reips & M. Bosnjak (Hrsg.), *Online Social Sciences* (S. 151-180). Seattle, WA: Hogrefe & Huber.

Wilhelm, O., Witthöft, M. & Größler, A. (1999). Comparisons of paper-and-pencil and internet administrated ability and achievement tests. In P. Marquet, S. Mathey, A. Jaillet & E. Nissen (Hrsg.), *Proceedings of IN-TELE 98* (S. 439-449). Berlin: Peter Lang.

Wilson, R. S. (1983). The Louisville twin study: Developmental synchronies in behavior. *Child Development, 54*, 298-316.

Wittman, W. W. & Süß, H.-M. (1996). Vorhersage und Erklärung von Schulnoten durch das Berliner Intelligenzstrukturmodell. In B. J. Ertlet & M. Hofer (Hrsg.), *Theorie und Praxis der Beratung in Schule, Familie, Beruf und Betrieb* (S. 161-184). Festschrift zum 80. Geburtstag von Frau Prof. Dr. Elfriede Höhn. Nürnberg: Bundesanstalt für Arbeit.

Wittmann, W. W., Süß, H.-M., Oberauer, K., Schulze, R. & Wilhelm, O. (1995). *Der Zusammenhang von Arbeitsgedächtniskapazität und Konstrukten der Intelligenzforschung.* Unveröffentlichter DFG-Bericht.

Abkürzungsverzeichnis

BIS	…	Berliner Intelligenzstrukturmodell
BIS-4	…	Berliner Intelligenzstrukturtest
BIS-B	…	Operationsklasse Bearbeitungsgeschwindigkeit des Berliner Intelligenzstrukturmodells
BIS-E	…	Operationsklasse Einfallsreichtum des Berliner Intelligenzstrukturmodells
BIS-K	…	Operationsklasse Verarbeitungskapazität des Berliner Intelligenzstrukturmodells
BIS-M	…	Operationsklasse Merkfähigkeit des Berliner Intelligenzstrukturmodells
BK	…	Bestätigung der Konsequenz
EFPA	…	European Federation of Psychologists' Associations
IRT	…	Item-Response-Theorie
k.A.m.	…	keine Aussage möglich
KKS	…	Kurztest zum Konditionalen Schlussfolgern
KTT	…	Klassische Testtheorie
LCA	…	Analyse latenter Klassen
LCM	…	Latente-Klassen-Modell
LR-Test	…	Likelihood Ratio Test
LST	…	Latent-State-Trait
MP	…	Modus Ponens
MT	…	Modus Tollens
NA	…	Negation des Antezedens
NCB	…	Negative Conclusion Bias
OR	…	Odds Ratio
SKS	…	Sophistiziertheit Konditionalen Schlussfolgerns
WMT	…	Wiener Matrizen-Test

Anhang

Anhang A: Der KKS – erste Version und überarbeitete Version

Anhang A.1: Instruktion des KKS (überarbeitete Version)

Sehr geehrte Teilnehmerin, sehr geehrter Teilnehmer,
im nachfolgenden Test werden Ihnen 16 kurze Aufgaben präsentiert. Es geht dabei um elektrische Schaltungen, allerdings brauchen Sie zum Lösen der Aufgaben keinerlei Wissen über Elektrotechnik. Bitte nehmen Sie sich Zeit, die Aufgaben vollständig und gewissenhaft zu bearbeiten. Dies wird voraussichtlich 5-10 Minuten in Anspruch nehmen. Die Aufgaben bestehen jeweils aus einer Regel und einer gegebenen Aussage.

Hier ein Beispiel:
Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2.
Aussage: Schaltung 1 funktioniert.

Außerdem werden Ihnen verschiedene Schlussfolgerungen vorgegeben. Sie sollen nun entscheiden, welche dieser Schlussfolgerungen sich logisch eindeutig aus Regel und Aussage ableiten lässt.

Für dieses sehr einfache Beispiel würden Ihnen folgende Schlussfolgerungen vorgegeben:

- ☐ Schaltung 2 funktioniert.
- ☐ Schaltung 2 funktioniert nicht.
- ☐ Es ist keine logisch eindeutige Schlussfolgerung über das Funktionieren von Schaltung 2 ableitbar.

Markieren Sie jetzt bitte das Kästchen neben der korrekten Schlussfolgerung für dieses Beispiel und klicken Sie danach auf den „Weiter“-Button.

[Wenn Item *nicht* korrekt gelöst, dann:]
Diese Lösung ist leider nicht richtig. Bitte versuchen Sie es noch einmal, indem Sie das Kästchen neben einer der anderen Schlussfolgerungen markieren.

[Wenn Item korrekt gelöst (Schaltung 2 funktioniert.), dann:]
Richtig, denn wenn Schaltung 1 funktioniert, muss entsprechend dieser Regel auch Schaltung 2 funktionieren.

In den nun folgenden 15 Testaufgaben steht Ihnen auch die Antwortoption „*Ich weiß es nicht.*“ zur Verfügung. Wählen Sie diese Option, wenn Sie sich entweder nicht zwischen den anderen drei Antwortoptionen entscheiden können oder wenn Sie die Lösung der jeweiligen Testaufgabe aufgeben möchten.

Im Gegensatz dazu stellt die Antwortoption „*Es ist keine logisch eindeutige Schlussfolgerung über das Funktionieren der Schaltung ableitbar.*“ eine echte Antwortoption dar, die für einige Aufgaben richtig sein kann.

Anhang A.2: Items des KKS

Es folgt ein Überblick über die 16 Items des KKS. Tabelle Anhang A.2 gibt die Position in der ersten Version an, die Position in der überarbeiteten Version, das Item selbst, die jeweils zugehörigen Antwortalternativen (siehe Legende) sowie eine kurze Beschreibung des Items, um die Einordnung in das Negationsparadigma (entsprechend der Nomenklatur in Abschnitt 2.3.4) zu erleichtern.

Legende

Antwortalternativen - Set I:

Schaltung 1 funktioniert.

Schaltung 1 funktioniert **nicht**. *

Es ist keine logisch eindeutige Schlussfolgerung über das Funktionieren von Schaltung 2 ableitbar.

Ich weiß es nicht. **

Antwortalternativen - Set II:

Schaltung 2 funktioniert.

Schaltung 2 funktioniert **nicht**. *

Es ist keine logisch eindeutige Schlussfolgerung über das Funktionieren von Schaltung 2 ableitbar.

Ich weiß es nicht. **

* … Hervorhebung auch im Test

** … Diese Antwortoption wird nur in der überarbeiteten Version vorgegeben.

Itembeschreibung:

MP … Schlussfigur: Modus Ponens

NA … Schlussfigur: Negation des Antezedens

BK … Schlussfigur: Bestätigung der Konsequenz

MT … Schlussfigur: Modus Tollens

1 … Hauptprämisse ohne Negation

2 … Hauptprämisse mit Negation in der Konsequenz

3 … Hauptprämisse mit Negation im Antezedens

4 … Hauptprämisse mit Negation sowohl im Antezedens als auch in der Konsequenz

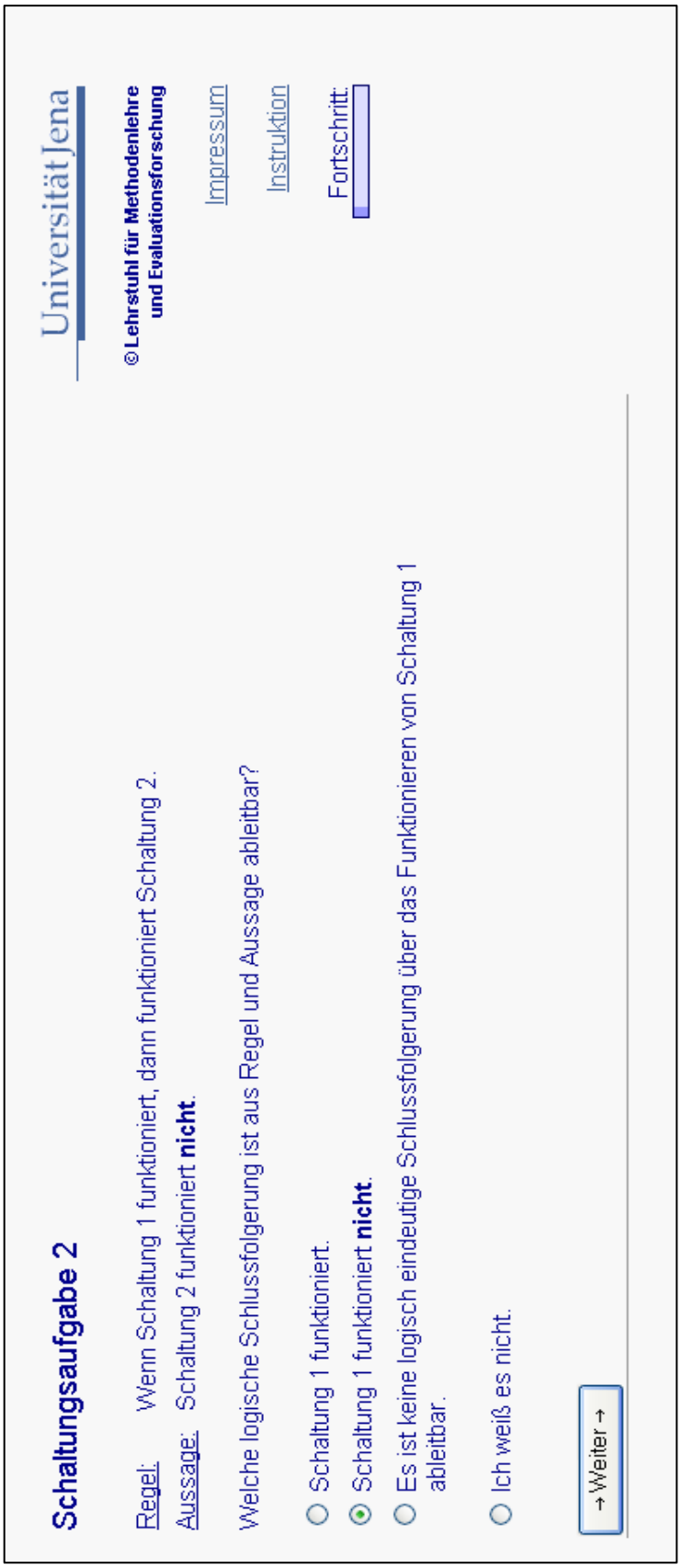
Tabelle Anhang A.2: Überblick über die 16 Items des KKS

Position im KKS (erste Version)	Position im KKS (überarbeitete Version)	Item	Set mit Antwortalternativen	Itembeschreibung (Nomenklatur s. Abschnitt 2.3.4)
Probe-Item (Instruktion)	Probe-Item (Instruktion)	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 1 funktioniert.	II	MP1
2	8	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 1 funktioniert **nicht**.	II	NA1
3	6	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 2 funktioniert.	I	BK1
4	3	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 2 funktioniert **nicht**.	I	MT1
5	13	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 1 funktioniert.	II	MP2
6	2	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 1 funktioniert **nicht**.	II	NA2
7	14	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 2 funktioniert **nicht**.	I	BK2
8	10	Regel: Wenn Schaltung 1 funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 2 funktioniert.	I	MT2

Tabelle Anhang A.2: Überblick über die 16 Items des KKS (Fortsetzung)

Position im KKS (erste Version)	Position im KKS (überarbeitete Version)	Item	Set mit Antwortalternativen	Itembeschreibung (Nomenklatur s. Abschnitt 2.3.4)
9	5	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 1 funktioniert **nicht**.	II	MP3
10	15	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 1 funktioniert.	II	NA3
11	4	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 2 funktioniert.	I	BK3
12	7	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2. Aussage: Schaltung 2 funktioniert **nicht**.	I	MT3
13	9	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 1 funktioniert **nicht**.	II	MP4
14	12	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 1 funktioniert.	II	NA4
15	11	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 2 funktioniert **nicht**.	I	BK4
16	16	Regel: Wenn Schaltung 1 **nicht** funktioniert, dann funktioniert Schaltung 2 **nicht**. Aussage: Schaltung 2 funktioniert.	I	MT4

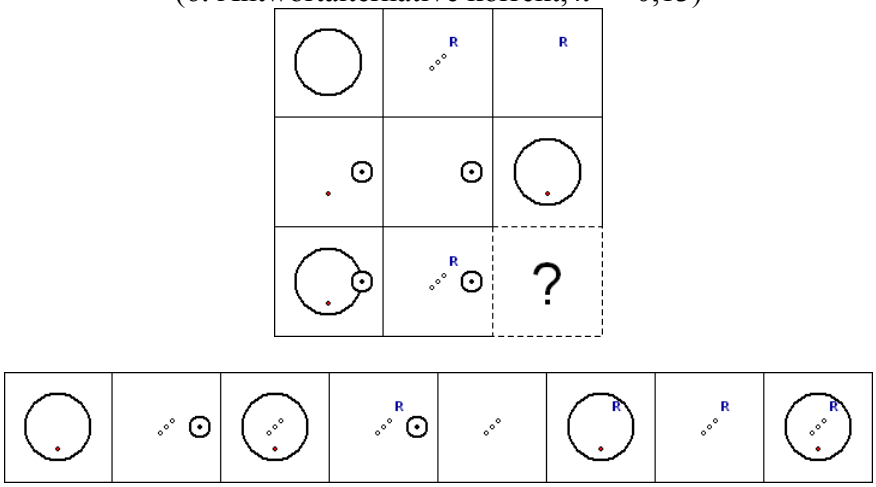
Anhang A.3: Beispiel-Screenshot eines Items des KKS (hier: MT1)



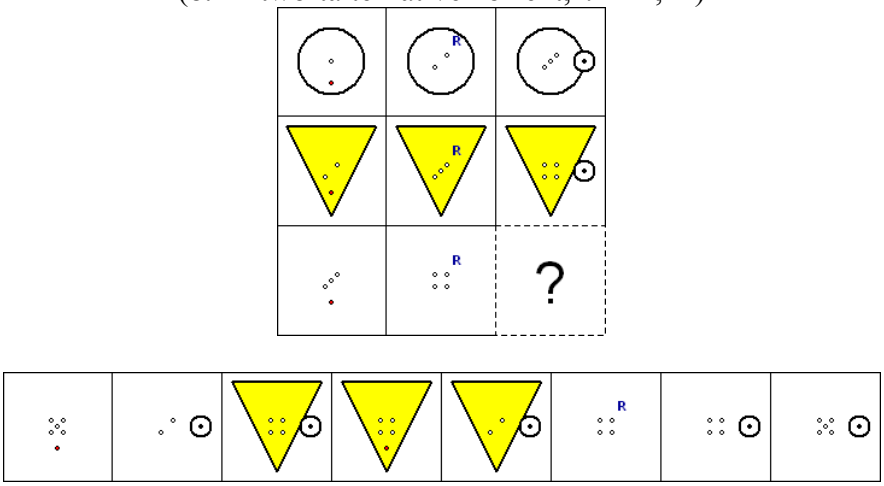
Anhang B: Matrizenitems aus der ersten empirischen Erprobung des KKS

Es folgt eine Darstellung der 10 Matrizenitems, die bei der ersten empirischen Erprobung des KKS zur Überprüfung von *Arbeitshypothese 2* (siehe Abschnitte 2.3.2.2 und 2.4.1.5) des erweiterten Stufen-Modells eingesetzt wurden. Die 10 Items wurden mit dem Programm ITEMGENERATOR (siehe Abschnitt 2.4.1.2) konstruiert. In Klammern sind jeweils die korrekte Antwortalternative und die Itemschwierigkeit κ angegeben. (Zur Schätzung der Itemschwierigkeiten κ wurde in der in Abschnitt 2.4.1.4 beschriebenen Stichprobe die Skala des Personenparameters ξ auf einen Mittelwert von $M_\xi = 0$ und eine Streuung von $Sd_\xi = 1$ fixiert.)

Matrizenitem 1
(6. Antwortalternative korrekt, κ = -0,13)

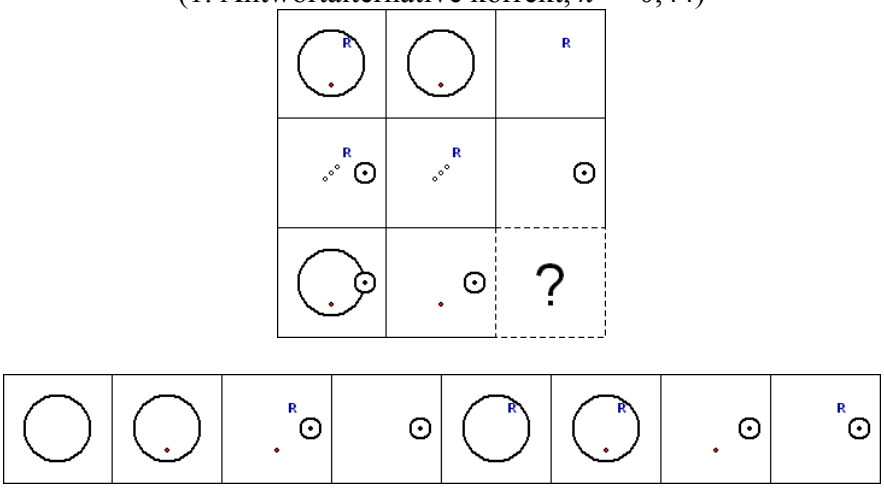


Matrizenitem 2
(8. Antwortalternative korrekt, κ = -1,24)



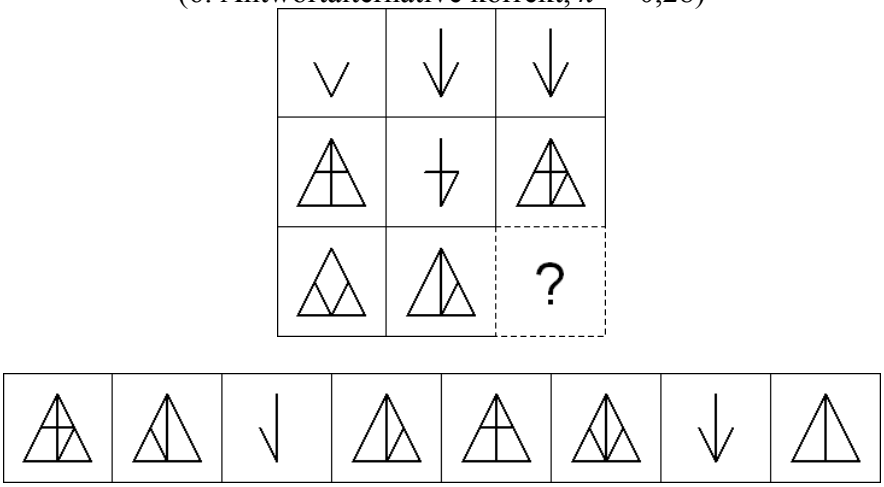
Matrizenitem 3

(1. Antwortalternative korrekt, κ = -0,44)



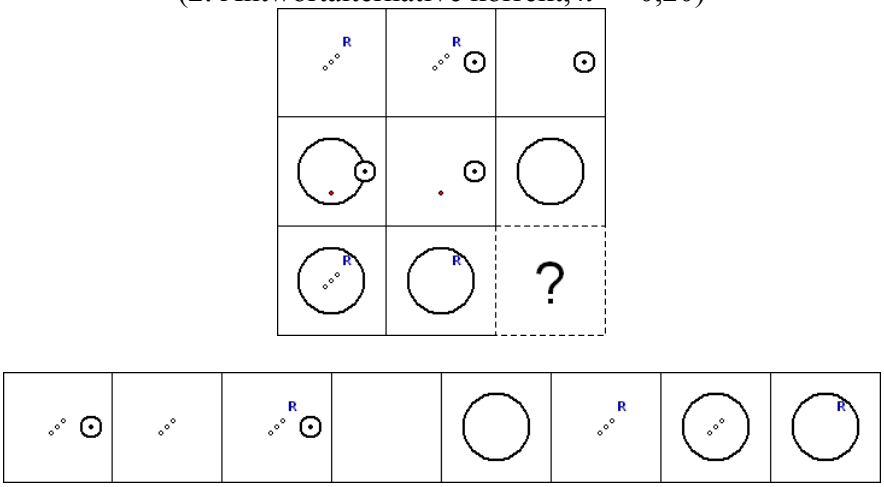
Matrizenitem 4

(6. Antwortalternative korrekt, κ = -0,28)



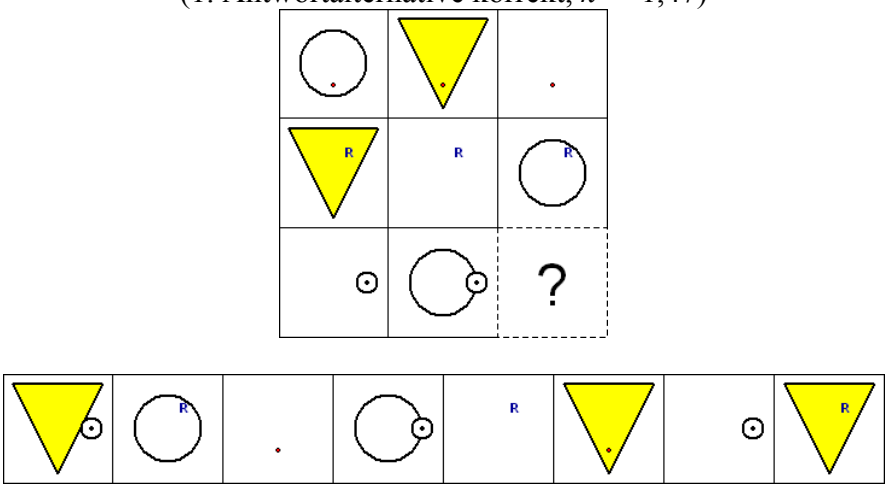
Matrizenitem 5

(2. Antwortalternative korrekt, κ = -0,20)



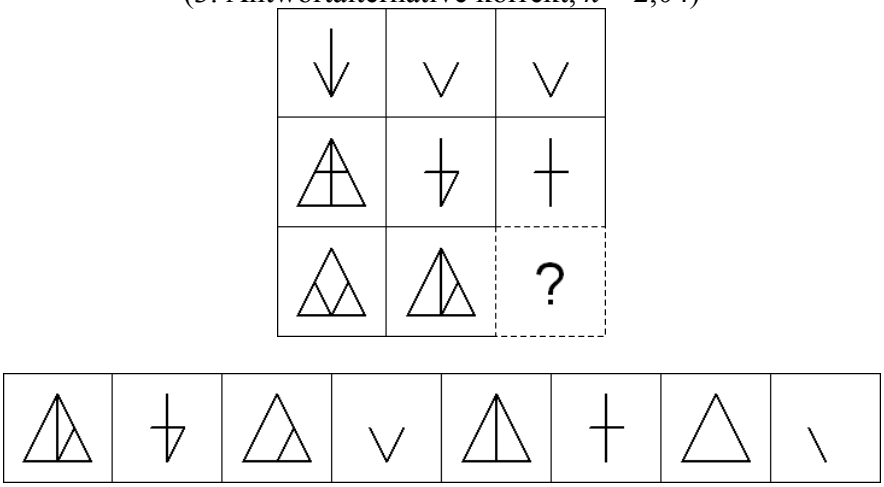
Matrizenitem 6

(1. Antwortalternative korrekt, κ = -1,47)



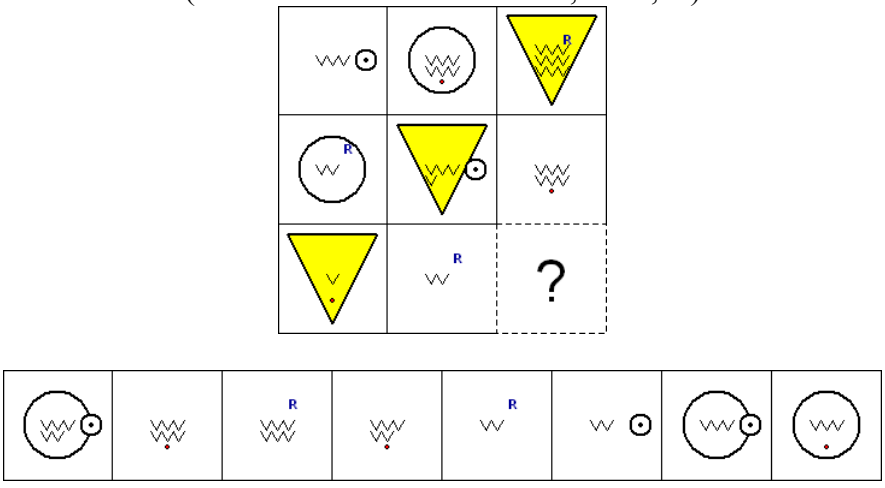
Matrizenitem 7

(3. Antwortalternative korrekt, κ = 2,04)



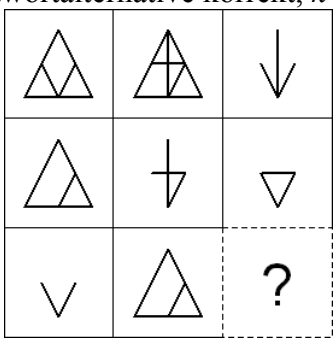
Matrizenitem 8

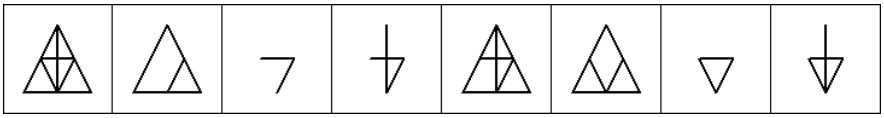
(7. Antwortalternative korrekt, κ = 0,02)



Matrizenitem 9

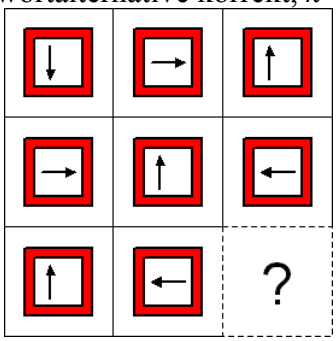
(4. Antwortalternative korrekt, $\kappa = 3{,}22$)

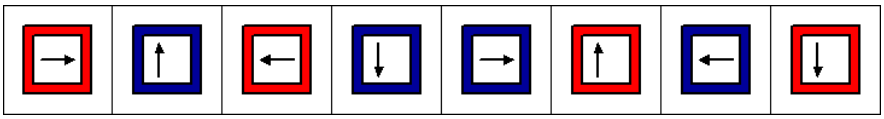




Matrizenitem 10

(8. Antwortalternative korrekt, $\kappa = -1{,}52$)





Anhang C: Berechnungsformel für Nagelkerkes R-Quadrat

$$R^2_{Nagelkerke} = \frac{1 - \left(\frac{LogL_0}{LogL_1}\right)^{\frac{2}{n}}}{1 - LogL_0^{\frac{2}{n}}}$$

mit:

$LogL_0$ … Log-Likelihood des restringierten Modells

$LogL_1$ … Log-Likelihood des unrestringierten Modells

n … Stichprobenumfang

Anhang D: M*plus*-Inputfiles zu den multinomialen logistischen Regressionen

Anhang D.1: SKS auf Reasoning (vgl. Abschnitt 2.4.1.5)

```
title:      Multinomiale Logistische Regression SKS auf Reasoning
            (10 Matrizenitems)

data:       file is datenbasis_studie1.dat;

variable:   names are mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2 bk3 bk4
            mt1 mt2 mt3 mt4 matr1 matr2 matr3 matr4 matr5 matr6 matr7
            matr8 matr9 matr10;

            usevariables mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2 bk3
            bk4 mt1 mt2 mt3 mt4 matr1 matr2 matr3 matr4 matr5 matr6
            matr7 matr8 matr9 matr10;

            categorical are mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2
            bk3 bk4 mt1 mt2 mt3 mt4 matr1 matr2 matr3 matr4 matr5
            matr6 matr7 matr8 matr9 matr10;

            CLASSES = c (5);

analysis:   type = mixture;
            algorithm = integration;

model:      %OVERALL%

            !Messmodell der 10 Matrizenitems
            reason BY matr1@1 matr2@1 matr3@1 matr4@1 matr5@1 matr6@1
            matr7@1 matr8@1 matr9@1 matr10@1;

            !Modell mit frei geschaetzten Anstiegskoeffizienten
            c ON reason;

            !Modell mit auf null fixierten Anstiegskoeffizienten
            !c ON reason@0;

            %c#1%
            [mp1$1-mt4$1];

            %c#2%
            [mp1$1-mt4$1];

            %c#3%
            [mp1$1-mt4$1];

            %c#4%
            [mp1$1-mt4$1];

            %c#5%
            [mp1$1-mt4$1];

output:     tech1 tech8;
```

Anhang D.2: SKS im Längsschnitt (Stabilitätsstudie; vgl. Abschnitt 3.2.2.2)

```
title:     Daten Stabilitaetsstudie - SKS zu t1 und t2;

data:      file is datenbasis_studie2.dat;

variable:  names are mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2 bk3 bk4
           mt1 mt2 mt3 mt4 t2_mp1 t2_mp2 t2_mp3 t2_mp4 t2_na1 t2_na2
           t2_na3 t2_na4 t2_bk1 t2_bk2 t2_bk3 t2_bk4 t2_mt1 t2_mt2
           t2_mt3 t2_mt4;

           usevariables mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2 bk3
           bk4 mt1 mt2 mt3 mt4 t2_mp1 t2_mp2 t2_mp3 t2_mp4 t2_na1
           t2_na2 t2_na3 t2_na4 t2_bk1 t2_bk2 t2_bk3 t2_bk4 t2_mt1
           t2_mt2 t2_mt3 t2_mt4;

           categorical mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2 bk3
           bk4 mt1 mt2 mt3 mt4 t2_mp1 t2_mp2 t2_mp3 t2_mp4 t2_na1
           t2_na2 t2_na3 t2_na4 t2_bk1 t2_bk2 t2_bk3 t2_bk4 t2_mt1
           t2_mt2 t2_mt3 t2_mt4;

           CLASSES = c1 (4) c2 (4);

analysis:  type = mixture;

model:     %OVERALL%

model c1:  %c1#1%
           [mp1$1-mt4$1*1] (1-16);

           %c1#2%
           [mp1$1-mt4$1*1] (17-32);

           %c1#3%
           [mp1$1-mt4$1*1] (33-48);

           %c1#4%
           [mp1$1-mt4$1*1] (49-64);

model c2:  %c2#1%
           [t2_mp1$1-t2_mt4$1*1] (1-16);

           %c2#2%
           [t2_mp1$1-t2_mt4$1*1] (17-32);

           %c2#3%
           [t2_mp1$1-t2_mt4$1*1] (33-48);

           %c2#4%
           [t2_mp1$1-t2_mt4$1*1] (49-64);

output:    tech1 tech8;
```

Anhang D.3: SKS auf Wiener Matrizentest (vgl. Abschnitt 3.3.2.2)

```
title:     Multinomiale Logistische Regression SKS auf WMT

data:      file is konstruktvalid_wmt.dat;

variable:  names are wmt1 wmt2 wmt3 wmt4 wmt5 wmt6 wmt7 wmt8 wmt9
           wmt10 wmt11 wmt12 wmt13 wmt14 wmt15 wmt16 wmt17 wmt18
           wmt19 wmt20 wmt21 wmt22 wmt23 wmt24 mp1 mp2 mp3 mp4 na1
           na2 na3 na4 bk1 bk2 bk3 bk4 mt1 mt2 mt3 mt4;

           usevariables wmt1 wmt2 wmt3 wmt4 wmt5 wmt6 wmt7 wmt8 wmt9
           wmt10 wmt11 wmt12 wmt13 wmt14 wmt15 wmt16 wmt17 wmt18
           wmt19 wmt20 wmt21 wmt22 wmt23 wmt24 mp1 mp2 mp3 mp4 na1
           na2 na3 na4 bk1 bk2 bk3 bk4 mt1 mt2 mt3 mt4;

           categorical are wmt1 wmt2 wmt3 wmt4 wmt5 wmt6 wmt7 wmt8
           wmt9 wmt10 wmt11 wmt12 wmt13 wmt14 wmt15 wmt16 wmt17 wmt18
           wmt19 wmt20 wmt21 wmt22 wmt23 wmt24 mp1 mp2 mp3 mp4 na1
           na2 na3 na4 bk1 bk2 bk3 bk4 mt1 mt2 mt3 mt4;

           CLASSES = c (4);

analysis:  type = mixture;
           algorithm = integration;

model:     %OVERALL%

           !Rasch-Modell als Messmodell fuer den WMT
           WMT BY wmt1@1 wmt2@1 wmt3@1 wmt4@1 wmt5@1 wmt6@1 wmt7@1
           wmt8@1 wmt9@1 wmt10@1 wmt11@1 wmt12@1 wmt13@1 wmt14@1
           wmt15@1 wmt16@1 wmt17@1 wmt18@1 wmt19@1 wmt20@1 wmt21@1
           wmt22@1 wmt23@1 wmt24@1;

           !Modell mit frei geschaetzten Anstiegskoeffizienten
           c ON WMT;

           !Modell mit auf null fixierten Anstiegskoeffizienten
           !c ON WMT@0;

           %c#1%
           [mp1$1-mt4$1];

           %c#2%
           [mp1$1-mt4$1];

           %c#3%
           [mp1$1-mt4$1];

           %c#4%
           [mp1$1-mt4$1];

output:    tech1 tech8;
```

Anhang D.4: SKS auf die vier Operationsklassen des BIS (K, M, B, E; vgl. Abschnitt 3.3.2.2)

```
title:     Multinomiale Logistische Regression SKS auf BIS-K,M,B,E

data:      file is konstruktvalid_bis.dat;

variable:  names are bis_m_og bis_e_ef bis_b_xg bis_e_lo bis_m_zp
           bis_b_bd bis_m_st bis_e_dr bis_b_tg bis_k_an bis_k_ch
           bis_k_sc bis_k_tm bis_k_wa bis_k_zn mp1 mp2 mp3 mp4 na1
           na2 na3 na4 bk1 bk2 bk3 bk4 mt1 mt2 mt3 mt4;

           usevariables
           bis_k_an bis_k_ch bis_k_sc bis_k_tm bis_k_wa bis_k_zn
           !bis_m_og bis_m_zp bis_m_st
           !bis_b_xg bis_b_tg bis_b_bd
           !bis_e_ef bis_e_lo bis_e_dr
           mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2 bk3 bk4 mt1 mt2
           mt3 mt4;

           categorical are mp1 mp2 mp3 mp4 na1 na2 na3 na4 bk1 bk2
           bk3 bk4 mt1 mt2 mt3 mt4;

           CLASSES = c (4);

analysis:  type = mixture;
           algorithm = integration;

model:     %OVERALL%

           !Messmodelle fuer die BIS-Skalen
           K BY bis_k_an bis_k_ch bis_k_sc bis_k_tm bis_k_wa
           bis_k_zn;
           bis_k_zn WITH bis_k_sc;
           !M BY bis_m_og@1 bis_m_zp@1 bis_m_st@1;
           !B BY bis_b_xg@1 bis_b_bd@1 bis_b_tg;
           !E BY bis_e_ef@1 bis_e_lo@1 bis_e_dr@1;

           !Modelle mit frei geschaetzten Anstiegskoeffizienten
           c ON K; !c ON B; !c ON M; !c ON E;

           !Modelle mit auf null fixierten Anstiegskoeffizienten
           !c ON K@0; !c ON B@0; !c ON M@0; !c ON E@0;

           %c#1%
           [mp1$1-mt4$1];

           %c#2%
           [mp1$1-mt4$1];

           %c#3%
           [mp1$1-mt4$1];

           %c#4%
           [mp1$1-mt4$1];

output:    tech1 tech8;
```

Ehrenwörtliche Erklärung

Die Promotionsordnung der Fakultät für Sozial- und Verhaltenswissenschaften der Friedrich-Schiller-Universität in der geltenden Fassung ist mir bekannt.

Ich habe diese Dissertation selbst angefertigt und dabei insbesondere die Hilfe eines Promotionsberaters nicht in Anspruch genommen. Alle von mir benutzten Quellen und Hilfsmittel habe ich kenntlich gemacht und an den entsprechenden Stellen angegeben.

Erik Sengewald hat mich im Rahmen seiner Tätigkeit als studentische Hilfskraft am Lehrstuhl für Methodenlehre und Evaluationsforschung entgeltlich unterstützt, und zwar in Form von Programmierarbeiten zur Umsetzung des KKS als Online-Test. Anna Grohmann, Anna Wollny und Juliane Graf haben mich im Rahmen ihrer Tätigkeit als studentische Hilfskräfte am Lehrstuhl für Methodenlehre und Evaluationsforschung entgeltlich bei der Vorbereitung und Durchführung der empirischen Studien unterstützt. Christiane Fiege, Steffi Pohl, Norman Rose und Tim Loßnitzer haben unentgeltlich Vorabversionen einzelner Teile des Manuskripts gelesen und mich auf Fehler und Inkonsistenzen aufmerksam gemacht. Dawn Girlich hat mich entgeltlich bei einzelnen englischen Formulierungen im Abstract unterstützt. Weitere entgeltliche oder unentgeltliche Unterstützung bei der Auswahl und Auswertung des Materials und bei der Herstellung des Manuskripts habe ich nicht erhalten.

Darüber hinaus haben Dritte von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Ich habe diese Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

Ich habe weder die gleiche noch eine in wesentlichen Teilen ähnliche noch eine andere Arbeit bei einer anderen Hochschule oder Fakultät als Dissertation eingereicht.

Ich versichere, dass die oben gemachten Angaben nach meinem besten Wissen der Wahrheit entsprechen und ich nichts verschwiegen habe.