Fakultät für Psychologie und Sportwissenschaft
der Universität Bielefeld

Validität und faktorielle Invarianz einer neuropsychologischen Testbatterie zur Intelligenzprüfung bei Patienten mit Epilepsie

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Martin T. Lutz, Dipl.-Psych.
geboren am 18. November 1973
wohnhaft Eschenstraße 3, 01097 Dresden

Dresden, im Juni 2006

Gutachter

Betreuer und Erstgutachter: Prof. Dr. Wolfgang Hartje

Zweitgutachter: Prof. Dr. Christoph Helmstaedter

Dank

An dieser Stelle möchte ich allen danken, die mich bei der Erstellung dieser Arbeit begleitet haben.

Mein herzlichster Dank geht an meine Eltern Renate und Walter Lutz. Sie förderten mich während meines Studiums und haben damit diese Arbeit erst ermöglicht. In der Zeit der Anfertigung der Promotion haben sie mich verständnisvoll begleitet und aufgemuntert.

Herrn Professor Hartje möchte ich für die warme und stets vertrauensvolle Unterstützung und den notwendigen Freiraum beim Erstellen der Arbeit danken. Ihm als engagierten Lehrer verdanke ich mein nachhaltiges Interesse an der klinischen Neuropsychologie und die vielseitige neuropsychologische Ausbildung, die er mir während meines Studiums und während der Promotionsphase mit auf den Weg gegeben hat. Täglich profitiere ich davon.

Mein besonderer Dank gilt ebenso Herrn Professor Helmstaedter. Danken möchte ich ihm für die freundliche Überlassung des Themas und die unkomplizierte Betreuung. Er stand mir bei dieser Arbeit ebenso wie bei meinen anderen Bonner Projekten jederzeit mit Rat und Tat zur Seite. Seine herausfordernde, kreative und immer unterstützende Anleitung hat mein wissenschaftliches Arbeiten ebenso wie meine klinische Tätigkeit geprägt.

Auch den Bonner Kolleginnen und Kollegen, die maßgeblich an der Datenerhebung mitbeteiligt waren, möchte ich für ihre Unterstützung danken. Der gesamten Arbeitsgruppe danke ich herzlich für die freundschaftliche und hilfsbereite Arbeitsatmosphäre.

Ich danke meiner Freundin Henrike Marlow für viele kleine und große Hilfen und den umfangreichen und unermüdlichen Einsatz bei der Durchsicht des Manuskripts.

Mein Dank gilt natürlich auch den Probandinnen und Probanden, die an den Untersuchungen teilgenommen haben.

Inhaltsverzeichnis

Zusammenfassung

Die neuropsychologische Diagnostik gehört neben der neuropsychologischen Therapie zu den wesentlichen Aufgabenfeldern der klinischen Neuropsychologie. Das Ziel der neuropsychologischen Diagnostik ist die Erfassung und Objektivierung von kognitiven Funktionsstörungen nach einer Hirnschädigung oder Hirnfunktionsstörung. Speziell in der neuropsychologischen Diagnostik bei Patienten mit Epilepsie sind Fragestellungen zur Lateralisation und Lokalisation der epileptischen Funktionsstörung und zur Qualitätskontrolle bei konservativen und operativen Epilepsiebehandlungen zu beantworten. Im Rahmen der psychometrischen Testdiagnostik werden den Patienten verschiedene Testaufgaben, häufig zusammengestellt zu festen Testbatterien, vorgegeben. Testbatterien decken ein breites Spektrum kognitiver Teilleistungen ab und ermöglichen die Errechnung breiter Fähigkeitsfaktoren. Zur Auswertung und Ergebnisinterpretation werden die Leistungen der Patienten auf Einzeltest- oder Faktorenebene zumeist mit einer hirngesunden Normierungsstichprobe verglichen. Dieser Vergleich einer Patientenstichprobe mit einer Normierungsstichprobe ist allerdings nur bei Erfüllung bestimmter psychometrischer Voraussetzungen zulässig, da nur dann eine unverzerrte normative Auswertung der Ergebnisse und eine stichprobenunabhängige Interpretation der Faktoren gewährleistet ist.

Die Überprüfung dieser Voraussetzungen ist Inhalt dieser Arbeit und erfolgt in drei Schritten: Zunächst wird anhand der Normierungsstichprobe untersucht, welches Intelligenzstrukturmodell sich am besten zur Beschreibung der faktoriellen Struktur der Testbatterie eignet. In einem zweiten Schritt wird überprüft, ob sich diese faktorielle Struktur auch in einer Stichprobe von Patienten mit Epilepsie wiederfindet. Schließlich werden im dritten Schritt zunehmend stringente Hypothesen bezüglich der Invarianz verschiedener Parameter des aufgestellten Modells getestet. Die Ergebnisse dieser Invarianztests geben Auskunft darüber, inwieweit die psychometrischen Eigenschaften der Faktoren und der zugehörigen Tests stichprobenunabhängig sind. Nur wenn hinreichende Invarianz bezüglich der Stichprobe nachgewiesen ist, kann zur Ergebnisinterpretation eine stichprobenunabhängige Validität der Faktoren angenommen werden und die Auswertungskriterien aus der Normstichprobe auf die klinische Stichprobe übertragen werden.

Die Basis für die empirische Untersuchung bilden 190 im Rahmen einer Normierungsstudie untersuchte Probanden und 190 Patienten mit Epilepsie. Beide Stichproben wurden mit der neuropsychologischen Testbatterie der Klinik für Epileptologie in Bonn getestet. Die vergleichende Untersuchung der Eignung der

Intelligenzstrukturmodelle sowie die Überprüfung der Invarianzhypothesen erfolgten mittels konfirmatorischer Faktorenanalysen.

Es kann gezeigt werden, dass sowohl die Daten der Normierungsstichprobe als auch die Daten der Patientenstichprobe am besten durch das Intelligenzstrukturmodell der *Cattell-Horn-Carroll Theory of Cognitive Abilities* beschrieben werden können. Dabei lassen sich folgende fünf Faktoren abbilden: kristalline Intelligenz (Gc), visuo-räumliche Fähigkeiten (Gv), Kurzzeitgedächtnis (Gsm), Langzeitgedächtnis und Abruf (Glr) sowie kognitive Verarbeitungsgeschwindigkeit (Gs). Durch Testung der Invarianzhypothesen konnte die Annahme *konfiguraler* und *schwacher metrischer* Invarianz bestätigt werden. Das bedeutet, dass für beide Gruppen die gleiche Anzahl an Faktoren vorliegt und auch die Faktorladungen in den Gruppen numerisch gleich sind. Die Annahme *starker metrischer* Invarianz, die zusätzlich in den Gruppen gleiche Itemkonstanten (Höhenlagen) erfordert, kann nicht bestätigt werden. Die nachgewiesene Invarianz der Faktorenkovarianzen verweist auf gleiche Beziehungen zwischen den latenten Faktoren.

Die Ergebnisse dieser Untersuchung haben verschiedene Implikationen für die neuropsychologische Testpraxis und für die klinische Forschung: Zunächst konnte gezeigt werden, dass die Testbatterie in beiden Gruppen die gleichen Leistungsdimensionen, bzw. die gleichen Konstrukte erfasst. Der dabei erbrachte Nachweis der Vereinbarkeit der Daten mit dem Intelligenzmodell der *Cattell-Horn-Carroll Theory of Cognitive Abilities* ermöglicht es, die in dieser Theorie beschriebenen Angaben zur Konstruktvalidität der Faktoren auf die Faktoren der neuropsychologischen Testbatterie zu übertragen und daraus verschiedene Untersuchungshypothesen oder klinische Schlussfolgerungen abzuleiten. Darüber hinaus ermöglicht der Nachweis schwacher metrischer Invarianz, gruppenspezifische Unterschiede in Kriterienkorrelationen zu untersuchen. Der Bezug auf das Cattell-Horn-Carroll-Modell kann die Kommunikation von Ergebnissen zwischen den verschiedenen Forschungszentren erleichtern und für die Weiterentwicklung der neuropsychologischen Testbatterie wichtige Impulse liefern. Weil die Hypothese der starken metrischen Invarianz nicht bestätigt werden konnte, ist ein unverzerrter Vergleich der Skalenmittelwerte zwischen den Gruppen nicht möglich. Dementsprechend ist es nicht statthaft, die Daten der Patientenstichprobe anhand der Normierungsstichprobe zu standardisieren. Dies ergibt sich daraus, dass ohne den erbrachten Nachweis starker metrischer Invarianz nicht davon ausgegangen werden kann, dass Probanden aus beiden Stichproben bei gleichen latenten Fähigkeiten auch die gleichen Testwerte auf Faktorenebene erhalten würden.

1 Einleitung

Im Zentrum der vorliegenden Arbeit steht die Frage nach der Invarianz. Diese Frage wird insbesondere dann wichtig, wenn eine aus mehreren Einzeltests oder Items bestehende Skala Probanden unterschiedlicher Populationen vorgegeben wird. Neben einzelfalldiagnostischen Fragestellungen könnte beispielsweise untersucht werden, ob sich die Skalenmittelwerte zwischen den Gruppen unterscheiden oder ob es gruppenspezifische Unterschiede in den Kriterienkorrelationen der Skala gibt. Für solche Fragestellungen ist es eine notwendige Voraussetzung, dass die Skala unabhängig der Gruppenzugehörigkeit für alle Probanden das gleiche Konstrukt in *gleicher* Weise erfasst und *gleiche* psychometrische Eigenschaften hat. Ist dies gegeben, ermöglicht die dann vorliegende *Messäquivalenz* oder *Invarianz* der Skala eine sinnvolle Interpretation der Ergebnisse (MacCallum, 2003).

Einleitend soll exemplarisch eine Untersuchung zur Invarianz des amerikanischen Wechsler-Gedächtnistests (Wechsler Memory Scale, WMS-III, Wechsler, 1997b) dargestellt werden (Wilde et al., 2003): Die Gedächtnistestbatterie WMS-III besteht aus zehn Untertests. Aus diesen können anhand vorgegebener Berechnungsvorschriften fünf zusammenfassende Indexwerte für globale Gedächtniskonstrukte ermittelt werden (sofortiges und verzögertes auditives Gedächtnis, sofortiges und verzögertes visuelles Gedächtnis, Arbeitsgedächtnis). Für die klinische Arbeit eröffnet dies die Möglichkeit der Profilinterpretation: So könnte beispielsweise eine Dissoziation zwischen dem auditiv-verbalen und dem visuellen Gedächtnisindex als Hinweis auf eine lateralisierte, temporale Dysfunktion betrachtet werden (siehe z. B. Helmstaedter & Kurthen, 2001). Wilde et al. (2003) haben die faktorielle Struktur der Wechsler-Gedächtnisskala in einer Stichprobe von Patienten mit überwiegend lateralisierten Temporallappenepilepsien mittels konfirmatorischer Faktorenanalysen untersucht und konnten zeigen, dass die Daten am besten und sparsamsten durch ein Modell aus nur zwei Faktoren (Arbeitsgedächtnis und allgemeines Gedächtnis) beschrieben werden konnten. Dieses Ergebnis widerspricht nicht nur der von den Testautoren vorgegebenen Faktorstruktur und den davon abgeleiteten Indexwerten, sondern auch der Erwartung, in dieser Stichprobe separate Faktoren für das verbale und visuelle Gedächtnis zu erhalten. Dazu schreiben Wilde et al. (2003):

> *At least in temporal lobe epilepsy, the findings of this study suggest caution in interpretation of the WMS-III auditory and visual index scores as representing separable dimensions. Similarly, any clinical interpretations of differences between the scores must remain tenuous – it is difficult to interpret patterns of scores that are*

> *unrelated to the test's factor structure, particularly when they do not discriminate between contrasting groups.* (S. 62)

Derartige Interpretationsprobleme ergeben sich aus der üblichen Vorgehensweise der Entwicklung und Anwendung faktorieller Testbatterien: Anhand einer hirngesunden Normstichprobe wird zunächst durch zumeist explorative Faktorenanalysen die faktorielle Struktur der Testbatterie untersucht. In einem weiteren Schritt werden Berechnungsvorschriften aufgestellt, mit denen Indexwerte ermittelt werden können, die die Faktoren repräsentieren sollen. Anhand dieser Indexwerte können dann im Rahmen der Testanwendung klinische Interpretationen zur Differenzierung klinischer Symptome vorgenommen werden. Problematisch an diesem Vorgehen ist, dass dabei Invarianz der latenten (Ko-)Variabilität der Untertests zwischen Normgruppe und Patientengruppe impliziert wird, ohne dass eine explizite Überprüfung dieser Annahme durchgeführt wurde. Ohne solch eine Überprüfung ist jedoch die Konstruktvalidität der Indexwerte bezüglich der Patientenpopulation als unbekannt anzusehen (Burton, Ryan, Axelrod & Schellenberger, 2002). Eine explizite Untersuchung der Äquivalenz der Konstruktvalidität von Indexwerten in verschiedenen Stichproben ermöglicht die Testung spezieller *Invarianzhypothesen*. Wird dabei der Nachweis vollständiger Invarianz erbracht, kann davon ausgegangen werden, dass die Indexwerte in den fraglichen Stichproben äquivalent bezüglich der faktoriellen Validität und der Messeigenschaften sind.

In einer Übersichtsarbeit zum Thema Invarianz schreiben Vandenberg und Lance (2000): „*If not tested, violations of measurement equivalence assumptions are as threatening to substantive interpretations as is an inability to demonstrate reliability and validity*" (S. 6). Trotz der Wichtigkeit der Invarianz liegen insgesamt nur wenige explizite Untersuchungen dazu vor. Insbesondere bezüglich kognitiver Fähigkeiten sind derartige Untersuchungen sehr selten (Bowden, Cook, Bardenhagen, Shores & Carstairs, 2004), etwas häufiger finden sich Untersuchungen zur Invarianz von Fragebögen. Eine Ursache für diesen empirischen Mangel liegt darin, dass entsprechende Invarianzhypothesen nur mit relativ großen Stichproben getestet werden können. Allerdings sind – von den Standardisierungsstichproben abgesehen – große Stichproben gesunder Probanden selten. Ein weiterer Grund ist, dass die Methodik zur Überprüfung der Invarianz recht komplex und noch relativ unbekannt ist. Erst seit neuerer Zeit sind geeignete Programme zur Analyse von Kovarianzstrukturmodellen in nutzerfreundlichen Versionen verfügbar (Byrne, 2001).

Trotz der grundsätzlichen Testbarkeit von Invarianzhypothesen und der Fragwürdigkeit der Ergebnisinterpretationen ohne erbrachtem Nachweis von Invarianz,

werden in der klinischen Praxis ebenso wie in der Forschung bewusst oder unbewusst zumeist *implizite* Annahmen über die Messäquivalenz getroffen (Vandenberg & Lance, 2000). Implizite Annahmen können die ungeprüfte Annahme von Invarianz beinhalten, aber auch die ungeprüfte Annahme fehlender Invarianz (vgl. Bowden et al., 2004; Wilde et al., 2003): Implizite Annahmen über *vorhandene* Invarianz finden sich dabei deutlich häufiger. Hier ist die oben dargestellte übliche klinische Praxis der ungeprüften Übertragung von Indexwerten auf unterschiedlichste klinische Stichproben zu nennen. Dies gilt beispielsweise für die Wechsler-Gedächtnistests (Härting, Markowitsch, Neufeld, Calabrese & Deisinger, 2000) oder für den Intelligenztest nach Wechsler (Tewes, 1991). Bezüglich impliziter Annahmen über *fehlende* Invarianz wird teilweise ungeprüft vorausgesetzt, dass die Interkorrelationsmuster und die Faktorenstrukturen an idiosynkratische Stichprobencharakteristika gebunden sind. So könnte spekuliert werden, dass sich schon geringe Unterschiede in soziodemographischen Variablen auf das Faktorenmuster auswirken könnten und folglich wäre mit viel größerer Wahrscheinlichkeit anzunehmen, dass sich die Faktorenstrukturen zwischen Gesunden- und Patientenkollektiven sowie zwischen verschiedenen Patientengruppen untereinander unterscheiden. So wären beispielsweise Art und Anzahl der Teilkomponenten des Gedächtnisses, die mit einer Testbatterie differenzierbar sind, als Funktion der zugrunde liegenden Hirnfunktionsstörungen anzusehen (Millis et al., 1999, zitiert nach Wilde et al., 2003).

Aus der Erkenntnis heraus, dass bestimmte Schädigungen des Gehirns die Struktur der kognitiven Fähigkeiten beeinflussen, wurden spezielle „neuropsychologische“ Testbatterien entwickelt, mit denen die kognitiven Auswirkungen neuropsychologischer Störungen erfassbar gemacht werden sollen. Hierauf beruhen die so genannten exekutiven Tests, die zwar für die klinische Neuropsychologie ein wichtiger Baustein der Diagnostik sind, aber keinen Eingang in typische Intelligenztestbatterien gefunden haben (Ardila, 1999). Die exekutive Hypothese beinhaltet somit die zumeist ungetestete Annahme über fehlende Invarianz. Wilde et al. (2003) stellen in diesem Zusammenhang fest, dass implizite Annahmen häufig die Basis zentraler kognitiver Untersuchungsparadigmen darstellen.

Ist von Invarianz die Rede, sind grundsätzlich weitere Präzisierungen nötig: Eine Aussage bezüglich der Invarianz einer Testbatterie oder einer Skala ist bedeutungslos, solange nicht expliziert wird, bezüglich *welches* Aspektes Invarianz vorliegt. Daher ist immer anzugeben, auf welche gruppierende Variable sich Invarianz bezieht (beispielsweise Geschlecht, Alter oder Pathologie). Invarianz im Sinne der vorliegenden

Arbeit bezieht sich auf den Vergleich zwischen Gesunden und Patienten mit Epilepsie. Zweitens ist Invarianz eine graduelle Eigenschaft auf einem Kontinuum zwischen vollständig fehlender Invarianz und strengster Invarianz. Der Grad der Invarianz wird über die Teilmenge der konstanten und somit als gruppenunabhängig konzipierten Parameter des Mess- und Strukturmodells definiert. In einem faktorenanalytischen Modell wären solche Parameter beispielsweise die Anzahl der Faktoren, das Muster aus Nullladungen und von Null verschiedenen Ladungen, die Ladungen an sich oder die Faktorinterkorrelationen (siehe z. B. Vandenberg & Lance, 2000). In direkter Abhängigkeit von den unterschiedlich strengen Invarianzannahmen stehen die Möglichkeiten und Beschränkungen der Testauswertung und Testinterpretation. Mittels der Methode der *konfirmatorischen Faktorenanalyse* kann der Grad der Generalisierbarkeit der Faktorenlösungen empirisch festgestellt werden. Diese Methode wird in Kapitel 3 dargestellt.

Das Ergebnis einer konfirmatorischen Faktorenanalyse könnte beispielsweise der Nachweis sein, das sich die Regressionsgleichungen, mit denen Faktorwerte aus beobachteten Testwerten vorhergesagt werden können, zwischen den fraglichen Gruppen unterscheiden. Was aber steckt psychologisch hinter dieser fehlenden Invarianz? Das soll anhand eines Beispiels verdeutlicht werden: Fujii, Lloyd und Miyamoto (2000) untersuchten die Leistung im Abzeichnen einer komplexen geometrischen Figur (Osterreith, 1944) in zwei Gruppen hirngesunder Probanden mit unterschiedlichem Intelligenzniveau („*average*“ vs. „*high*“). Vollkommen erwartungskonform zeigte sich dabei zunächst, dass sich die Abzeichenleistung zwischen den beiden Gruppen *nicht* unterschied. Mittels Regressionsanalysen wurde in einem nächsten Analyseschritt überprüft, durch welche Leistungen in zusätzlich durchgeführten Testverfahren die Leistung im Abzeichnen vorhergesagt werden kann. In der durchschnittlich intelligenten Gruppe hat sich dafür ein Testwert, der die Organisation bzw. Fragmentierung beim Abzeichnen erfasst, als am besten geeignet erwiesen. In der Gruppe mit überdurchschnittlicher Intelligenz hatte ein Testwert, der die visuo-konstruktive Fähigkeit erfasst, die höchste Vorhersagekraft.

Die Autoren beziehen die Interpretation der Ergebnisse auf den Zusammenhang zwischen Intelligenz und Informationsverarbeitungsgeschwindigkeit: Die intelligentere Gruppe könne aufgrund einer höheren Informationsverarbeitungsgeschwindigkeit die komplexe Figur schnell und effizient wahrnehmen und im Arbeitsgedächtnis verarbeiten. Somit sei für diese Probanden die Komplexität relativ geringer und die Abzeichenleistung in geringerem Maße von einem organisierten, planvollen Vorgehen abhängig. Die Gruppe mit durchschnittlicher Intelligenz sei dagegen aufgrund der

ineffizienteren unmittelbaren Aufnahme von Gestalt und Teilkomponenten in stärkerem Maße von einer organisierenden Strategie beim Abzeichnen abhängig. Insgesamt scheint die Intelligenz also eine wichtige Moderatorvariable für die Abzeichen-Aufgabe zu sein – auch wenn das rein numerische Testergebnis letztlich nicht unterschiedlich ist.

Die Ergebnisse werden von den Autoren der Studie in den Rahmen des „*Boston Process Approach*“ (Kaplan, 1988) gestellt, einem Ansatz, der dem Weg zur Aufgabenlösung eine hohe Bedeutung zukommen lässt. Dies ist gut mit dem Konzept der Invarianz vereinbar: Der gleiche Punktwert für die Abzeichenleistung ließe zunächst erwarten, dass in beiden Gruppen die latente Variable, für die diese Aufgabe ein Indikator ist, auch gleichermaßen ausgeprägt ist. Tatsächlich unterscheiden sich aber die Prozesse, die die Abzeichenleistung bedingen. Das bedeutet, dass es gruppenspezifische Unterschiede im Testkonstrukt gibt, und der Test somit nicht unabhängig von der Gruppenzugehörigkeit das Gleiche erfasst. Da also nicht gruppenunabhängig von der Testleistung auf die zugrunde liegende latente Leistung geschlossen werden darf, ist ohne weitere Analysen schwer interpretierbar, was Gleichheit der Testscores in beiden Gruppen bedeutet. Für einen Vergleich bezüglich einer spezifischen kognitiven Funktion oder einer aus Faktorwerten geschlossenen latenten Eigenschaft reicht es also nicht aus, in beiden Gruppen denselben Test einzusetzen. Vielmehr muss gewährleistet sein, dass Messäquivalenz vorliegt, also dass ein Test den gleichen zugrunde liegenden Faktor in den Vergleichsgruppen in gleicher Weise erfasst. Nur dann haben zwei aus unterschiedlichen Gruppen stammende Probanden bei gleichem Testergebnis auch die gleiche Ausprägung der zugehörigen latenten Eigenschaft. Dann erst kann von einer gleichen Konstruktvalidität des Tests und der zugehörigen latenten Eigenschaft ausgegangen werden.

Es sei angemerkt, dass sich Invarianz als übergeordnetes Konzept für verschiedene testdiagnostische Konzepte eignet, so z.B. für die Konzepte der Moderatorvariablen, der Testfairness oder des Testbias. In der diagnostischen Literatur werden diese zumeist isoliert dargestellt (z. B. Amelang & Zielinski, 1997).

Die vorliegende Arbeit untersucht drei zentrale Fragestellungen: Zunächst soll die Konstruktvalidität der Testverfahren, die in der neuropsychologischen Testbatterie der Epileptologie der Universitätsklinik Bonn zusammengestellt sind, untersucht werden. Für eine rationale klinische Diagnostik ist die genaue Kenntnis darüber, was der eingesetzte Test misst, also der Beziehung zwischen Konstrukt und Testwert, unabdingbar, da nur dann das Testergebnis interpretierbar ist (Daniel, 2000). Der klinische Nutzen jeder neuropsychologischen Diagnostik hängt unmittelbar von der Validität des eingesetzten

Testverfahrens in Hinblick auf die jeweilige diagnostische Fragestellung ab. Zur Klärung der Beziehung eines Tests zu einem Konstrukt stehen verschiedene Ansätze zur Verfügung (Lienert & Raatz, 1998). Grundlegend sind zunächst eine genaue Inhalts- und Aufgabenanalyse sowie die Untersuchung der Validität des Tests bezüglich externer Kriterien. Für epileptologische Fragestellungen liegen für die in der Bonner Testbatterie zusammengestellten Verfahren viele Studien vor (z. B. Gleissner, Helmstaedter & Elger, 1998; Helmstaedter, Brosch, Kurthen & Elger, 2004; Helmstaedter, Gleissner, Zentner & Elger, 1998; Müller, Hasse-Sander, Horn, Helmstaedter & Elger, 1997). Von übergeordnetem Interesse zur weiteren Klärung der Konstruktvalidität ist die Einordnung der Testkonstrukte in einen theoretischen Rahmen. Wichtig hierfür ist die Kenntnis der faktoriellen Struktur der Testbatterie. Faktorielle Strukturmodelle geben Auskunft über die Beziehung des Tests zu verschiedenen Modellfaktoren, über die Bedeutung der Faktoren und der Beziehung der Faktoren untereinander. Zur Klärung der Konstruktvalidität der mit der Bonner Testbatterie erfassbaren Leistungsdimensionen werden in der vorliegenden Arbeit faktorielle Intelligenzstrukturmodelle herangezogen. Diese Modelle zeichnen sich durch eine gute empirische Fundierung und durch differenzierte Annahmen über die Faktoren und die dahinter stehenden Prozesse aus. Sofern nachgewiesen werden kann, dass die mit der Testbatterie erhobenen Daten durch ein bestimmtes Intelligenzstrukturmodell beschreibbar sind, können auch die Faktorinterpretationen dieses Modells auf die Testbatterie übertragen werden. Neben der Frage, ob die Normdaten mit einem etablierten Intelligenzstrukturmodell vereinbar sind und welches der konkurrierenden Intelligenzstrukturmodelle sich zur Beschreibung des Interkorrelationsmusters am besten eignet, wird außerdem untersucht, ob dieses Modell auch in der klinischen Zielpopulation Gültigkeit besitzt. Um zu untersuchen, welches Intelligenzstrukturmodell sich am besten zur Beschreibung der Normdaten eignet, sind die Daten der Normstichprobe mit den Vorhersagen verschiedener theoretischer Modelle abzugleichen. Die Fragestellung, ob dieses Modell sich auch zur Beschreibung der Daten der Patientenstichprobe eignet, könnte grundsätzlich auf gleichem Wege beantwortet werden. Methodisch günstiger als dieses sequenzielle Vorgehen ist jedoch ein mehrgruppenanalytisches Vorgehen, indem die Eignung des Modells zur Vorhersage der Stichprobendaten simultan für beide Stichproben überprüft wird. Die hinter beiden Fragestellungen stehenden Invarianzhypothesen sind Hypothesen *konfiguraler* und *schwacher metrischer* Invarianz.

Der dritte Teil der Untersuchung befasst sich mit der Überprüfung weitergehender Invarianzhypothesen. Für die übliche diagnostische Praxis, die Testauswertung auf

Basis hirngesunder Normkollektive vorzunehmen, ist es wichtig, dass nicht nur die faktoriellen Strukturen in der Norm- und in der Zielpopulation übereinstimmen, sondern auch, dass die Faktoren in beiden Populationen die gleichen psychometrischen Eigenschaften aufweisen (Hypothese starker metrischer Invarianz und Hypothese strenger metrischer Invarianz). Für wissenschaftliche Untersuchungen der Korrelationen der latenten Faktoren mit weiteren Variablen ist starke metrische Invarianz notwendig, für die Untersuchung von latenten Mittelwertsunterschieden zusätzlich strenge metrische Invarianz.

Zur Untersuchung der Faktorenstruktur gibt es verschiedene statistische Methoden. Die mittels explorativer Faktorenanalysen gefundenen Faktorenlösungen hängen in hohem Maße von der jeweiligen Probandenstichprobe und von der Zusammenstellung der Einzeltests ab. Entsprechend schwierig gestaltet sich eine theoretisch fundierte Interpretation der gefundenen Faktoren. Günstiger ist es, eine Testbatterie von vornherein in einen etablierten und empirisch fundierten theoretischen Rahmen zu stellen. Hierzu sind konfirmatorische Faktorenanalysen die Methode der Wahl. Bevor diese Methodik dargestellt wird (Kapitel 3), sollen im nächsten Kapitel relevante Intelligenzstrukturmodelle dargestellt werden (Kapitel 2). Kapitel 4 dient der Darstellung der verschiedenen Invarianzhypothesen. Diese sind mit der Methode der konfirmatorischen Faktorenanalyse überprüfbar und in der Notation faktorieller Modelle formulierbar. Kapitel 5 schließlich stellt die wenigen vorliegenden empirischen Untersuchungen zur Überprüfung verschiedener Invarianzannahmen kognitiver Testbatterien dar.

2 Konzeptionen der Intelligenz

Die folgende Darstellung wichtiger Konzeptionen der Intelligenz fokussiert auf eine kurze Beschreibung der faktoriellen Intelligenzstrukturmodelle, die in dieser Studie untersucht werden. Zunächst soll aber das Konstrukt „Intelligenz“ definitorisch eingegrenzt werden.

2.1 Intelligenzdefinitionen

Ein schwerwiegendes Problem der Psychologie ist, dass Intelligenz zwar gemessen werden kann, aber nur schwer zu definieren ist. Das ist aber keinesfalls eine Besonderheit des Konstruktes Intelligenz: Auch andere mentale Fähigkeitsbereiche oder neuropsychologische Konstrukte mit hohem Allgemeinheitsgrad (z. B. „exekutive Funktionen“, „Kurzzeitgedächtnis“) bleiben häufig unscharf definiert. Bezogen auf Intelligenzdefinitionen wird insbesondere die Gefahr sinnfreier Tautologien, in denen lediglich der Begriff „intelligent“ durch verwandte Begriffe wie „zweckvoll“ oder „vernünftig“ ersetzt wird, betont (Amelang & Bartussek, 2001). Dies geschickt umgehend hat die American Psychological Association folgende Definition von Intelligenz vorgeschlagen (Neisser et al., 1996):

> *“Individuals differ from one another in their ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought. Although these individual differences can be substantial, they are never entirely consistent: a given person’s intellectual performance will vary on different occasions, in different domains, as judged by different criteria. Concepts of ‘intelligence’ are attempts to clarify and organize this complex set of phenomena.”*

Diese Definition ist eine vorsichtige inhaltliche Charakterisierung der Dimensionen, auf denen sich Menschen unter Anwendung des Intelligenzkonzeptes unterscheiden lassen. Humphreys (1994, zitiert nach Amelang & Bartussek, 2001) umschreibt Intelligenz als das *„erworbene Repertoire von intellektuellen (kognitiven) Fertigkeiten und Wissensbeständen, die einer Person zu einem gegeben Zeitpunkt verfügbar sind“.* Diese intellektuellen Fertigkeiten und Wissensbestände können als Voraussetzung für das Verständnis komplexer Sachverhalte und für die effektive Adaptation an die Umwelt etc. (siehe oben) angesehen werden. Während die erste Definition eher die Resultate intelligenten Verhaltens beschreibt, umreißt die zweite Definition den Umfang der für dieses Verhalten notwendigen Fähigkeiten.

In Hinblick auf eine Reihe bekannter Intelligenzdefinitionen stellt Mackintosh (1998) fest: „*Are they really of any great practical value? Are they necessary to direct the*

course of scientific enquiry? Are they even helpful? The answer to all these questions is: probably not." (S. 4). Unberührt vom Fehlen einer allgemein anerkannten Definition gehen aus der psychometrischen Intelligenzforschung vielfältige für Theorie und Praxis relevante Ergebnisse vor. Von besonderer Bedeutung ist die Aufstellung einer Terminologie und Taxonomie zur Klassifikation kognitiver Fähigkeiten, während gleichzeitig die Bedeutung, Probanden durch globale Aussagen über die intellektuelle Leistungsfähigkeit zu klassifizieren, in den Hintergrund rückt. Die Beschreibung der verschiedenen Facetten der Intelligenz und der Beziehung zueinander ist für die Testauswahl und Interpretation ebenso hilfreich wie für die Entwicklung psychometrischer Testbatterien, und stellt somit eine Brücke zwischen Theorie und Praxis dar. Dies ist auch von höchster Relevanz für die klinische Neuropsychologie, für die die Psychometrie ein zentraler Bestandteil ist.

2.2 Struktur- und Prozessmodelle der Intelligenz

Theorien zur Intelligenz bestehen aus zwei sich wechselseitig ergänzenden Teilen, einer Struktur- und einer Prozesskomponente. Strukturmodelle der Intelligenz enthalten zentrale Annahmen über die Struktur der menschlichen Intelligenz. Ein Beispiel einer vieldiskutierten Frage hierzu wäre, ob Intelligenz besser als eine einzige, allgemeine Fähigkeit zu konzipieren ist, oder sich aus mehreren, zumindest teilweise unabhängigen Fähigkeiten zusammensetzt. Methodisch basieren Strukturmodelle zumeist auf faktorenanalytischen Auswertungen kognitiver Testbatterien. Die dabei identifizierten Faktoren werden als latente Quelle interindividueller Unterschiede der Intelligenz betrachtet.

Faktorielle Strukturmodelle beschreiben letztlich jedoch nur die Zusammenhänge zwischen verschiedenen Untertests einer Testbatterie, was nicht mit der Struktur menschlicher Fähigkeiten gleichgesetzt werden darf (Mackintosh, 1998). Dies wird schon durch die Tatsache deutlich, dass Art und Anzahl der durch explorative Faktorenanalysen ermittelten Faktoren immer von Art und Anzahl der in die Analyse eingehenden Tests abhängt. Näherer Aufschluss über die Struktur menschlicher Fähigkeiten kann nur durch Entwicklung psychologischer Theorien erlangt werden. Die Faktorenanalyse kann somit als Voranalyse betrachtet werden, mit der komplexe Datensätze zusammengefasst werden. In weiteren Schritten werden die gefundenen Faktoren (bzw. die zugeordneten Konstrukte) zur Entwicklung psychologischer Theorien psychologisch getestet (Mackintosh, 1998). Die postulierten Faktoren bekommen ihre Rechtfertigung niemals allein durch ihren faktorenanalytischen Nachweis, sondern müssen durch weitere wissenschaftliche Methoden Bestätigung finden (siehe z. B.

Institute for Applied Psychometrics, 2002; Sternberg, 2000). Beispiele solcher Methoden sind:

- Kriterienvalidierung: Z. B. Untersuchung der differenziellen Kovariation zwischen den verschiedenen Faktoren eines Intelligenzmodells einerseits und bestimmten Outcome-Kriterien andererseits.
- Neurokognitive Forschung: Z. B. Untersuchung der Beziehungen zwischen den Faktoren und (neuro-)physiologischen und neurologischen Maßen.
- Genetik: Z. B. Untersuchung der Vererblichkeitsraten verschiedener Faktoren.
- Entwicklungspsychologie: Z. B. Untersuchung des Wachstums und des Abbaus der Leistungen verschiedener Faktoren über die Lebensspanne hinweg.

Neben der Strukturkomponente einer Theorie ist auch die Komponente, die die prozeduralen Aspekte der Intelligenz berücksichtigt, bedeutsam. Dabei steht die Frage im Vordergrund, welche kognitiven Prozesse für die Unterschiede in Intelligenztestleistungen verantwortlich gemacht werden können. Hängen die Leistungen in allen Tests mehr oder minder an einem einzigen Prozess, oder sprechen – und diesbezüglich scheint sich ein Konsens abzuzeichnen – verschiedene Tests verschiedene, aber überlappende Fähigkeiten an, von denen jedoch keine für alle Tests relevant ist (Mackintosh, 1998)?

Der Zusammenhang zwischen den Gruppenfaktoren und den zugrunde liegenden kognitiven Prozessen ist auch für die klinische Neuropsychologie hoch relevant. Wie Burton et al. (2002) jedoch anmerken, gibt es bisher wenig Übereinstimmung zwischen der klinischen und nicht-klinischen Literatur. Dies ist aus den verschiedenen Traditionen der Forschungsrichtungen heraus erklärbar: Die nomothetische Intelligenzforschung fokussierte traditionell auf die Untersuchung *inter*individueller Leistungsunterschiede; die klinische Forschung hingegen befasst sich vorwiegend mit der Untersuchung von *intra*individuellen Leistungsunterschieden.

2.3 Faktorenanalytische Intelligenzstrukturmodelle

Faktorenanalytische Intelligenzstrukturmodelle müssen der empirischen Gegebenheit einerseits durchweg positiver Interkorrelationen zwischen den kognitiven Tests und andererseits Clusterbildungen aus hoch und niedrig interkorrelierten Tests gerecht werden. Trotz dieser allgemeinen Anforderung liegen analog zur Vielzahl möglicher Arten der Varianzaufspaltung viele verschiedene Intelligenzstrukturmodelle vor. So entsteht der Eindruck, dass diese Modelle gleichwertig nebeneinander stehen und allenfalls über eine Bewertung der Adäquatheit der zugrunde liegenden

faktorenanalytischen Methode beurteilt werden können. Aber auch eine genaue Untersuchung der Distinktheit der den einzelnen Faktoren zugeordneten Prozesse und die Überprüfung der Kriterienvalidität der Faktoren kann Entscheidungen für oder gegen eines der konkurrierenden Strukturmodelle herbeiführen: „*This confirmation from a relatively independent source provides gratifying evidence that psychometric or factorial theories of human intelligence do not operate in a complete vacuum*" (Mackintosh, 1998, S. 214). Seit relativ kurzer Zeit liegt die konfirmatorische Faktorenanalyse als weitere Methode zur Entscheidung für oder gegen ein Strukturmodell vor. Mit dieser Methode können die verschiedenen Modelle einheitlich nachmodelliert werden und ihre Eignung zur Beschreibung empirischer Daten wird vergleichbar.

In einer Vorauswahl sind aus der hohen Modellvielfalt die Modelle zu identifizieren, die sich am besten zur Beschreibung der Daten der neuropsychologischen Testbatterie eignen. Ein wichtiges Selektionskriterium sollte die Relevanz der Modelle für aktuelle Intelligenztestverfahren sein. Zudem müssen die Modelle die Tatsache interkorrelierter Gruppenfaktoren berücksichtigen. Daher wird Thurstones Modell der *primary mental abilities* (Thurstone, 1938), welches faktoriell reine Tests zur Erfassung unabhängiger Fakultäten erfordert, nicht untersucht. Weiterhin nicht berücksichtigt werden Modelle mit einem sehr breiten Gültigkeitsanspruch (z. B. Gardner, 1993; Sternberg, 1985). Dies ist nicht als Ausdruck ihrer fehlenden Relevanz zu verstehen, sondern ist der Tatsache geschuldet, dass die neuropsychologische Testbatterie keinesfalls als gültige Operationalisierung dieser breit angelegten Intelligenzmodelle betrachtet werden kann. Auch die PASS-Theorie (Das, Naglieri & Kirby, 1994; Naglieri, 1997), die vor allem in der Testbatterie CAS (*cognitive assessment system* (Naglieri & Das, 1996)) und teilweise auch in der *Kaufman-assessment battery for children* (Kaufman, Kaufman, Melchers & Preuß, 2001) für die aktuelle Testpraxis operationalisiert wurde, soll hier keine Rolle spielen. Die PASS-Theorie besteht aus den vier kognitiven Komponenten *Planning, Attention, Simultaneous und Successive Processes*. Da in dieser Theorie die *Prozesse* und weniger die *Struktur* der Intelligenz im Vordergrund stehen, erfordert eine angemessene Operationalisierung sehr spezielle Tests, die in der hier zu untersuchenden neuropsychologischen Testbatterie nicht enthalten sind.

In der folgenden Darstellung faktorieller Intelligenzstrukturmodelle sind zwei Aspekte von übergeordnetem Interesse: Der Inhaltsaspekt umfasst die Frage, welche Fähigkeitsfaktoren die Modelle berücksichtigen, der zweite Aspekt fokussiert auf die Beziehung der Faktoren untereinander (Strukturfrage oder formaler Aspekt nach Amthauer, Brocke, Liepmann & Beauducel, 2001). Die konkrete Operationalisierung der

Intelligenzmodelle in Hinblick auf die neuropsychologische Testbatterie wird im Methodenteil (Kapitel 7.3.1) dargestellt.

2.3.1 Das Generalfaktormodell

Die folgende Darstellung der Intelligenzstrukturmodelle soll mit dem Generalfaktormodell von Spearman (1927) beginnen. Obwohl dessen klassische Form inzwischen als widerlegt gelten kann, ist das Konzept des Generalfaktors im Rahmen der später dargestellten hierarchischen Modelle bedeutsam.

Die klassische Zwei-Faktoren-Theorie der Intelligenz besagt, dass es einen einheitlichen, fundamentalen Prozess der allgemeinen Intelligenz, den General- oder g-Faktor, gibt, der alle intellektuellen Fähigkeiten durchdringt und somit die Leistung in allen Intelligenztests bestimmt. Das Hauptindiz für den g-Faktor ist, dass jede Korrelationsmatrix kognitiver Leistungstests nur positive Korrelationen aufweist (*positive manifold*). Zur Erfassung des g-Faktors eignet sich eine möglichst breite Auswahl verschiedener Einzeltests, da sich dadurch die idiosynkratischen Züge der einzelnen Tests gegenseitig aufheben. Der Name „Zwei-Faktoren-Theorie“ bezieht sich auf die Aufspaltung der Varianz eines Tests in einen auf die allgemeine Intelligenz zurückgehenden Anteil und einen testspezifischen Anteil (*uniqueness;* vergleiche Abbildung 2-A). Bezogen auf die Interkorrelationsmatrix einer Testbatterie würde eine aus dieser Theorie abgeleitete Vorhersage lauten, dass alle positiven Korrelationen auf den g-Faktor zurückgeführt werden können und nach deren Auspartialisierung nur noch nicht-signifikante Korrelationen verbleiben.

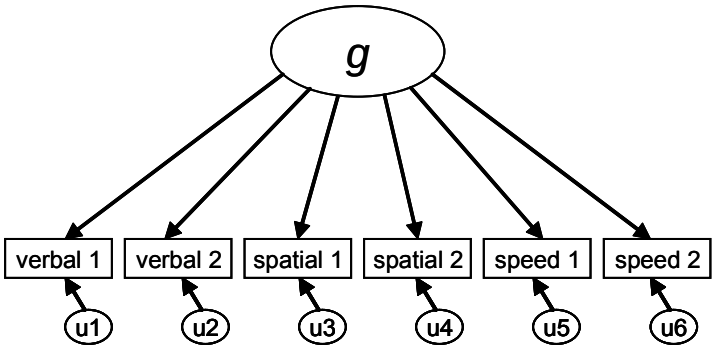


Abbildung 2-A: Klassische Spezifikation des Generalfaktormodells (nach Keith, 1997).

Der g-Faktor drückt das Gemeinsame aller kognitiven Test aus, nämlich die *allgemeine Intelligenz*. Sie unterscheidet sich von der *Intelligenz im Allgemeinen*, also einem Durchschnittsniveau der kognitiven Leistungen (Amelang & Bartussek, 2001; Ardila, 1999). Auch wenn Intelligenztests darauf abzielen, die allgemeine Intelligenz zu

erfassen, so wird dies häufig über Durchschnittsbildung oder Summation einer breiten Auswahl verschiedener Tests operationalisiert. Der g-Faktor an sich ist als faktorenanalytisches Konstrukt nicht direkt messbar.

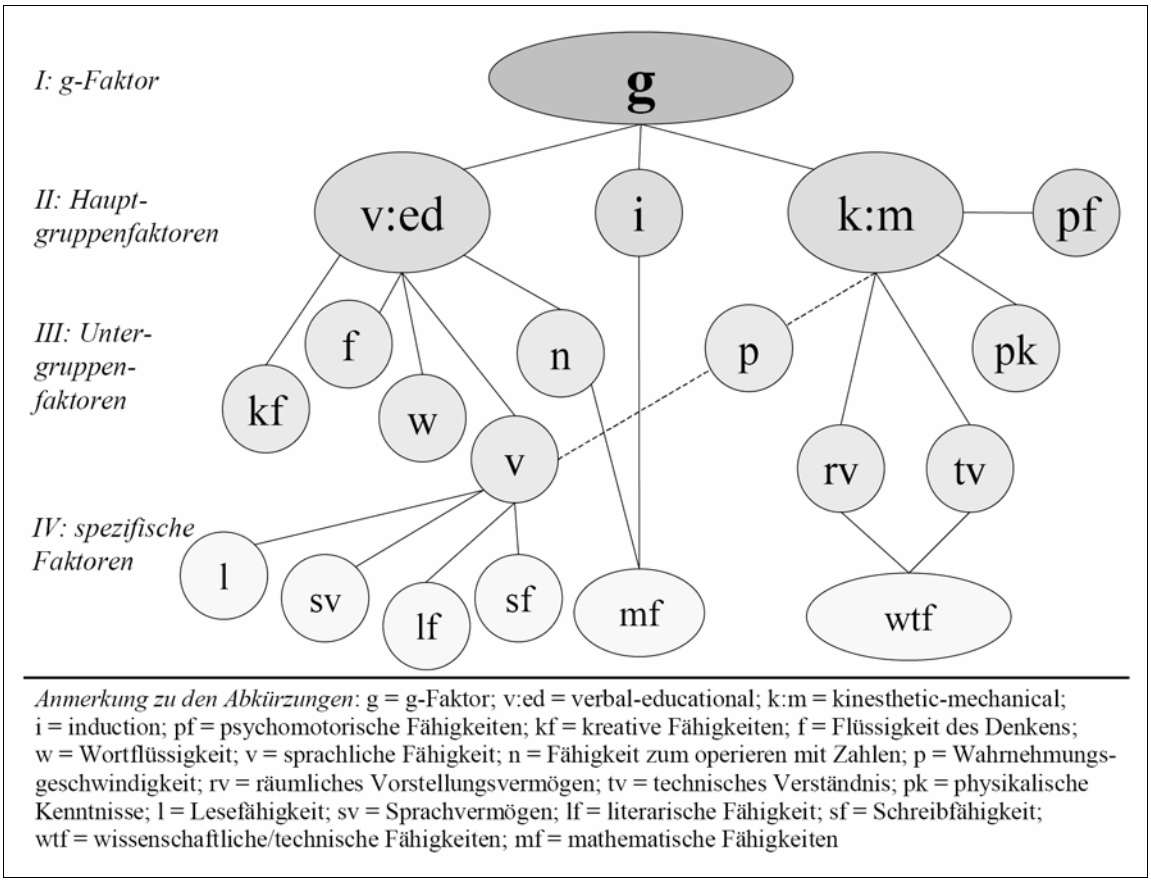
Die reine Zwei-Faktoren-Theorie kann als widerlegt gelten. Es zeigt sich, dass auch nach Extraktion der auf den g-Faktor zurückgehenden Kovarianz aus einer Interkorrelationsmatrix kognitiver Tests weiterhin ein (clusterartiges) Muster positiver Interkorrelationen übrig bleibt. Die aktuelle Forschung zum g-Faktor (nun zumeist in Rahmen hierarchischer Intelligenzstrukturmodelle) beschäftigt sich mit der Detektion und Erklärung von Gruppenunterschieden (zur vielfach rassistischen Ausrichtung dieser Forschung siehe Schonemann, 1997; Schonemann, 2001; Stöcker, 2005) und mit den kognitiven Prozessen, die dem g-Faktor zugrunde liegen. Dabei wird angenommen, dass es einen (wie auch immer gearteten und zu nennenden) einheitlichen Verarbeitungsmechanismus gibt (fluide Intelligenz, Arbeitsgedächtnis, mentale Energie etc.), der wiederum durch eine Anzahl spezifischer Prozessoren oder Module ergänzt wird. Kritiker des Generalfaktormodells argumentieren, dass das Muster positiver Interkorrelationen auch anders erklärbar sei, beispielsweise durch verschiedene überlappende Prozesse oder durch Umwelteinflüsse, die auf alle Teilfähigkeiten gleichermaßen wirken (Gould, 1997). Ein weiterer Kritikpunkt ist die dem g-Faktor inhärente Annahme, dass der g-Faktor immer das Gleiche repräsentieren soll, unabhängig davon, aus welcher konkreten Testzusammenstellung er extrahiert wurde (Mackintosh, 1998).

Die im Folgenden besprochenen Intelligenzstrukturmodelle gehen nun im Gegensatz zum Generalfaktormodell davon aus, dass Intelligenz aus vielen, hierarchisch organisierten Facetten besteht.

2.3.2 Das hierarchische Intelligenzmodell von Vernon

Eine klassische Dichotomie der Neuropsychologie ordnet den beiden Großhirnhemisphären unterschiedliche Aufgabenschwerpunkte zu (*cerebral specialization theory* (siehe Kaufman, 2000)): Inhaltlich wird dabei zwischen verbalen und nonverbalen Funktionen unterschieden, prozedural zwischen analytisch-sequenzieller und ganzheitlicher Verarbeitung (für eine ausführliche Diskussion siehe Hartje, 2002). Insbesondere die erste Unterscheidung hat sich lange Zeit als prädominantes Modell für Intelligenztestungen erwiesen; so können entsprechend der Auswertungsrichtlinie der revidierten Version der Intelligenztestbatterie von Wechsler (Wechsler, 1981) die Faktoren *verbale Intelligenz* und *Handlungsintelligenz* ermittelt

werden[1]. Aufgrund der für die prächirurgische Diagnostik hohen Wichtigkeit der Fragestellung nach der Lateralisierung kognitiver Dysfunktionen soll im Rahmen dieser Arbeit auch ein dichotomes Modell überprüft werden, namentlich das hierarchische Modell von Vernon (Vernon, 1950, 1971). Dieses Modell enthält zwei Faktoren von hohem Allgemeinheitsgrad, die der verbal-nonverbal Dichotomie sehr nahe kommen. Das Modell besitzt an der Spitze einen g-Faktor und auf der zweiten Ebene die breiten Gruppenfaktoren *verbal-numerical-educational* (v:ed) und *spatial-practical-mechanical-physical* bzw. *kinesthetic-mechanical* (k:m). Auf der dritten Ebene sind Untergruppenfaktoren und darunter aufgabenspezifische Residualfaktoren angeordnet. Abbildung 2-B gibt das Modell von Vernon wieder.



Anmerkung zu den Abkürzungen: g = g-Faktor; v:ed = verbal-educational; k:m = kinesthetic-mechanical; i = induction; pf = psychomotorische Fähigkeiten; kf = kreative Fähigkeiten; f = Flüssigkeit des Denkens; w = Wortflüssigkeit; v = sprachliche Fähigkeit; n = Fähigkeit zum operieren mit Zahlen; p = Wahrnehmungsgeschwindigkeit; rv = räumliches Vorstellungsvermögen; tv = technisches Verständnis; pk = physikalische Kenntnisse; l = Lesefähigkeit; sv = Sprachvermögen; lf = literarische Fähigkeit; sf = Schreibfähigkeit; wtf = wissenschaftliche/technische Fähigkeiten; mf = mathematische Fähigkeiten

Abbildung 2-B: Das hierarchische Intelligenzmodell von Vernon (aus: Nettelnstroth, 2003).

2.3.3 Das Berliner Intelligenzstrukturmodell

Die Lösungen von Faktorenanalysen und folglich auch die darauf aufbauenden Intelligenzstrukturmodelle sind von den Aufgaben abhängig, die in die Analyse eingehen. Um diesen Einfluss zu minimieren, wurde zur Entwicklung des Berliner

[1] Die neueste Version dieses Intelligenztests (WAIS-III, Wechsler, 1997a) geht über diese einfache Dichotomie hinaus und repräsentiert die Faktoren verbales Verständnis, Wahrnehmungsorganisation, Arbeitsgedächtnis und Verarbeitungsgeschwindigkeit.

Intelligenzstrukturmodells (BIS, Jäger, 1982; Jäger, 1984) ein möglichst umfassender und repräsentativer Aufgabenpool zusammengestellt. Hierzu wurden um die 2000 Aufgaben gesichtet, reduziert und analysiert. Das gefundene Modell zeichnet sich im Unterschied zu den bisher dargestellten Modellen durch *Bimodalität* aus: Intelligenzleistungen werden dabei unter den Aspekten *Operationen* (das sind für die Aufgabenbearbeitung wichtige Prozesse) und *Inhalte* klassifiziert. In Übereinstimmung mit den hierarchischen Modellen werden die Fähigkeitskonstrukte auch im BIS hierarchisch konzipiert und stehen in gegenseitig abhängiger Beziehung zueinander. An der Spitze der Fähigkeitshierarchie steht als Integral aller Fähigkeiten die allgemeine Intelligenz, auf der Ebene darunter sind sieben hochgradig generelle Fähigkeitskonstrukte der Modalitäten *Operationen* und *Inhalte* angeordnet (Abbildung 2-C). Die operativen Fähigkeiten sind *Verarbeitungskapazität, Einfallsreichtum, Bearbeitungsgeschwindigkeit* und *Merkfähigkeit*, die inhaltsgebundenen Fähigkeiten sind *sprachgebundenes Denken*, *zahlengebundenes Denken*, *anschauungsgebundenes bzw. figural-bildhaftes Denken*. Eine Besonderheit des BIS ist, dass *Kreativität* (Skala Einfallsreichtum) als gleichwertiges Fähigkeitskonstrukt neben den eher klassischen Leistungsdimensionen aufgenommen wurde.

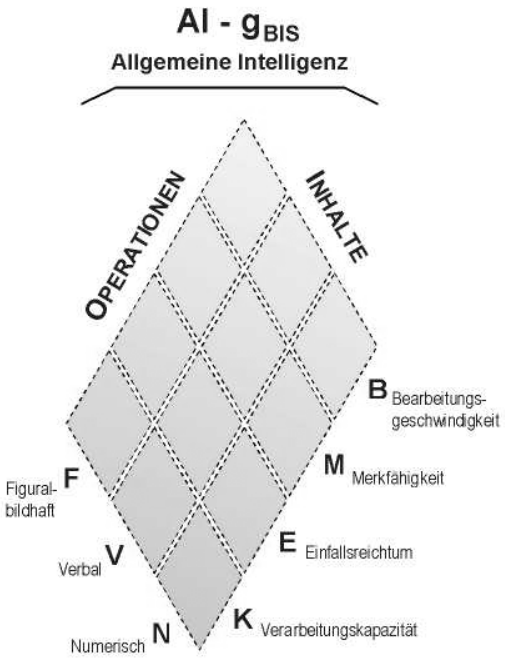


Abbildung 2-C: Das Berliner Intelligenzstrukturmodell (aus: Jäger, Süß & Beauducel, 1997).

Im Berliner Intelligenzstruktur-Test (Jäger et al., 1997) wird jede Aufgabe zur Messung einer operativen und einer inhaltsgebundenen Fähigkeit verwendet. Die mehrfache

Verwendung der Tests ist Abbild der theoretischen Annahmen des Modells, steht aber im Gegensatz zu den klassischen hierarchischen Modellen. Diese basieren auf faktorenanalytischen Techniken, die die Einfachstruktur als Rotationskriterium anwenden. Dieses Kriterium optimiert Lösungen, bei denen eine Variable auf einem Faktor hoch und auf allen weiteren Faktoren niedrig lädt. Eine direkte Bestätigung des Berliner Intelligenzstrukturtests mittels explorativen Faktorenanalysen ist also nicht zu erwarten. Nur durch spezielle methodische Zwischenschritte („Bifaktormethode“) konnte Evidenz für die Gültigkeit des Berliner Intelligenzstrukturmodells gefunden werden (Jäger, 1982). Schmidt sowie Bucik und Neubauer bestätigten später das Modell mit der wesentlich besser geeigneten konfirmatorischen Faktorenanalyse (Bucik & Neubauer, 1996; Schmidt, 1984).

Abschließend seien für den aktuellen Zusammenhang noch folgende drei Punkte betont: Erstens konnte Schmidt (Schmidt, 1993) das Berliner Intelligenzstrukturmodell auch mit anderen Testaufgaben, die einer anderen Intelligenztesttradition entstammen, replizieren (mit Thurstones Test zur Erfassung der *primary mental abilities*) und hat somit einen Hinweis auf die Generalisierbarkeit des Modells erbracht. Zweitens merken Beauducel und Kersting (2002) an, dass das Berliner Intelligenzstruktur grundsätzlich ein offenes Modell ist, in das je nach Evidenzlage noch weitere Kategorien eingefügt werden können - ein möglicher weiterer Faktor wäre ein Wissensfaktor. Drittens steht das Berliner Intelligenzstrukturmodell in guter Übereinstimmung mit der in der klinischen Neuropsychologie wohlbekannten Tatsache der Abhängigkeit kognitiver Leistungen von den inhaltlichen Modalitäten.

2.3.4 Das Cattell-Horn-Carroll-Modell und dessen Vorläufer

Der Tatsache der durchweg positiven Interkorrelationen zwischen kognitiven Tests kann entweder durch hierarchische Modelle oder durch Modelle mit obliquen Gruppenfaktoren Rechnung getragen werden. Hierarchische Modelle bestehen aus mehreren Ebenen unterschiedlich breit angelegter Faktoren. Häufig, aber nicht immer, steht dabei ein g-Faktor mit höchstem Allgemeinheitsgrad an der Spitze. Anders als in der Zwei-Faktoren-Theorie nach Spearman (1927, siehe oben), wird in hierarchischen Theorien der g-Faktor jedoch als hierarchische Fähigkeit angesehen, der weitere untergeordnete Ebenen einschließt. Unter der Ebene des g-Faktors befinden sich Gruppenfaktoren, die spezifische Fähigkeitsbereiche widerspiegeln. Auf Gruppenfaktoren lädt jeweils nur eine Untergruppe von Variablen. Die Einführung von Gruppenfaktoren wurde schon früh vorgeschlagen (Burt, 1909). Auf der niedrigsten Hierarchieebene finden sich sehr enge Fähigkeitskonstrukte, die häufig mit spezifischen Einzeltests korrespondieren.

Hierarchische Modelle können beispielsweise über Faktorenanalysen mit obliquer Rotation erhalten werden, wobei aus den obliquen Faktoren durch Faktorenanalysen höherer Ordnung sukzessive ein g-Faktor ermittelt werden kann.

Cattell und Horn haben eines der einflussreichsten Gruppenfaktorenmodelle aufgestellt, dessen noch gebräuchlicher Name „*Gf-Gc-Theorie*“ an die frühe Form der zugrunde liegenden Theorie von Cattell (1963) erinnert. Aber auch heute noch wird dieses Modell häufig mit der Dichotomie *fluide* versus *kristalline* (Gf, Gc) Intelligenz gleichgesetzt. *Fluide Intelligenz* beinhaltet die Fähigkeit, Beziehungen in Stimulus-Mustern zu erkennen, Implikationen zu verstehen und Inferenzen aus Beziehungen zu erschließen. Die fluide Intelligenz wird als die biologische Vorbedingung von Intelligenz betrachtet, die sich in der kristallinen Intelligenz manifestiert. Die *kristalline Intelligenz* ist eher von Erfahrung und kulturspezifischer Bildung abhängig und umfasst über die Lebensspanne hinweg angeeignete Fähigkeiten und Wissensinhalte. Bei der modernen erweiterten Version dieser Theorie handelt es sich um ein Zwei-Schichten-Modell, welches durch folgende Aspekte charakterisiert werden kann (Horn, 1991; Sternberg, 2000):

- Auf der untersten Schicht (niedrigste Hierarchieebene) befinden sich ca. vierzig Faktoren erster Ordnung. Diese engen Fähigkeitsfaktoren entsprechen häufig einzelnen Subtests.
- Auf der übergeordneten Ebene werden acht bis zehn breite Fähigkeitskonstrukte postuliert, neben Gf und Gc noch folgende Faktoren: *Short-Term Acquisition and Retrieval* (Gsm), *Visual Intelligence* (Gv), *Auditory Intelligence* (Ga), *Long-Term Storage and Retrieval* (Glr), *Cognitive Processing Speed* (Gs), *Correct Decision Speed* (CDS), *Quantitative Knowledge* (Gq).
- Ein Faktor höherer Ordnung wird nicht angenommen.

Eine explizit auf diesem Modell basierende kognitive Testbatterie ist die Batterie von Woodcock und Johnson (WJ-R, Woodcock & Johnson, 1989). Hierin werden sieben der oben genannten Faktoren operationalisiert.

Carroll (1993) hat eine vollständige Kartierung aller bekannten kognitiven Fähigkeiten versucht und darauf aufbauend eine dreischichtige Intelligenztheorie postuliert (*three stratum theory of intelligence*). Zu diesem Zweck hat Carroll 460 veröffentlichte Datensätze der letzten 60 bis 70 Jahre mit einer einheitlichen faktorenanalytischen Methodik reanalysiert (Schmid-Leiman-Transformation, Schmid & Leiman, 1957). Ein zentrales Element der Drei-Schichten-Theorie ist das Ordnungsschema, das die verschiedenen Fähigkeitsdimensionen als Funktion ihrer Breite gliedert. Dabei

kennzeichnet der Schichtbegriff (*stratum)* die Enge bzw. Breite der einzelnen Faktoren. Alle kognitiven Leistungsfaktoren können auf einer von drei Schichten eingeordnet werden: Auf der dritten Schicht ist der g-Faktor angesiedelt, auf der zweiten Schicht befinden sich wenige (ca. acht bis zehn) breite Faktoren [*fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception, broad auditory perception, broad retrieval ability, broad cognitive speediness, processing speed*] und auf der ersten Schicht finden sich (knapp 70) sehr enge Fähigkeitsfaktoren. Die Ähnlichkeit zu Horns Gf-Gc-Theorie ist offensichtlich; ein wichtiger Unterschied besteht in der expliziten Berücksichtigung eines g-Faktors in der Theorie von Carroll.

Das *Institute for Applied Psychometrics* hat das *Cattell-Horn-Carroll (CHC) Definition Project* als Teil des *Carroll Human Cognitive Abilities Project* initiiert (Institute for Applied Psychometrics, 2002; McGrew, 2003). Im Rahmen dieses Projektes wurde eine Taxonomie kognitiver Fähigkeiten entwickelt und das hierarchische Cattell-Horn-Carroll-Modell (CHC-Modell) aufgestellt. Dieses Intelligenzstrukturmodell stellt eine Synthese aus den sehr ähnlichen Theorien von Horn und Cattell (Horn, 1991) einerseits und Carroll (Carroll, 1993) andererseits dar. Beide Theorien definieren die breiten Fähigkeitsfaktoren sehr ähnlich, Unterschiede finden sich in der Annahme eines g-Faktors und in der Zuordnung einzelner enger Faktoren zu den breiteren Faktoren. Darüber hinaus nimmt Carroll nur einen Gedächtnisfaktor an, Horn unterscheidet einen Kurzzeit- und einen Langzeitgedächtnisfaktor. Im CHC-Modell ist über die Existenz eines g-Faktors noch nicht endgültig entschieden.

Abbildung 2-D wurde aus den im Internet verfügbar gemachten Informationen des *Institute for Applied Psychometrics* entnommen (McGrew, 2003). Sie stellt zusammenfassend die Vorläufermodelle des CHC-Modells dar: das Spearman'sche g-Faktormodell (1a), die Strukturtheorie nach Thurstone mit unabhängigen Faktoren (primäre mentale Fähigkeiten - 1b) und die zwei konkurrierenden neueren Ansätze, nämlich das hierarchische Modell von Cattell und Horn mit interkorrelierten Faktoren (1c) sowie das Modell von Carroll (1d). Die vielen Pfade im Modell von Carroll sind auf die hierarchische Schmid-Leiman-Faktorenanalyse zurückzuführen, die dem Modell zugrunde liegt: Das daraus resultierende Modell enthält Faktoren, die mit jedem untergeordneten Faktor und Test gemeinsame Pfade (Ladungen) aufweisen. Teil 1e der Abbildung schließlich zeigt das hierarchische Cattell-Horn-Carroll-Modell, an dessen Spitze ein g-Faktor steht.

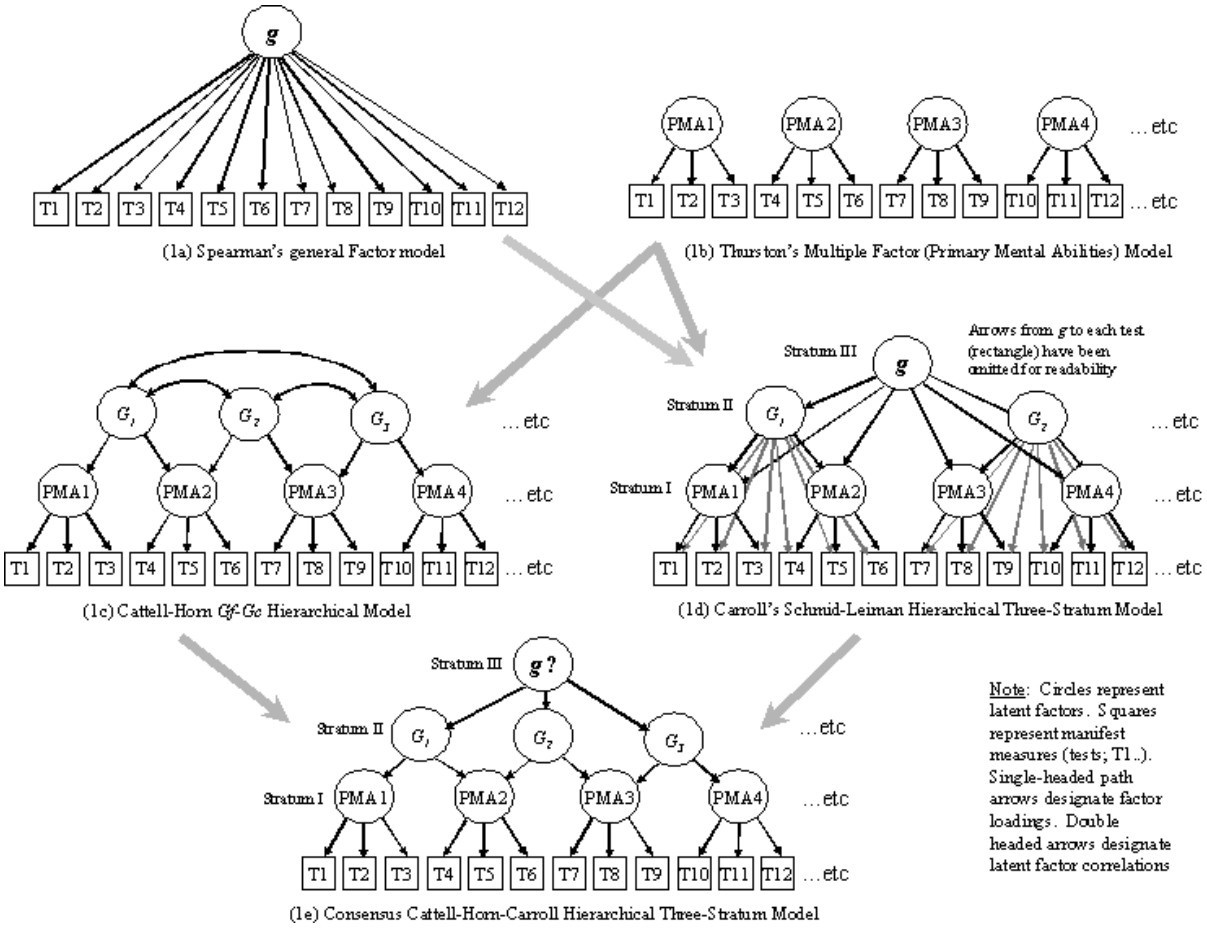


Abbildung 2-D: Das Cattell-Horn-Carroll-Modell (1e, unten) und dessen Vorläufer (nach McGrew, 2003).

Bei der Entwicklung des CHC-Modells wurde versucht, eine Entscheidung über die unterschiedlichen Modellannahmen der Vorgängertheorien mittels *konfirmatorischer Faktorenanalysen* herbeizuführen. Bezüglich der Gedächtnisleistungen konnte so beispielsweise die Unterscheidung eines Kurzzeit- und eines Langzeitgedächtnisfaktors bestätigt werden. Weiterhin ungelöst blieb allerdings die Frage nach der Existenz eines g-Faktors. Neben der Aufstellung eines Intelligenzstrukturmodells soll im Rahmen des *Carroll Human Cognitive Abilities* Projektes auch eine umfassende Taxonomie der kognitiven Fähigkeiten entwickelt werden. Der neueste Stand der Konsensdefinitionen der wichtigsten engen und breiten Fähigkeiten (erste und zweite Schicht) findet sich im Internet veröffentlicht (McGrew, 2003). Die Definitionen basieren auf den Arbeiten von Carroll (1993) und McGrew (1997). Anhand des breiten Fähigkeitsfaktors der zweiten Schicht (visuell-räumliche Fähigkeiten) soll die Art der Definition und die Charakterisierung der hierarchischen Struktur verdeutlicht werden (McGrew, 2003):

> *Visual-spatial abilities (Gv): 'The ability to generate, retain, retrieve, and transform well-structured visual images' (Lohman, 1994, p.1000). The Gv domain represents a collection of different abilities each that emphasize a different process involved in the*

generation, storage, retrieval and transformation (e.g., mentally reverse or rotate shapes in space) of visual images. Gv abilities are measured by tasks (figural or geometric stimuli) that require the perception and transformation of visual shapes, forms, or images and/or tasks that require maintaining spatial orientation with regard to objects that may change or move through space.

Die diesem Faktor zugeordneten engen Fähigkeitsfaktoren der ersten (untersten) Schicht werden im einzelnen aufgeführt und definiert, wobei hier exemplarisch nur die Definitionen für die Faktoren *visualization* und *spatial relations* wörtlich wiedergegeben werden sollen:

Visualization (Vz): The ability to apprehend a spatial form, object, or scene and match it with another spatial object, form, or scene with the requirement to rotate it (one or more times) in two or three dimensions. Requires the ability to mentally imagine, manipulate or transform objects or visual patterns (without regard to speed of responding) and to "see" (predict) how they would appear under altered conditions (e.g., parts are moved or rearranged). Differs from Spatial Relations primarily by a deemphasis on fluency.

Spatial Relations (SR): Ability to rapidly perceive and manipulate (mental rotation, transformations, reflection, etc.) visual patterns or to maintain orientation with respect to objects in space. SR may require the identification of an object when viewed from different angles or positions.

Die zum Faktor Gv gehörenden weiteren engen Fähigkeitsfaktoren der ersten (untersten) Schicht sind:

- *Closure Speed* (CS)
- *Flexibility of Closure* (CF)
- *Visual Memory* (MV)
- *Spatial Scanning* (SS)
- *Serial Perceptual Integration* (PI)
- *Length Estimation* (LE)
- *Perceptual Illusions* (IL)
- *Perceptual Alternations* (PN)
- *Imagery* (IM)

2.4 Intelligenz und Neuropsychologie

Die vorliegende Arbeit untersucht, inwieweit eine neuropsychologische Testbatterie auch zur Intelligenzprüfung valide und geeignet ist. Die Beantwortung dieser Fragestellung erfordert eine nähere Bestimmung der Beziehung zwischen

Intelligenztests und neuropsychologischen Tests beziehungsweise zwischen Intelligenzprüfungen und neuropsychologischen Untersuchungen. Inwieweit kann eine neuropsychologische Testbatterie, in der genuine Intelligenztests die Minderheit darstellen, überhaupt zur Untersuchung von Intelligenz herangezogen werden?

Zunächst ist es in diesem Zusammenhang notwendig, die Unterschiede zwischen Intelligenztestaufgaben und neuropsychologischen Testaufgaben zu erörtern. Russ (2003) definiert einen neuropsychologischen Test anhand überwiegend *psychometrischer* Kriterien wie folgt:

> *Ein neuropsychologischer Test ist ein psychologischer Test, der den spezifisch neuropsychologischen Messgegenstand objektiv, valide und differenziert erfasst und für die Anwendung im klinischen Bereich besonders überprüft worden ist. ... Der Messgegenstand [ist] das individuelle Leistungsdefizit, also eine negative Differenzgröße, die aufgrund neuropsychologischer Theorie und Anwendungserfahrung auf eine krankhafte organische Veränderung des Gehirns mit hinlänglicher Sicherheit rückführbar ist. Ein guter neuropsychologischer Test sollte die im klinischen Bereich möglichen Differenzwerte in der ganze Breite auf einer Skala quantifizieren können, die sich zwischen dem Normalbereich (Messwert 0) und dem maximal möglichen Differenzwert erstreckt. ... . Eine solche Skala kann naturgemäß im Normalbereich nicht mehr fein differenzieren; dazu sind konventionelle, bevölkerungsrepräsentativ normierte Verfahren besser geeignet. Für die klinische Anwendung ist es ausreichend, wenn die Grenze zum Normalbereich – bezogen auf das individuelle prämorbide Niveau – grob miterfasst wird. (S. 11).*

Entsprechend dieser Definition ist ein Intelligenztest nicht per se als neuropsychologischer Test geeignet, da zum einen dessen Differenzierung im niedrigsten Leistungsbereich nicht uneingeschränkt gewährleistet ist und zum anderen der Messgegenstand nicht das Leistungsdefizit sondern die Intelligenzleistung ist. Andererseits wäre ein so definierter neuropsychologischer Test nicht als Intelligenztest geeignet, da eine Differenzierung im überdurchschnittlichen Bereich nicht gegeben ist. Es sei jedoch angemerkt, dass es durchaus vielerlei bedeutende neuropsychologische Fragestellungen gibt, die eine Differenzierung der Untertestleistungen im durchschnittlichen oder überdurchschnittlichen Niveau erfordern. Ein Beispiel wäre hier die Frühdiagnostik von Demenzen bei prämorbid überdurchschnittlichem Leistungsniveau. Hierzu haben klassische neuropsychologische Verfahren eine zu niedrige Sensitivität.

Wie sieht es bezüglich der inhaltlichen Gemeinsamkeiten und Unterscheide zwischen Intelligenz- und neuropsychologischen Tests aus? Bei der Betrachtung der Untertests nach Wechsler (Tewes, 1991) fällt auf, dass sie keine Operationalisierungen

der typischen und häufig erwarteten Intelligenzkonzeptionen wie beispielsweise abstraktes oder schlussfolgerndes Denken darstellen. Vielmehr sind sie inhaltlich sehr breit und heterogen angelegt. Dahinter liegt die Annahme, dass sich Intelligenz in vielen verschiedenen Weisen manifestieren kann. Daher erhöht eine große Vielzahl unterschiedlicher Subtests die Chance, dass der Gesamttest fair ist und ein abgerundetes Bild der Intelligenz ergibt (Mackintosh, 1998). Die schon oben aufgeführte Definition von Intelligenz als das *„erworbene Repertoire von intellektuellen (kognitiven) Fertigkeiten und Wissensbeständen, die einer Person zu einem gegeben Zeitpunkt verfügbar sind“* (Humphreys, 1994, zitiert nach Amelang & Bartussek, 2001) widerspricht in keiner Weise der Verwendung von neuropsychologischen Tests zur Beantwortung von Fragestellungen nach Intelligenz. Auch die hierarchischen und multifaktoriellen Intelligenzmodelle, die in Entwicklung und Konzeption verschiedenste kognitive Testaufgaben berücksichtigt haben (siehe Kapitel 2.3), geben keinen Anhalt dafür, dass neuropsychologische Aufgaben von Inhalt und Struktur zur Intelligenztestung ungeeignet sein könnten.

Aus inhaltlichen Gesichtspunkten kann zusammenfassend eine grundsätzliche Eignung neuropsychologischer Testverfahren zur Intelligenzprüfung angenommen werden. Bezüglich der erwähnten psychometrischen Eigenschaften sind allerdings sicherlich solche neuropsychologischen Tests, die von gesunden Probanden vollständig gemeistert werden und in Stichproben gesunder Probanden keinerlei Varianz aufweisen, zur Beantwortung der Fragestellungen dieser Arbeit nicht geeignet. Letztlich verbleibt somit die Frage nach der Validität neuropsychologischer Testverfahren zur Intelligenztestung. Zentrale Validitätsaspekte von Intelligenztests sind nach Flanagan und McGrew (1997) die Gültigkeit von Testwerten als Indikatoren wichtiger kognitiver Konstrukte (Konstruktvalidität) und die Eignung der Testwerte zur Vorhersage gegenwärtiger und zukünftiger Leistungen (Kriteriumsvalidität). Die Validität einer neuropsychologischen Testbatterie zur Intelligenzprüfung soll in Bezug auf den Teilaspekt der Konstruktvalidität im empirischen Teil dieser Arbeit untersucht werden.

2.4.1 Intelligenz als Organismusvariable

In allgemeinen Lehrbüchern zur klinischen Neuropsychologie (z. B. Hartje & Poeck, 2002; Sturm, Herrmann & Wallesch, 2000) findet Intelligenz nur beiläufig Erwähnung. Daraus kann geschlossen werden, dass dem Konzept der Intelligenz in der klinischen Neuropsychologie lediglich eine Nebenrolle zugeordnet wird. In dem Kapitel über „Aufgaben und Strategien neuropsychologischer Diagnostik“ geht Sturm (2000) auf die Erfassung des „intellektuellen Niveaus und Leistungsprofils“ ein:

> *Eine Erfassung des intellektuellen Niveaus eines Patienten dient in der Regel zur Beurteilung des intellektuellen Hintergrunds, vor dem spezifischere Funktionen wie Gedächtnis und Aufmerksamkeit berücksichtigt werden sollen. Wichtiger als die Bestimmung des allgemeinen Intelligenzniveaus ist es für die neuropsychologische Diagnostik jedoch, auch speziellere kognitive Leistungen des Patienten zu analysieren. Aufschlüsse darüber ergeben sich bereits aus einer Profilanalyse der verschiedenen Untertestleistungen in den gängigen Intelligenz-Testverfahren (z. B. HAWIE-R, LPS, LPS 50+, IST-70) und in einigen anderen, wie dem Berufseignungstest (BET) oder dem Wilde-Intelligenz-Test (WIT). Die meisten dieser Testverfahren enthalten Untertests, die spezielle Funktionen, wie ‚visuelle Auffassungsgeschwindigkeit', ‚räumliche Orientierungs- und Vorstellungsfähigkeit', ‚sprachgebundenes und sprachunabhängiges logisches Denken', ‚Wortflüssigkeit', ‚Form- oder Gestalterfassung' prüfen, welche durch fokale Hirnschädigungen selektiv beeinträchtigt sein können. (S. 267)*

Hieraus lassen sich typische Grundannahmen für die klinisch-neuropsychologische Diagnostik entnehmen:

- Das intellektuelle Niveau bildet den Hintergrund für die Interpretation spezifischerer Funktionen. Intelligenz wird also nicht als Gegenstand einer neuropsychologischen Untersuchung an sich, so wie beispielsweise das Gedächtnis, betrachtet.
- Stattdessen nimmt Intelligenz die Rolle einer Hintergrundvariablen ein, die eine relativ stabile Eigenschaft einer Person widerspiegelt und potenziell andere Messergebnisse beeinflussen kann. Somit ähnelt Intelligenz anderen Organismusvariablen wie Alter oder Geschlecht.
- Das „allgemeine Intelligenzniveau" wird implizit im Sinne eines Generalfaktors konzipiert und steht somit in Widerspruch zu der in der Neuropsychologie sonst üblichen modularen Sichtweise und in Widerspruch zu den aktuellen hierarchischen Gruppenfaktormodellen der Intelligenz.
- Zwischen „spezifischeren Funktionen" und „allgemeinem Intelligenzniveau" wird unterschieden. Unter spezifischeren Funktionen werden speziellere kognitive Leistungen, die durch fokale Hirnschädigungen selektiv beeinträchtigt werden können, verstanden. Somit wird die selektive Störbarkeit der Funktionen durch fokale Hirnschädigungen als Validitätskriterium für die Eignung von Testverfahren für neuropsychologische Untersuchungen impliziert.
- Intelligenztestverfahren sind insbesondere dann interessant, wenn sie Untertests enthalten, die diese spezielleren Funktionen abbilden.

Anderson (2005) diskutiert im Editorial zum Schwerpunktthema *general intelligence* der Zeitschrift *Cortex*, warum Intelligenz in der Neuropsychologie so selten thematisiert wird. Intelligenz werde allgemein als die Summe der zur Kognition beitragenden Teile aufgefasst. Aber nicht die Summe, sondern die Teile an sich seien von neuropsychologischem Interesse. So stehe der in der Neuropsychologie dominante modulare Ansatz im Gegensatz zur Annahme eines einheitlichen Konstruktes (*general intelligence*). Aktuell scheine es allerdings zu einer Wiederentdeckung des Konstruktes der globalen Intelligenz zu kommen. Diese Wiederentdeckung wird mit der Erkenntnis begründet, dass es neben den speziellen Mechanismen, die spezifische kognitive Funktionen stützen, auch allgemeine Eigenschaften von Denken und Problemlösen gibt. Dabei bezieht sich Anderson insbesondere auf die fluide Intelligenz. Die Sichtweise gebe es in der Neuropsychologie allerdings schon seit über zwanzig Jahren, "*masquerading as ‚executive functioning'*" (Anderson, 2005, S. 10).

2.4.2 Fluide Intelligenz und exekutive Funktionen

Indem gezeigt wurde, dass die Beziehung zwischen fluider Intelligenz und globaler Intelligenz so stark ist, dass diese beiden Konstrukte kaum voneinander zu trennen sind (Gustafsson, 1988), ist das Konstrukt der fluiden Intelligenz in den Mittelpunkt der Intelligenzforschung getreten. Verglichen mit der globalen Intelligenz im Sinne eines g-Faktors kann die fluide Intelligenz definitorisch deutlich besser eingegrenzt und somit besser operationalisiert werden. Die klassische Auffassung nach Cattell (1971) konzipiert fluide Intelligenz als die Fähigkeit, sich neuen Problemen oder Situationen anzupassen, ohne dass es dazu im wesentlichen Ausmaß früherer Erfahrungen bedürfe (Amelang & Bartussek, 2001). Die Grundkapazitäten sind dabei schlussfolgerndes Denken und Problemlösen. Inhalte typischer Tests zur Erfassung der fluiden Intelligenz wären Analogieaufgaben, Reihenvervollständigungen oder Aufgaben zum abstrakten schlussfolgernden Denken; ein typisches Testverfahren wäre der progressive Matrizentest von Raven (Raven, 1998). Kognitionspsychologisch kann fluide Intelligenz mit Metakognition (Wissen und Reflektion über die eigenen mentalen Prozesse) und Arbeitsgedächtnis (aktives Aufrechterhalten bereichsspezifischer Informationen und bereichsübergreifende Aufmerksamkeitskontrolle der Informationsverarbeitung) in Verbindung gebracht werden (Gray, Chabris & Braver, 2003). Diese kognitive Dekomposition des Konstruktes der fluiden Intelligenz legt die Gleichstellung von fluider Intelligenz mit dem klassischen neuropsychologischen Konstrukt der „exekutiven Funktionen" nahe. Neuropsychologisch werden exekutive Funktionen als kognitive Prozesse beschrieben, die dazu dienen, „Handlungen über mehrere Teilschritte hinweg

auf ein übergeordnetes Ziel zu planen, die Aufmerksamkeit auf relevante Informationen zu fokussieren und ungeeignete Handlungen zu unterdrücken“ (Karnath & Sturm, 2002, S. 393) Subsumiert werden beispielsweise die Prozesse Problemlösen, mentales Planen und Initiieren und Inhibition von Handlungen.

Trotz dieser Nähe des zentralen Konzeptes der fluiden Intelligenz zu neuropsychologisch beschreibbaren Prozessen erweist es sich als schwierig, mittels neuropsychologischer Testbatterien einen Faktor der exekutiven Funktionen zu erfassen, dessen Stärke und Dominanz vergleichbar ist mit dem über Intelligenztests erfassbaren Faktor der fluiden Intelligenz (Miyake et al., 2000): Dies liegt zum einen an der Heterogenität der exekutiven Funktionen. Diesen Funktionen werden so unterschiedliche Teilkomponenten wie Konzeptwechsel, Monitoring oder Inhibition zugeordnet, ohne dass entschieden wäre, ob diese als Einheit oder als Repräsentation eines gemeinsamen Mechanismus betrachtet werden können. Zudem kommt der exekutiven Kontrolle eine Sonderrolle zu, weil sie über verschiedene Leistungsbereiche und Modalitäten hinweg operiert. Aufgrund dieser Modalitätsheterogenität besteht die Gefahr, dass die Unterschiedlichkeit der exekutiven Tests größer ist als ihre Gemeinsamkeit. Diese Eigenschaft der exekutiven Funktionen benennt das „*impurity problem*“: *”Commonly used executive tasks are highly complex and typically place heavy demands on not just executive processes of interest, but also nonexecutive processes within which the executive processing requirement is embedded*.“ (Miyake et al., 2000, Seite 90).

Auch Duncan (2003) betrachtet globale Intelligenz nicht als diffuse neurophysiologische Eigenschaft, sondern verweist auf die Zentralität exekutiver Funktionen für Problemlöseprozesse (*problem solving*): So verbinden die fluide Intelligenz und die generelle Intelligenz nicht nur gemeinsame Prozesse, wie Aufmerksamkeit, Inhibition und Arbeitsgedächtnis, sondern auch gemeinsame Strukturen, um diese Prozesse zu kontrollieren: Insbesondere der präfrontale Kortex ist für die fluide Intelligenz relevant. Duncan begründet dies damit, dass dessen Strukturen und Neurone das ideale Arbeitsumfeld für Verarbeitung neuer Aufgaben, unabhängig von der Aufgabenart darstellen.

Andererseits wurde schon früh dargestellt, dass selbst ausgedehnte frontale Läsionen Intelligenztestleistungen kaum beeinflussen (z. B. Hebb, 1939, zitiert nach Ardila, 1999). Ardila (1999) berichtet von sehr niedrigen Korrelationen zwischen Intelligenztests der Wechsler-Familie und exekutiven Testaufgaben. Dies steht im Widerspruch zum ansonsten robusten Befund substantieller Interkorrelationen zwischen allen kognitiven Leistungstests. Er schließt daraus, dass in den klassischen verfügbaren

Intelligenztests die exekutiven Funktionen nicht erfasst werden. Da die exekutiven Funktionen ein zentrales, für intelligentes Verhalten unumgängliches Konzept sind, eignen sich diese Tests nicht zur Intelligenzerfassung, folgert Ardila weiter. Ähnlich schreibt Kaufman (2000), dass der verbale Intelligenzquotient der Wechsler-Intelligenztests zwar ein gutes Maß für die kristalline Intelligenz darstellt, fluide Intelligenz in den älteren Versionen jedoch nicht erfasst wurde. Die Handlungsintelligenz ist eher ein Maß für visuo-räumliche Intelligenz (Stone, 1992, zitiert nach Kaufman, 2000). Allenfalls enthält die Aufgabe Mosaiktest eine klare Problemlösungskomponente. In die neueste Version des Wechsler-Intelligenztests (Wechsler, 1997a) wurde eine Matrizenaufgabe zur Erfassung fluider Intelligenz zusätzlich aufgenommen.

Zusammenfassend geht aus dem bisher Gesagtem hervor, dass aus neuropsychologischer Sicht die globale Intelligenz eine vergleichsweise uninteressante Information darstellt. Interessanter wird das Konstrukt der Intelligenz bei Anwendung neuerer Intelligenzkonzeptionen (z. B. Horn, 1991), die verschiedene Teilkomponenten erfassen, darunter die fluide Intelligenz. Bezüglich solcher Intelligenzkonzeptionen ist die Nähe zwischen hierarchischen Gruppenfaktoren der Intelligenz und zentralen neuropsychologischen Fähigkeitskonstrukten offensichtlich. Ähnliches gilt für die Testung von Intelligenz: Indem neuere Versionen der Intelligenztests die aktuellen hierarchischen Intelligenzmodelle operationalisieren und somit auch beispielsweise die fluide Intelligenz erfassen, steigt der Nutzen der Intelligenztests für die Neuropsychologie, da als Ergebnis einer aufwändigen Intelligenztestung nicht nur ein Gesamtwert, sondern auch detaillierte, klinisch verwertbare und sehr gut validierte Informationen über relevante Fähigkeitskonstrukte herauskommen.

2.4.3 Intelligenz als abhängige Variable

In neuropsychologischen Studien kommt der Intelligenz häufig die Rolle einer abhängigen Variablen zu. Zumeist definieren die Autoren jedoch nicht, was sie konkret unter Intelligenz verstehen und auf welches Intelligenzmodell sie sich beziehen. So gilt also das wohlbekannte Diktum, dass Intelligenz das ist, was der Intelligenztest misst (Boring, 1923). Deshalb sind viele Studien nicht vergleichbar. Ein weiteres Problem ergibt sich aus der nahezu monopolistischen Stellung der Wechsler-Intelligenztests in epileptologischen Studien zur Intelligenz, da weder die zweifaktorielle Struktur der älteren Testversionen noch die vier Faktoren der neueren Version den modernen Intelligenzstrukturmodellen entsprechen. Während eine Literaturrecherche in der elektronischen Datenbank MEDLINE ("Pubmed", 2006) für die Suchabfrage „*epilepsy intelligence Wechsler*“ (jeweils in Titel oder Zusammenfassung) immerhin 110 Treffer

lieferte, finden sich für die verknüpfte Suche „*epilepsy intelligence*“ mit *Horn*, *Cattell* oder *Carroll* keine Treffer. Eine als Abstract veröffentlichte Vorläuferstudie dieser Arbeit scheint die einzige Studie zu sein, die die Frage der Intelligenz bei Patienten mit Epilepsie explizit in Bezug auf aktuelle Intelligenzstrukturmodelle zu beantworten versucht. Die Studie setzt die Methode der explorativen Faktorenanalyse ein (Lutz & Helmstaedter, 2004).

Inhaltlich geht es in epileptologischen Studien zur Untersuchung von Intelligenz fast ausschließlich um die Frage, ob und inwieweit krankheitsspezifische Faktoren (Anfälle, Status, Medikamente, Operationen, strukturelle Veränderungen, Medikation, Krankheitsbeginn, Epilepsiesyndrome) die Intelligenzleistungen bei Erwachsenen und Kindern mit Epilepsie, oder bei Kindern von Müttern mit Epilepsie, beeinflussen.

Da die aktuellen Modelle sich um eine umfassende Berücksichtigung und Taxonomie aller kognitiven Teilleistungen bemühen, impliziert deren konsequente Anwendung die Abkehr von einer Sichtweise, in der Intelligenz lediglich als relativ zeitkonstante Organismusvariable betrachtet und von spezielleren kognitiven Teilfunktionen abgegrenzt wird. Dadurch lassen sich verschiedene methodische Probleme lösen, so zum Beispiel das Interpretationsproblem, welches sich bei hohen Streuungen in einzelnen Intelligenzuntertests ergeben kann und die Interpretation globaler Fähigkeitskonstrukte erschwert bzw. verhindert. Auch wird dadurch die Schwierigkeit der definitorischen Abgrenzung ähnlicher Konstrukte wie *Intelligenz* oder *globales kognitives Leistungsniveau* überbrückt.

2.5 Der „*cross-battery approach*“

In Kapitel 2.3 wurden relevante Intelligenzstrukturmodelle beschrieben. Insbesondere die Modelle von Horn, Cattell und Carroll sowie das neuere Cattell-Horn-Carroll-Modell sind empirisch gut fundiert und spiegeln den aktuellen Stand der Intelligenzstrukturforschung wider. Allerdings ist keine Intelligenztestbatterie verfügbar, die die relevanten Gruppenfaktoren dieser Modelle angemessen operationalisiert. Der “*cross-battery approach*” soll die zwischen Theorie und Praxis bestehende Lücke überwinden. Die folgende kurze Einführung in diesen Ansatz zur Intelligenztestung basiert im Wesentlichen auf den grundlegenden Texten von Flanagan und McGrew (1997), Ortiz und Falangan (2002) sowie auf der im Internet verfügbaren Informationssammlung (Flanagan & Ortiz, 2006)[2].

[2] Hier finden sich vielfältige Informationen zur Theorie und Praxis dieses testpsychologischen Ansatzes, unter anderem PowerPoint-Präsentation sowie Auswertungsbögen und -tabellen für die konkrete Anwendung.

Dieser testpsychologische Ansatz überschreitet die Grenzen zwischen den herkömmlichen Testbatterien und ermöglicht eine individuelle Zusammenstellung von Einzeltests zur besseren Erfassung relevanter Fähigkeitsdimensionen. Durch entsprechende Studien wird gewährleistet, dass die Zusammenstellung einer neuen Testbatterie aus den verschiedenen Intelligenztests systematisch und empirisch basiert vollzogen werden kann. Im Wesentlichen sind folgende Schritte auszuführen:

- Zunächst sind anhand eines Intelligenzstrukturmodells (namentlich des CHC-Modells) die relevanten breiten Fähigkeitsdimensionen, die erfasst werden sollen, auszuwählen.
- Zur Operationalisierung dieser Faktoren sind Einzeltests so zusammenzustellen, dass die konstruktirrelevante Varianz minimiert wird und gleichzeitig eine Unterrepräsentation des Konstruktes vermieden wird. Um ersteres zu erreichen, wurden die verschiedenen Einzeltests der verfügbaren Intelligenztestbatterien nach dem erfassten Konstrukt und nach der Stärke, mit der die Tests Indikatoren dieses Konstruktes sind, klassifiziert. Um eine Unterrepräsentation zu vermeiden, sind mindestens zwei Indikatoren für unterschiedliche enge Faktoren, die dem breiten Fähigkeitsfaktor untergeordnet sind, notwendig.
- Anhand dieser Testauswahl können dann die Tests durchgeführt werden und die Ergebnisse auf eine gemeinsame Skala (Mittelwert 100, Standardabweichung 15) gebracht werden.
- Zur Ergebnisdarstellung werden die Einzelergebnisse gegliedert nach den jeweiligen Fähigkeitsdimensionen in ein Diagramm eingetragen. Anhand eines geschätzten Standardfehlers wird zusätzlich das Konfidenzintervall angegeben. Ferner wird pro Fähigkeitsdimension der Mittelwert errechnet.
- Zur Ergebnisinterpretation sind die dimensionsweise zusammengefassten Ergebnisse und nicht die Einzelwerte heranzuziehen.

Dieses Vorgehen fügt Tests zusammen, die zu verschiedenen Zeiten anhand unterschiedlicher Stichproben normiert wurden. Aus diesem Vorgehen sollten sich keine methodischen Probleme ergeben, sofern jede Testbatterie für sich anhand einer sorgfältig zusammengestellten repräsentativen Stichprobe normiert wurde. Das einzige relevante Problem könnte der Flynn-Effekt darstellen, also die ständige Steigerung der Intelligenztestleistungen einer Population über die Zeit hinweg (Flynn, 1987). Dieser kann aber durch Ausschluss von Verfahren, deren Normierung älter als zehn Jahre ist, minimiert werden. Auch aufgrund potenzieller Reihenfolgeeffekte sind nicht mehr Verzerrungen zu erwarten als bei jeder anderen, aus verschiedenen Einzelverfahren zusammengestellten (neuro-)psychologischen Untersuchung.

3 Die konfirmatorische Faktorenanalyse

Die konfirmatorische Faktorenanalyse ist die zentrale Analysemethode dieser Arbeit. Konfirmatorische Faktorenanalyen sind theoriegeleitet und hypothesentestend und überwinden somit verschiedene Schwächen der explorativen Faktorenanalysen. Das folgende Kapitel dient der Darstellung dieses relativ neuen Verfahrens.

Es ist hinreichend bekannt und beschrieben, dass die gewählte Methodik der explorativen Faktorenanalyse die Faktorenlösung und damit auch die postulierten Intelligenzstrukturmodelle beeinflusst (z. B. Amelang & Bartussek, 2001; Bortz, 1993). Mit der konfirmatorischen Faktorenanalyse steht ein Verfahren zur Verfügung, mit dem die Strukturmodelle, die ursprünglich mit unterschiedlichen explorativen Faktorenanalysen entwickelt wurden, in einheitlicher Weise nachmodelliert und anhand empirischer Daten miteinander verglichen werden können. Konfirmatorische Faktorenanalysen sind sehr flexibel einsetzbar: Im einfachsten Falle können die Vorhersagen eines Modells mit den Daten einer Stichprobe abgeglichen werden. Darüber hinaus lassen sich auch Daten mehrerer Stichproben (Mehr-Gruppen-Fall der konfirmatorische Faktorenanalyse) oder Daten einer Stichprobe, die zu unterschiedlichen Zeitpunkten erfasst worden sind, auf Modellkonformität überprüfen (Rietz, 1996).

Der empirische Untersuchungsgegenstand konfirmatorischer Faktorenanalysen ist im Wesentlichen der Zusammenhang verschiedener Testleistungen untereinander, also die empirische Varianz-Kovarianzmatrix[3]. Darin unterscheiden sich konfirmatorische Faktorenanalysen nicht von explorativen Faktorenanalysen. Der Unterschied liegt im Ausgangspunkt: Konfirmatorische Faktorenanalysen basieren auf Theorien, die Angaben zur Faktorenstruktur (z. B. hierarchisch, oblique oder orthogonal), zur Faktorenanzahl und möglichst auch zur Zuordnung der auf den Faktoren ladenden Variablen machen. Anhand dieser Angaben können faktorielle Modelle spezifiziert werden, deren Gültigkeit durch einen Vergleich der empirischen Datenmatrix und der aufgrund der Modelleigenschaften vorhergesagten (impliziten oder reproduzierten) Matrix getestet und mittels Indizes der Anpassungsgüte quantitativ bewertet werden kann. Dabei muss das Modell die Daten nicht exakt reproduzieren, vielmehr werden minimale Abweichungen zwischen empirischer und vorhergesagter Datensituation als Stichprobenfehler toleriert. Die grundlegende Logik der konfirmatorischen Faktorenanalyse ist, dass die spezifizierten Modelleigenschaften bestimmte

[3] Werden zusätzlich Annahmen bezüglich der latenten Mittelwerte getroffen, so erweitert sich der empirische Untersuchungsgegenstand um die Mittelwerte der beobachteten Variablen.

Datenkonstellationen der empirischen Varianz-Kovarianzmatrix erwarten lassen. Finden sich die theoretisch angenommenen Modelleigenschaften nicht in der empirischen Realität wieder, kann nicht von Modellgültigkeit ausgegangen werden. Beispielsweise sollten sich bei einem zweifaktoriellen Modell in der empirischen Datenmatrix zwei Cluster mit höheren Interkorrelationen zeigen.

Die Trias *Theorieauswahl, Modellaufstellung* und *empirische Überprüfung des Modells* ist kennzeichnend für konfirmatorische Faktorenanalyse, das konkrete Vorgehen wird von Bollen und Long (1993, zitiert nach Kelloway, 1998) in die Phasen *Modellspezifikation, Modellidentifikation, Schätzung, Testung der Anpassungsgüte* und *Respezifikation* gegliedert. Diese fünf grundlegenden Analyseschritte sollen im Folgenden kurz dargestellt werden. Ausführliche Darstellungen linearer Strukturgleichungsmodelle, zu denen als Spezialfall die konfirmatorische Faktorenanalyse gehört, finden sich in den Lehrbüchern von Raykov (2000) oder Kelloway (1998).

3.1 Theorieauswahl und Modellspezifikation

Wie erwähnt, steht am Anfang jeder konfirmatorischen Faktorenanalyse eine Theorie. Ein entscheidender Zwischenschritt, um die fragliche Theorie empirisch überprüfbar zu machen, ist die Modellspezifikation. Dabei wird die Theorie in ein faktorielles Modell überführt, indem man die Beziehungen der Faktoren untereinander und die Beziehungen der Faktoren zu den Indikatoren definiert. Da allerdings die Übersetzung einer Theorie in ein Modell aufgrund häufig unpräzise formulierter psychologischer Theorien schwierig sein kann, haben Jöreskog und Sörbom (1993) drei Situationen unterschieden (zitiert nach Raykov & Marcoulides, 2000): Im streng konfirmatorischen Vorgehen wird ein einzelnes Modell aufgestellt und dieses anhand der Ergebnisse der konfirmatorischen Faktorenanalyse entweder akzeptiert oder verworfen (*strictly confirmatory situation*). Weniger strikt ist ein Vorgehen, bei dem verschiedene Modelle aufgestellt und getestet werden und anhand der Ergebnisse die Entscheidung für eines dieser Modelle getroffen wird (*alternative-models or competing-models situation*). Schließlich kann ein anfänglich aufgestelltes Modell sukzessive modifiziert und getestet werden, bis eine hinreichende Passung besteht (*model-generating situation*).

Im engeren Sinne ist nur die erste Situation als konfirmatorisch zu werten, die zweite und dritte Situation haben mehr oder weniger starke explorative Komponenten – auch wenn die zugrunde liegende Methodik jeweils die konfirmatorische Faktorenanalyse ist. Keith (1997) merkt an, dass auch explorative Faktorenanalysen theoriegeleitet und hypothesentestend eingesetzt werden können und konfirmatorische Faktorenanalysen

explorativ. Im Allgemeinen gelte aber für konfirmatorische Faktorenanalysen (*confirmatory factor analysis*, CFA): „*Nevertheless, the simple fact that CFA require the specification of a model – and thus knowledge about the probable structure of the characteristic being measured – means that some sort of theory, formal or informal, strong or weak, is required*." (S. 374)

Wichtige Elemente eines faktoriellen Modells sind die Faktoren (latente Variablen) und deren Beziehung untereinander sowie die Testindikatoren, die zur Messung der Faktoren verwendet wurden (manifeste Variablen). Die Unterscheidung zwischen latenten und manifesten Variablen ist zentral für Strukturgleichungsmodelle: Latente Variablen repräsentieren hypothetische oder theoretische Konstrukte. Sie sind nicht direkt beobachtbar, aber ihre Manifestation kann durch Testung bestimmter Konstrukteigenschaften mit geeigneten Instrumenten erfasst werden (Raykov & Marcoulides, 2000). Die manifesten Variablen (Indikatoren, Tests) sind im Gegensatz zu den Konstrukten fehleranfällig und unreliabel. Daher ist jedes Konstrukt möglichst durch mehrere Indikatoren zu operationalisieren.

Pfaddiagramme geben grafisch die relevanten Elemente eines Faktorenmodells wieder. Ihre Darstellung ist durch eine Notationskonvention vereinheitlicht. Wichtige grafische Elemente sind zunächst Ellipsen und Vierecke, die latente bzw. manifeste Variablen repräsentieren. Beziehungen zwischen den Variablen werden durch Pfeile ausgedrückt. Pfeile mit einer Pfeilspitze zeigen einen gerichteten Zusammenhang an (Regressionskoeffizient, Ladung). Das heißt, die Variable, die den Ausgangspunkt des Pfeils darstellt, beeinflusst die Variable, auf die der Pfeil zeigt. Im konfirmatorischen faktorenanalytischen Modell werden die latenten Variablen, also die Faktoren, als Ursache für die Ausprägung der manifesten Variablen betrachtet. Pfeile mit zwei Spitzen drücken Kovariationen (Korrelationen) zwischen zwei Variablen aus und zeigen an, dass eine ungerichtete Beziehung zwischen diesen Variablen besteht. Doppelpfeile sind nur zwischen latenten Variablen zulässig. Eine durch einen Doppelpfeil modellierte Beziehung zwischen zwei Faktoren entspricht der Annahme obliquer Faktoren. Orthogonale Faktorenmodelle werden ohne Doppelpfeile zwischen den latenten Variablen dargestellt. Korrelative Beziehungen zwischen beobachteten Variablen (Tests) dürfen nicht mit Doppelpfeilen angezeigt werden, sondern müssen durch andere Modellspezifikationen wie Kovarianzen zwischen den zugehörigen Faktoren oder gemeinsame Ladungen auf einem Faktor ausgedrückt werden. Doppelpfeile können aber auch korrelierte Fehlerterme (Residualvarianzen) der Tests indizieren, da die Fehlerterme – entsprechend der Annahmen der klassischen Testtheorie – keine direkt beobachtbaren Größen sind und somit latente Variablen darstellen. Inhaltlich drücken

solche korrelierten Messfehler gemeinsame Methodenvarianz aufgrund identischer Testmethodik aus. Beispielsweise könnten die Indikatoren der sofortigen und verzögerten Wiedergabe eines Gedächtnistests über korrelierte Fehlerterme miteinander verbunden sein, da spezifische Varianzquellen, die den sofortigen Abruf beeinflussen, aufgrund des identischen Testmaterials und des identischen Testsettings sehr wahrscheinlich auch den verzögerten Abruf beeinflussen werden. Fehlerkovarianzen sollten sehr sparsam und nur mit einer guten theoretischen Begründung postuliert werden. Rietz (1996) merkt an, dass die Aufnahme korrelierter Messfehler geeignet ist, durch gezielte Modellmodifikationen die Modelle zu verbessern – entsprechend häufig seien korrelierte Messfehler in Publikationen zu finden: "*Fast jedes Strukturgleichungsmodell kann angepasst werden, wenn hinreichend Kovarianzen ... freigesetzt* werden" (S. 45). Eine derartige Überanpassung eines Modells an die Stichprobendaten birgt aber die Gefahr einer zu hohen Stichprobenspezifität und einer zu geringen Generalisierbarkeit.

Abbildung 3-A soll exemplarisch anhand eines obliquen und eines orthogonalen Faktorenmodells die relevanten Elemente faktorenanalytischer Pfaddiagramme darstellen.

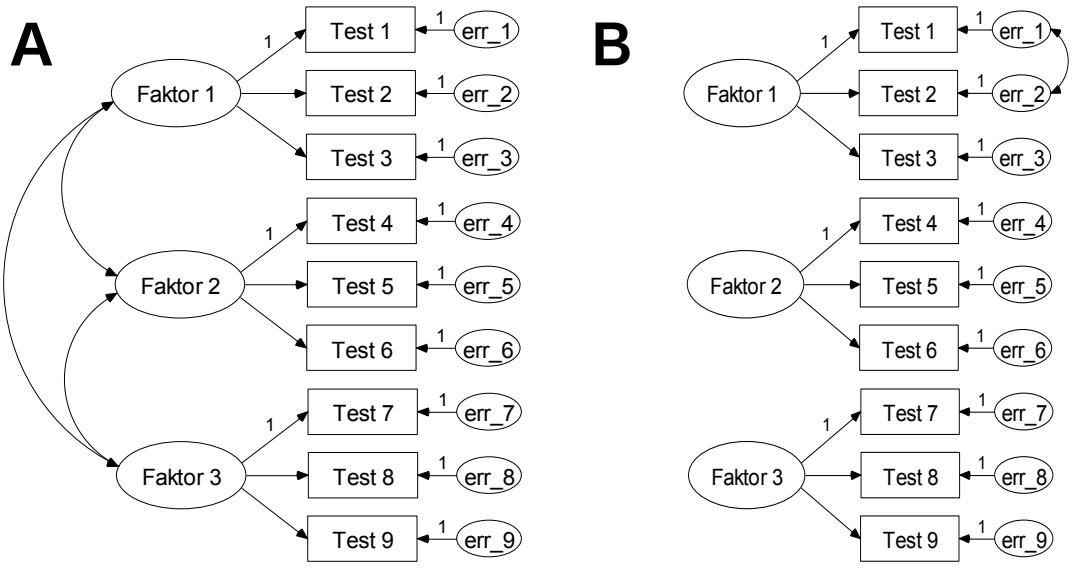


Abbildung 3-A: Zwei faktorenanalytische Modelle, dargestellt als Pfaddiagramme. Modell A beinhaltet drei oblique Faktoren (Ellipsen), die durch jeweils drei Tests (Vierecke) operationalisiert werden. Modell B impliziert drei orthogonale Faktoren. In Modell B wurde zusätzlich eine Fehlerkovarianz zwischen Test 1 und 2 aufgenommen.

Faktorenanalytische Modelle setzten sich aus einem Messmodell und einem Strukturmodell zusammen. Das Messmodell beinhaltet die Verknüpfung von latenten

Variablen mit ihren Indikatoren und entspricht der Operationalisierung der Konstrukte. Strukturmodelle beinhalten Verknüpfungen der latenten Variablen untereinander. Modellparameter sind die Parameter des Mess- oder Strukturmodells, die durch konfirmatorische Faktorenanalysen geschätzt werden. Bezogen auf das Messmodell sind das die Ladungen, Höhenlagen (engl.: *intercepts*), Fehlervarianzen und Fehlerkovarianzen, und bezogen auf das Strukturmodell die Faktorinterkorrelationen, Faktorvarianzen und Faktorenmittelwerte.

Um zu untersuchen, ob das Faktorenmodell die empirische Kovarianzmatrix genau reproduziert, muss das Pfaddiagramm in ein lineares Gleichungssystem transformiert werden. Strukturgleichungen drücken die Zerlegung der Varianzen und Kovarianzen der manifesten (beobachteten) Variablen in hypothetische Varianzquellen latenter Variablen aus. Die unbekannten Größen in diesem Gleichungssystem sind die Modellparameter. Diese werden anhand der empirischen Daten geschätzt. Die notwendige Transformation des Pfaddiagramms in ein Gleichungssystem übernehmen in der praktischen Anwendung entsprechende Computerprogramme (z. B. das Programm AMOS von Arbuckle, 1997). Hier werden daher nur überblicksartig in Anlehnung an Lubke, Dolan, Kelderman und Mellenbergh (2003a) und Rietz (1996) die wichtigsten Gleichungen dargestellt. Dies ist nicht verzichtbar, da aus diesen Gleichungen im zentralen Teil dieser Arbeit die zu überprüfenden Invarianzhypothesen abgeleitet werden.

Der beobachtete Testwert einer Person kann im Rahmen eines faktoriellen Modells in drei Teilgrößen zerlegt werden: Die erste Teilgröße ist die testspezifische Konstante ν_i. Gäbe es in den Testleistungen keinerlei Varianz zwischen den Probanden einer Stichprobe, so wäre diese Konstante allein hinreichend, um die Testwerte jeder Person vorherzusagen. Bei interindividueller systematischer Varianz im Test kann ein faktorielles Modell den Anteil der Testleistung, der nicht auf die testspezifische Konstante zurückzuführen ist, voraussagen. Der Anteil wird mithilfe der Produktsumme aus den personenspezifischen Faktorwerten η_{jl} der Person j bezüglich der l Faktoren und der zugehörigen personenunabhängigen Faktorladung λ_{il} ausgedrückt. Die Faktorwerte repräsentieren das Leistungsniveau des Probanden im jeweiligen für die Testleistung relevanten Konstruktbereich. Schließlich ist als dritter Anteil der individuellen Testleistung ein Residualterm ε_{ij} anzunehmen, der den testspezifischen Fehler und den zufälligen Messfehler einer Person umfasst. Folgende lineare Strukturgleichung (1) beschreibt die Zerlegung des im Test i erreichten Wert y der Person j.

$$y_{ij} = \nu_i + \sum_{l=1}^{L} \lambda_{il} \eta_{jl} + \varepsilon_{ij} \quad (1)$$

Die Interpretation des Residualwertes ε_{ij} ist allerdings kontrovers: Teilweise wird er als Summe aus rein testspezifischen Fehlern und zufälligen Messfehlern interpretiert und somit als Funktion der Reliabilität des Test aufgefasst. Teilweise gilt er auch als nicht modellierte Varianz, die durch Aufnahme weiterer Modellfaktoren grundsätzlich noch reduzierbar wäre. Auf dieses Problem wird später im Zusammenhang mit den Invarianzhypothesen (strenge metrische Invarianz) noch einzugehen sein. Die Möglichkeit, Messfehler explizit zu berücksichtigten sowie Methodenvarianz infolge gemeinsamer Messmethoden verschiedener Untertests in den Modellspezifikationen über korrelierte Messfehler gezielt zu modellieren (Wilde et al., 2003), ist ein Vorteil der konfirmatorischen Faktorenanalyse im Vergleich zur explorativen Faktorenanalyse.

Das faktorenanalytische Modell zur Beschreibung von Gruppendaten geht nicht mehr von den individuellen Rohwerten, sondern von den Mittelwerten und Kovarianzen der beobachteten Variablen aus. Gleichung (2) gibt die Formel für den Mittelwert des Tests i über alle J Personen einer Gruppe wieder. Der Messfehler (siehe Gleichung (1)) mittelt sich entsprechend der Axiome der klassischen Testtheorie hinaus, so dass sich der Mittelwert aus der testspezifischen und daher personenunabhängigen Konstante und aus der Produktsumme aus den Faktorladungen λ_{il} und den mittleren Faktorwerten α_l zusammensetzt:

$$\mu_i = \nu_i + \sum_{l=1}^{L} \lambda_{il} \alpha_l \quad (2)$$

Bei *I* Tests einer Testbatterie, liegen *I* solcher Gleichungen vor, die vereinfachend in Matrizenschreibweise umformuliert werden können:

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \\ \vdots \\ \nu_I \end{bmatrix} + \begin{bmatrix} \lambda_{1l} & \cdots & \lambda_{1L} \\ \lambda_{2l} & \cdots & \lambda_{2L} \\ \vdots & \ddots & \vdots \\ \lambda_{Il} & \cdots & \lambda_{IL} \end{bmatrix} \times \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_L \end{bmatrix} \quad (3)$$

bzw. $\mu_s = \nu_s + \Lambda_s \alpha_s$ (4)

In Gleichung (4) wurde für den Mehr-Gruppen-Fall der Gruppenindex s eingeführt. Λ ist die Faktorladungsmatrix. Ähnlich dieser Aufspaltung der Testmittelwerte können auch die Varianzen und Kovarianzen der beobachteten Variablen durch die zugrunde liegenden Faktoren und Residuen ausgedrückt werden. Die Gleichung für die empirische Varianz-Kovarianzmatrix Σ aller Tests lautet:

$$\Sigma_s = \Lambda_s \Psi_s \Lambda_s^t + \Theta_s \qquad (5)$$

Dabei ist Ψ die Varianz-Kovarianzmatrix der Faktoren, Θ ist die Varianz-Kovarianzmatrix der Residuen und Λ ist die Faktorladungsmatrix.

Lubke et al. (2003a) weisen darauf hin, dass sowohl die Gleichung für die Mittelwerte als auch die Gleichung für die Kovarianzen von der gleichen Ausgangsgleichung (1) abgeleitet werden können. Gleichung (1) gibt die Regression der beobachteten Testwerte auf die zugrunde liegenden Faktoren wieder, Gleichung (4) beschreibt die Mittelwerte der Tests anhand der Faktorenmittelwerte und Gleichung (5) beschreibt die (Ko-)Varianzen der Tests anhand der Kovarianzen der Faktoren. Die Matrix der Faktorladungen Λ ist jeweils gleich.

3.2 Modellparameter, Identifizierbarkeit und Parameterschätzung

Modellparameter drücken die unbekannten Aspekte eines linearen Strukturgleichungsmodells aus. Das gilt zumindest für den Anfang der Auswertung, später werden sie im Modelltest auf Basis der empirischen Daten mittels Optimierungskriterien und entsprechender Computerprogramme geschätzt. Die Qualität der Parameterschätzung kann anschließend bewertet werden, indem man die Kovarianzmatrix, die auf der Basis der geschätzten Parameter reproduziert wurde, mit der empirischen Kovarianzmatrix vergleicht.

Die oben dargestellten Gleichungen zeigen, welche Modellparameter zur Reproduktion der empirischen Daten notwendig sind. Für die empirische Varianz-Kovarianzmatrix sind dies die Faktorvarianzen und Faktorenkovarianzen (Matrix Ψ), die Faktorladungen (Matrix Λ) sowie die Fehlervarianzen und Fehlerkovarianzen (Matrix Θ); für die Reproduktion des empirischen Mittelwertvektors sind zusätzlich die Faktorenmittelwerte (Vektor α) und die testspezifischen Konstanten (Vektor ν) relevante Modellparameter. Abbildung 3-B stellt sämtliche Parameter, Matrizen und Vektoren eines dreidimensionalen Strukturmodells dar.

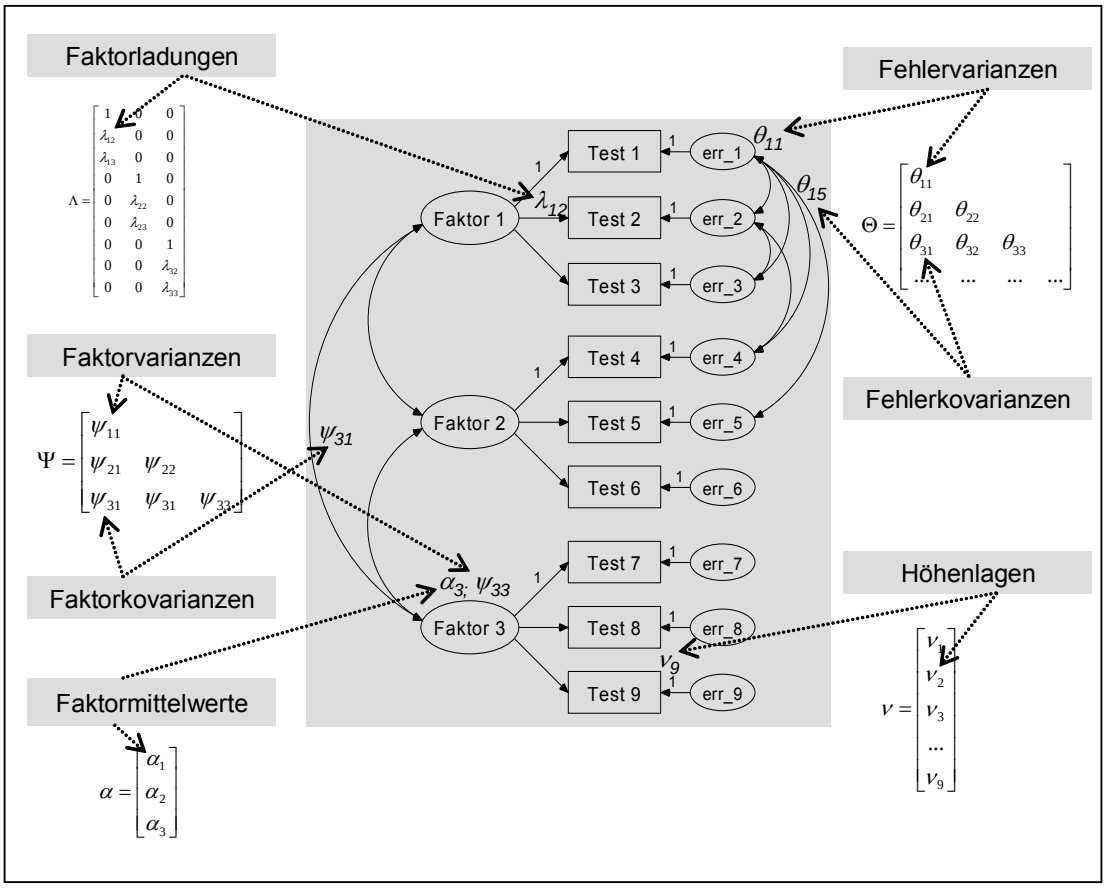


Abbildung 3-B: Überblick über die Modellparameter und die zugehörigen Matrizen und Vektoren eines dreifaktoriellen obliquen Strukturmodells.

Strukturgleichungsmodelle erfordern die Schätzung unbekannter Parameter anhand empirischer Daten und das ist nur möglich, wenn das komplexe lineare Gleichungssystem identifizierbar ist. Die Identifizierbarkeit eines Modells entspricht der Lösbarkeit des Gleichungssystems: Ein Modell ist nur dann lösbar, wenn genügend bekannte Größen vorhanden sind, also wenn die Anzahl der unbekannten Größen kleiner oder gleich der Anzahl der bekannten Größen ist. Bekannte Größen sind die empirischen Varianzen und Kovarianzen und Mittelwerte der Tests, unbekannt sind Modellparameter. Um Identifizierbarkeit zu erreichen, werden häufig Modellparameter fixiert, um die Anzahl der zu schätzenden Modellparameter zu verringern. Gibt es weniger zu schätzende Parameter als verfügbare empirische Größen, auf denen die Schätzung beruht, so ist das Modell überidentifiziert und besitzt Freiheitsgrade (FG).

Tatsächlich verhält es sich zumeist so, dass manche Modellparameter schon durch theoretische Annahmen festgelegt sind und nicht geschätzt werden müssen. Dies verbessert das Verhältnis zwischen bekannten und unbekannten Größen und begünstigt die Identifizierbarkeit eines Modells. Beispiele wären Nullladungen zwischen einem Faktor und einem konstruktfremden Test oder orthogonale Faktoren, bei denen die

Faktoreninterkorrelation gleich Null ist. Eine zweite Art festgelegter Parameter sind Parameter, denen nicht aus theoretischen, sondern aus psychometrischen Gründen zur Standardisierung und Normierung a priori ein konstanter Wert zugeordnet wird. Dies ist notwendig, da latente Variablen nicht direkt beobachtbar sind und somit auch keine definierte metrische Skala besitzen. Dabei ist es üblich, einen Pfadkoeffizienten (Ladung) pro Faktor auf den Wert 1 zu fixieren und damit der latenten Variablen die Skala des verknüpften Indikators zuzuordnen. Die Variable, deren Ladung auf den Wert 1 fixierte wurde, heißt Referenzvariable. Häufig wird dazu diejenige mit der höchsten Reliabilität ausgewählt. Außerdem können latente Variablen skaliert werden, indem man ihre Varianz auf 1 setzt (Standardisierung). Dieses Vorgehen bedeutet allerdings, dass bei Gruppenvergleichen die Varianz eines Faktors in beiden Gruppen identisch ist. Da dies nicht unbedingt angenommen werden kann, wird bei konfirmatorischen Faktorenanalysen für die Skalierung der Faktoren und der Fehlervarianzen der Indikatoren zumeist die erste Möglichkeit eingesetzt. Hauptziel bei diesem Vorgehen ist es, den latenten Variablen eine Skala zuzuordnen, gleichzeitig wird dadurch aber auch die Identifizierbarkeit eines Modells begünstigt.

Neben den Modellparametern, die im Rahmen der Modellformulierung aus theoretischen oder psychometrischen Gründen a priori festgelegt wurden, gibt es noch zwei weitere Arten von Modellparametern: die freien Parameter und die beschränkten Parameter. Sie werden anhand der empirischen Daten geschätzt. Freie Parameter können ohne jegliche Vorbedingungen frei geschätzt werden. Die Schätzungen der beschränkten Parameter müssen bestimmte numerische Bedingungen erfüllen: Ein Beispiel hierfür sind die Untersuchungen zur Invarianz bestimmter Parameter zwischen verschiedenen Gruppen. Dabei müssen bestimmte Parameter in beiden Gruppen gleiche Werte annehmen (das heißt, sie sind invariant zwischen den Gruppen). Die numerische Ausprägung dieses Wertes allerdings ist paarweise frei schätzbar. Durch den Vergleich zwischen einem mehr und einem weniger restringierten Modell kann überprüft werden, ob diese Restriktionen empirisch haltbar sind. Genau dies ist der Kern der im nächsten Kapitel vorzustellenden Invarianzhypothesen. Sie entsprechen den Restriktionen bestimmter Untergruppen der oben genannten Parameter.

Die Modellparameter werden so geschätzt, dass das Modell die empirische Varianz-Kovarianzmatrix und die Matrix der Stichprobenmittelwerte möglichst exakt reproduziert. Als Ausdruck des Grades der Genauigkeit der Reproduktion können Anpassungsfunktionen (*fit functions*) ermittelt werden. Sie sind Funktionen der Modellparameter sowie der empirischen Stichprobenmatrizen und drücken die Distanz der beobachteten und implizierten (das heißt aufgrund der Modellparameter

reproduzierten) Matrizen aus. Die Parameterschätzung kann mit unterschiedlichen Methoden, die jeweils an unterschiedliche Bedingungen geknüpft sind, erfolgen. Zumeist wird eine *Maximum-Likelihood*-Schätzung durchgeführt. Diese Methode erfordert multivariat-normalverteilte Daten, ist aber – insbesondere wenn die Analyse auf den Kovarianzmatrizen basiert – robust gegenüber moderaten Verletzungen dieser Bedingungen (Byrne, 2001; Cheung & Rensvold, 2000).

3.3 Bewertung der Modellgüte

Ein Modell ist gültig, wenn die empirischen Daten und die vom Modell vorgegebene Datenstruktur möglichst weitgehend übereinstimmen. Die empirische Überprüfung der Modellgültigkeit basiert auf den Implikationen, die die Beziehungen der latenten Variablen untereinander und die Beziehungen der latenten Variablen zu den manifesten Variablen für die Schätzung der Varianzen und Kovarianzen der beobachteten Variablen haben. Anhand der oben dargestellten Gleichungen können die Varianzen und Kovarianzen der beobachteten Variablen vorhergesagt werden. So sollte bei Modellgültigkeit die Varianz einer manifesten Variablen gleich der zugehörigen quadrierten Faktorladung plus Residuum sein. Die Kovarianz von zwei auf einem Faktor ladenden manifesten Variablen entspricht dem Produkt aus beiden Ladungen, die Kovarianz zweier auf verschiedenen Faktoren ladenden Variablen entspricht dem Produkt ihrer Faktorladungen multipliziert mit der Kovarianz zwischen den entsprechenden Faktoren. Mittels solcher Umformungen kann jedes Element der empirischen Varianz-Kovarianzmatrix anhand der Modellparameter vorhergesagt werden und ist demnach als Funktion der Modellparameter zu verstehen.

Die sich daraus ergebenden Gleichungen für jedes Element dieser Matrix werden in der *Sigma-Matrix* Σ (reproduzierte oder implizierte Matrix) zusammengefasst. Jedes Element von Σ hat als Gegenstück einen numerischen Wert der beobachteten Varianz-Kovarianzmatrix (S). Zur Parameterschätzung werden diese beiden Matrizen gleich gesetzt. Dieses Gleichungssystem wird über den iterativen mathematischen Prozess der Parameterschätzung geschätzt, indem durch wiederholtes Verändern der Parameterschätzungen die Abweichung zwischen Σ und S minimiert wird, bis ein Optimierungskriterium für die Distanz zwischen den Matrizen erfüllt ist. Wie gut die Modelle zu den Daten passen, kann durch Indikatoren der Passungsgüte (*goodness of fit*) ausgedrückt und im Vergleich mit etablierten Standards bemessen werden. Die verschiedenen Indikatoren der Passungsgüte (siehe unten) basieren auf dem Minimum der empirischen Modellanpassungsfunktion (F_{min}). F_{min} ist der Unterschied zwischen der

Kovarianzmatrix der Stichprobe (S) und der durch das Modell implizierten Kovarianzmatrix (Σ).

Der „klassische“ Indikator für die Übereinstimmung zwischen Modell und Stichprobendaten ist der χ^2-Anpassungstest. Das gewünschte Ziel des Tests ist ein nicht-signifikantes Ergebnis, das auf eine perfekte Übereinstimmung verweist. Der χ^2-Anpassungstest ist stark von der Stichprobengröße abhängig und reagiert bei großen Stichproben sehr schnell auf kleine, zu vernachlässigende Unterschiede. Das ist insbesondere deshalb ungünstig, weil konfirmatorische Faktorenanalysen Verfahren sind, die an große Stichproben gebunden sind. Aufgrund dieses Nachteils wurde zur Bewertung der Modellgüte die Forderung aufgegeben, dass ein Modell die empirische Kovarianzmatrix perfekt reproduzieren muss (Raykov & Marcoulides, 2000). Auch theoretisch ist das gut begründbar, da ein Modell zur Beschreibung des interessierenden Phänomens eine deutliche Vereinfachung bei maximaler Annäherung an die Realität bieten soll. Daher kann ein Modell niemals korrekt sein - wäre es dies, so wäre es eine exakte Kopie der Realität und somit nutzlos. Dementsprechend wurde die Bedeutung des χ^2-Anpassungstest herabgestuft.

Ausgehend von obiger Kritik am klassischen χ^2-Anpassungstest wurden Alternativen entwickelt. Zunächst kann das Problem der Abhängigkeit von der Stichprobengröße durch die Einführung des Verhältnisses χ^2 zur Anzahl der Freiheitsgrade (χ^2/FG) abgeschwächt werden. Es sollte zumindest kleiner als 3 sein, besser aber kleiner als 2 (Byrne, 2001). Andere Indizes zur Erfassung und Bewertung der Modellgüte entfernen sich deutlich weiter vom χ^2-Anpassungstest. Allen Kriterien liegt das Minimum der empirischen Modellanpassungsfunktion (F_{min}) zugrunde. Die Unterschiede zwischen den Indizes liegen in der jeweils unterschiedlichen Gewichtung verschiedener Aspekte des Modells (Modellkomplexität versus Parsimonität, Stichprobengröße, Modellparameter). Ein bisher ungelöstes Problem linearer Strukturgleichungsmodelle ist der fehlende Standard darüber, welche Anpassungsindizes zur Modellbewertung herangezogen werden sollten und welche Grenzwerte dieser Indizes für eine hinreichende Modellpassung sprechen. Einklang besteht lediglich darüber, dass es zum jetzigen Zeitpunkt keinen einzelnen universal einsetzbaren Index gibt und dass somit eine Zusammenstellung verschiedener Indizes zu berücksichtigen ist.

Grundsätzlich unterschieden werden absolute und vergleichende Indizes. Erstere (*absolute fit indices*) drücken die absolute Güte der Modellanpassung aus und erfassen die Eignung des Modells, die empirische Kovarianzmatrix abzubilden. Vergleichende Indizes (*comparative indices*) vergleichen, mit oder ohne Berücksichtigung der

Parsimonität, d.h. des Verhältnisses zwischen Anpassungsgüte und Anzahl der Modellparameter, konkurrierende Modelle miteinander, wobei auch „kein Modell“ als Modell gilt.

Im Folgenden werden fünf ausgewählte Indizes kurz dargestellt. Die Auswahl basiert auf Empfehlungen von Byrne (2001), Kelloway (1998), Raykov und Marcoulides (2000) sowie Vandenberg und Lance (2000). Die Grenzwerte dieser Indizes der Anpassungsgüte finden sich in Tabelle 3-A. Anhand der Grenzwerte kann entschieden werden, inwieweit sich das Modell zur Beschreibung der Daten eignet.

Tabelle 3-A: Indizes der Anpassungsgüte mit zugehörigen Grenzwerten.

Name	statistische Grenzwerte
Comparative fit index (CFI)	Möglichst nahe 1; .90 als untere Grenze, ≥.95: gute Modellanpassung
Expected cross-validation index (ECVI)	Möglichst niedrig; Der Wert sollte gleich oder kleiner des ECVI-Wertes des gesättigten Modells sein. Zum Vergleich verschiedener Modelle kann eine Rangreihe der Werte gebildet werden, wobei das Modell mit dem kleinsten ECVI am besten passt.
Non-normed fit index (NNFI)	Möglichst nahe 1; .90 als untere Grenze für eine gute Modellanpassung, ≥ .95: hohes Vertrauen in Modellanpassung.
Root mean squared error of approximation (RMSEA)	Möglichst nahe 0; > .1: schlechte Modellanpassung; .08 bis .1: mittelmäßig; < .08 akzeptabel; < .05: sehr gut; < .01: hervorragend. Zusätzlich wird das 90%-Konfidenzintervall für die Punktschätzung angegeben. Dabei sollte das linke (niedrigere) Ende kleiner als .05 sein; das Intervall sollte möglichst schmal sein.
Goodness of fit index (GFI)	Möglichst nahe 1; mindestens ≥ .9, ideal > .95.

Anmerkung: Die Angaben basieren auf: Byrne (2001), Kelloway (1998), Raykov und Marcoulides (2000), Vandenberg und Lance (2000) .

Der c*omparative fit index* (CFI) ist ein absoluter Index und schätzt, wie gut das Modell auf die Population passt. Die Anpassungsgüte wird in Relation zu einem Nullmodell ausgedrückt. Im Nullmodell bestehen keinerlei Beziehungen zwischen den manifesten Variablen und damit ist seine Anpassungsgüte normalerweise schlecht. Der *expected cross-validation index* (ECVI) gibt die Wahrscheinlichkeit an, mit der das spezifizierte Modell in einer ähnlich großen Stichprobe der gleichen Population repliziert werden kann. So wird insbesondere die Generalisierbarkeit der Ergebnisse erfasst. Zusätzlich wird berücksichtigt, dass komplexe Modelle schlecht an sehr kleine Stichproben angepasst werden können (siehe French & Tait, 2004). Der *nonnormed fit index (NNFI), vormals Tucker-Lewis Index* (TLI) reagiert besonders sensitiv auf die Fehlspezifikationen bezüglich der Faktorladungen. Der Index r*oot mean squared error of approximation* (RMSEA) basiert auf den Residuen, die bei guter Modellanpassung möglichst klein sein sollten. Dazu wird das aktuelle Modell mit einem idealen, freilich unbekannten Modell

verglichen. Auch dieser Index reagiert sehr sensitiv auf Fehlspezifikationen, insbesondere bezüglich der Faktorladungen und ist ferner nicht von der Stichprobengröße abhängig. Der *goodness of fit index* (GFI) drückt den prozentualen Anteil der beobachteten Kovarianzen aus, die durch das Modell impliziert werden und ist ein Maß für den Fehler bei der Reproduktion der empirischen Varianz-Kovarianzmatrix.

Die befriedigende Ausprägung der Indizes der Anpassungsgüte ist eine notwendige, aber keine hinreichende Bedingung für die Validität eines Modells. In der Praxis kann es passieren, dass ein Modell trotz guter Modellanpassungsindizes falsch spezifiziert ist. Aus der Betrachtung der Parameterschätzungen lassen sich Hinweise über die statistische Plausibilität eines Modells ableiten. Diese wurden zumeist als Daumenregeln formuliert (Kelloway, 1998): Im Allgemeinen sollten die Parameterschätzungen (insbesondere die Faktorladungen und Faktorvarianzen) reliabel, das heißt, signifikant von Null verschieden sein. Die Ladungen (Regressionsgewichte) sollten einen Wert von möglichst > .60 aufweisen und signifikant von Null verschieden sein ($p < .01$). Ist dies der Fall, unterstreicht dies die Reliabilität der vorhergesagten Beziehung zwischen Indikator und Faktor. Als Daumenregel für die Parameterschätzungen gilt allgemein, dass sie mindestens doppelt so groß wie die zugehörigen Standardfehler sein sollten. Die Korrelationen zwischen Faktoren sollten von 1 mindestens einen, besser drei Standardfehler entfernt sein. Sehr hohe Korrelationen legen unbefriedigende Modellspezifikationen nahe. Die quadrierten multiplen Korrelationen zwischen den Faktoren und den zugeordneten Indikatoren zeigen an, welcher Anteil der Varianz der Faktoren auf die Indikatoren rückführbar ist. Die multiplen Korrelationen sollten > .60 sein. Standardisierte Residualkovarianzen entsprechen den standardisierten Differenzen zwischen den Stichprobenkovarianzen und den Kovarianzen, die das Modell vorhergesagt. Als Daumenregel sollte der absolute Wert jeder standardisierten Residualvarianz kleiner als 2 sein (kritischer Wert bei α = .05: +/- 1.96.). Auch sollten die Residuen möglichst gleichmäßig verteilt sein und keine Ausreißer aufweisen (Heijden & Donders, 2003). Ein klarer Modellspezifikationsfehler sind zum Beispiel negative Fehlervarianzen.

Abschließend sei angemerkt, dass die konfirmatorischen Faktorenanalysen Modelle nicht bestätigen, sondern nur widerlegen können. Wenn eine Faktorenstruktur die Daten gut beschreibt, heißt dies nicht, dass es nicht bessere oder genauso gute Faktorenstrukturen geben könnte, die nur noch nicht untersucht sind. Auch geben die Indizes kaum Informationen darüber, wie nützlich, plausibel, bedeutsam oder sparsam ein Modell ist.

3.4 Respezifikation

Respezifikation sind Modifikationen eines Modells, um seine Anpassungsgüte zu erhöhen. Die Zulässigkeit von Respezifikationen wird kontrovers diskutiert, weil der konfirmatorische und theoriegeleitete Charakter der Untersuchung in einen explorativen übergeht, wenn aufgrund der stichprobenbasierten Analyseergebnisse Veränderungen am Modell vorgenommen werden, um die Anpassungsgüte zu verbessern (Kelloway, 1998). Post-hoc-Modelländerungen bergen die Gefahr einer bloßen Anpassung des Modells an die Stichprobendaten, so dass sie allenfalls unter Einhaltung bestimmter inhaltlicher und statistischer Regeln zulässig erscheinen. Zunächst sind Veränderungen nur dann zulässig, wenn sie theoretisch gut untermauert sind. Eleganter wäre allerdings, gleich von vorne herein konkurrierende Modelle zu formulieren und vergleichend zu testen. Zur statistischen Beurteilung, ob Post-hoc-Modelländerungen psychometrisch gerechtfertigt sind, können zunächst die Signifikanzniveaus der Parameterschätzungen herangezogen werden und nicht-signifikante Parameter eliminiert werden. Somit kann die Parsimonität des Modells erhöht werden.

Modifikationsindizes geben darüber hinaus Auskunft, welche weiteren Parameter in das Modell aufgenommen werden sollten um die Güte der Anpassung zu erhöhen. Statistisch geben die Modifikationsindizes die Veränderung der Modellanpassungsgüte, also die Veränderung des χ^2-Wertes, wieder, wenn ein weiterer Parameter in das Modell aufgenommen wird.

4 Konzeptionen der Invarianz

Systematische Überprüfungen von Invarianzhypothesen finden sich in der Literatur nur selten. Dies steht im Widerspruch zur Häufigkeit, mit der Stichproben im Rahmen des gruppenmethodischen Ansatzes anhand psychometrischer Tests verglichen werden. Diesem Ansatz können verschiedene Vergleichskonstellationen zugeordnet werden, beispielsweise der Vergleich einer Patientengruppe mit einer Normgruppe oder der Vergleich zweier Patientengruppen untereinander. In der Neuropsychologie der Epileptologie ist das Vorgehen sowohl für Theoriebildung als auch für die Klinik relevant, etwa zur Aufklärung von Ort-Funktions-Zusammenhängen oder zur Beschreibung von Nebenwirkungsprofilen antiepileptischer Medikamente. Auch der Vergleich der Leistungen eines individuellen Patienten mit einer Normgruppe (z. B. zur Lokalisationsdiagnostik) ebenso wie intraindividuelle Vergleiche zweier Testungen im zeitlichen Verlauf (z. B. zur Bewertung der Folgen epilepsiechirurgischer Eingriffe oder medikamentöser Umstellungen) sind dem gruppenmethodischen Ansatz zuzuordnen. Die Zuordnung einzelfalldiagnostischer Fragestellungen zum gruppenmethodischen Ansatz ergibt sich daraus, dass wichtige Schlussfolgerungen dabei nur auf Basis von Stichprobenparametern getroffen werden können: Ein Vergleich der Testleistung eines Patienten mit einer adäquaten Vergleichsgruppe basiert auf der Kenntnis der gruppenspezifischen Mittelwerte und Standardabweichungen, ein intraindividueller Profilvergleich bedarf der Kenntnis der gruppenspezifischen Reliabilitäten und Testinterkorrelationen.

Zum Beispiel ist die Untersuchung der Gedächtnisleistungen vor einem epilepsiechirurgischen Eingriff zentraler Bestandteil jeder prächirurgischen neuropsychologischen Untersuchung. Dabei ist es wichtig, verbale und visuo-räumliche Gedächtnisleistungen zuverlässig voneinander abgrenzen zu können (Bowden et al., 2004; Helmstaedter, 2000). Die eingesetzten Gedächtnistests müssen nicht nur in der Normierungsstichprobe die Konstrukte „verbales Gedächtnis“ und „visuo-räumliches Gedächtnis“ adäquat operationalisieren, sondern es muss zudem gewährleistet sein, dass die Konstruktvalidität der Tests auch auf die klinische Zielpopulation generalisierbar ist. Dies zu überprüfen ist mittels konfirmatorischer Faktorenanalysen gut machbar und höchst bedeutsam, da Verletzungen wichtiger Invarianzhypothesen schwerwiegende Probleme für die Testauswertung und -interpretation mit sich bringen. In der folgenden Darstellung der Invarianzkonzeptionen wird gezeigt, dass nur bei erbrachtem Nachweis von Invarianz bedeutungsvolle und zwischen verschiedenen

Untersuchungsgruppen generalisierbare Ergebnisse erreicht werden können. Vandenberg und Lance (2000) drücken dies wie folgt aus:

> *The crux is that cross-group comparisons require prerequisite assumptions of invariant measurement operations across the groups being compared. … . Demonstration of measurement equivalence is a logical prerequisite to the evaluation of substantive hypotheses regarding group differences, regardless of whether the comparison is as simple as a between-group mean differences test or as complex as testing whether some theoretical structural model is invariant across groups" (S. 9)*

In Hinblick auf die häufig fehlende Überprüfung der Invarianzhypothesen stellen Hoyle und Smith fest: „*The use of a measure that has not demonstrated measurement invariance to compare groups is effectively a worthless exercise*" (1994, zitiert nach French & Tait, 2004). Ähnlich drastisch schreiben Bear und Kollegen: "*It is a dubious process to extend theories and their associated constructs to other groups without determining that the measurement operations yield measures of the same attributes*" (Bear et al., 1992, S. 117, zitiert nach Steenkamp & Baumgartner, 1998).

4.1 Hierarchie der Invarianzhypothesen

In der Einleitung zur konfirmatorischen Faktorenanalyse wurde das faktorenanalytische Mess- und Strukturmodell mit den zugehörigen Modellparametern dargestellt. Invarianzhypothesen beinhalten Gleichheitsannahmen bezüglich dieser Parameter. Die systematische Überprüfung der Invarianz erfolgt in sequenzieller und hierarchischer Weise über zunehmend stringentere Hypothesentests, wobei sich die zunehmende Stringenz aus der steigenden Anzahl von zwischen den Gruppen konstanten Parametern ergibt. Dabei bedeutet Konstanz Gleichheit der geschätzten Populationsparameter.

Im ersten Abschnitt der Überprüfung auf Invarianz werden die Hypothesen des Messmodells untersucht. Das Messmodell bestimmt die Beziehung zwischen den manifesten und latenten Variablen; sie können konfigural invariant oder schwach, stark und streng metrisch invariant sein. Im zweiten Abschnitt werden Invarianzhypothesen, die sich auf die Parameter des Strukturmodells beziehen, überprüft. Sie beinhalten die Beziehung der latenten Variablen untereinander und überprüfen die *strukturelle* Invarianz. An dieser Stelle sei schon angemerkt, dass die Gleichheit der latenten Mittelwerte als Parameter des Strukturmodells keine notwendige Bedingung für den Nachweis der psychometrischen Qualität des Messinstrumentes ist. Für die relevanten klinischen Fragestellungen reicht es also, die Invarianz des Messmodells nachzuweisen.

Die im Folgenden vorgestellten Konzeptionen der Invarianz werden in Anlehnung an Bowden et al. (2004), Meredith (1993) und Widaman und Reise (1997) vorgestellt, die Notation der Hypothesen erfolgt weitest gehend nach Cheung und Rensvold (1999). Eine umfassende Darstellung der Invarianzhypothesen und deren Überprüfung findet sich beispielsweise bei Vandenberg und Lance (2000), Meredith (1993) und Byrne (2001). Tabelle 4-A gibt vorab einen Überblick über die Invarianzhypothesen, ihre symbolische Schreibweise sowie ihre konzeptionelle Bedeutung. Ein allgemeines Prinzip bei der Untersuchung auf Invarianz ist, dass für die nachfolgend überprüften Hypothesen die vorhergehende Stufe der Invarianz bestätigt sein muss.

Tabelle 4-A: Überblick über die Invarianzhypothesen, angelehnt an Cheung und Rensvold (2000).

	Hypothese	Symbol	Konzeptionelle Bedeutung
Messmodell	konfigurale Invarianz	$\Lambda_{form}^{(1)} = \Lambda_{form}^{(2)}$	In beiden Gruppen sind die gleichen Item-Untergruppen mit den gleichen Konstrukten verbunden (kognitive Bereiche sind identisch).
	schwache metrische Invarianz	$\Lambda^{(1)} = \Lambda^{(2)}$	In beiden Gruppen ist die Stärke der Beziehung zwischen den Items und den zugehörigen Konstrukten identisch (die Konstrukte manifestieren sich in gleicher Weise).
	starke metrische Invarianz	$\tau_i^{(1)} = \tau_i^{(2)}$	Die Gruppenunterschiede in den Items sind identisch über alle Items hinweg. Alternativ: Alle Items zeigen denselben Gruppenunterschied an.
	strenge metrische Invarianz	$\Theta_\delta^{(1)} = \Theta_\delta^{(2)}$	Die Items haben dieselbe internale Konsistenz in den beiden Gruppen. Alternativ: Items haben in beiden Gruppen die gleiche Genauigkeit als Maße des zugrunde liegenden Konstruktes.
Strukturmodell – strukturelle Invarianz	Äquivalenz der Konstruktvarianzen	$\Phi_{ii}^{(1)} = \Phi_{ii}^{(2)}$	In beiden Gruppen liegt dieselbe Streubreite der Itemantworten vor. Alternativ: Das Ausmaß der Variabilität bezüglich der Konstrukte ist in beiden Gruppen gleich.
	Äquivalenz der Konstruktkovarianzen	$\Phi_{ij}^{(1)} = \Phi_{ij}^{(2)}$	Die Beziehungen zwischen den Konstrukten (d.h. die Korrelationen) sind in den Gruppen gleich.
	Äquivalenz der Konstruktmittelwerte	$\kappa_j^{(1)} = \kappa_j^{(2)}$	Die Mittelwerte aller Konstrukte sind in den Gruppen identisch.

Anmerkungen: Eigene Übersetzung. Hochgestellte Zahlen indizieren die Gruppen für den Zwei-Gruppen-Fall, wobei jede Hypothese grundsätzlich auch auf mehrere Gruppen generalisierbar ist. Λ: Ladungsmatrix, τ: Vektor der Höhenlagen, Θ: Varianz-Kovarianzmatrix der Messfehler, Φ: Varianz-Kovarianzmatrix der Konstrukte, κ: latente Mittelwerte.

4.1.1 Konfigurale Invarianz

Die konfigurale Invarianz (auch Forminvarianz oder schwache faktorielle Invarianz genannt) ist die schwächste Form der Invarianz. Sie besagt lediglich, dass in beiden Untersuchungsgruppen eine gleiche Anzahl von Faktoren und gleiche Ladungsmuster vorliegen, also dass in beiden Gruppen die gleichen Items mit den gleichen Faktoren verbunden sind. Abbildung 4-A gibt ein Beispiel für konfigurale Invarianz zwischen zwei Gruppen wieder.

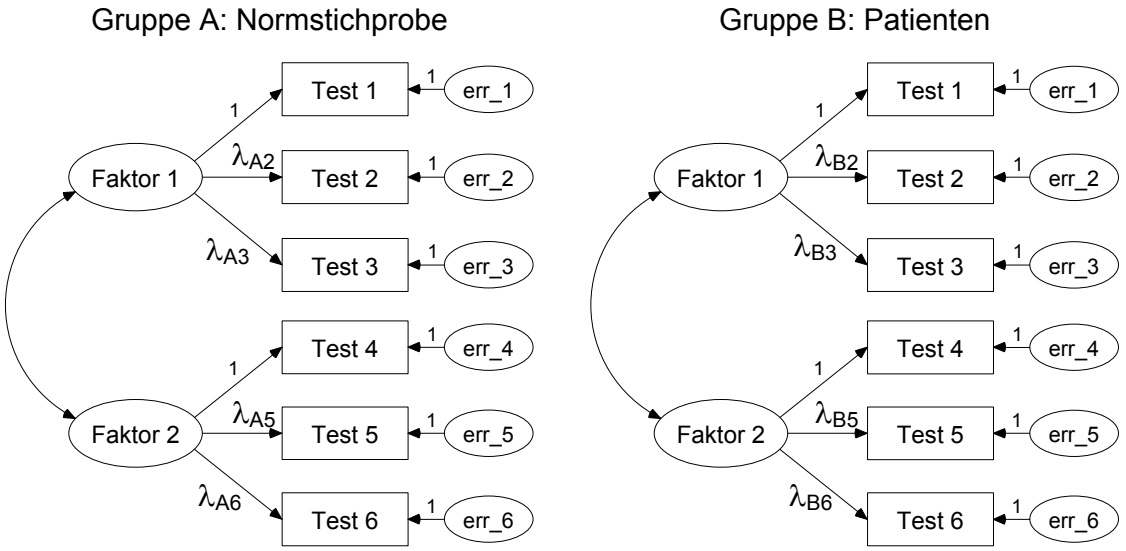


Abbildung 4-A: Konfigurale Invarianz. Sowohl in der Normstichprobe, als auch in der Patientenstichprobe findet sich das gleiche faktorielle Modell.

Konfigurale Invarianz liegt vor, wenn folgende Nullhypothese beibehalten werden kann:

Hypothese: $\Lambda_{form}^{(1)} = \Lambda_{form}^{(2)}$

Λ ist die Matrix der Faktorladungen. Gilt die Nullhypothese, findet sich in den Gruppen ein gleiches Muster aus Nullladungen und salienten, also von Null verschiedenen Ladungen. Der Wert der salienten Ladungen darf zwischen den Gruppen variieren. Gruppenunterschiede in den beobachteten Mittelwerten, Standardabweichungen und Kovarianzen können aus Gruppenunterschieden in allen Modellparametermatrizen herrühren (numerische Werte der Ladungsmatrix Λ, Faktorenkovarianzen Ψ, Fehlerkovarianzen Θ, Höhenlagenvektor ν, Faktorenmittelwerte α). Dies wird in den unten dargestellten Gleichungen für die Kovarianzen Σ und Mittelwerte μ dadurch verdeutlicht, dass Gruppenindizes für sämtliche Parametermatrizen und –vektoren vorliegen:

$$\Sigma_s = \Lambda_s \Psi_s \Lambda_s^t + \Theta_s$$

bzw. $\mu_s = \nu_s + \Lambda_s \alpha_s$

Konfigurale Invarianz ist ein erster Hinweis darauf, dass sich die entsprechenden Faktoren zur Beschreibung der Unterschiede zwischen den Gruppen eignen, da die Faktoren in beiden Gruppen durch dieselben Tests definiert sind. Somit ist die Struktur der Testinstrumente in beiden Gruppen gleich und die Testinstrumente erfassen gleiche kognitive Bereiche. Erste Hinweise auf konfigurale Invarianz können sich schon aus

explorativen Faktorenanalysen ergeben. Dazu sind explorative Faktorenanalysen in jeder Gruppe einzeln durchzuführen und die erhaltenen Faktorladungsmatrizen auf Gleichheit im oben genannten Sinne zu analysieren. Nach Bowden et al. (2004) beschränken sich die meisten faktorenanalytischen Studien zur Untersuchung zu Invarianz auf einen derartigen Vergleich der Faktorenstrukturen (z. B. Bachetzky & Jahn, 2005). Ein Problem dabei ist, dass anhand eines zumeist willkürlich gesetzten Kriteriums zwischen substanziellen und weniger substanziellen Ladungen unterschieden wird, eine inferenzstatistische Unterscheidung zwischen Null- und salienten Ladungen erfolgt nicht.

Konfigurale Invarianz ist die notwendige Minimalbedingung für Gruppenvergleiche, sie ist letztlich aber nicht hinreichend. Ein Faktor wird durch das Gemeinsame aller zugehörigen Indikatoren definiert. Liegt konfigurale Invarianz vor, können die Faktoren ähnlich, wenn auch nicht notwendigerweise gleich, interpretiert werden. Fehlende konfigurale Invarianz hingegen bedeutet, dass sich die Gruppen in den die Faktoren konstituierenden Indikatoren unterscheiden und die Faktoren nur gruppenspezifisch interpretiert werden können. Da faktorielle Validität ein Teilaspekt der Konstruktvalidität ist (Lienert & Raatz, 1998), ist bei fehlender Invarianz von einer gruppenspezifischen Konstruktvalidität der Faktoren bzw. der Tests auszugehen.

Die bisherigen und auch die folgenden Ausführungen zur Invarianz beziehen sich auf den Vergleich zweier Stichproben. Speziell die konfigurale Invarianz ist darüber hinaus für eine zweite Untersuchungssituation wichtig, nämlich für Datensituationen, in denen Daten aus nur einer Stichprobe vorliegen und mit einer Populationslösung verglichen werden (Rietz, 1996). Konfigurale Invarianz liegt vor, wenn nachgewiesen werden kann, dass sich ein theoretisch begründetes Faktorenmodell zur Beschreibung der Stichprobendaten eignet. Diese theoriegeleitete Untersuchung der Konstruktvalidität einer Testbatterie kann durch folgende Hypothese überprüft werden:

$$\text{Hypothese: } \Lambda_{form}^{(1)} = \Lambda_{form}^{(Population)}$$

Hierzu wird das theoretische Modell zunächst über ein Muster aus salienten und Nullladungen operationalisiert und die Vereinbarkeit mit den Stichprobendaten anschließend mittels konfirmatorischer Faktorenanalysen überprüft. Psychologische Theorien beinhalten zumeist nur Annahmen über *Ladungsmuster*, detailliertere Annahmen, etwa über die Höhe der Ladungen, sind dann nicht ableitbar. Diese Form der Invarianz, also die Frage der Vereinbarkeit zwischen den Daten einer Stichprobe

und theoretischen Modellen, wird im ersten empirischen Teil dieser Arbeit untersucht werden.

4.1.2 Schwache metrische Invarianz

Für einen sinnvollen Vergleich zweier Gruppen bezüglich der Leistungsparameter (Testwerte, Faktorwerte) ist die Gleichheit des Ladungsmusters (konfigurale Invarianz) notwendig, aber nicht hinreichend. Die schwache metrische Invarianz (auch faktorielle Invarianz, metrische Invarianz oder starke faktorielle Invarianz genannt) erfordert neben gleichen Ladungsmustern auch gleiche Ladungshöhen und somit gleich starke Beziehungen zwischen Subtest und Faktor. Schwache metrische Invarianz kann angenommen werden, wenn folgende Hypothese nicht widerlegt wird:

Hypothese: $\Lambda^{(A)} = \Lambda^{(B)}$

Dabei kennzeichnet Λ die Faktorladungsmatrizen. Abbildung 4-B gibt analog zu Abbildung 4-A (siehe oben) die faktorielle Konstellation bei schwacher metrischer Invarianz als Pfaddiagramm mit zwei Faktoren und jeweils drei Indikatoren wieder:

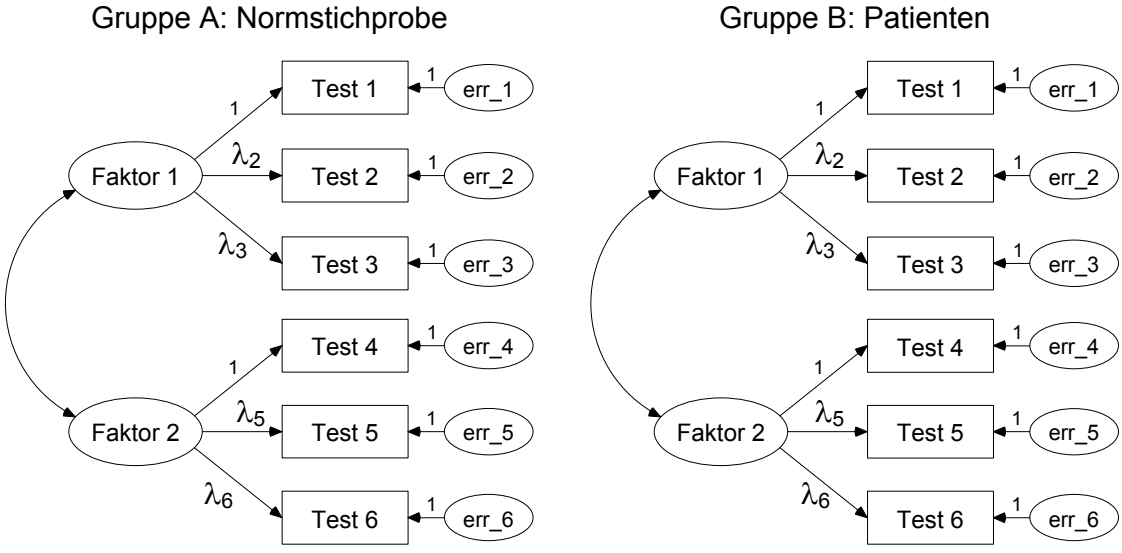


Abbildung 4-B: Schwache metrische Invarianz. In den beiden Untersuchungsgruppen entsprechen sich nicht nur die faktoriellen Strukturen, sondern auch sämtliche Ladungshöhen.

Oben genannte Hypothese ($\Lambda^{(A)} = \Lambda^{(B)}$) auf Abbildung 4-B bezogen ist in Matrizenschreibweise folgendermaßen darzustellen:

$$\begin{bmatrix} 1 & 0 \\ \lambda_{A2} & 0 \\ \lambda_{A3} & 0 \\ 0 & 1 \\ 0 & \lambda_{A5} \\ 0 & \lambda_{A6} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \lambda_{B2} & 0 \\ \lambda_{B3} & 0 \\ 0 & 1 \\ 0 & \lambda_{B5} \\ 0 & \lambda_{B6} \end{bmatrix}$$

Hierbei sind die Gruppenspezifikatoren A und B eigentlich entbehrlich, da sich die Ladungen in beiden Gruppen bei schwacher metrischer Invarianz ohnehin nicht unterscheiden dürfen. Die Fehlervarianzen und -kovarianzen sowie Faktorvarianzen und -kovarianzen können dabei durchaus gruppenspezifische Werte annehmen. Zusammenfassend sind Gruppenunterschiede in den beobachteten Mittelwerten, Standardabweichungen und Kovarianzen nur auf Gruppenunterschiede in den Faktorenkovarianzen Ψ, den Fehlerkovarianzen Θ, dem Höhenlagenvektor ν und den Faktorenmittelwerten α rückführbar. Dies wird in den unten dargestellten Gleichungen für die Kovarianzen Σ und Mittelwerte μ dadurch verdeutlicht, dass Gruppenindizes verglichen mit der konfiguralen Invarianz nun nicht mehr für alle Parametermatrizen und -vektoren vorliegen:

$$\Sigma_s = \Lambda \Psi_s \Lambda^t + \Theta_s \quad \text{bzw.} \quad \mu_s = \nu_s + \Lambda \ \alpha_s$$

Gruppenunabhängigkeit der Faktorladungen ist für Gruppenvergleiche wichtig, da die Interpretation eines Faktors nicht nur von den Inhalten der zugehörigen Tests abhängt, sondern auch von der Stärke der Ladungen (Bortz, 1993; Lubke et al., 2003a). Gibt es zwischen den Gruppen ordinale Unterschiede in den Faktorladungen, so impliziert dies, dass die Indikatoren in den Gruppen für die Konstruktdefinition unterschiedliche Bedeutung haben. Folglich müssen die Faktoren in den Gruppen unterschiedlich interpretiert und möglicherweise auch unterschiedlich benannt werden. Abbildung 4-C verdeutlicht das am Beispiel eines hypothetischen Gedächtnisfaktors:

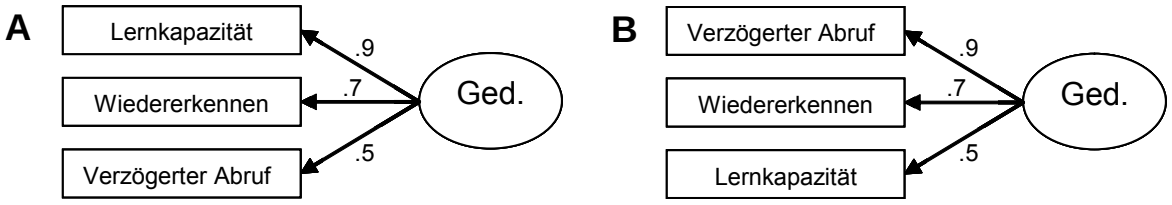


Abbildung 4-C: Ordinale Unterschiede in den Faktorladungen eines Faktors implizieren Gruppenunterschiede in der Faktoreninterpretation. Die Fehlerterme wurden zur Vereinfachung ausgelassen.

Während in Gruppe A die Lernkapazität der Indikator mit der höchsten Ladung auf dem Gedächtnisfaktor ist, hat in Gruppe B der verzögerte Abruf die stärkste Ladung. Für die schwächsten Ladungen sind die Verhältnisse umgekehrt. Dieser Unterschied schlägt sich in der Interpretation nieder: In Gruppe A ist der Gedächtnisfaktor eher ein Kurzzeitgedächtnisfaktor; in Gruppe B ist er als Komponente des Langzeitgedächtnisses interpretierbar. Läge schwache metrische Invarianz vor, wäre die Rangreihe der Ladungen gruppenunabhängig und somit würden sich die Konstrukte in den Gruppen nicht nur in ähnlicher, sondern in gleicher Art und Weise manifestieren.

Die Faktorladung drückt aus, wie stark sich die Testwerte bei Veränderungen der Faktorwerte verändern. Ändert sich der Faktor bzw. das Konstrukt um eine Einheit, so sollten sich bei schwacher metrischer Invarianz in allen Gruppen die zugehörigen Testwerte gleichermaßen (nämlich um λ_j) ändern. Ist dagegen zum Beispiel in einem einfaktoriellen Modell in Gruppe B die Faktorladung doppelt so hoch wie die entsprechende Faktorladung der Referenzgruppe A, so könnte das mit folgenden Gleichungen ausgedrückt werden:

$$\text{Gruppe A: } x_{1A} = \lambda_A * F + e$$
$$\text{Gruppe B: } x_{1B} = \lambda_B * F + e;$$
$$\text{mit } \lambda_B = 2 * \lambda_A$$

Vergleicht man zwei Probanden miteinander, die zwar den gleichen Faktorwert F (d. h. die gleiche „wahre“ Fähigkeitsausprägung) haben, aber aus unterschiedlichen Gruppen stammen, so zeigt sich, dass der Proband aus Gruppe B einen um 2*F-F höheren Testwert erreicht. Anders gesagt muss der Proband aus Gruppe A einen um F höheren Faktorwert aufweisen, um das gleiche Testergebnis wie der Proband aus Gruppe B zu erlangen. Der Test ist also den Probanden der Gruppe A gegenüber unfair (zum Konzept der Testfairness siehe z. B. Amelang & Zielinski, 1997). Unterscheiden sich die Ladungen zwischen den Gruppen, sind folglich auch die Skalierungseinheiten unterschiedlich. Das heißt, der Test auf schwache metrische Invarianz ist auch ein Test auf Gleichheit der Skalierung in den Gruppen (Vandenberg & Lance, 2000).

Zudem ist schwache metrische Invarianz eine notwendige Bedingung, um strukturelle Beziehungen zwischen den Konstrukten (Faktorinterkorrelationen, siehe unten) in den verschiedenen Gruppen zu vergleichen (Finney & Davis, 2003; Steenkamp & Baumgartner, 1998). Nach Bowden et al. (2004) garantiert schwache metrische Invarianz auch, dass Gruppenunterschiede bei Varianzen und Kovarianzen der latenten Variablen von Umskalierungen der manifesten Variablen unberührt bleiben.

Eine typische Skalierung ist zum Beispiel die Standardisierung der Testwerte auf Basis der Mittelwerte und Standardabweichungen einer Referenzgruppe, um die diagnostische Interpretation zu erleichtern. Nur wenn die Varianzen und Kovarianzen der latenten Variablen von Transformationen unbeeinflusst bleiben, können auch die konvergenten und divergenten Validitätsmuster stichprobenübergreifend interpretiert werden.

Während die schwache metrische Invarianz unter anderem eine gleiche Konstruktvalidität in beiden Gruppen anzeigt, können die einzelnen Subtestleistungen in ihrer Höhe trotzdem verzerrt sein. Ein Vergleich der beobachteten Mittelwerte ist also nicht interpretierbar und die wahren latenten Mittelwertsdifferenzen können nicht erschlossen werden (Finney & Davis, 2003).

Daher ist im nächsten Schritt zu untersuchen, ob die beobachteten Variablen in beiden Gruppen unverzerrt sind oder ob ein Bias vorliegt. Für diesen nächsten Schritt, bei dem die starke metrische Invarianz untersucht wird, ist in der Hierarchie der Invarianzhypothesen der Nachweis der schwachen metrischen Invarianz eine notwendige Bedingung.

4.1.3 Starke metrische Invarianz

Liegt starke metrische Invarianz (auch Höhenlagen- oder Skalarinvarianz genannt) vor, so erzielen Probanden mit gleicher Ausprägung des latenten Merkmals auch gleiche Werte in der Testvariable, unabhängig von ihrer Gruppenzugehörigkeit (Finney & Davis, 2003). Somit spiegeln Gruppenunterschiede in den beobachteten Mittelwerten wahre Differenzen zwischen den Gruppen auf der Ebene der latenten Eigenschaft wieder. Wird folgende Hypothese über die Vektoren der Höhenlageparameter (y-Achsenabschnitte, engl.: *intercepts*) nicht widerlegt, so kann starke metrische Invarianz angenommen werden:

Hypothese: $\tau_i^{(1)} = \tau_i^{(2)}$

Gruppenunterschiede in den beobachteten Mittelwerten, Standardabweichungen und Kovarianzen sind bei starker metrischer Invarianz nur noch auf Gruppenunterschiede in den Faktorenkovarianzen Ψ, den Fehlerkovarianzen Θ und den Faktorenmittelwerten α rückführbar. In den hier erneut dargestellten Gleichungen für die Kovarianzen und Mittelwerte ändert sich die Gleichung der Kovarianzstruktur Σ nicht. In der Gleichung für die Mittelwertstruktur verliert der Höhenlagenvektor jedoch seinen Gruppenindex:

$$\Sigma_s = \Lambda \Psi_s \Lambda^t + \Theta_s \quad \text{bzw.} \quad \mu_s = \nu + \Lambda \ \alpha_s$$

Höhenlagen entsprechen dem für eine beobachtete Variable vorhergesagten Wert, wenn die latente Variable Null ist. Die Hypothese starker metrischer Invarianz besagt, dass die Höhenlagen paarweise für alle Indikatoren in beiden Gruppen gleich sind. Bezogen auf das obige einfaktorielle Beispiel würde folgende Konstellation auf fehlende starke (aber immerhin schwache) metrische Invarianz hinweisen:

Gruppe A: $x_{1A} = \nu_A + \lambda * F^1 + e$

Gruppe B: $x_{1B} = \nu_B + \lambda * F^1 + e$;

mit $\nu_A \neq \nu_B$

Der Höhenlagenindex A bzw. B verweist darauf, dass die Höhenlagen gruppenspezifische Ausprägungen haben. Auch bei gruppenunabhängigen Ladungen λ und gleichen Konstruktausprägungen würden zwei Probanden aus unterschiedlichen Gruppen unterschiedliche Testergebnisse erlangen.

Starke metrische Invarianz ist Voraussetzung für die Untersuchung und Interpretation von Differenzen in den beobachteten Werten. Fehlt diese Form von Invarianz, kann nicht entschieden werden, ob Unterschiede in den beobachteten Variablen Verzerrungen der Messung widerspiegeln oder tatsächliche Gruppenunterschiede in den latenten Eigenschaften. Existieren systematische Gruppenunterschiede über alle Items hinweg und sind diese konsistent mit den Unterschieden in den latenten Variablen, können die Höhenlagen als invariant angesehen werden (Finney & Davis, 2003). Zusätzlich können bei starker metrischer Invarianz latente Mittelwertdifferenzen mittels z-Tests und Effektgrößen beschrieben und interpretiert werden. Das hat den Vorteil, dass Gruppenunterschiede ohne Messfehler betrachtet werden können, da Konstrukte als fehlerfreie Schätzungen angesehen werden (Finney et al. 2003). Des Weiteren ist es bei starker metrischer Invarianz möglich, Gruppenunterschiede in den latenten Varianzen der Konstrukte und der Korrelationen zwischen den Konstrukten zu untersuchen.

Eine sehr gute zusammenfassende grafische Veranschaulichung schwacher und starker Invarianz findet sich bei (Rensvold, 2002). Diese gibt Abbildung 4-D wieder.

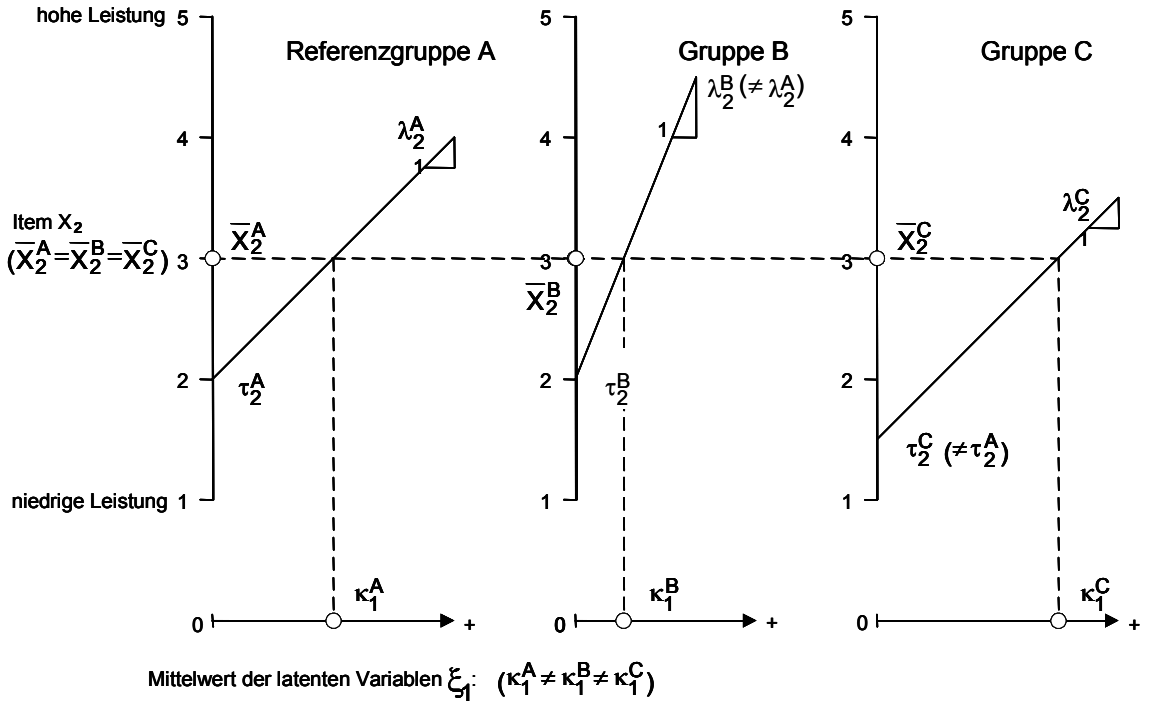


Abbildung 4-D: Veranschaulichung der Folgen fehlender schwacher (Gruppe B) und fehlender starker metrischer Invarianz (Gruppe C), nach Renvsvold (2002).

Dargestellt sind drei Gruppen mit jeweils gleichem Mittelwert in Item X_2. Ohne eingehende Prüfung der Invarianzhypothesen würde sicherlich auf eine gleiche Leistungsfähigkeit im Fähigkeitskonstrukt ξ, das diesem Item zugeordnet ist, geschlossen werden. Ausgehend von Gruppe A als Referenzgruppe zeigt der Vergleich mit Gruppe B, dass trotz gleichen beobachteten Werten Gruppe B eine niedrigere tatsächliche Ausprägung κ in der latenten Variablen ξ_1 aufweist. Die Grafik macht deutlich, dass das Item in beiden Gruppen zwar die gleiche Höhenlage τ, aber unterschiedliche Steigungen λ (Ladungen) hat. Das ist ein Beispiel fehlender schwacher Invarianz.

Der Vergleich zwischen der Referenzgruppe A und Gruppe C zeigt, dass wieder beide Gruppen trotz des gleichen beobachteten Mittelwerts unterschiedliche latente Variablen aufweisen. Das ist auf Unterschiede in der Höhenlage τ bei hier allerdings gleichen Steigungsparametern λ rückführbar. Dieses Beispiel verdeutlicht fehlende starke metrische Invarianz.

4.1.4 Strenge metrische Invarianz

Strenge metrische Invarianz (oder Messfehlerinvarianz) besagt, dass alle Elemente der Varianz-Kovarianzmatrix der Messfehler (Θ) in den Gruppen identisch sind. Bei strenger metrischer Invarianz gilt folgende Hypothese:

Hypothese: $\Theta_{\delta}^{(1)} = \Theta_{\delta}^{(2)}$

Finney und Davis (2003) merken an, dass dieser Test von manchen Forschern als exzessiv streng angesehen wird (z. B. Byrne, 2001; Cheung & Rensvold, 1999). Die Bedeutung der strengen metrischen Invarianz und die von den Autoren zugeordnete Wichtigkeit hängen von der Interpretation der Fehlerterme ab. Gegen eine allzu hohe Bedeutung der strengen metrischen Invarianz spricht nach Byrne et al. (1989), dass die Fehlervarianzen und -kovarianzen stichprobenspezifisch sind. Deren Gleichheit kann keine notwendige Bedingung für die Äquivalenz des Messmodells sein. In dieser Konzeption wird der Fehlerterm als ein rein zufälliger Residualterm aufgefasst. Andere Ansätze gehen davon aus, dass Fehlerterme nicht nur zufälliges Rauschen beinhalten (dessen Größe von der Reliabilität des Tests abhängig ist), sondern auch einen mehr oder weniger großen Anteil nicht-modellierter Varianz, die prinzipiell durch Aufnahme weiterer Faktoren erfassbar ist. Die Aufnahme weiterer Faktoren würde dann die Fehlervarianzen und auch die Interkorrelationen der Messfehler verringern (Levine et al., 2003). Nicht-modellierte äußere Quellen der Fehlerkovarianzen können beispielsweise Methodenvarianz oder Methodeneffekte sein, nicht-lineare Beziehungen zwischen den Items oder andere Konstrukte, die im Modell nicht erfasst sind, aber einzelne Itemuntergruppen beeinflussen. Selten wird ein Modell alle latenten Einflussgrößen erfassen. Die korrelations- und faktorenanalytische Grundannahme der unkorrelierten Residuen ist also kaum haltbar. Daher wäre es theoretisch stringent, interkorrelierte Messfehler in das Modell aufzunehmen. Tatsächlich ist das aus methodischen Gründen meist nicht möglich, da solche Modelle zu wenige Freiheitsgrade aufweisen und nicht identifizierbar sind.

Die Höhe der Fehlerterme der Tests (Fehlervarianzen) hängt letztlich sowohl von den test- und stichprobenspezifischen Reliabilitäten als auch von der Varianz ab, die durch die modellierten Faktoren erklärt wird. Die Varianz-Kovarianzmatrix der Messfehler ist deshalb nur bei gleichzeitiger Berücksichtigung der Faktorvarianzen interpretierbar: Sind die Faktorvarianzen zwischen den Gruppen äquivalent, verweisen Gruppenunterschiede in den Residualvarianzen auf unterschiedliche Reliabilitäten der Untertests. Somit ist unter der Bedingung identischer Faktorvarianzen der Test auf strenge metrische Invarianz ein Test der Invarianz der Untertestreliabilitäten. Liegt

hingegen keine Gleichheit der Faktorvarianzen vor, so können Unterschiede in den Reliabilitäten nur nach Kontrolle der Faktorvarianzen untersucht werden (Vandenberg & Lance, 2000); ohne Nachweis oder Kontrolle der Gleichheit der Faktorvarianzen ist der Test auf strenge metrische Invarianz als bloßer Test der Gleichheit der Fehlervarianzen zu betrachten. Unterscheiden sich die Residualvarianzen, ist auch ein Unterschied in den zugehörigen Kovarianzen nicht überraschend.

Die Gruppenunterschiede in den Fehlertermmatrizen weisen bei gleichzeitiger Invarianz der Faktorvarianz auf gruppenweise unterschiedliche Untertestreliabilitäten hin. Die herkömmliche klinische Praxis, die Untertestergebnisse für sich genommen zu interpretieren, ist höchst fraglich, da Gruppenunterschiede in der Konstruktvalidität der Untertests nicht auszuschließen sind (Bowden et al., 2004). Die Konstruktvalidität der Einzeltests hängt unter anderem von den Validitätskorrelationen dieser Untertests ab und die Korrelationen eines Untertests mit einem zweiten Maß (externes Kriterium, weitere Untertests) hängen ihrerseits von der Untertestreliabilität ab. Messfehler können Validitätskoeffizienten stark beeinflussen (siehe Schmidt & Hunter, 1999). Unproblematisch ist dagegen die Interpretation der Faktoren, da die Zuordnung der Untertests zu den Faktoren nicht von Gruppenunterschieden in den Reliabilitäten beeinflusst ist. Weder die Invarianz der Residualvarianzen, noch die Invarianz der Faktorvarianzen sind notwendige Voraussetzungen für die Interpretation von Gruppenunterschieden in den Faktorwerten. Wurde starke Invarianz, aber keine strenge Invarianz gezeigt, ist es somit am sichersten, die Interpretation auf der Faktorenebene anhand von Faktorwerten vorzunehmen (Bowden et al., 2004), weil deren Mittelwerte und Korrelation von Messfehlern in den beobachteten Variablen weniger beeinflusst werden.

4.1.5 Strukturelle Invarianz

Wurde mindestens starke metrische Invarianz gezeigt, so ist die Voraussetzung geschaffen, im Rahmen des vollständigen Messmodells[4] die Gruppenunterschiede der latenten Variablen direkt zu vergleichen (Bowden et al., 2004). Strukturelle Invarianz bezieht sich im Gegensatz zu den bisher besprochenen Invarianzhypothesen nicht auf die Komponenten des Messmodells, sondern auf die Komponenten des Strukturmodells. Die vollständige strukturelle Invarianzannahme besagt, dass die Strukturparameter Faktorvarianzen, -kovarianzen und -mittelwerte in allen Gruppen gleich sind. Dieser

[4] Für die Untersuchung aller Strukturparameter („mean structure approach“) müssen als Datenbasis neben den Untertestvarianzen und -kovarianzen auch die Untertestmittelwerte vorliegen.

Hypothesentest kann global oder für jeden Parameter einzeln ausgeführt werden, wobei folgende drei Unterhypothesen zu prüfen sind:

- Hypothese A (Äquivalenz der Konstruktvarianzen): $\Phi_{ii}^{(1)} = \Phi_{ii}^{(2)}$
- Hypothese B (Äquivalenz der Konstruktkovarianzen): $\Phi_{ij}^{(1)} = \Phi_{ij}^{(2)}$
- Hypothese C (Äquivalenz der latenten Mittelwerte): $\kappa_j^{(1)} = \kappa_j^{(2)}$

Im Gegensatz zu den oben dargestellten Invarianzhypothesen ist die Reihenfolge der Überprüfung der strukturellen Invarianzhypothesen beliebig, da die Schlussfolgerungen der Invarianztests unabhängig von einander bestehen und nicht in hierarchischen Beziehungen logisch aufeinander aufbauen. Die Bedeutung dieser Stufe der Invarianz ist wenig kontrovers: Für eine gruppenunabhängige und einheitliche Auswertung einer Testbatterie (sei es auf der Stufe der beobachteten Werte oder auf der Stufe der Faktorwerte) ist die Invarianz der Strukturparameter keine notwenige Voraussetzung. Vielmehr ist sie gerade Gegenstand des Gruppenvergleichs, dessen Zulässigkeit durch Nachweis der (starken) metrischen Invarianz bestätigt wurde.

Von besonderem theoretischen als auch praktischen Interesse ist die Frage nach Gruppenunterschieden in den Faktorwerten. Ein Mittelwertsvergleich auf Ebene der Faktoren hat – verglichen mit den Rohwerten – den Vorteil, dass Messfehler auspartialisiert wurden. Unterschiede in den Faktorwerten drücken Unterschiede in den zugrunde liegenden Fähigkeitsgraden der verschiedenen Gruppen aus. In der Literatur zu kognitiven Beeinträchtigungen bei Epilepsie wurde hinreichend beschrieben, dass Patienten mit Epilepsie in den kognitiven Leistungen im Allgemeinen unterhalb den Leistungen der Normgruppe liegen (siehe z. B. Helmstaedter, Kurthen, Lux, Reuber & Elger, 2003; Thompson & Duncan, 2005). Entsprechend wäre bei einem Vergleich zwischen Patienten und gesunden Probanden Invarianz in den Faktorenmittelwerten zu erwarten und nicht gesondert zu überprüfen.

Die Faktorvarianzen geben Auskunft über Gruppenunterschiede in der Faktorwertstreuung. Liegt Invarianz der Faktorvarianzen vor, so ist die Voraussetzung für einen Gruppenvergleich der Korrelationen der latenten Variabeln gegeben. Wie oben beschrieben (siehe 4.1.5) ist der Test auf Gleichheit der Faktorvarianzen Voraussetzung für die Untersuchung der Reliabilitäten auf Testebene.

Die Untersuchung der Faktorenkovarianzen ist eher von theoretischem Interesse: Faktorenkovarianzen geben Auskunft über konvergente und diskriminative Validität auf Konstruktebene (Finney & Davis, 2003). So könnte in einer Gruppe eine starke Beziehung zwischen zwei Konstrukten darauf verweisen, dass beide Konstrukte in

ähnlicher Weise konzeptualisiert sind, während in der Vergleichsgruppe eine höhere Distinktheit der Konstruktdefinition vorliegt (niedrigere Kovarianzen). Anhand der Muster der Faktorenkovarianzen können Hypothesen über die stichprobenspezifischen Charakteristika in der Intelligenzstruktur aufgestellt werden. Liegt Invarianz der Faktorenkovarianzen und der Faktorvarianzen vor, so kann auf Gleichheit der Faktorkorrelationen geschlossen werden.

4.2 Zusammenfassung und Fazit

Die Darstellung der verschiedenen Invarianzhypothesen zeigt, dass sinnvolle Gruppenvergleiche nur möglich sind, wenn zumindest der Nachweis starker metrischer Invarianz gelingt und somit die Metrik der Messinstrumente (d.h. die Regressionsparameter Steigung bzw. Ladung und Höhenlage) gruppenunabhängig ist. Das soll zusammenfassend am Beispiel zweier Probanden aus unterschiedlichen Populationen verdeutlicht werden, deren Fähigkeiten bezüglich eines Konstruktes gleich sind (Lubke et al., 2003a). Diese Fähigkeit wird von einem konstruktvaliden Test erfasst. Erreichen beide Probanden unabhängig von ihrer Gruppenzugehörigkeit den gleichen Testwert, so liegt metrische Invarianz vor. Mittels bedingter Wahrscheinlichkeiten kann das anhand folgender Gleichung ausgedrückt werden (Lubke, Dolan, Kelderman & Mellenbergh, 2003b; Meredith, 1993):

$$f(Y \mid \eta, s) = f(Y \mid \eta)$$

Dabei werden die Populationen bzw. Gruppen über eine Gruppenvariable (Selektionsvariable) mit den Realisationen s definiert (z. B. Kontrollgruppe vs. Patienten). Y steht für die beobachteten Testwerte der Probanden und η für deren Faktorwerte. Die Gleichung besagt, dass der Wert Y bei Vorliegen der Fähigkeit η unabhängig der Gruppenzugehörigkeit erreicht wird. Sie besagt nicht, dass die Gruppen den gleichen beobachteten Mittelwert oder Faktorenmittelwert haben müssen. Andersherum wird natürlich bei Gruppen mit gleichem Faktorenmittelwert auch der gleiche Y-Mittelwert erwartet. Für die klinische Forschung und Praxis kann zusammenfassend festgehalten werden, dass der Nachweis von zumindest strenger metrischer Invarianz eine notwendige und hinreichende Bedingung dafür ist, dass:

- Gruppenvergleiche bezüglich beobachteter oder latenter Werte zulässig und eindeutig sowie aussagekräftig und interpretierbar sind, damit Gleiches mit Gleichem verglichen wird;
- Messinstrumente und Items über Gruppen oder Zeitpunkte hinweg in gleicher Weise funktionieren;

- die Wahrscheinlichkeit, ein Testitem richtig zu beantworten, für alle Probanden mit der gleichen getesteten Fähigkeit gleich ist und nicht von anderen Charakteristika wie Gruppenzugehörigkeit oder Geschlecht abhängt. Nur dann sind Unterschiede in beobachteten Variablen tatsächlich eine Funktion eines wahren quantitativen Gruppenunterschiedes auf Konstruktebene und nicht durch unterschiedliche Konzeptualisierungen des Konstruktes oder weitere konfundierende Variablen erklärbar.

Die Betrachtung der latenten Mittelwerte bringt ferner den Vorteil der Messfehlerfreiheit. Von theoretischer Bedeutung ist, dass durch den Prozess der Invarianztestungen Hypothesen über die Struktur der Intelligenz getestet und weiterentwickelt werden können und wichtige Hinweise für eine etwaige Stichprobenabhängigkeit der Intelligenzmodelle gefunden werden können. Für die neuropsychologische Testpraxis bedeutet dies, dass vor Anwendung eines Tests oder einer Testbatterie in Populationen, die von der Standardisierungsstichprobe abweichen, die Invarianz zu überprüfen ist.

4.3 Statistische Bewertung der Invarianzhypothesen

Auch wenn der χ^2-Anpassungstest – wie in Kapitel 3.3 erläutert – für den Vergleich eines Modells mit den Stichprobendaten kaum von Bedeutung ist, spielt er für eine vergleichende Bewertung der Güte konkurrierender Modelle im Rahmen von Invarianzuntersuchungen eine etwas wichtiger Rolle. Dabei wird für die aufgestellten zunehmend strengen Modelle überprüft, ob die Anpassungsgüte auch nach den schrittweise eingeführten Gleichheitsbeschränkungen weiterhin zufrieden stellend bleibt oder ob es zu einem signifikanten Abfall der Modellpassung kommt. Die konkurrierenden Modelle können direkt mithilfe des χ^2-Differenztests geprüft werden, dessen Teststatistik sich aus der Differenz des χ^2-Wertes des restriktiveren Modells (mit Invarianzbeschränkungen) und des χ^2-Wertes des freieren Modells ergibt. Da χ^2 jedoch nur für kleine Stichproben eine sensitive Teststatistik ist und bei größeren Stichproben zu unspezifisch, soll den anderen Kennwerten zur vergleichenden Abschätzung von Spezifikationsfehlern mehr Relevanz zugeordnet werden (Bowden et al., 2004). Die Differenzen zwischen den anderen Indizes der Anpassungsgüte können zumeist nicht herangezogen werden, da sie keiner bekannten Verteilung folgen. Auch die Verteilung des *comparative fit index* ist unbekannt. Trotzdem wurden für ihn zumindest empirisch sinnvolle Grenzwerte der Differenzen ermittelt (Cheung & Rensvold, 2000). Bei Differenzen ≤ -.01 kann die Invarianzannahme beibehalten werden, Differenzen

zwischen -.01 und -.02 lassen die Invarianzannahme fraglich erscheinen und Differenzen > -.02 sprechen sicher gegen Invarianz.

Als dritte Möglichkeit zur vergleichenden Bewertung zweier Modelle bei der Überprüfung der Invarianzhypothesen können die für den RMSEA-Index berechenbaren Konfidenzintervalle herangezogen werden. Liegt der Wert des strengeren Modells noch innerhalb des Konfidenzintervalls des liberaleren Modells, so ist die eingetretene Verringerung der Anpassungsgüte durch die Restriktion weiterer Parameter nicht von Bedeutung und die überprüfte Invarianzannahme ist haltbar.

4.4 Fehlende Invarianz – Ursachen und Folgen

Die verschiedenen Invarianzkonzeptionen sind oben ordinal nach ihrer Strenge geordnet. Beim Testen auf Invarianz kann es grundsätzlich auf jeder Stufe zum Nachweis fehlender Invarianz kommen. Es wurde gezeigt, dass Invarianz auf struktureller Stufe nicht notwendig ist. Auf der nächst niedrigeren Stufe, der strengen metrischen Invarianz wurde die kontroverse Diskussion dargestellt. Nach Bowden et al. (2004) sind invariante Residualvarianzen keine notwendigen Voraussetzungen, so dass Invarianz auf dieser Stufe nicht den Vergleich der Faktorwerte zwischen den Gruppen beschränkt. Auf allen niedrigeren Stufen bringt das Verwerfen der Nullhypothesen gewichtige psychometrische und interpretatorische Schwierigkeiten aufgrund der gruppenspezifisch unterschiedlichen Messmodelle mit sich.

Fehlende Invarianz kann durch vielerlei Faktoren verursacht werden (Finney & Davis, 2003): Auf der Stufe der konfiguralen Invarianz können Gruppenunterschiede durch unterschiedliche Konzeptionalisierungen der Faktoren zustande kommen, ebenso durch externe Variablen, die die Messungen konfundieren. Ursache fehlender starker metrischer Invarianz können Gruppenunterschiede im Antwortstil sein (d.h. eine Tendenz, systematisch höhere oder niedrigere Antworten zu geben, unterschiedliche Testmotivation und Anstrengungsbereitschaft, unterschiedlich wahrgenommener Leistungsdruck etc.) oder Unterschiede in der Bedeutung des Tests für das entsprechende Konstrukt.

Eine zentrale und auch per se theoretisch interessante Fragestellung bei der Überprüfung der Invarianzhypothesen ist die Identifikation der nicht-invarianten Items oder Tests. Da vollständige metrische Invarianz schwer zu erreichen ist, wurde vorgeschlagen, sie als notwendige Bedingung für Gruppenvergleiche zu relativieren. Horn (1991) schreibt dazu: „*Metric invariance is a reasonable ideal...a condition to be striven for, not one expected to be fully realized*“ (S. 125, zitiert nach Baer, Prince & Velez, 2004). Als Kompromiss zwischen vollkommener Invarianz und komplettem

Fehlen von Invarianz wurde das Konzept der partiellen Invarianz eingeführt. Dieses Konzept soll in loser Anlehnung an Levine et al. (2003) für den Zwei-Gruppen-Fall verdeutlicht werden: Ausgangspunkt ist der Vergleich der Parametervektoren zweier Gruppen. Dies kann beispielsweise der Höhenlagenvektor sein, prinzipiell ist partielle Invarianz aber ebenso auf alle anderen Parametermatrizen übertragbar. Zur Überprüfung, ob diese Parameter invariant sind, werden sie zunächst simultan in beiden Gruppen als paarweise invariant angenommen und durch eine konfirmatorische Faktorenanalyse geschätzt. Drei Ergebniskonstellationen können auftreten: Liegt komplette Invarianz vor, so nehmen die einzelnen Parameter des Messmodells in beiden Gruppen bei guter Modellanpassungsgüte paarweise identische Werte an. Somit kann das Messmodell sparsam über P Parameter gruppenunabhängig beschrieben werden (Abbildung 4-E, Teil A). Eine unzureichende Güte der Anpassung bei Annahme invarianter Parameter zeigt Gruppenunterschiede in den Parametern an und bedeutet, dass die Messmodelle beider Gruppen nicht mehr mit dem gleichen Satz von P Parametern beschrieben werden können. Stattdessen sind 2xP Parameter notwendig, damit die einzelnen Parameter gruppenspezifische Werte annehmen können. Dieses Ergebnis verweist auf komplett fehlende Invarianz (Abbildung 4-E, Teil C). Im dritten möglichen Fall sind einige, aber nicht alle Parameter invariant. Die nicht-invarianten Parameter werden schrittweise befreit und gruppenspezifisch geschätzt. Wird ein Parameter gruppenabhängig frei geschätzt, so kann er gruppenspezifische Werte annehmen, wodurch sich die Anzahl der zu schätzenden Parameter pro befreiten Parameterpaar von P auf P+1 erhöht (Abbildung 4-E, Teil B1 und B2). Im diesem Falle liegt weder perfekte Invarianz vor, noch komplette Unterschiedlichkeit – sondern eben partielle Invarianz. Welche Parameter des Messmodells zwecks gruppenspezifischer Schätzung zur Erhöhung der Güte der Modellpassung freigesetzt werden, kann anhand der „Modifikationsindizes“, die die Statistikprogramme ausgeben, entschieden werden (Baer et al., 2004).

$$\text{A}\begin{bmatrix} x_{11} & = & x_{12} \\ x_{21} & = & x_{22} \\ x_{31} & = & x_{32} \end{bmatrix} \qquad \text{B1}\begin{bmatrix} x_{11} & = & x_{12} \\ x_{21} & = & x_{22} \\ x_{31} & \neq & x_{32} \end{bmatrix} \qquad \text{B2}\begin{bmatrix} x_{11} & = & x_{12} \\ x_{21} & \neq & x_{22} \\ x_{31} & \neq & x_{32} \end{bmatrix} \qquad \text{C}\begin{bmatrix} x_{11} & \neq & x_{12} \\ x_{21} & \neq & x_{22} \\ x_{31} & \neq & x_{32} \end{bmatrix}$$

Abbildung 4-E: Drei Ergebniskonstellationen einer konfirmatorischen Faktorenanalyse. A: vollständige Invarianz; B1 und B2: partielle Invarianz; C: komplett fehlende Invarianz.

Somit kann Invarianz als graduelles Konzept aufgefasst werden (Levine et al., 2003), wobei der Anteil der äquivalenten Parameter an der Gesamtzahl der Modellparameter

ein einfacher Index für den Grad der Invarianz ist. Zusätzlich muss berücksichtigt werden, welche Items in welchem Ausmaß nicht-invariant sind. Das Konzept der partiellen Messinvarianz kann genutzt werden, wenn wenigstens konfigurale Invarianz vorliegt (Baer et al., 2004). Darüber hinaus legt keine verbindliche Regel fest, bis zu welchem Grad Verletzungen der Invarianzannahmen noch akzeptabel sind. Praktisch hilfreich, aber wenig rational sind Minimalanforderungen, die davon ausgehen, dass ein kleiner Anteil nicht-invarianter Items Gruppenvergleiche nur unmaßgeblich beeinflusst. Für eine weitere Minimalanforderung reicht das Vorliegen konfiguraler Invarianz und zumindest einer nicht-invarianten Faktorladung pro Faktor, um eine Testbatterie als invariant einzustufen. Ein konservativerer Zugang zur partiellen Invarianz wurde von Steenkamp und Baumgartner (1998) formuliert: Grundsätzlich sollte nur eine Minderheit von Parametern frei geschätzt werden, das Freisetzen von Parametern sollte auf starker theoretischer Basis beruhen und durch Kreuzvalidierungen überprüft werden. Klare statistische Kriterien zur Bestimmung eines akzeptablen Maßes an partieller Invarianz fehlen jedoch (Vandenberg & Lance, 2000).

Ist eine korrekte Identifikation der nicht-invarianten Tests oder Items erfolgt, so ergeben sich mehrere Möglichkeiten (Cheung & Rensvold, 1999): Ist es oberstes Ziel, möglichst alle Items zu erhalten, können die Faktorwerte der nicht-invarianten Subtests angepasst werden (Bowden et al., 2001). In einer klinischen Gruppe kann beispielsweise eine Faktorladung höher sein als in der Normgruppe. Das heißt, ein hoher Wert in diesem Untertest führt zusammen mit der höheren Ladung zu einer höheren Ausprägung im zugehörigen Faktor in dieser Gruppe. Bezieht sich die Testauswertung nun auf die Normgruppe, in der hohe Werte im entsprechenden Test nicht mit gleicher Enge auch eine hohe Ausprägung im Faktor bedeuten, werden die latenten Leistungen der Patienten unterschätzt. Die Unterschätzung ist zu korrigieren. Nicht invariante Items oder Tests können zweitens aus der Testbatterie entfernt werden, weil sie für einen Gruppenvergleich faktisch wertlos sind und ihre Validität nicht eindeutig erklärbar ist. Gruppenunterschiede sind dann anhand der verbliebenen und tatsächlich invarianten Items und zugehörigen Faktoren zu interpretieren. Die nicht-invarianten Tests können zur Beschreibung von Gruppenunterschieden eine eigenständige Bedeutung haben. Drittens können über die Selektionsvariable hinaus weitere Moderatorvariablen eingeführt werden.

4.5 Kritik und Alternativen

Die Überprüfung der Invarianzhypothesen mittels konfirmatorischer Faktorenanalysen stellt hohe Anforderungen an Datensituation und Datenanalyse. Daher sollen nun

alternative Zugänge betrachtet werden. Eine Möglichkeit ist es, für jede Gruppe einzeln explorative Faktorenanalysen durchzuführen und die gruppenspezifischen Ladungsmuster als Grundlage des Gruppenvergleichs einzuführen. Damit solch ein Vergleich über die Untersuchung konfiguraler Invarianz hinausgehen kann, sind inferenzstatistische Methoden der Bewertung der Gleichheit und Unterschiedlichkeit notwendig (siehe z. B. Rietz, 1996). Diese Methoden erfordern jedoch zum einen eine anspruchsvolle Methodik, zum anderen sind sie verglichen mit der konfirmatorischen Faktorenanalyse schwerer handhabbar, weniger empirisch fundiert und ermöglichen keine explizite Berücksichtigung der Messfehler (Raykov & Marcoulides, 2000). Aufgrund dieser Schwierigkeiten hat die konfirmatorische Faktorenanalyse die dominante, wenn nicht alleinige Rolle als Methode zur Überprüfung Invarianzhypothesen inne, zumal nur mit ihr hochgradig komplexe Intelligenzmodelle testbar werden.

Eine Alternative ist die komplette Abwendung vom psychometrischen und faktorenanalytischen Paradigma. Ein Haupteinwand gegen dieses Paradigma ist, dass durch den Bezug auf Norm- oder heterogene Patientengruppen wichtige Informationen, die den Einzelfall und homogene Patientengruppen charakterisieren, verloren gehen (siehe z. B. Delis, Jacobson, Bondi, Hamilton & Salmon, 2003). Als weiterer Einwand wird die Frage nach der klinischen Relevanz von gruppenanalytischen Ergebnissen vorgebracht. Die klinisch gut fundierte Unterscheidung der Gedächtniskomponenten entlang der Zeitachse ist beispielsweise faktorenanalytisch nicht durchgehend replizierbar (Wilde et al., 2003). Das prozessorientierte Vorgehen löst derartige Schwierigkeiten zumindest teilweise, weil dabei die Art und Weise der Aufgabenlösung ebenso relevant ist wie das sich ergebende Testprofil (Kaplan, 1988). Auch das deskriptiv-einzelfallorientierte Vorgehen nach Luria (1973) oder das kriteriumsorientierte Testen bieten Lösungsmöglichkeiten. Diese Ansätze und das faktorenanalytische Vorgehen sollten jedoch nicht gegeneinander ausgespielt werden; vielmehr haben sie unterschiedliche Relevanz für die verschiedenen Ebenen der Theoriebildung und der Konstruktvalidierung. Auch der kognitiv-experimentelle Ansatz des Vergleichs homogener Patientengruppen anhand sparsam ausgewählter Variablen, den Delis et al. (2003) zur Überwindung der Schwierigkeiten der Faktorenanalyse vorschlagen, basiert auf der Grundannahme, dass die Tests in beiden Gruppen das gleiche Konstrukt erfassen. Diese Annahme ist aber ganz ohne gruppenstatistische und korrelationsanalytische Methoden nicht hinreichend zu beantworten.

5 Klinische Studien zur Invarianz

Relativ häufig wird die Invarianz der Faktorenstruktur von Fragebögen untersucht (z. B. Baer et al., 2004; Finney & Davis, 2003; French & Tait, 2004; Levine et al., 2003; Schmitt, Maes & Seiler, 1999). Fragebogenstudien haben den Vorteil, schnell die für konfirmatorische Faktorenanalysen benötigten Stichprobengrößen zu erreichen. Klinische Studien zur Untersuchung der Invarianz von kognitiven Testbatterien sind dagegen sehr selten, am ehesten finden sich Studien, die im Rahmen der Normierung und Validierung von Standardtestbatterien durchgeführt wurden und dabei auf die entsprechenden Normstichproben zurückgreifen können (z. B. Taub, McGrew & Witta, 2004; Weiss & Price, 2001). Der häufige Einsatz von Standardtestbatterien ist einerseits mit deren höherer publikatorischer Relevanz aufgrund des allgemeinen Interesses zu erklären. Zum anderen bieten Standardtestbatterien als Vergleichsgruppe eine zumeist sehr große Normierungsstichprobe, so dass die zusätzliche Erfassung von Normdaten entbehrlich wird. Für den deutschsprachigen Raum fehlen Studien zur Invarianz nahezu vollkommen. Eine aktuelle Studie von Bachetzky und Jahn (2005) überprüft zwar Invarianzhypothesen, setzt dazu jedoch lediglich explorative Faktorenanalysen ein.

Tabelle 5-A (siehe nächste Seite) gibt einen Überblick über die hier näher vorgestellten klinisch-neuropsychologische Studien zur Invarianz. Bei den eingesetzten Standardtestbatterien handelt es zumeist um Testbatterien der Wechsler-Familie, die einzeln oder kombiniert eingesetzt werden (Wechsler, 1981, 1987, 1997a, 1997b). Dargestellt werden ausschließlich Studien, die konfirmatorische Faktorenanalysen einsetzen.

Müller, Hasse-Sander, Horn, Helmstaedter und Elger (1997) haben die faktorielle Struktur des verbalen Gedächtnistests VLMT (neue Version von Helmstaedter, Lendt & Lux, 2001) untersucht. Dabei wurde ein zweifaktorielles Modell mit einem Langzeit- und einem Kurzzeitgedächtnisfaktor aufgestellt und mittels konfirmatorischer Faktorenanalysen für zwei Stichproben untersucht: eine psychiatrische mit 232 Patienten mit überwiegend beginnenden demenziellen Erkrankungen oder Depressionen sowie eine Stichprobe mit 872 Epilepsiepatienten. Die konfirmatorischen Faktorenanalysen wurden für jede Stichprobe separat durchgeführt, so dass lediglich konfigurale Invarianz des postulierten zweifaktoriellen Modells bestätigt werden konnte.

Bowden et al. (2001) haben anhand einer Stichprobe mit 289 alkoholabhängigen Patienten die Faktorenstruktur einer Testbatterie überprüft, die aus der WMS-R und der WAIS-R bestand. Mit konfirmatorischen Faktorenanalysen wurde untersucht, inwieweit die empirische Varianz-Kovarianzmatrix der Rohwerte mit konkurrierenden

Faktorenmodellen vereinbar ist. Die Faktorenmodelle bestanden aus vier bis sieben obliquen Faktoren, zudem wurden weitere Modelle mit einem zusätzlichen Generalfaktor operationalisiert. Das Modell, welches am besten mit der empirischen Datenstruktur vereinbar war, bestand aus den fünf Faktoren *verbal comprehension, perceptual organization, attention-concentration, verbal memory* und *visual memory*. Die Invarianz des Faktorenmodells wurde mittels Mehr-Gruppen-Analyse mit einer zusätzlichen Stichprobe mit 399 gesunden Kontrollprobanden untersucht. Die Invarianzuntersuchungen bezogen sich auf Faktorladungen, Faktorenkovarianzen und Faktorvarianzen. Für die Faktorladungen konnte partielle Invarianz gezeigt werden, wobei sich lediglich 14 von 22 Faktorladungen als invariant erwiesen. Deutliche Gruppenunterschiede wurden in den Faktorenkovarianzen und den Faktorvarianzen festgestellt.

Tabelle 5-A: Überblick über die Methodik klinisch-neuropsychologischer Studien zu Invarianz.

Studie	Tests	Kovarianzstruktur	Mittelwertstruktur	Stichproben	Invarianz bestätigt? (j = ja; n = nein; p = partial)	Konfiguration	Faktorladungen	Itemhöhenlagen	Messfehler	Faktorenkovarianzen	Faktorvarianzen	Faktorenmittelwerte
Müller et al., 1997	RAVLT	ja	nein	zwei Patienten-gruppen		j	-	-	-	-	-	-
Bowden et al., 2001	WMS-R, WAIS-R	ja	nein	Patienten, Kontrollgruppe		j	p	-	-	n	n	-
Burton et al., 2002	WAIS-III	ja	nein	Patienten, Normgruppe		j	-	-	-	-	-	-
Wilde et al., 2003	WMS-III	ja	nein	Patienten		j	-	-	-	-	-	-
Heijden et al., 2003	WAIS-III	ja	nein	Patienten		j	-	-	-	-	-	-
Bowden et al., 2004	WMS-R, WAIS-R	ja	ja	Patienten, Normgruppe		j	j	j	n	-	-	-
Banos et al., 2004	CVLT	ja	nein	Patienten		j	-	-	-	-	-	-

Anmerkung: Literaturangaben siehe Text. Abkürzungen: WMS-R = Wechsler Memory Scale – Revised; WAIS-R: Wechsler Adult Intelligence Scale – Revised; WAIS-III: Wechsler Adult Intelligence Scale – Third Edition; RAVLT: Rey Auditory Verbal Learning Test; CVLT: California Verbal Learning Test.

Burton et al. (2002) haben die Faktorenstruktur des Wechsler-Intelligenztests für Erwachsene (WAIS-III, Wechsler, 1997a) anhand einer heterogenen Stichprobe mit 328 neurologischen Patienten untersucht und das Ergebnis mit der originalen Standardisierungsstichprobe dieses Tests kreuzvalidiert (n = 2450). Wahrscheinlich

aufgrund der unterschiedlichen Gruppengrößen wurde nicht das methodisch stringentere Vorgehen paralleler konfirmatorischer Faktorenanalysen in beiden Gruppen gewählt. Stattdessen wurden zur Überprüfung der neun ausgewählten Faktorenmodelle in jeder Gruppe separate konfirmatorische Faktorenanalysen durchgeführt. Dabei wurde pro Gruppe versucht, durch Vergleiche zwischen den ausgewählten Modellen und zusätzlichen Modellvarianten eine Entscheidung für eines dieser Modelle herbeizuführen. Das Modell, welches in beiden Gruppen die beste Anpassungsgüte aufwies, bestand aus folgenden sechs Faktoren: *semantic memory, verbal reasoning, constructional praxis, visual reasoning, working memory* und *processing speed*.

Gänzlich auf eine zweite Stichprobe verzichteten Wilde et al. (2003): Anhand einer klinischen Stichprobe mit 254 Patienten mit Temporallappenepilepsie wurde untersucht, welche Faktorenstruktur die empirische Varianz-Kovarianzmatrix der Wechsler-Gedächtnistestbatterie am besten beschreibt (Wechsler, 1997b). Hierfür wurden fünf konkurrierende Modelle miteinander verglichen, die beste Passung bei maximaler Parsimonität wies ein zweifaktorielles Modell mit einem Arbeitsgedächtnisfaktor und einem allgemeinen Gedächtnisfaktor auf. Diese Studie mit nur einer Stichprobe kann im engeren Sinne kaum mehr als Untersuchung zur faktoriellen Invarianz gewertet werden: Im üblichen Ein-Gruppen-Fall der konfirmatorischen Faktorenanalyse wird ein theoretisch gut begründetes Modell mit empirischen Daten einer Stichprobe abgeglichen. Im hier vorliegenden Falle wurden allerdings verschiedene, im Testhandbuch der WMS-III vorgestellte Modelle überprüft, ohne deren theoretische und empirische Fundierung zu belegen. Eine solcherart eingesetzte konfirmatorische Faktorenanalyse kann im Vergleich zu einer explorativen Faktorenanalyse nur wenig neue Erkenntnisse bringen.

Ein grundsätzlich ähnliches Vorgehen, bei etwas stringenterer theoretischer Fundierung der Modellauswahl, haben Heijden und Donders gewählt (2003). Gegenstand der Untersuchung war hier wiederum der Wechsler-Intelligenztest für Erwachsene (WAIS-III), dessen Faktorenstruktur anhand einer Stichprobe von 166 Patienten mit traumatischen Hirnschäden untersucht wurde. Zu diesem Zweck wurden vier Modelle miteinander verglichen. Das Modell mit der besten Anpassungsgüte bestand aus den vier Faktoren *verbal comprehension, perceptual organization, working memory* und *processing speed*.

Eine deutlich umfassendere Untersuchung zur Invarianz faktorieller Modelle wurde von Bowden et al. (2004) anhand einer neurologischen Stichprobe (n = 277) und einer Kontrollstichprobe (n = 399) vorgelegt. Gegenstand dieser Untersuchung war nicht nur die empirische Varianz-Kovarianzmatrix, sondern auch der empirische Mittelwertvektor

einer Testbatterie aus der WMS-R und der WAIS-R. So konnte neben der konfiguralen Invarianz auch die metrische Invarianz untersucht werden. Besonders zu erwähnen ist, dass die neurologische Stichprobe 177 Patienten mit Epilepsie beinhaltete. Weil sich die Itemhöhenlagen zwischen den Gruppen nicht unterschieden, konnte starke metrische Invarianz für ein Modell mit den fünf Faktoren *verbal comprehension, perceptual organization, working memory, verbal memory, visual memory* und *processing speed* bestätigt werden.

In der letzten hier vorzustellenden Studie wird die faktorielle Struktur des California Verbal Learning Tests (Delis, Kramer, Kaplan & Ober, 1987) untersucht (Banos et al., 2004). Die Untersuchung erfasst eine Stichprobe mit 388 Epilepsiepatienten. Von acht überprüften Faktorenmodellen wies ein dreifaktorielles Modell mit den Faktoren *auditory attention, verbal learning* und *inaccurate recall* die beste Anpassungsgüte auf.

Zusammenfassend ist zu sagen, dass in den hier vorgestellten Studien zumeist nur konfigurale Invarianz untersucht wurde. Diese schwächste Form der Invarianz wurde auch überwiegend bestätigt. Eine vollständige Invarianzuntersuchung anhand der Kovarianz- und Mittelwertstruktur wurde lediglich von Bowden et al. (2004) durchgeführt.

6 Fragestellungen

In dieser Arbeit werden fünf Fragestellungen behandelt. Die Datengrundlage stellen umfangreiche neuropsychologische Testungen einer gesunden Kontrollgruppe und einer Gruppe von Patienten mit Epilepsie dar.

Fragestellung A

Zunächst soll mittels konfirmatorischer Faktorenanalysen untersucht werden, welches faktorielle Intelligenzstrukturmodell die empirische Kovarianzstruktur der gesunden Kontrollprobanden am besten erklärt. Hierzu werden konfirmatorische Faktorenanalysen für das Generalfaktormodell, das hierarchische Intelligenzstrukturmodell von Vernon, das Berliner Intelligenzstrukturmodell und für das Cattell-Horn-Carroll-Modell durchgeführt. Zusätzlich wird die Güte der Anpassung eines typischen neuropsychologischen Faktorenmodells untersucht. Es wird vorhergesagt, dass die Normdaten am besten durch das Cattell-Horn-Carroll-Modell beschreibbar sind. Dieses Modell zeichnet sich durch eine gute empirische Fundierung und klare Faktorendefinitionen aus. Zudem postulieren die mit dem CHC-Modell verknüpften Aussagen zur diagnostischen Praxis, dass die Faktorenstruktur hochgradig unabhängig von den eingesetzten Testverfahren sein sollte. Auch hat sich dieses Modell in einer früheren Studie zur Beschreibung der Faktoren, die aus einer explorativen Faktorenanalyse resultieren, als geeignet erwiesen (Lutz & Helmstaedter, 2004).

Fragestellung B

Ebenfalls mittels konfirmatorischer Faktorenanalyse, nun aber im Zwei-Gruppen-Fall, soll untersucht werden, ob das Modell, das die Daten der Normstichprobe am besten beschreibt, gleichzeitig auch in einer Stichprobe aus Patienten mit Epilepsie Gültigkeit hat. Dies entspricht der Frage, inwieweit die Testbatterie auch bei Patienten die Konstrukte erfasst, die aufgrund der Befundlage in der Normstichprobe erwartet werden. Methodisch entspricht dies der Frage nach konfiguraler Invarianz.

Fragestellung C

Wenn Fragestellung B positiv beantwortet ist, soll durch weitere konfirmatorische Faktorenanalysen der Grad der Invarianz des Messmodells zwischen der Normierungs- und der Patientenstichprobe untersucht werden. Im Speziellen werden dabei die Hypothesen schwacher, starker und strenger metrischer Invarianz sequenziell überprüft. Aufgrund der unzureichenden empirischen Datenlage lassen sich anhand von

Vorstudien keine fundierten Vorhersagen über den zu erwartenden Grad der Invarianz ableiten.

Fragestellung D

Sofern die Gültigkeit eines faktoriellen Modells für beide Stichproben nicht widerlegt wird und zumindest schwache metrische Invarianz gezeigt werden kann, können im nächsten Schritt die Faktorvarianzen und -kovarianzen der Stichproben miteinander verglichen werden. Wird darüber hinaus zumindest starke metrische Invarianz bestätigt, können zusätzlich Faktorenmittelwerte beider Stichproben miteinander verglichen werden. Da Patienten mit Epilepsie als Gruppe grundsätzlich niedrigere kognitive Leistungen als gesunde Normprobanden aufweisen, sind Unterschiede in den Strukturparametern, insbesondere in den latenten Faktorenmittelwerten, zu erwarten. Vollständige strukturelle Invarianz wird also nicht angenommen.

Fragestellung E

In Abhängigkeit von den Ergebnissen der oben genannten Fragestellungen sollen einzelne Untersuchungen zur Konstruktvalidität des faktoriellen Modells durchgeführt werden. Sie haben rein explorativen Charakter, da grundsätzlich die Untersuchung der Faktorenstruktur und die Validierung des faktoriellen Modells nicht mit ein und derselben Stichprobe durchgeführt werden können.

7 Methode

Die folgende Darstellung der Methode, die dieser Arbeit zugrunde liegt, beginnt mit der Beschreibung der Stichproben. Hierbei wird zwischen Ausgangs- und Analysestichproben unterschieden. Die Unterscheidung wurde notwendig, da aufgrund zu großer Unterschiede im kognitiven Grundniveau beider Stichproben eine angemessene Korrektur vorgenommen werden musste. Im nächsten Abschnitt werden die Testverfahren der neuropsychologischen Testbatterie und die Indikatoren, die für die Analyse ausgewählt wurden, vorgestellt. Schließlich wird das Vorgehen bei den konfirmatorischen Faktorenanalysen spezifiziert. Hierbei wird insbesondere auf die konkrete Operationalisierung der Modelle fokussiert. Für die Fragestellung A werden im Ein-Gruppen-Fall der konfirmatorischen Faktorenanalyse verschiedene faktorielle Modelle operationalisiert und miteinander verglichen. Zur Beantwortung der Fragestellungen B, C und D werden im Zwei-Gruppen-Fall der konfirmatorischen Analyse die Hypothesen metrischer und struktureller Invarianz operationalisiert.

7.1 Stichproben

Die folgende Beschreibung der Stichproben unterscheidet zwischen den Ausgangsstichproben und den Analysestichproben. Die Ausgangsstichproben spiegeln den vollständigen Umfang der jeweiligen Datenbank zum Zeitpunkt der Abfrage wieder. Da die erhaltene Normierungsstichprobe aufgrund des klar überdurchschnittlichen mittleren Intelligenzniveaus jedoch nicht als repräsentativ gelten kann, wurde mit einer noch näher zu beschreibenden Strategie eine Teilstichprobe erstellt, um die Unterschiedlichkeit der Intelligenzverteilungen zu minimieren. Diese Stichprobe wird als Analysestichprobe für die Untersuchungen auf Invarianz zwischen Patienten und Gesunden herangezogen.

7.1.1 Die Ausgangsstichproben

Die Normierungsstichprobe besteht aus hirngesunden Probanden, die im transregionalen Sonderforschungsbereich "Mesiale Temporallappen-Epilepsien" (SFB/TR3) freiwillig an einer Studie zur Normierung der Bonner neuropsychologischen Testbatterie teilgenommen haben. Zum Zeitpunkt der Datenbankabfrage (13.07.2005) umfasste die Stichprobe 256 Probanden. Die Probanden wurden in einem einleitenden Anamnesegespräch nach neurologischen und psychiatrischen Vorerkrankungen sowie nach weiteren Faktoren, die die Eignung der Probanden für eine Normierungsstudie in Frage stellen könnten, befragt. Zehn Probanden wurden ausgeschlossen: Neun davon

hatten nicht-näher bestimmte neurologische Erkrankungen angegeben und deutliche Leistungsdefizite in einzelnen Testverfahren aufgewiesen und ein Proband war mit den Tests schon vertraut. Elf weitere gaben zwar zurückliegende Erkrankungen an, da aber in der neuropsychologischen Untersuchung weder fokale noch globale kognitive Funktionsbeeinträchtigungen vorlagen, wurde auf deren Ausschluss aus der Normierungsstichprobe verzichtet. Sieben dieser Probanden hatten ein nicht näher bestimmtes Schädel-Hirn-Traumata erlitten (14 bis 34 Jahre zurückliegend), vier Probanden schilderten psychiatrische oder psychotherapeutische Vorbehandlungen.

Als nächster Schritt wurden die verbliebenen n = 246 Datensätze auf fehlende Werte überprüft: Ein Datensatz mit drei fehlenden Untertests – der Proband hatte einen Testabbruch gewünscht – wurde ausgeschlossen. So verblieb schließlich eine Stichprobengröße von n = 245. Weitere fehlende Werte waren äußerst selten: Insgesamt haben nur sechs Untertestwerte gefehlt, bei keinem Probanden fehlte mehr als ein Untertest und nur ein Untertest (Labyrinthtest, Chapuis, 1992) fehlte zweimal. Weil somit extrem wenige Daten fehlten, stellte sich die Frage nach deren Nicht-Zufälligkeit nicht. Da zur Überprüfung auf faktorielle Invarianz die Datensätze vollständig sein müssen, wurden mittels EM-Analyse (SPSS, 2003) Mittelwerte, Korrelationen und Kovarianzen geschätzt und die fehlenden Werte durch abgeleitete Werte ersetzt. Somit resultiert eine Ausgangsstichprobe von n = 245 kompletten Datensätzen hirngesunder Kontrollprobanden.

Die Patienten wurden in der Klinik für Epileptologie in Bonn auf die Möglichkeit eines epilepsiechirurgischen Eingriffes hin untersucht. Es handelt sich nicht um eine vollkommen repräsentative Stichprobe von Patienten mit Epilepsie, vielmehr finden sich in der Stichprobe insbesondere pharmakoresistente und zumeist langjährige und schwere Formen der Epilepsie. Trotz dieser Selektivität sind die Ätiologien, Pathologien oder epileptogene Läsionen heterogen. Die Datenabfrage erfolgte am 03.08.2005. Aus der neuropsychologischen Datenbank der Epileptologie in Bonn wurden die Patienten ausgewählt, für die eine vollständige Intelligenztestung vorlag und die präoperativ getestet wurden. Die Abfrage lieferte 322 Datensätze, von denen 17 wieder ausgeschlossen wurden: Neun Patienten hatten nur ungenügende Deutschkenntnisse, bei drei Datensätzen waren die Rohwerte der durchgeführten Intelligenztestung nicht verfügbar, bei zwei Patienten erfolgte zwischen zwei Testteilen eine klinische Intervention (Medikamentenumstellung beziehungsweise Implantation von Tiefenelektroden), ein Patient hatte laut Datenbankkommentar eine ungenügende Sehleistung, ein Patient, der zu einer Callosotomie anstand, wies einen globalen

Bodeneffekt in den Testleistungen auf und schließlich ergab sich bei einem weiteren Patienten der Verdacht auf eine Medikamentenintoxikation. Somit verblieb eine Stichprobe von 305 Patienten.

Im nächsten Schritt wurden die fehlenden Werte analysiert. Für 251 Patienten lagen komplette Datensätze vor, bei 41 Patienten fehlte ein Testwert, bei 13 Patienten fehlten zwei oder mehr Testwerte. Letztere 13 Patienten wurden ausgeschlossen. Bei den 41 anderen Patienten fehlten die Werte überproportional häufig in den Verfahren *Boston Naming Test*, *semantische Wortflüssigkeit* und *Mehrfachwahl-Wortschatz-Intelligenztest* (14, 11 beziehungsweise 10 fehlende Werte; Testbeschreibungen siehe 7.2.1). Da man deshalb nicht ohne weiteres von der Nicht-Zufälligkeit der fehlenden Testwerte ausgehen kann, wurden diese 35 Patienten von den folgenden Analysen ausgeschlossen. Die restlichen sechs fehlenden Werte wurden wiederum mittels EM-Analyse (SPSS, 2003) geschätzt und ersetzt. So resultierte eine Stichprobengröße von 257 Patienten.

7.1.2 Alters-, Geschlechts- und Intelligenzunterschiede

Tabelle 7-A gibt die allgemeinen Stichprobencharakteristika für die Normierungs- und die Patientenstichprobe wieder. Dabei zeigt sich, dass sich signifikant mehr Frauen als Männer an der Normierungsstudie beteiligt haben, während das Geschlechterverhältnis bei den Patienten ausgeglichen ist. Die Patienten waren mit 36.86 Jahren insgesamt etwas jünger als die im Schnitt 40.49 Jahre alten Normprobanden. Der geschätzte Intelligenzquotient der Normierungsstichprobe lag mit 119.92 Punkten deutlich über dem anzunehmenden Populationsmittelwert von 100 und weit über dem geschätzten Intelligenzquotienten der Patienten von 93.74; auch haben die Normprobanden ein höheres Bildungsniveau erreichen können als die Patienten. Diese Charakteristika (insbesondere der hohe Frauenanteil und das hohe intellektuelle Grundniveau) spiegeln eine typische Tendenz solcher auf Selbstselektion beruhenden Stichproben wider. Die Normstichprobe kann *nicht* als repräsentativ für die Grundpopulation angesehen werden.

Tabelle 7-A: Allgemeine Stichprobencharakteristika der Ausgangsstichproben.

	Normstichprobe (n = 245)	**Patienten (n = 257)**	
Geschlecht (w/m)	141/104	123/134	χ^2 [1] = 4.73; p = .032
Alter M (SD)	40.49 (13.94)	36.86 (12.27)	F [1] = 9.60; p = .002
Mittlerer Intelligenzquotient[a]	119.92 (12.06)	93.74 (16.87)	F [1] = 396.75; p < .001
Schulbildung			χ^2 [4] = 148.29; p < .001
- kein Abschluss oder Sonderschule	0	19	
- Haupt-/Volksschule	20	104	
- Mittlere Reife	56	83	
- Abitur	98	39	
- Hochschule	71	12	
Händigkeit (R/L/Ambidexter)	209/6/30	217/16/24	χ^2 [2] = 5.078; p = .078

Anmerkung: [a] geschätzt nach HAWIE-R (Tewes, 1991), Kurzform (Schwarzkopf-Streit, 2000)

Deutliche Gruppenunterschiede in den demografischen Variablen stehen der Untersuchung der Invarianz bezüglich der Gruppenzugehörigkeit im Wege, da dann über die reine Gruppenzugehörigkeit hinaus weitere potenzielle Quellen fehlender Invarianz vorliegen. Daher sind diese Faktoren vor Beginn der eigentlichen Analysen zu kontrollieren. Für die verschiedenen in Tabelle 7-A dargestellten Variablen wird ein differenzielles Vorgehen gewählt: Die Unterschiede in der Geschlechts- und Altersverteilungen sollen – sofern laut Normstichprobe erforderlich – über eine Alters- und Geschlechtskorrektur der Rohwerte ausgeglichen werden (siehe 8.1). Eine Bildungskorrektur wird dagegen nicht vorgenommen, da ein hoher Anteil der Varianz, die durch die Gruppenunterschiede im Bildungsniveau entsteht, schon durch die Korrektur der demografischen Variablen Alter, Geschlecht und insbesondere Intelligenz abgefangen wird. Außerdem steht das Bildungsniveau bei Patienten mit Epilepsie in einem loseren Zusammenhang zu Intelligenz und den weiteren kognitiven Testleistungen als bei den Kontrollprobanden, weil die Krankheit häufig den Zugang zu Ausbildung und Beruf erschwert.

Der große Gruppenunterschied im Intelligenzniveau ist problematisch, insbesondere da die Intelligenzleistungen in den folgenden Analysen – anders als Alter und Geschlecht – nicht nur Hintergrundvariablen sind, sondern Gegenstand der Analysen. Damit können die Unterschiede im Intelligenzniveau nicht einfach durch Normierung ausgeglichen werden. Im Folgenden werden deshalb besser geeignete Unterstichproben erstellt, um die Gruppenunterschiede im Intelligenzniveau zu vermindern. Zur Auswahl einer gesunden Unterstichprobe, die das Kriterium der Intelligenz-Repräsentativität erfüllt, wird zunächst als Zielvorgabe ein durchschnittlicher

Intelligenzquotient von 105 gesetzt[5]. Tabelle 7-B gibt die unter Normalverteilungsannahme erwarteten prozentualen Häufigkeiten zusammen mit den tatsächlichen Häufigkeiten in der Normierungsstichprobe (n = 245) wieder.

Tabelle 7-B: Erwartete und tatsächliche Häufigkeiten für die Intelligenzquotienten in der Normierungsstichprobe.

Intelligenz-quotient	erwartete Häufigkeiten[a] (%)	tatsächliche Häufigkeiten (%)	tatsächliche Häufigkeiten
≤ 89	15.87	1.22	3
90-94	9.27	2.45	6
95-99	11.93	2.86	7
100-104	12.93	4.49	11
105-109	12.93	8.98	22
110-114	11.93	8.16	20
115-119	9.27	14.29	35
120-124	6.69	19.59	48
125-129	4.43	13.47	33
130-134	2.47	13.47	33
135-139	1.29	8.16	20
≥ 140	0.99	2.86	7
Gesamt	100	100	245

Anmerkung: [a] unter Normalverteilungsannahme

Es ist ersichtlich, dass die Intelligenzbänder oberhalb des Mittelwertes von 105 deutlich überrepräsentiert sind, während der untere Intelligenzbereich unterrepräsentiert ist. Somit ist die Zusammenstellung einer bezüglich des Intelligenzniveaus repräsentativen Unterstichprobe nicht umsetzbar. Sollten beispielsweise die sieben Probanden aus dem Intelligenzbereich [95 bis 99] ca. 12 % einer Unterstichprobe entsprechen, würde dies eine Reduktion der Gesamtstichprobe auf n = 59 erzwingen.

Daher ist ein anderes Vorgehen zu wählen, dessen Ausgangspunkt die minimale notwendige Stichprobengröße für die später einzusetzenden Analyseverfahren ist. Die minimale Stichprobengröße hängt bei Invarianzanalysen von der Anzahl der Indikatoren ab. In die konfirmatorischen Faktorenanalysen sollen 19 Indikatoren aufgenommen werden, so dass die Stichprobengröße auf n = 190 begrenzt werden kann. Die Auswahl

[5] Der Wechsler-Intelligenztest HAWIE-R (Tewes, 1991) wurde 1991 veröffentlicht. Die Daten wurden sicherlich deutlich früher erhoben, ohne dass der genaue Zeitraum im Handbuch angegeben ist. Es kann also geschätzt werden, dass der HAWIE-R inzwischen auf mindestens 15 Jahre alten Normdaten beruht. Dies bedeutet zum einen, dass die Normierung für ein Intelligenztestverfahren zu alt ist, zum anderen impliziert es eine Fehleinschätzung der tatsächlichen Intelligenz aufgrund des „Flynn-Effektes“. Er besagt, dass die Populationsmittelwerte in den letzten 60 Jahren um drei IQ-Punkte pro Dekade angestiegen sind (z. B. Flynn, 1987). Das bedeutet, dass unter Anwendung des HAWIR-R bei einer aktuell erhobenen repräsentativen Stichprobe ein durchschnittlicher Intelligenzquotient von ca. 105 zu erwarten ist. (100+1.5*3). Eine neue Version des Hamburg-Wechsler-Intelligenztests ist kürzlich erschienen (Tewes, Neubauer & Aster, 2006).

dieser Indikatoren wird im nächsten Kapitel noch ausführlich dargestellt. Die Häufigkeitsverteilung (Abbildung 7-A) der Intelligenzquotienten beider Stichproben zeigt, dass es an den Rändern der Verteilungen kaum Überlappungen des Streubereiches gibt. Daher werden zur Angleichung beider Intelligenzverteilungen und zur Reduktion des mittleren Intelligenzquotienten in der Normierungsstichprobe von beiden Stichproben die Extreme der Intelligenzverteilung gekappt.

Für die Patientenstichprobe mit ursprünglich 257 Patienten bedeutet dies, dass die 67 Patienten mit dem niedrigsten Intelligenzquotienten von den späteren Analysen ausgeschlossen werden. Aus der Normierungsstichprobe mit ursprünglich 245 Probanden werden die 54 Probanden mit den höchsten Intelligenzquotienten ausgeschlossen; zusätzlich wurde aus der Normierungsstichprobe ein Proband mit einem Intelligenzquotienten von 80 ausgeschlossen, da es sich hierbei um einen Ausreißer handelt. Bezogen auf den Intelligenzquotienten bedeutet dieses Vorgehen, dass Patienten mit Werten kleiner als 82.20 ausgeschlossen wurden, während aus der Normierungsstichprobe Probanden mit einem geschätzten Intelligenzquotienten ab 130.15 ausgeschlossen wurden (Abbildung 7-B).

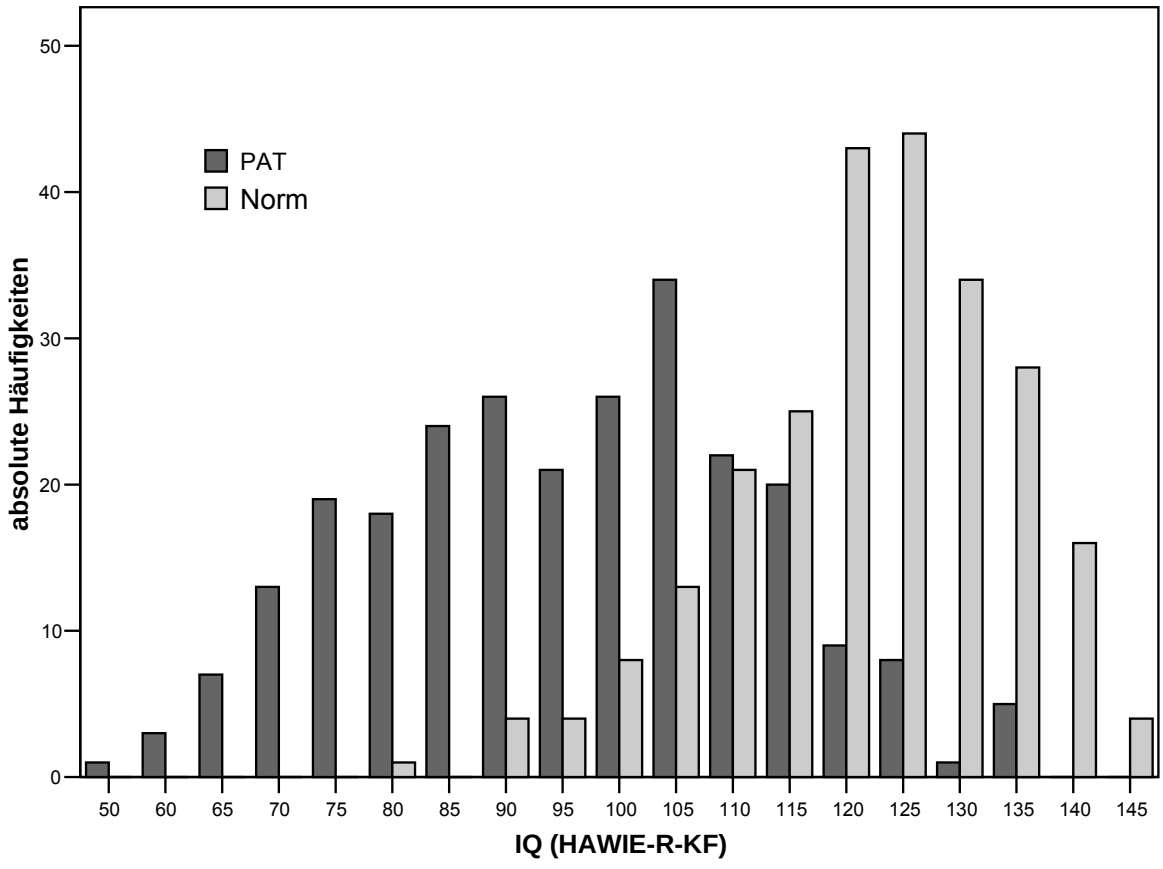


Abbildung 7-A: Verteilung der Intelligenzquotienten in den Ausgangsdaten der Normierungsstichprobe (Norm, n = 245) und der Patientenstichprobe (PAT, n = 257).

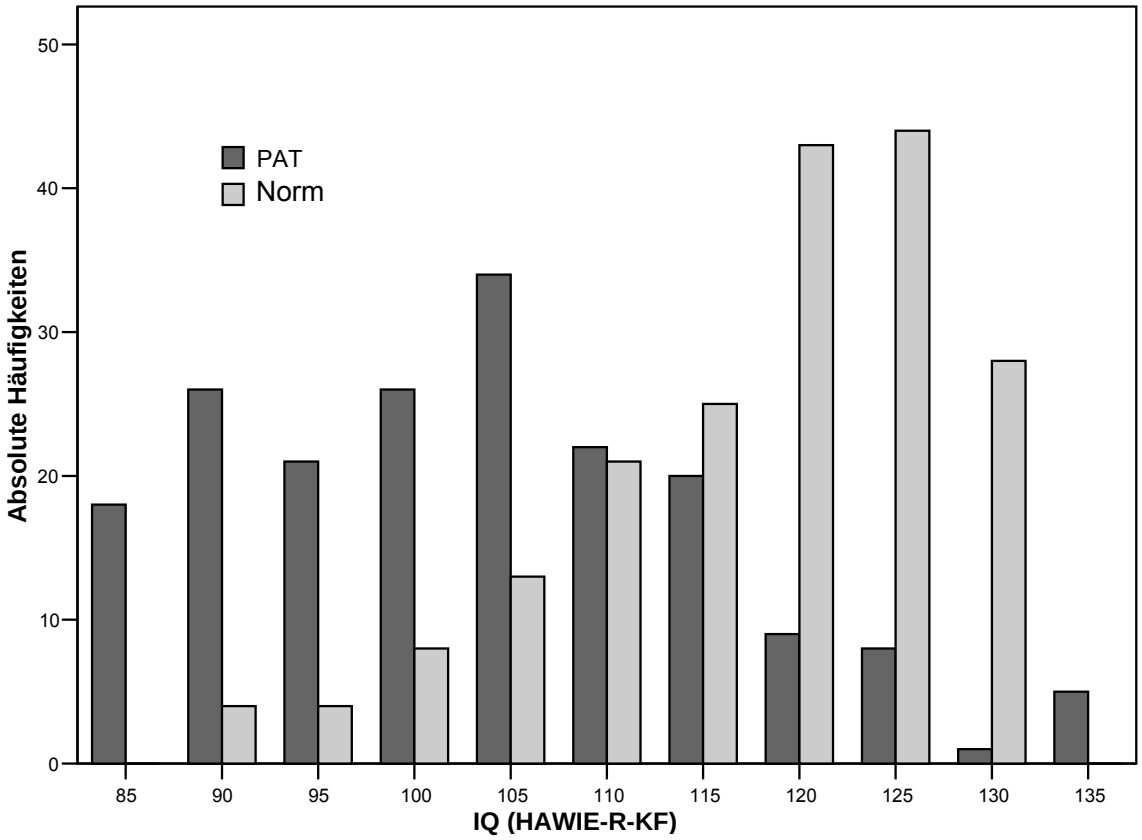


Abbildung 7-B: Resultat der Angleichung der Intelligenzverteilungen mittels Beschränkung der Spannbreite der Intelligenzquotienten der Normierungsstichprobe (Norm) nach oben, und der Patientenstichprobe (PAT) nach unten (n jeweils = 190).

Tabelle 7-C stellt vergleichend die Mittelwerte der Intelligenzquotienten beider Stichproben vor und nach Angleichung der Verteilungen dar. Dabei zeigt sich, dass der Unterschied zwischen den Patienten und der Normierungsstichprobe erwartungsgemäß kleiner geworden ist und mit 14.59 Punkten ungefähr einer Standardabweichung entspricht, aber auch nach der Angleichungsprozedur hoch signifikant ist ($p < .001$). Zwar ist eine Leistungsbeeinträchtigung von einer Standardabweichung ein typisches Ergebnis repräsentativer Studien zu kognitiven Leistungen bei Epilepsie (Bowden et al., 2004), die Mittelwerte beider Stichproben sind in den vorliegenden Daten jedoch um ca. eine Standardabweichung nach rechts verschoben.

Das Kaiser-Meyer-Olkin-Maß (KMO-Maß) für die Angemessenheit der Stichproben überprüft, ob die partiellen Korrelationen zwischen Variablen klein sind. Dieses Maß entspricht dem Varianzanteil in den Variablen, der durch gemeinsame Faktoren verursacht wird. Durch die Angleichungsprozedur fällt das KMO-Maß etwas geringer aus, liegt in beiden Stichproben aber weiterhin deutlich über dem Cut-Off-Wert von .5 (Hatcher, 1994) (.865 in der Patientenstichprobe, .857 in der Normierungsstichprobe).

Somit sind die abhängigen Variablen auch nach Kappung der Varianz gut für eine Faktorenanalyse geeignet.

Tabelle 7-C: Vergleich der mittleren Intelligenzquotienten beider Stichproben vor und nach der Angleichung der Intelligenzverteilungen anhand verschiedener statistischer Parameter (siehe Text).

Stichprobe	Gruppe	n	M (SD)	Mittelwertsvergleich	KMO
Ausgangsstichprobe	Patienten	257	93.74 (16.87)	F [1] = 396.752;	.914
	Normgruppe	245	119.92 (12.06)	p < .001	.889
Analysestichprobe	Patienten	190	101.32 (12.12)	F [1] = 166.253;	.864
	Normgruppe	190	115.91 (9.82)	p < .001	.857

Dieses Vorgehen ist als pragmatische Lösung des Problems der Verschiedenheit beider Stichproben zu verstehen. Das Vorliegen repräsentativer Stichproben, insbesondere einer repräsentativen Normierungsstichprobe, wäre freilich wünschenswert. Relevante Faktoren werden schwerer aufzufinden sein, weil die verminderte Varianz auch die zu erwartenden Korrelationen verringert. Inhaltlich ist die Generalisierbarkeit der Ergebnisse reduziert, insbesondere auf Patientenstichproben mit sehr niedrigem Intelligenzquotienten.

7.1.3 Die Analysestichproben

Tabelle 7-D zeigt die allgemeinen Stichprobencharakteristika der Analysestichproben. Über die oben beschriebenen Unterschiede im geschätzten Intelligenzniveau hinaus zeigen sich weiterhin Unterscheide in der Geschlechts-, Alters- und Bildungsstruktur. Die Normierungsstichprobe beinhaltet überproportional mehr Frauen, auch sind die Probanden etwas älter als die Patienten (40.92 Jahre vs. 36.58 Jahre) und haben ein höheres Bildungsniveau.

Tabelle 7-D: Allgemeine Stichprobencharakteristika der Analysestichproben.

	Normstichprobe (n = 190)	**Patienten (n = 190)**	**statistische Kennwerte**
Geschlecht (w/m)	122/68	85/105	χ^2 [1] = 14.527; p < .001
Alter M (SD)	40.92 (14.49)	36.58 (12.11)	F [1] = 9.993 ; p = .002
Mittlerer Intelligenzquotient[a]	115.91 (9.82)	101.32 (12.12)	F [1] = 166.253; p < .001
Schulbildung - kein Abschluss oder Sonderschule	0	3	Exakter Test nach Fisher 62.271; p < .001
- Haupt-/Volksschule	19	62	
- Mittlere Reife	53	76	
- Abitur	73	37	
- Hochschule	45	12	
Händigkeit (R/L/Ambidexter)	164/4/21	158/13/19	χ^2 [2] = 5.016; p < .087

Anmerkung: [a] geschätzt nach HAWIE-R (Tewes, 1991), Kurzform (Schwarzkopf-Streit, 2000)

Tabelle 7-E gibt die epilepsiebezogenen Stichprobencharakteristika der Analysestichprobe wieder. Die Dauer der Erkrankung beträgt im Schnitt 19.32 Jahre (SD = 12.41), begonnen hat die Erkrankung im Schnitt mit 17.26 Jahren (SD = 12.97). Bei 149 Patienten konnten mittels MRT eine oder mehrere strukturelle Läsionen gesichert werden. Über Ort und Lateralisation der Läsionen gibt Tabelle 7-F Auskunft.

Tabelle 7-E: Epilepsiebezogene Stichprobencharakteristika der Analysestichprobe (n = 190).

Klinische Charakteristika	**statistische Kennwerte**
Antiepileptische Medikamente (AEDs) [M (SD)]	1.97 (0.93)
- Häufigkeiten	11 x 0 AEDs 41 x 1 AED 91 x 2 AEDs 37 x 3 AEDs 10 x 4 AEDs
Dauer der Erkrankung [M (SD)]	19.32 (12.41)
Beginn der Erkrankung [M (SD)]	17.26 (12.97)
Fokalität (lt. MRT)	
- keine fokale Läsion	41 (21.6%)
- unifokal	119 (62.6%)
- multifokal	30 (15.8%)
Lateralisation (lt. MRT; n = 149)	
- rechts	61 (40.9%)
- links	71 (47.7%)
- bilateral	17 (11.4%)
Depressionsscore (BDI[a], n = 177) [M (SD)]	10.95 (8.92)

Anmerkung: [a] Beck-Depressions-Inventar (Hautzinger, Worall & Keller, 1995)

Tabelle 7-F: Häufigkeiten der Lokalisationen und Lateralisationen der MRT-gesicherten Läsionen bei Patienten mit symptomatischen Epilepsien.

Region	Ort	links	bilateral	rechts
Temporallappen	Hippocampus	29	5	31
	temporo-lateral	9	-	13
	Temporalpol	5	-	3
	Amygdala	3	-	-
	temporal medial und lateral	8	-	5
	temporo-basal	2	1	2
Frontallappen	frontal	13	1	7
	Zentralregion	-	-	1
	Gyrus frontalis inferior	1	-	-
	fronto-mesial	1	-	1
	anteriorer Balken	-	1	-
Weitere	parietal	2	-	2
	occipital	-	1	-
	temporo-parietal	1	-	1
	parieto-occipital	2	-	2
	temporo-occipital	3	-	3
	Insel	-	-	1
	hemisphärisch	4	-	2
	Hypothalamus	1	1	-
	Gyrus cinguli	-	1	-
	multiple nicht-kortikale Läsionen	2	2	-

7.2 Tests und Indikatoren

Im Folgenden werden die Indikatoren, die in die konfirmatorischen Faktorenanalysen eingehen, beschrieben. Zunächst wird die Bonner neuropsychologische Testbatterie vorgestellt. Da diese Testbatterie eine Vielzahl von Untertests und Testparametern umfasst, ist vor dem Hintergrund der begrenzten Stichprobengrößen eine Auswahl einiger Testindikatoren unumgänglich.

7.2.1 Die neuropsychologische Testbatterie

Die Untertests entstammen dem neuropsychologischen präoperativen Testprotokoll der Klinik für Epileptologie in Bonn (Helmstaedter, 2000). Die Testzusammenstellung hat sich für die Beantwortung einer Vielzahl typischer Fragestellungen der epileptologischen Neuropsychologie bewährt; insbesondere können anhand des Testprofils wichtige lateralisations- und lokalisationsdiagnostische Hinweise für die präoperative Diagnostik gewonnen werden. Die Testbatterie besteht aus 18 Untertests. Einige davon liefern

mehr als ein relevantes Teilergebnis (etwa die Lern- und Gedächtnistests) oder bestehen aus mehreren Teilaufgaben (z. B. der *Trail Making Test* (TMT, Reitan, 1979)). Aufmerksamkeit, Gedächtnis, Sprache, Intelligenz, exekutive Funktionen und visuell-räumliche Leistungen werden als zentrale kognitive Funktionsbereiche erfasst. In Hinblick auf die theoretische Interpretation der Ergebnisse dieser Arbeit ist anzumerken, dass es sich bei dieser – wie auch bei jeder anderen – Testauswahl um eine kleine und hoch selektive Stichprobe aus dem Universum kognitiver Tests handelt. Sie stellt keine auch nur annähernd umfassende Abbildung aller wichtigen Konstrukte menschlicher Intelligenz dar.

Zwei im präoperativen Testprotokoll vorgesehene Untertests wurden von vorneherein von allen weiteren Analysen ausgeschlossen und werden auch im Folgenden nicht näher besprochen: Im *Token-Test* (Orgass, De Renzi & Vignolo, 1982) zum Sprachverständnis- und Aphasiescreening erbrachten 97.6% der hirngesunden Probanden eine fehlerfreie Leistung. Entsprechend weist dieser Test nahezu keine Varianz auf. In der Aufgabe zum *Fingertapping* (klinikinterne Version von C. Hoppe) fehlen häufig Werte, weil die computerisierte Aufgabe wegen des hohen apparativen Aufwandes nur unregelmäßig eingesetzt wurde. Deshalb wird dieser Test nicht berücksichtigt. Die verbleibenden Tests werden im Folgenden hinsichtlich des Testprinzips, der erfassten Leistungsbereiche und der relevanten Indikatoren beschrieben.

Verbaler Lern- und Merkfähigkeitstest (VLMT)

Der Verbale Lern- und Merkfähigkeitstest (VLMT, Helmstaedter et al., 2001) ist ein Test zum seriellen verbalen Listenlernen mit nachfolgender Distraktion, Abruf nach Distraktion und halbstündiger Verzögerung und Wiedererkennen. Zunächst werden 15 Wörter vorgelesen. Anschließend soll der Proband alle noch erinnerbaren Wörter wiedergeben. Dies wird insgesamt fünfmal durchgeführt. Daran schließt sich die Vorgabe einer zweiten Liste (Liste B), bestehend aus 15 anderen Wörtern, an. Nachdem diese vom Probanden reproduziert wurde, soll nochmals die erste Liste wiedergegeben werden (Durchgang 6). Nach ca. einer halben Stunde werden ein freier Abruf der ersten Liste (Durchgang 7) und eine Wiedererkennensaufgabe durchgeführt. Der zentrale Indikator für die Lernkapazität ist die Lernleistung über alle fünf Lerndurchgänge hinweg (Summe richtig erinnerter Wörter), als Indikatoren für die langfristige Enkodierungs- bzw. Abrufleistung gelten die Leistung im verzögerten Abruf (Durchgang 7), der Verlust beim verzögerten Abruf (Differenzwert: Durchgang 5 minus Durchgang 7) sowie die

fehlerkorrigierte Rekognitionsleistung. Weitere Indikatoren sind die Leistung im Durchgang 1, im Durchgang 5 sowie die Leistung im Abruf der Interferenzliste (Liste B).

Diagnosticum für Cerebralschädigungen (DCS; abgewandelte Bonner Version)

Das Diagnosticum für Cerebralschädigungen (DCS, Helmstaedter, Pohl, Hufnagel & Elger, 1991; Weidlich, Lamberti & Hartje, 2001) ist ein Lern- und Gedächtnistest für figurales Material. Der Test ist so aufgebaut, dass der Proband sukzessive neun geometrische Figuren sieht und dann die behaltenen Figuren mit fünf Holzstäbchen nachlegen muss, wobei in der Bonner Version fünf Lerndurchgänge vorgesehen sind. Nach ca. einer halben Stunde erfolgt eine Rekognitionsbedingung. Der Test erfasst in erster Linie mnestische Hirnfunktionsstörungen. In die Testleistung gehen aber auch Faktoren wie Gestaltwahrnehmung, Gestaltspeicherung und -reproduktion sowie selektive Aufmerksamkeitszuwendung ein.

Zahlen- und Blockspanne in Anlehnung an WMS-R

Aus dem Wechsler-Gedächtnistests (WMS-R, Härting et al., 2000) entstammen drei Untertests zum Kurzzeit- und Arbeitsgedächtnis: In der Aufgabe *Zahlennachsprechen vorwärts* hat der Proband sukzessiv immer länger werdende Zahlenfolgen nachzusprechen, bei der Aufgabe *Zahlennachsprechen rückwärts* werden ebenfalls Zahlenfolgen zunehmender Länge vorgegeben, die anschließend in umgekehrter Reihenfolge reproduziert werden sollen. In der Aufgabe *Blockspanne* (*Corsi block-tapping*) wird eine zunehmend umfangreichere Auswahl von auf einen Brett befestigten Holzblöcken angetippt. Der Proband soll diese Tipp-Sequenz reproduzieren. In Abweichung zur Originalversion der WMS-R wurde als abhängige Variable die längste erreichte Zahlen- bzw. Blockfolge erfasst (diese wurde in mindestens einem von zwei Durchgängen richtig reproduziert).

Kurztest für cerebrale Insuffizienz (c.I.-Test)

Im Untertest *Symbolezählen* des c.I.-Tests (Lehrl & Fischer, 1997) soll der Proband alle Quadrate auf einer Tafel mit drei verschiedenen Symbolen zählen. Im Untertest *Interferenz* soll eine unregelmäßige Folge der Buchstaben A und B in invertierter Form vorgelesen werden (statt ABAAB soll BABBA vorgelesen werden). Beide Aufgaben sind möglichst schnell zu bearbeiten. Der Test erfasst selektive Aufmerksamkeit und kognitive Geschwindigkeit, im Interferenztest ist konzentrativer Widerstand gegenüber dominierenden Reaktionstendenzen, ähnlich wie im bekannten Stroop-Paradigma, gefordert.

Trail Making Test (TMT; Teil A und B)

Beim *Trail Making Test* (Reitan, 1979) sollen in der ersten Version (*TMT A*) die Zahlen von 1 bis 25 in aufsteigender Reihenfolge miteinander verbunden werden. *TMT B* erfordert das alternierende Verbinden von Zahlen und Buchstaben (1-a-2-b etc.). Der Test erfasst kognitive und psychomotorische Geschwindigkeit, visuelles Scanning, kognitives Tracking und in der B-Version zusätzlich kognitive Flexibilität und Aufmerksamkeitswechsel.

Aufmerksamkeits-Belastungs-Test d2

Der Test d2 (Brickenkamp, 2002) ist ein Durchstreichtest und erfordert das möglichst schnelle Durchstreichen bestimmter Zeichen („d“ mit zwei flankierenden Strichen). Der Test misst Tempo und Sorgfalt des Arbeitsverhaltens bei der Unterscheidung ähnlicher visueller Reize (Detaildiskrimination) und ermöglicht damit die Beurteilung individueller Aufmerksamkeits- und Konzentrationsleistungen. Zentrale Testwerte sind die Gesamtzahl, die fehlerkorrigierte Gesamtzahl, der Fehlerprozentsatz und die Schwankungsbreite als Maß der Konsistenz von Sorgfalt.

Motorische Sequenzierung nach A.R. Luria

Hierbei wird die Durchführung verschiedener handmotorischer Bewegungsabfolgen (Sequenzen) in Adaptation einer von A. R. Luria beschriebenen Aufgabe überprüft (Luria, 1973). Die verschiedenen Bewegungen werden zunächst vom Testleiter vorgeführt und sollen anschließend vom Probanden nachgemacht werden. Vorgegeben werden eine unimanuelle Bewegungsabfolge, erst mit der rechten, dann mit der linken Hand sowie zwei bimanuelle Bewegungssequenzen. Erfasst wird die Fähigkeit der sequenziellen motorischen Organisation. In die Analysen geht jeweils die Bewertung der Richtigkeit der Reproduktion auf einer Skala von 1 (richtige Reproduktion) bis 4 (stark auffällig) ein.

Perdue Pegboard Test

Der *Perdue Pegboard Test* (Tiffin, 1968) ist eine Aufgabe zur Hand- und Armmotorik. Der Proband hat dabei uni- und bimanuell kleine Metallstifte in die Löcher eines Steckbrettes (*Pegboard*) zu stecken (Aufgabe zur Grobmotorik von Händen, Fingern und Armen) und auch bimanuell komplexere Montagen mittels mehrerer Einzelteile vorzunehmen (Aufgabe zur Feinmotorik bzw. „Fingerspitzengefühl“). Die Rohwerte beziehen sich auf die Einzelteile pro vorgegebenem Zeitintervall (30 bis 60 Sekunden).

Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE-R), Kurzform

Beim HAWIE-R (Tewes, 1991) handelt es sich um einen Intelligenztests für Erwachsene (deutsche Adaptation der *Wechsler Adult Intelligence Scale - revised* von D. Wechsler (Wechsler, 1981)). Die hier eingesetzte Kurzform besteht aus fünf Untertests, wobei die Korrelationen zwischen HAWIE-R und HAWIE-R Kurzform r = .97 (Schwarzkopf-Streit, 2000) beträgt. In die Analysen gehen die Untertest-Summenwerte der nachfolgend kurz beschriebenen fünf Untertests ein: Im Untertest *Bilderergänzen* hat der Proband fehlende Elemente in konkreten Strichzeichnungen zu entdecken. Erfasst werden soll hier die Fähigkeit, zwischen wichtigen und unwichtigen Details bei visuellen Vorlagen zu unterschieden, ferner Wahrnehmungsgenauigkeit und Begriffsbildung. Im *Wortschatztest* sind kurze Definitionen für vorgegebene Begriffe abzugeben. Dieser Test erfasst laut Handbuch allgemeine Intelligenz, zusätzlich Lernfähigkeit und verbale Informationsbreite. Im *Mosaiktest* soll der Proband mittels vier bzw. neun Mosaiksteinen mehrere vorgegebene Muster möglichst rasch nachbauen. Der Test erfasst die Fähigkeit, Formen wahrzunehmen und sie zu analysieren und das Ganze in seine Komponenten zu zerlegen; zusätzlich wird die Belastung unter Zeitdruck erfasst. Im Untertest *Gemeinsamkeitenfinden* soll der Proband zu jeweils zwei vorgegebenen Wörtern das Gemeinsame bzw. einen passenden Oberbegriff nennen. Erfasst werden sprachliche Fähigkeiten, Wortschatz und vor allem das sprachliche Abstraktionsvermögen. Im Test *Rechnerisches Denken* sind kurze mündlich vorgegebene Textaufgaben zu lösen. Der Test ist in starkem Maße von schulischer und beruflicher Erfahrung und vom Konzentrationsvermögen abhängig.

Mehrfachwahl-Wortschatz-Intelligenztest (MWT-B)

Der MWT-B (Lehrl, 1999) dient der Bestimmung des allgemeinen Intelligenzniveaus und der kristallinen Intelligenz, im Speziellen der Bestimmung der Fähigkeit, Bekanntes von Unbekanntem zu unterscheiden. Der Test besteht aus 37 Wortreihen, wobei in jeder Zeile ein umgangsbildungs- oder wissenschaftssprachlich bekanntes Wort unter vier sinnlosen Wörtern steht. Der Proband soll das bekannte Wort herauszufinden. In die folgenden Analysen geht die Gesamtzahl richtig erkannter Wörter ein.

Wortflüssigkeit (semantisch und phonematisch)

Bei den Wortflüssigkeitsaufgaben hat der Proband innerhalb von jeweils einer Minute zu vorgegebenen phonematischen oder semantischen Kategorien möglichst viele Wörter zu generieren. Bei der semantischen Wortflüssigkeitsaufgabe (siehe z. B. Spreen & Strauss, 1998) ist innerhalb von einer Minute eine mündliche Aufzählung möglichst

vieler Tiere gefordert. Bei der phonematischen Aufgabe ist die Notation möglichst vieler, mit einem vorgegebenen Anfangsbuchstaben beginnender Wörter, gefordert. Die relevanten Anfangsbuchstaben sind L, P und S (entsprechend Untertest 6 „Wortflüssigkeit“ des Leistungsprüfsystems (LPS, Horn, 1983)). Pro Buchstabe hat der Proband wieder eine Minute Zeit. Der Gesamtrohwert der phonematischen Teilaufgabe ist die Summe der korrekten Nennungen über die drei Minuten hinweg. Erfasst wird die Schnelligkeit des Zugriffs auf den lexikalischen Speicher, strategisches Vorgehen und semantisches Altgedächtnis. Aufgrund der Unterschiede in der Abrufkategorie (semantisch vs. phonematisch) und der Antwortmodalität (mündlich vs. schriftlich) sollen beide Teilaufgaben der Wortflüssigkeit in die Analysen eingehen.

Boston-Naming-Test (BNT)

Beim Boston-Naming-Test (Kaplan, Goodglass & Weintraub, 1976) werden dem Probanden bis zu 60 konkrete Strichzeichnungen (z. B. von Gegenständen oder Tieren) vorgegeben, die er benennen muss. Erfasst werden Wortfindung bzw. Wortabruf und Lexikon. In die folgenden Analysen geht die Summe korrekt benannter Abbildungen ein.

Räumliches Rotieren (Leistungsprüfsystem LPS)

Beim Untertest 7 des Leistungsprüfsystems (Horn, 1983) soll der Proband innerhalb von zwei Minuten aus mehreren Zeilen mit jeweils gleichen, aber in der Ebene gedrehten Buchstaben oder Zahlen dasjenige Zeichen herausfinden, das nicht nur gedreht, sondern auch gespiegelt ist. Hierbei werden räumliches Vorstellungsvermögen und die Fähigkeit zur mentalen Rotation erfordert. Gegenstand der folgenden Analysen ist die Anzahl richtig erkannter Zeichen in der vorgegebenen Zeit.

Labyrinthtest

Beim Labyrinthtest (Chapuis, 1992) soll der Proband den aus drei zunehmend schwierigeren (größeren) Labyrinthen führenden Weg einzeichnen. Gefordert wird dabei neben selektiver Aufmerksamkeit die Fähigkeit zur visuellen Antizipation. Rohwerte sind der Zeitscore (Gesamtbearbeitungszeit für alle drei Labyrinthe) als auch die Gesamtfehlersumme. Zwar wirken sich Fehler negativ auf die Zeit aus, da dadurch Zurückziehen erforderlich wird. Andererseits kann mit einer etwas „mutigeren“ Strategie ein höheres Tempo erreicht werden, eine fehlerfreie Leistung jedoch nur durch überlegtes, planendes und daher zeitintensiveres Vorgehen.

7.2.2 Auswahl der Modellindikatoren

Für robuste Ergebnisse konfirmatorischer Faktorenanalysen sollte eine minimale Stichprobengröße nicht unterschritten werden und gleichzeitig ein bestimmtes Verhältnis zwischen Anzahl der Variablen und Probandenanzahl eingehalten werden. Für die Bestimmung der notwendigen minimalen Stichprobengröße existieren lediglich Daumenregeln (Kelloway, 1998): So wird für Modelle mittlerer Komplexität eine Stichprobengröße von n = 200 gefordert (Boomsma, 1983; zitiert bei Kelloway, 1998). Auch für Mehr-Gruppen-Analysen scheinen 200 Probanden pro Gruppe eine hinreichende Gruppengröße darzustellen (Hoelter, 1983; Levine et al., 2003). Bentler und Chou (1987, ebenfalls zitiert bei Kelloway, 1998) schlugen vor, dass das Verhältnis von Stichprobengröße und geschätztem Parameter zwischen 5:1 und 10:1 sein sollte. Eine weitere Daumenregel besagt, dass 10 Probanden pro Testvariable oder 5 Probanden pro freiem Parameter notwendig sind. Für die aktuelle Fragestellung ist es andererseits wichtig, möglichst viele Indikatoren in das Modell aufzunehmen, damit die Faktoren reliabel operationalisiert sind und breite Intelligenzstrukturen überprüft werden können.

Bei Aufnahme von 24 Testvariablen wäre für die Ausgangsstichproben (245 bzw. 257 Probanden) die 10:1 Relation bezüglich der Indikatoren erfüllt. Je nach Modellkomplexität wird die 5:1 Relation für die freien Parameter allerdings leicht unterschritten: Beispielsweise wären in einem typischen konfirmatorischen Modell der Kovarianzstruktur mit sechs interkorrelierten Faktoren, die jeweils durch vier Untertests operationalisiert sind, 63 Parameter frei zu schätzen[6] und somit 63 x 5, also 315 Probanden pro Gruppe notwendig. Ein analoges Modell mit fünf Faktoren besäße 40 frei schätzbare Parameter und würde 200 Probanden erfordern. Um möglichst robuste Modellschätzungen zu erreichen, sollen als Kompromiss 20 Indikatoren in die Analyse eingehen. Das macht eine Selektion der Modellindikatoren unumgänglich.

Die Selektion einer geeigneten Untermenge von Modellindikatoren erfolgt in zwei Schritten. Die erste Selektionsphase ist von folgenden Gesichtspunkten geleitet: Zum einen sollten die in die konfirmatorischen Faktorenanalyse eingehenden Untertests konzeptionell unabhängig voneinander sein (Kriterium der Nicht-Redundanz). Zweitens sollten pro Untertest nicht mehr als zwei Indikatoren gewählt werden, da ansonsten aufgrund gemeinsamer Methodenvarianz ein eigener Faktor entstehen könnte, der aber theoretisch bedeutungslos wäre. Um die Indikatoren anhand dieser Kriterien empirisch fundiert zu selektieren, wurden in der Normierungsstichprobe verschiedene

[6] Dies wären, um es explizit zu machen, 15 Faktorenkovarianzen, 6 Faktorvarianzen, 18 Ladungen und 24 Fehlervarianzen. Bei zusätzlicher Schätzung von Fehlerkovarianzen oder im Rahmen der Mittelwertstruktur der Faktorenmittelwerte und der Itemhöhenlagen wären entsprechend mehr Parameter zu schätzen.

Faktorenanalysen berechnet. Deren Ergebnisse und die daraus resultierenden Schlussfolgerungen für die Auswahl der Indikatoren werden kurz erläutert. Schließlich sind für die Auswahl der in die Analysen eingehenden Variablen auch theoretisch-inhaltliche und psychometrische Überlegungen zu berücksichtigen.

Der Verbale Lern- und Merkfähigkeitstest (VMLT) stellt zur Ergebnisinterpretation sehr viele Testvariablen zur Verfügung. Eine Faktorenanalyse mit diesen Parametern erbrachte zwei Komponenten (siehe Tabelle 7-G; Mustermatrix, Ladungen ≤.4 unterdrückt).

Tabelle 7-G: Oblique Faktorenanalyse des VLMT.

Parameter	**Komponenten**	
	1	**2**
Differenz 5 minus 7	-.924	
Durchgang 7	.914	
Durchgang 6	.812	
Rekognition (fehlerkorrigiert)	.689	
Interferenzliste		.871
Durchgang 1		.821
Summe Durchgang 2 bis 5	.518	.576
% Varianzaufklärung	61%	17%

Anmerkungen: Hauptkomponentenanalyse mit obliquer Rotation; redundante Parameter wurden ausschlossen, z. B. Durchgang 5 UND Summe Durchgang 1 bis 5

Die zwei Komponenten korrelieren mit .4. Die Annahme ihrer Unabhängigkeit ist also nicht haltbar und somit ist die Wahl der obliquen Rotation gerechtfertigt. Komponente 1 klärt 61% der Gesamtvarianz auf und kann als Langzeitgedächtnis-Komponente interpretiert werden. Die Markervariablen sind der Verlust vom fünften zum siebten Durchgang (Differenz 5 minus 7) und die absolute Leistung im siebten Durchgang (verzögerter freier Abruf). Komponente 2 mit lediglich 17% Varianzaufklärung kann als Kurzzeitgedächtnis-Komponente gelten, wobei hier die unmittelbare Wiedergabeleistungen der Liste A und B Markervariablen sind. Die Lernleistung (nicht-redundante Summe der Durchgänge zwei bis fünf) lädt auf beiden Komponenten gleichermaßen. Dies verweist darauf, dass zu dieser Leistung sowohl Langzeit- als auch Kurzzeitgedächtnisprozesse beitragen (vergleiche Müller et al., 1997). Zur Untermauerung dieses Komponentenmusters sei angemerkt, dass bei Einschluss der Gesamtlernleistung (Summe über alle fünf Lerndurchgänge) diese statt der Lernleistung über die letzten vier Durchgänge hinweg nur auf der Kurzzeitgedächtniskomponente lädt. Dies allerdings drückt lediglich die Kontamination aufgrund der doppelten

Berücksichtigung des ersten Lerndurchgangs aus. Eine Faktorenanalyse mit der Gesamtlernleistung (Durchgang 1 bis 5), aber ohne Durchgang 1 erbringt nur einen Gedächtnis-Generalfaktor und stützt nicht mehr die theoretisch gut begründete Annahme getrennter Langzeit- und Kurzzeitgedächtniskomponenten. Basierend auf den Ergebnissen der oben dargestellten Faktorenanalyse werden für die weiteren Analysen beide Komponenten über eine einfache additive Verknüpfung ihrer Markervariablen operationalisiert[7].

Auch die weiteren Tests ergeben jeweils mehrere Ergebnisvariablen und erfordern eine Selektion der Modellindikatoren: Beim *Diagnosticum für Cerebralschädigungen* indizierte eine Faktorenanalyse (Lernleistung im ersten Durchgang, Summe Durchgang zwei bis fünf und Rekognition (fehlerkorrigiert)) Eindimensionalität. Entsprechend geht nur die Gesamtlernleistung (Summe Durchgang 1 bis 5) in die weiteren Analysen ein. Für den *Aufmerksamkeits-Belastungs-Test d2* wird aus theoretischen Überlegungen heraus nur die fehlerkorrigierte Gesamtzahl für die Analysen verwendet. Eine oblique Faktorenanalyse der motorischen Sequenzen nach Luria enthüllt zwei weitgehend unabhängige Faktoren (r = .22). Der erste Faktor umfasst die beiden unimanuellen Aufgaben, der zweite Faktor die bimanuellen Aufgaben. In die Analysen gehen die beiden Summenwerte der quantitativen Bewertung der Reproduktion ein. Auch der *Perdue Pegboard Test* ist laut Faktorenanalyse eindimensional. Also geht die Gesamtsumme in die folgenden Analysen ein. Für den *Labyrinthtest* ist die Korrelation zwischen dem Zeit- und dem Fehlerscore mit r = .14 zwar signifikant (p = .025), aber insgesamt niedrig, so dass sowohl der Zeitscore als auch die Fehlersumme in die weiteren Analysen eingehen sollen.

Das Ergebnis dieses ersten Selektionsschrittes sind 26 inhaltlich und statistisch nicht-redundante Testvariablen. Diese werden in Tabelle 7-H mit den für alle folgenden Ausführungen relevanten Kürzeln benannt und beschrieben.

[7] Kurzzeitgedächtnis: Durchgang 1 plus Liste B; Langzeitgedächtnis: Durchgang 7 minus (Durchgang 5 minus Durchgang 7).

Tabelle 7-H: Bezeichnung und Beschreibung der 26 Testvariablen nach dem ersten Selektionsschritt.

Variable	Beschreibung (Berechnungsvorschrift)
VLMT_LZG	VLMT, Durchgang 7 minus (Durchgang 5 minus Durchgang 7)
VLMT_KZG	VLMT, Durchgang 1 plus Liste B
DCS_15	DCS, Summe Durchgang 1 bis 5
DIGITS_V	Zahlennachsprechen vorwärts, längste erreichte Zahlenfolge
DIGITS_R	Zahlennachsprechen rückwärts, längste erreichte Zahlenfolge
CORSI_V	Blockspanne vorwärts, längste erreichte Folge
CIT_SZ	C.I.-Test, Untertest Symbolezählen, Zeit (s)
CIT_INT	C.I.-Test, Untertest Interferenz, Zeit (s)
TMT_A	TMT A, Zeit (s)
TMT_B	TMT B, Zeit (s)
D2_GZF	Test d2, Gesamtzahl minus Fehler
LURIA_RL	Luria-Sequenzen, unimanuell (rechts, links)
LURIA_BI	Luria-Sequenzen, bimanuell
PEGBOARD	Perdue Pegboard Test, Gesamtsumme über alle Untertests
BE_HAWIE	Rohwert Untertest Bilderergänzen, HAWIE-R
WT_HAWIE	Rohwert Untertest Wortschatztest, HAWIE
MT_HAWIE	Rohwert Untertest Mosaiktest, HAWIE
GF_HAWIE	Rohwert Untertest Gemeinsamkeitenfinden, HAWIE
RD_HAWIE	Rohwert Untertest Rechnerisches Denken, HAWIE
MWT_B	Mehrfachwahl-Wortschatz-Intelligenztest, Rohwert
WFL_SEM	Semantische Wortflüssigkeit „Tiere“, Summe (1 min)
WFL_PHO	Phonematische Wortflüssigkeit „L, P, S“, Summe (3 min)
BNT	Summe richtiger Bennennungen im Boston Naming Test 2
LPS_7	Summe, Leistungsprüfsystem 7, „mentale Rotation“
LAB_FEHL	Chapuis-Labyrinthtest, Fehler
LAB_ZEIT	Chapuis-Labyrinthtest, Zeit (s)

Das Ziel von 20 Variablen ist mit dieser Auswahl noch nicht erreicht, deshalb ist eine zweite Selektionsphase notwendig. Die Anzahl der Modellindikatoren wird hierzu mithilfe *deskriptiv-statistischer* Überlegungen weiter reduziert. Als Selektionskriterien werden die Reliabilitäten (Retestreliabilität) und die Kommunalitäten (Varianzanteil jeder Variable, der durch den Rest aller anderen Variablen erklärbar ist) herangezogen. Diese Kennwerte finden sich in Tabelle 7-I. Als weitere Gütekriterien für die Eignung der Testvariablen als Modellindikatoren sind in der Tabelle die *Schiefe* (als Maß der Asymmetrie der Verteilung) und der *Exzess* (Schmal- bzw. Breitgipfligkeit der Verteilung) dargestellt. Diese Werte sind wichtig, da die konfirmatorische Faktorenanalyse eine multivariate Normalverteilung voraussetzt. Kline (1998) schlägt Cut-Off-Werte von 3.0 für die Schiefe und 8.0 für den Exzess vor, um einen möglichen

Verstoß grob abzuschätzen. In der ersten Wertespalte der Tabelle 7-I ist als Maß der *Eignung* jeder Variablen für die Faktorenanalyse das Kaiser-Meyer-Olkin-Maß dargestellt. Das Maß entspricht dem Varianzanteil in den Variablen, der durch gemeinsame Faktoren verursacht wird.

Tabelle 7-I: Eignung, Kommunalität, Reliabilität, Schiefe und Exzess der 26 Testvariablen.

Indikator	Eignung	Kommunalität[1]	Reliabilität[2]	Schiefe	Exzess
CIT_INT	.957	.443	.731	.875	1.071
LPS_7	.940	.467	.780	.107	-.724
DCS_15	.934	.588	.773	-.747	.052
D2_GZF	.934	.533	.851	-.076	-.339
TMT_B	.926	.585	.709	1.883	5.193
BE_HAWIE	.916	.366	.614	-1.898	5.243
VLMT_LZG	.914	.358	.663	-1.012	.537
VLMT_KZG	.913	.389	.477	.830	1.523
TMT_A	.910	.516	.715	1.177	1.333
PEGBOARD	.903	.461	.612	-.082	.209
CIT_SZ	.902	.319	.531	3.790	24.703
WFL_SEM	.890	.335	.589	.483	.320
MT_HAWIE	.890	.604	.712	-.448	-.431
GF_HAWIE	.883	.542	.662	-1.885	4.853
WFL_PHO	.879	.604	.802	-.170	.261
LURIA_BI	.858	.302	.606	2.635	6.977
WT_HAWIE	.856	.652	.812	-.622	-.412
LAB_ZEIT	.852	.442	.630	1.199	1.487
CORSI_V	.850	.295	.358	.248	-.005
LAB_FEHL	.837	.320	.579	2.080	7.810
DIGITS_R	.827	.446	.507	.447	-.605
BNT	.824	.331	.748	-1.991	8.091
RD_HAWIE	.823	.373	.723	-.866	.955
DIGITS_V	.793	.378	.682	.241	-.494
MWT_B[3]	.788	.511	.870	-1.206	1.944
LURIA_RL	.578	.241	.698	7.762	68.506

Anmerkungen: [1] Anfängliche Kommunalitäten; Extraktionsmethode: Hauptachsen-Faktorenanalyse; [2] Werte aus der Diplomarbeit von Mladenka Gresch (Gresch, 2005); [3] Reliabilität laut Testmanual. Grau unterlegt sind die anhand verschiedener Kriterien ausgeschlossenen Variablen (siehe Text).

Anhand dieser Parameter (Reliabilität, Kommunalität, Schiefe, Exzess) wurden folgende Variablen von den weiteren Analysen ausgeschlossen:

- Das Corsi-Block-Tapping (CORSI_V) hat eine sehr geringe Retestreliabilität (Rang 25 von 26) und gleichzeitig eine sehr niedrige Kommunalität (Rang 25).

- Die unimanuellen Sequenzen nach Luria (LURIA_RL) haben die niedrigste Reliabilität (Rang 26) und überschreiten mit einem Wert von 68.506 den Exzess-Cut-Off von 8.0 maximal. Zudem wurde diese Aufgabe in der Normierungsstichprobe zumeist fehlerfrei gemeistert.
- Die bimanuelle Sequenzierungsaufgabe (LURIA_BI) hat ebenfalls eine sehr geringe Reliabilität (Rang 24). Zwar sind die anderen Parameter weitgehend akzeptabel und erzwingen den Ausschluss dieser Variable nicht. Aber aus inhaltlichen Gründen soll auf eine Operationalisierung eines motorischen Faktors zugunsten der weiteren kognitiven Faktoren verzichtet werden, da die unimanuelle Aufgabe ohnehin ausgeschlossen wurde.
- Für den Pegboard-Test (PEGBOARD) gilt das Gleiche wie bei der Variable LURIA_BI.
- Der Test Symbolezählen (CIT_SZ) hat eine sehr niedrige Reliabilität (Rang 23) und einen zu hohen Exzess (24.703).
- Der Fehlerwert des Labyrinthtests hat eine niedrige Reliabilität (Rang 22) bei relativ hohem Exzess.

Die im zweiten Selektionsschritt ausgeschlossenen Variablen erscheinen verzichtbar: Drei der sechs entfernten Variablen sind motorische Testwerte, es entfällt also ein motorischer Faktor. Der Test *Symbolezählen* ist ein reiner Schnelligkeitstest. Der Verzicht darauf erscheint unproblematisch, da mit dem *Interferenztest*, dem *Trail Making Test*, dem *Labyrinthtest* und dem *Test d2* dieses Konstrukt breit abgedeckt ist. Der Fehlerwert des *Labyrinthtests* ist ein Subindex der Labyrinthaufgabe und tendenziell verzichtbar. Allenfalls der Ausschluss der *Blockspanne* könnte zu einer Verringerung der theoretischen Breite der Testkonstrukte führen. Doch wegen der mangelnden Reliabilität kann der Test nicht beibehalten werden.

Schließlich soll im letzten Schritt die Eignung der Indikatorenauswahl für eine faktorenanalytische Studie überprüft werden. Da dieser Bewertung die Test-Interkorrelationen zugrunde liegen, sind die Datensätze zunächst auf Extremwerte zu kontrollieren, da diese zu deutlichen Verzerrungen von Korrelationen führen können. Dazu wurden in der SPSS-Prozedur „Deskriptive Datenanalyse“ (SPSS, 2003) Boxplots zum Anzeigen von Ausreißern und Extremwerten ermittelt. Obwohl es sich dabei nicht zwingend um Messfehler handelt, sollen zumindest die Extremwerte, also die Werte, die mehr als drei Balkenlängen von der oberen oder unteren Kante der Box entfernt sind, durch den Wert, der diese Grenze markiert, ersetzt werden (Berechnungsvorschrift: $T = Q_3 + 3 * IQR$). Ausreißer (Werte von 1.5 bis 3 Boxlängen vom oberen oder unteren Rand

der Box entfernt) werden unverändert beibehalten. In der Normierungsstichprobe (n = 245) mussten 13 Extremwerte ersetzt werden, in der Patientenstichprobe neun Werte[8].

Nach der Korrektur der Extremwerte kann das Maß der Stichprobeneignung (*sampling adequacy)* nach Kaiser-Meyer-Olkin ermittelt werden. Das Maß entspricht dem Varianzanteil in den Variablen, der durch gemeinsame Faktoren verursacht wird. Es beträgt für die Normierungsstichprobe 0.891 und liegt über dem Cut-Off von 0.5 (Hatcher, 1994). Auf Variablenebene liegen alle Werte über .805. Somit sind alle Variablen für eine Faktorenanalyse geeignet. Ebenso bestätigt der Bartlett-Test auf Sphärizität die Eignung der Datenmatrix für eine Faktorenanalyse (ungefähres χ^2 = 1965.570; $p < .001$) und verweist auf eine hoch-signifikante Abweichung von der Einheitsmatrix. Des Weiteren wurde überprüft, inwieweit die Datenmatrix der Prämisse der positiven Mannigfaltigkeit entspricht (Spearman, 1927). Aus Tabelle 7-J ist für die Normierungsstichprobe ersichtlich, dass von den 190 nicht-redundanten Korrelationen 170 (89.5%) signifikant positiv sind; 17 Korrelationen (8.9%) weisen ein positives Vorzeichen auf, sind aber nicht signifikant und nur drei Korrelationen (1.5%) sind negativ oder gleich Null. Das bestätigt die Annahme der positiven Mannigfaltigkeit.

Weiterhin zeigt Tabelle 7-J, dass bei 11 der 20 nicht-signifikanten Korrelationsindizes der Untertest MWT-B involviert ist. Anders ausgedrückt sind 11 aller 19 möglicher Interkorrelationen mit dem Untertests MWT-B nicht-signifikant positiv. Dies deutet darauf hin, dass dieser Untertest für die faktorenanalytischen Auswertungen ungeeignet ist. Daher soll er aus den folgenden Analysen ausgeschlossen werden. Möglicherweise hängt die Testleistung positiv vom Alter ab; für viele der anderen Variablen (z. B. die Ergebnisse der Geschwindigkeitstests) sind eher negative Alterskorrelationen zu erwarten (zu weiteren Schwächen des MWT-B siehe Satzger, Fessmann & Engel, 2002).

[8] In der Normierungsstichprobe handelte es sich um folgende Werte: viermal GF_HAWIE [T = 18], dreimal BE_HAWIE [T = 8], zweimal TMT_B [T = 156], einmal TMT_A [T = 69], einmal MWT_B [T = 21], einmal LAB_ZEIT [T = 704], einmal BNT [T = 41]). In der Patientenstichprobe wurden folgende Werte ersetzt: dreimal TMT_B [T = 321], viermal TMT_A [T = 97] und zweimal CIT_INT [T = 55].

Tabelle 7-J: Interkorrelationsmatrix der 20 Testindikatoren.

	WT_HAWI	WFL_PHO	MT_HAWIE	DCS_15	TMT_B	GF_HAWIE	D2_GZF	TMT_A	MWT_B	LPS_7	DIGITS_R	CIT_INT	LAB_ZEIT	VLMT_KZ	DIGITS_V	RD_HAWIE	BE_HAWIE	VLMT_LZG	WFL_SEM
WFL_PHON	++ +	1																	
MT_HAWIE	++ +	++ +	1																
DCS_15	++ +	++ +	++ +	1															
TMT_B	++ +	++ +	++ +	++ +	1														
GF_HAWIE	++ +	++ +	++ +	++ +	++ +	1													
D2_GZF	++ +	++ +	++ +	++ +	++ +	++ +	1												
TMT_A	++ +	++ +	++ +	++ +	++ +	++ +	++ +	1											
MWT_B	++ +	++ +	+	+	+	++ +	++	-	1										
LPS_7	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	+	1									
DIGITS_R	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	1								
CIT_INT	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	+	++ +	++ +	1							
LAB_ZEIT	+	++	++ +	++ +	++ +	++	++ +	++ +	+	++ +	++	++ +	1						
VLMT_KZG	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	+	++ +	++ +	++ +	++ +	1					
DIGITS_V	++ +	++ +	++ +	++ +	++ +	++ +	++	+	++ +	++ +	++ +	++ +	++ +	++	1				
RD_HAWIE	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++	++ +	++ +	++ +	++ +	-	+	++ +	1			
BE_HAWIE	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	+	++ +	++ +	++ +	++ +	++ +	++	++	1		
VLMT_LZG	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	+	++	++ +	++ +	++ +	++ +	+	0	++ +	1	
WFL_SEM	++ +	++ +	++ +	++ +	++ +	++ +	++ +	++ +	+	++ +	++ +	++ +	++	++ +	++	+	++ +	++ +	1
BNT	++ +	++ +	++ +	++ +	++ +	++ +	++	++ +	++ +	++ +	+	++	+	++	++	++ +	++ +	++	++ +
n.s. positiv (von 19)	1	0	1	1	1	0	0	2	11	1	1	1	4	2	2	4	1	3	2

Anmerkungen: +++: Korrelation positiv und auf dem Niveau von .01 signifikant; ++: Korrelation positiv und auf dem Niveau von .05 signifikant; +: Korrelation positiv (n.s.); 0: Die Korrelation ist Null (n.s.); -: Korrelation ist negativ (n.s.). Variablen nach absteigender Kommunalität geordnet.

Zusammenfassend zeigt sich also, dass die Korrelationsmatrix für faktorenanalytische Auswertungen und oblique Faktorenrotationen geeignet erscheint. Durch verschiedene Selektionsprozesse wurde eine Auswahl von 19 Variablen für die weiteren Analysen zusammengestellt. Bei 19 Variablen ist eine minimale Stichprobengröße von 190 Probanden notwendig.

7.3 Überprüfung der konfiguralen Invarianz im Ein-Gruppen-Fall

Zur Überprüfung von Fragestellung A soll mittels konfirmatorischer Faktorenanalysen untersucht werden, welches faktorielle Intelligenzstrukturmodell die empirische Kovarianzstruktur der gesunden Kontrollprobanden am besten erklärt. Für mehrere Modelle werden unabhängige konfirmatorische Faktorenanalysen durchgeführt. Alle Modelle werden anhand der Kovarianzstruktur mittels Maximum-Likelihood-Schätzung durch das AMOS-Programm (Version 4.0 Arbuckle, 1997; Arbuckle & Wothke, 1999) untersucht. Die Operationalisierung der Modelle erfolgt anhand eines einheitlichen Satzes von 19 Testindikatoren und unter Zugrundelegung von Regeln der Ad-hoc-Spezifikation, die für alle Modelle gültig sind (siehe 7.3.1). Gegebenenfalls werden Post-hoc-Modifikationen zur Korrektur von Fehlspezifikationen im Rahmen der Ad-hoc-Modellaufstellungen vorgenommen. Auch diese entstammen einem vordefinierten und für alle Modelle gleichermaßen verbindlichen Satz möglicher Respezifikationen (siehe 7.3.2). Durch Vergleich der Anpassungsgüte wird schließlich eines der getesteten Modelle ausgewählt.

7.3.1 Ad-hoc-Modellspezifikation

Folgende Punkte skizzieren die allgemeinen Grundregeln für die Operationalisierungen der Modelle:

- Die Skalierung der Modelle wird über die Fixierung eines Pfades pro Faktor auf den Wert 1 vorgenommen. Hierzu wird der Pfad gewählt, dessen zugeordneter Untertest die höchste Reliabilität aufweist (siehe Tabelle 7-I).
- Jeder Faktor soll durch mindestens drei Indikatoren operationalisiert werden, da Faktoren mit nur zwei oder weniger Indikatorvariablen häufig Identifikations- oder Konvergenzprobleme verursachen.
- Indikatoren können auf mehreren Faktoren laden, wobei die empirisch günstigste Zuordnung von den Ergebnissen der Analysen abhängig gemacht wird.
- Korrelierte Messfehler werden ad hoc nicht operationalisiert.
- Es werden nur faktorielle Modelle erster Ordnung (ohne Generalfaktor) modelliert. Die Korrelationen zwischen den Faktoren werden nicht beschränkt.

7.3.1.1 Das Generalfaktormodell

Zur Operationalisierung des Generalfaktormodells (Zwei-Faktoren-Theorie, siehe 2.3.1) von Spearman (1927) wird die Varianz eines Tests in einen auf die allgemeine

Intelligenz zurückgehenden Anteil und einen testspezifischen Anteil (*uniqueness*) aufgespaltet. Abbildung 7-C gibt das einfache Modell wieder.

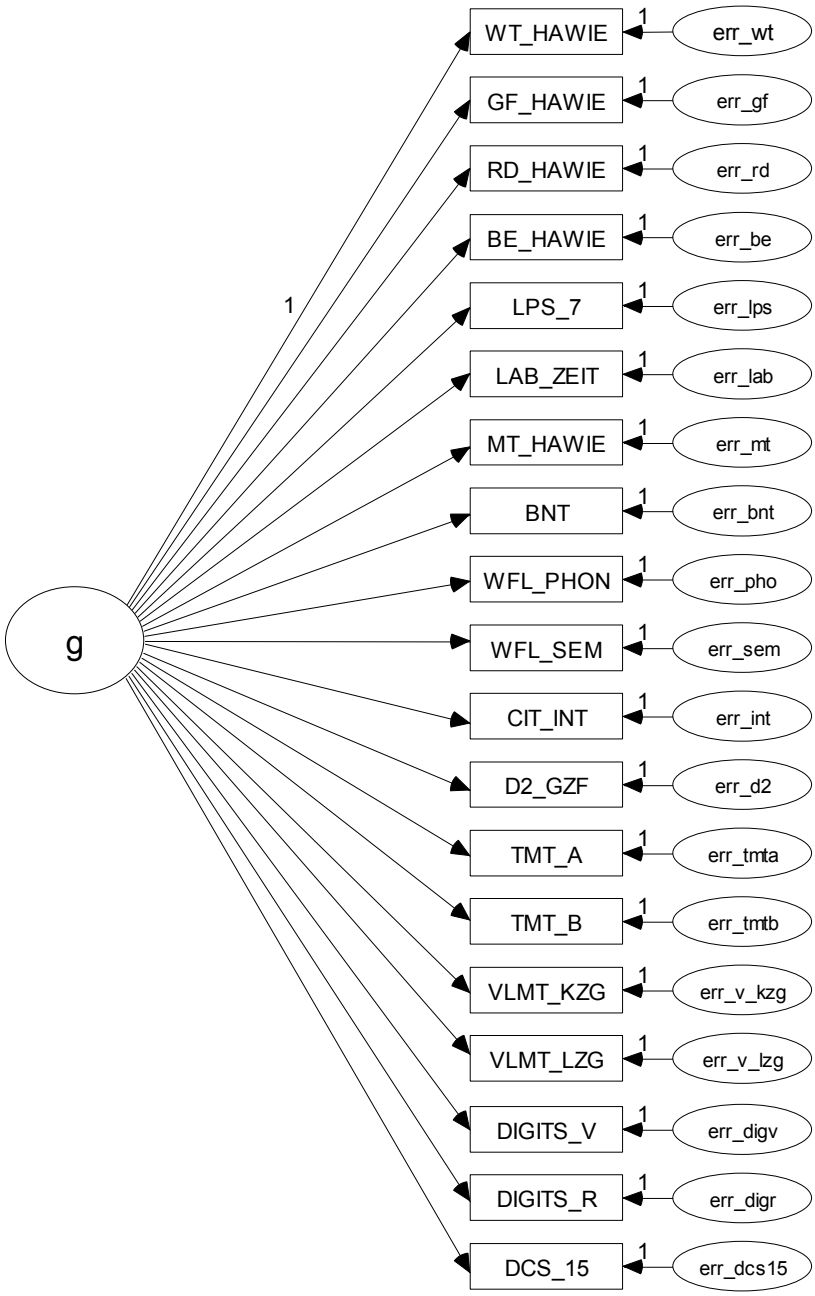


Abbildung 7-C: Operationalisierung des Generalfaktormodells nach Spearman.

7.3.1.2 Das hierarchische Intelligenzmodell von Vernon

Zentral im hierarchischen Intelligenzmodell nach Vernon (siehe 2.3.2) ist die Dichotomisierung der kognitiven Leistungen in einen verbal-edukationalen Faktor und einen kinesthetisch-mechanischen Faktor. Diesen Faktoren werden wiederum Faktoren

niedrigerer Generalität zugeordnet. Beim Versuch dieses Modell zu operationalisieren zeigt sich, dass zwar den beiden Hauptfaktoren jeweils mehrere Untertests zugeordnet werden können, aber nicht für jeden Unterfaktor Indikatoren verfügbar sind.

Für den verbal-edukationalen Faktor konnten die Konstrukte *kreative Fähigkeiten* (WFL_PHO, WLF_SEM, GF_HAWIE), *Wortflüssigkeit* (WFL_PHO, WFL_SEM), *sprachliche Fähigkeiten* (mit den Unterfaktoren *Lesefähigkeit, Sprachvermögen, literarische Fähigkeit, Schreibfähigkeit*: GF_HAWIE, WT_HAWIE, BNT) und *Fähigkeit zum Operieren mit Zahlen* (einschließlich *mathematischen Fähigkeiten*: DIGITS_V, DIGITS_R, RD_HAWIE) durch die genannten Indikatoren operationalisiert werden. Für den kinesthetisch-mechanischen Faktor (k:m) konnten den Unterfaktoren *Wahrnehmungsgeschwindigkeit* (D2_GZF, TMT_A), *räumliches Vorstellungsvermögen* (LPS_7, HAWIE_MT, LAB_Z) und *psychomotorische Fähigkeiten* (TMT_A) Indikatoren zugeordnet werden. Den Unterfaktoren *Flüssigkeit des Denkens* (v:ed) sowie *technisches Verständnis* (k:m) und *physikalische Kenntnisse* (k:m) konnten keine Tests zugeordnet werden.

Unklar bleiben die Zuordnungen folgender Untertests: Beim Bilderergänzen (BE_HAWIE) handelt es sich einerseits um einen Untertest, der eine deutliche visuo-perzeptive und psychomotorische Komponente aufweist, andererseits scheinen die Definitionen des kinesthetisch-mechanischen Faktors zu schmal. Daher werden zunächst Doppelladungen auf beiden Hauptfaktoren modelliert. Auch für eine klare Zuordnung des Interferenztests (CIT_INT) zum Faktor k:m (Unterfaktor *Wahrnehmungsgeschwindigkeit*) erscheint dessen Definition zu schmal, vor allem da die Fähigkeit zur Interferenzunterdrückung nicht berücksichtigt wird. Da dieser Test auch dem v:ed-Unterfaktor *Flüssigkeit des Denkens* zugeordnet werden könnte, sollen auch hier Doppelladungen auf beiden Hauptfaktoren modelliert werden. Die Konzeption nach Vernon berücksichtigt keinen Gedächtnisfaktor. Daher soll dieser zusätzlich eingeführt werden (mit VLMT_KZG, VLMT_LZG und DCS_15, DIGITS_V, DIGITS_R), wobei DCS_15 eingangs auch auf k:m (*räumliches Vorstellungsvermögen*) laden soll und die beiden Zahlennachsprechaufgaben sowohl auf dem verbalen (Fähigkeit zum Operieren mit Zahlen) als auch auf dem Gedächtnisfaktor laden können.

Abbildung 7-D gibt die genauen Spezifikationen der drei Faktoren (verbal-edukational, kinesthetisch-mechanisch, Gedächtnis) wieder. Die niedrig-generellen Faktoren können nicht operationalisiert werden, da ihnen zu wenige Indikatoren zugeordnet werden können. Doppelladungen werden für die Indikatoren BE_HAWIE, CIT_INT, DCS_15 und DIGITS_V und DIGITS_R operationalisiert. Die Ergebnisse der Analysen sollen Klarheit über die empirisch günstigste Zuordnung bringen.

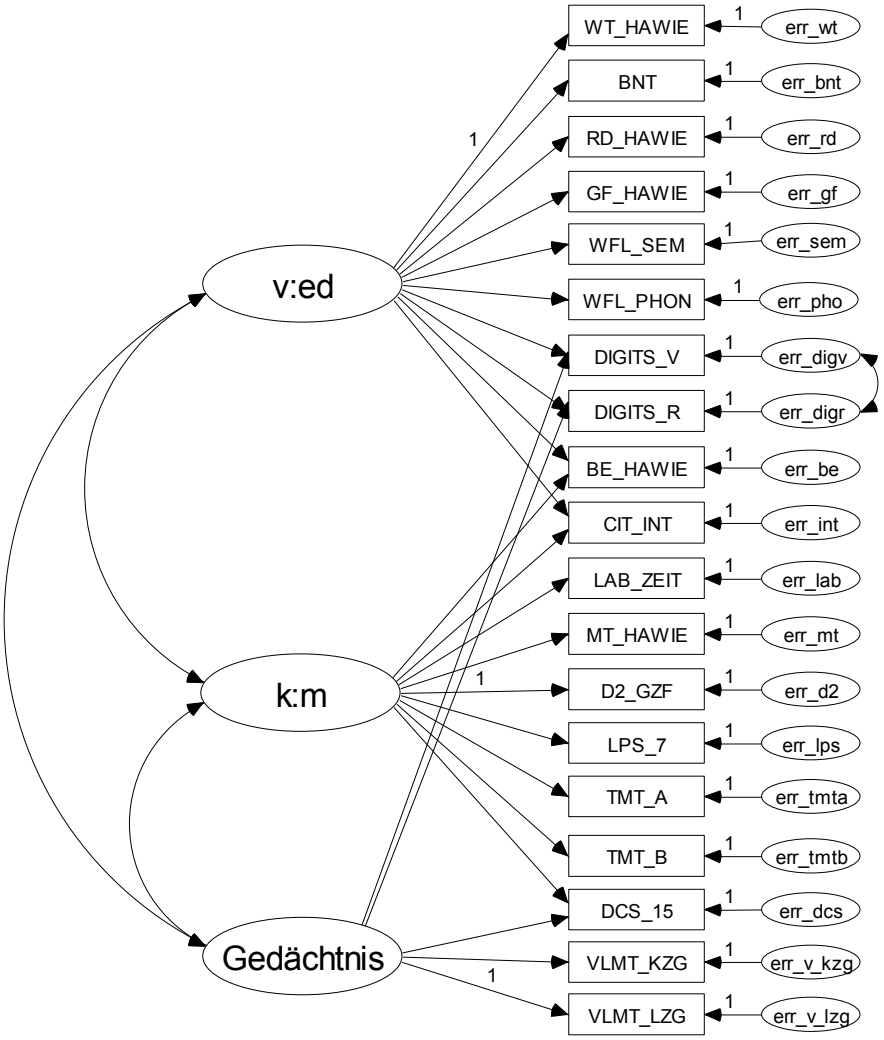


Abbildung 7-D: Operationalisierung des hierarchischen Intelligenzmodells von Vernon mit einem verbal-edukationalen Faktor (v:ed), einem kinesthetisch-mechanischen (k:m) Faktor und einem Gedächtnisfaktor.

7.3.1.3 Das Berliner Intelligenzstrukturmodell

Abbildung 7-E gibt die Operationalisierung des Berliner Intelligenzstrukturmodells nach Jäger wieder (Bucik & Neubauer, 1996; Jäger, 1982; Jäger, 1984; Schmidt, 1984). Folgende Charakteristika sind kennzeichnend:

- Entsprechend dem Originalmodell werden vier Operations- und drei Inhaltsfaktoren modelliert.
- Die Operations- und Inhaltsfaktoren sind jeweils interkorreliert.
- Jeder Test lädt auf einem Operations- und einem Inhaltsfaktor.

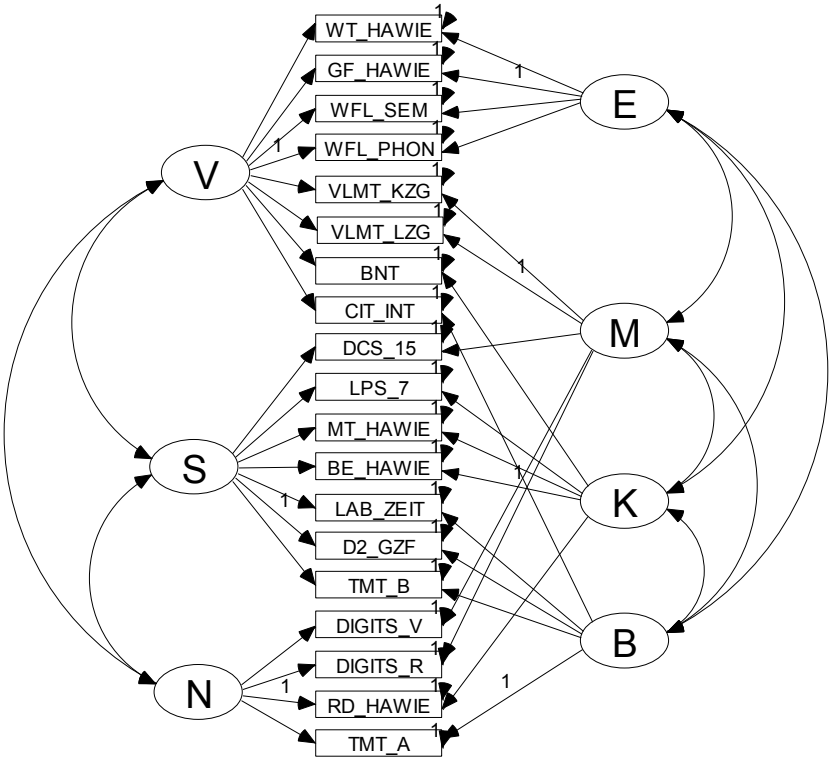


Abbildung 7-E: Operationalisierung des Berliner Intelligenzstrukturmodells. V: sprachgebundenes Denken, S: bildhaft-räumliches Denken, N: zahlengebundenes Denken; E: Einfallsreichtum, M: Merkfähigkeit, K: Verarbeitungskapazität. B: Bearbeitungsgeschwindigkeit.

Zur Operationalisierung ist anzumerken, dass die Zuordnung der Indikatoren zu den Inhaltsfaktoren (verbal, spatial, numerisch) weitgehend eindeutig ist. Bei der Zuordnung der Tests zu den Operationen zeigt sich, dass deren Definitionen relativ unscharf sind und tendenziell zur Beschreibung der vorliegenden Testbatterie ungeeignet erscheinen. Beispielsweise findet sich in der Testbatterie für den Faktor E (Einfallsreichtum) kaum ein Korrelat. Trotzdem konnte in größtmöglicher Anlehnung an die Faktorendefinitionen eine Zuordnung erreicht werden. Die Faktorenschreibungen für die operativen und inhaltsgebundenen Fähigkeiten lauten wie folgt (siehe z. B. Amthauer et al., 2001):

Operative Fähigkeiten:

- Verarbeitungskapazität (K): Verarbeitung komplexer Informationen bei Aufgaben, die nicht auf Anhieb zu lösen sind, sondern das Heranziehen und sachgerechtes Beurteilen von Informationen, vielfältiges Beziehungsstiften und formal-logisch exaktes Denken erfordern. K umfasst die Fähigkeiten zum induktiven und deduktiven Denken und wird in der englischsprachigen Literatur als "Reasoning" bezeichnet. Unterscheidung der K-Komponente in regelerkennende und regelanwendende Teilfähigkeiten.

- Einfallsreichtum (E): Flexible Ideenproduktion, die Verfügbarkeit vielfältiger Informationen, Reichtum an Vorstellungen und das Sehen vieler verschiedener Seiten, Varianten, Gründe und Möglichkeiten von Gegenständen und Problemen voraussetzt, wobei es um problemorientierte Lösungen geht, nicht um ein ungesteuertes Luxurieren der Fantasie. Flexible Ideenproduktion, die Verfügbarkeit vielfältiger Informationen und Perspektivenwechsel werden als zentrale Fähigkeit für erfolgreiche Problemlöseprozesse mit erfasst.
- Merkfähigkeit (M): Aktives Einprägen und kurzfristiges Wiedererkennen oder Reproduzieren von verschiedenartigem Material.
- Bearbeitungsgeschwindigkeit (B): Arbeitstempo, Auffassungsleichtigkeit und Konzentrationskraft beim Lösen einfach strukturierter Aufgaben von geringem Schwierigkeitsniveau.

Inhaltsgebundene Fähigkeiten:

- Sprachgebundenes Denken (V): Grad der Aneignung und der Verfügbarkeit des Beziehungssystems Sprache.
- Zahlengebundenes Denken (N): Grad der Aneignung und der Verfügbarkeit des Beziehungssystems Zahlen.
- Anschauungsgebundenes, figural-bildhaftes Denken (F): Einheitsstiftendes Merkmal scheint hier die Eigenart des Aufgabenmaterials zu sein, dessen Bearbeitung figural-bildhaftes und/oder räumliches Vorstellen erfordert.

Die Zuordnung der Indikatoren zu den Faktoren wurde aus diesen Faktorenbeschreibungen und anhand von Beispielaufgaben hergeleitet. Die Struktur des Modells ähnelt der im Rahmen einer konfirmatorische Faktorenanalyse von Bucik und Neubauer (1996) aufgestellten Struktur (siehe Abbildung 7-F):

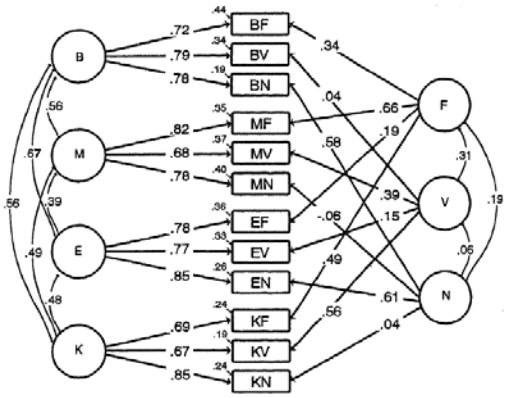


Abbildung 7-F: Abbildung 4 aus Bucik und Neubauer (, 1996).

7.3.1.4 Das Cattell-Horn-Carroll-Modell

Zur Operationalisierung des Cattell-Horn-Carroll-Modells wurden die 19 Indikatoren anhand der Fähigkeitsdefinitionen und empiriebasierter Empfehlungen (vgl. McGrew, 2003) den breiten Fähigkeitsfaktoren (*stratum II*) der Cattell-Horn-Carroll-Theorie (McGrew, 1997) zugeordnet. Hierzu waren insbesondere die Beschreibungen der schmaleren Fähigkeitsfaktoren des ersten Stratums hilfreich. Die vorgenommene Zuordnung gibt Tabelle 7-K (siehe nächste Seite) wieder.

Folgende weitere Charakteristika kennzeichnen die Ad-hoc-Modellspezifikation:

- Zur Zuordnung aller Indikatoren sind fünf Faktoren hinreichend (Gc, Gv, Gs, Gsm, Glr).
- Lediglich der Untertest „Rechnerisches Denken“ wäre laut Theorie einem weiteren Faktor zuzuordnen (Fluid reasoning und Quantitative Knowledge). Fähigkeitsfaktoren mit nur einem Indikator sind jedoch nicht operationalisierbar.
- Die Faktoren sind interkorreliert.
- Bei 14 der 19 Indikatoren konnte eine eindeutige Zuordnung in guter theoretischer Übereinstimmung mit den Faktorendefinitionen getroffen werden, so dass diese auf jeweils nur einem Faktor laden.
- Fünf Indikatoren konnten mit annähernd gleicher Plausibilität laut Theorie zwei Faktoren zugeordnet werden und wurden mit Doppelladungen auf beiden Faktoren modelliert.
- Der Untertest WFL_PHO wurde zunächst als auf dem Faktor Glr ladend operationalisiert, da auch WFL_SEM auf diesem Faktor lädt und eine einheitliche Faktorenzuordnung konsequent erscheint.

Tabelle 7-K: Zuordnung der Untertestindikatoren zu den breiten (*stratum II*) und engen (*stratum I*) Fähigkeitsdimensionen des Cattell-Horn-Carroll-Modells.

Indikator	**Zuordnung nach**[1]	**Stratum II**	**Stratum I (enges Fähigkeitskonstrukt)**
BNT	DEF	Gc oder Glr	VL (Lexical Knowledge) NA (Naming Facility)
WT_HAWIE	FA	Gc	LD (Language Development), VL (Lexical Knowledge)
GF_HAWIE	FA	Gc	LD (Language Development)
BE_HAWIE	FA	Gc oder Gv	KO (General Information) CF (Flexibility of Closure), VZ (Visualization)
LPS_7	DEF	Gv	VZ (Visualization), SR (Spatial Relations)
LAB_ZEIT	FA	Gv	SS (Spatial Scanning)
MT_HAWIE	FA	Gv	VZ (Visualization), SR (Spatial Relations)
DCS_15	DEF	Gv oder Glr	VM (Visual Memory) M6 (Free Recall Memory)
VLMT_LZG	DEF	Glr	M6 (Free Recall Memory)
WFL_SEM	DEF	Glr	FW (Word Fluency)
WFL_PHO	DEF	Glr oder Gs	FW (Word Fluency) WS (Writing speed)
D2_GZF	DEF	Gs	P (Perceptual Speed, Unterfunktion Ps (Scanning))
CIT_INT	DEF	Gs	P (Perceptual Speed, Unterfunktion: Pm (Memory))
TMT_A	DEF	Gs	P (Perceptual Speed, Unterfunktion Ps (Scanning))
TMT_B	keine passende DEF, daher gleiche Zuordnung wie TMT A	Gs	P (Perceptual Speed, Unterfunktion Ps (Scanning))
RD_HAWIE	FA, DEF	Gsm oder Gq oder Gf oder Gs	MW (Working Memory) KM (Mathematical Knowledge) RQ (Quantitative Reasoning) N (Number Facility)
DIGITS_V	FA	Gsm	MS (Memory Span)
DIGITS_R	DEF	Gsm	MW (Working Memory)
VLMT_KZG	DEF	Gsm	MS (Memory Span)

Anmerkungen: [1]FA: Zuordnung wurde explizit von McGrew (1997) aufgrund von Faktorenanalysen vorgeschlagen; DEF: Deduktive Zuordnung basierend auf Faktorenbeschreibungen (McGrew, 2003). Gc Crystallized Intelligence/Knowledge; Gs: Perceptual Speed; Gsm: Short-Term Memory; Gq: Quantitative Reasoning; Glr: Long-Term Retention; Gv: Visuo-Spatial Ability.

Abbildung 7-G gibt die Ad-hoc-Modellspezifikation des neuropsychologischen Faktorenmodells als Pfaddiagramm wieder. Die fünf Faktorendefinitionen werden im englischen Originalwortlaut wiedergegeben (McGrew, 2003), um Bedeutungsverschiebungen durch die Übersetzung ins Deutsche zu vermeiden.

- *Crystallized Intelligence/Knowledge* (Gc): “Can be thought of as the intelligence of the culture that is incorporated by individuals through a process of acculturation” (Horn, 1994, p.443). Gc is typically described as a person’s wealth (breadth and depth) of acquired knowledge of the language, information and concepts of specific a culture, and/or the application of this knowledge. Gc is primarily a store of verbal or language-based declarative (knowing “what”) and procedural (knowing “how”) knowledge acquired through the “investment” of other abilities during formal and informal educational and general life experiences.
- *Visual-Spatial Abilities* (Gv): “The ability to generate, retain, retrieve, and transform well-structured visual images” (Lohman, 1994, p.1000). The Gv domain represents a collection of different abilities each that emphasize a different process involved in the generation, storage, retrieval and transformation (e.g., mentally reverse or rotate shapes in space) of visual images. Gv abilities are measured by tasks (figural or geometric stimuli) that require the perception and transformation of visual shapes, forms, or images and/or tasks that require maintaining spatial orientation with regard to objects that may change or move through space.
- *Short-term Memory* (Gsm): The ability to apprehend and maintain awareness of elements of information in the immediate situation (events that occurred in the last minute or so). A limited-capacity system that loses information quickly through the decay of memory traces, unless an individual activates other cognitive resources to maintain the information in immediate awareness.
- *Cognitive processing Speed* (Gs): The ability to automatically and fluently perform relatively easy or over-learned cognitive tasks, especially when high mental efficiency (i.e., attention and focused concentration) is required. The speed of executing relatively over-learned or automatized elementary cognitive processes.
- *Long-term Storage and Retrieval* (Glr): The ability to store and consolidate new information in long-term memory and later fluently retrieve the stored information (e.g., concepts, ideas, items, names) through association. Memory consolidation and retrieval can be measured in terms of information stored for minutes, hours,

weeks, or longer. Horn (Horn & Masunaga, 2000) differentiates two major types of Glr--fluency of retrieval of information over minutes or a few hours (intermediate memory) and fluency of association in retrieval from storage over days, months or years. Ekstrom et al. (1979) distinguished two additional characteristic processes of Glr: “(1) reproductive processes, which are concerned with retrieving stored facts, and (2) reconstructive processes, which involve the generation of material based on stored rules” (p. 24). Glr abilities have been prominent in creativity research where they have been referred to as idea production, ideational fluency, or associative fluency.

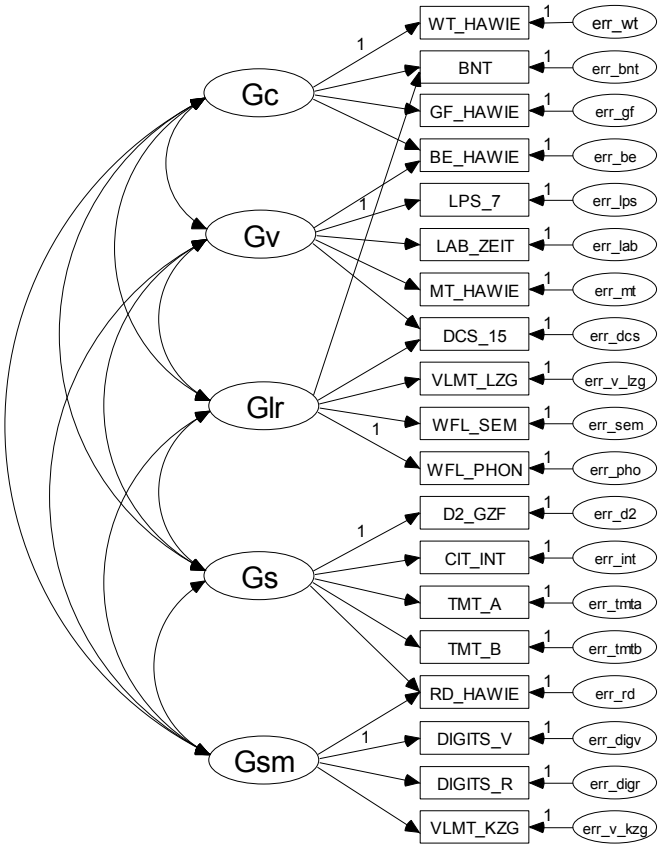


Abbildung 7-G: Operationalisierung des Cattell-Horn-Carroll-Modells. Gc: Crystallized Intelligence/Knowledge; Gv: Visuo-Spatial Ability; Glr: Long-Term Retention; Gs: Perceptual Speed; Gsm: Short-Term Memory.

7.3.1.5 Das neuropsychologische Modell

Das neuropsychologische Modell beruht auf einer Operationalisierung von typischerweise in der klinischen Neuropsychologie herangezogenen Funktionsbereichen. Wichtige Funktionsbereiche sind dabei (vgl. S. 25 in Hartje & Poeck, 2002): basale und höhere Wahrnehmungsleistungen, Aufmerksamkeitsleistungen, Gedächtnisfunktionen, intellektuelles Niveau, räumlich-perzeptive, räumlich-kognitive und räumlich-konstruktive Leistungen, Sprache,

Planungs- und Kontrollfunktionen (exekutive Funktionen), sensomotorische Leistungen und motorische Planung, Zahlenverarbeitung und Rechenleistung. Folgende Charakteristika kennzeichnen die Ad-hoc-Modellspezifikation:

- Die 19 Indikatoren werden fünf Faktoren zugeordnet. Im Speziellen sind dies *Gedächtnis, Aufmerksamkeit, visuell-räumliche Fähigkeiten, verbaler IQ* und *Sprache*.
- Der verbale Intelligenzfaktor und der Sprachfaktor unterscheiden sich darin, dass im Sprachfaktor grundlegende Sprachleistungen (*Benennen* und *Wortflüssigkeit*) erfasst werden, im verbalen Intelligenzfaktor dagegen höhere kortikale Funktionen subsumiert werden – im Speziellen die sprachlichen Intelligenzuntertests *Gemeinsamkeitenfinden* und *Wortschatz*. Der verbale Intelligenzfaktor stellt eine Operationalisierung des sprachlichen Teilaspekts im oben genannten Funktionsbereich des intellektuellen Niveaus dar.
- Der Indikator *Rechnerisches Denken* soll entsprechend der klassischen Konzeption nach Wechsler (Wechsler, 1981) dem verbalen Intelligenzfaktor zugerechnet werden.
- Die Faktoren sind interkorreliert.
- Jeder Indikator lädt auf nur einem Faktor, Doppelladungen werden nicht modelliert.

Abbildung 7-H gibt die Ad-hoc-Modellspezifikation des einfachen neuropsychologischen Faktorenmodells wieder. Zu den nicht operationalisierten Funktionsbereichen ist anzumerken:

- Basale und höhere Wahrnehmungsleistungen werden im visuell-räumlichen Faktor integriert.
- Das oben beschriebene *Impurity-Problem* (siehe Kapitel 2.4.2) der exekutiven Funktionen erfordert die Modellierung hoch komplexer Modelle mit modalitätsspezifischen und operationsspezifischen Faktoren. Hierfür sind jedoch weder Anzahl der Indikatoren noch Stichprobengröße ausreichend.
- Ein spezieller motorischer Faktor kann nach Ausschluss der motorischen Untertests der Testbatterie aufgrund deren schlechter psychometrischer Eignung (siehe oben) nicht operationalisiert werden. Verschiedene attentionale Untertests weisen aber eine hohe psychomotorische Komponente auf.

- Die Operationalisierung des Funktionsbereichs Zahlenverarbeitung und Rechenleistung soll aufgrund der Probleme bei Operationalisierungen von Faktoren mit nur einem Indikator vermieden werden.

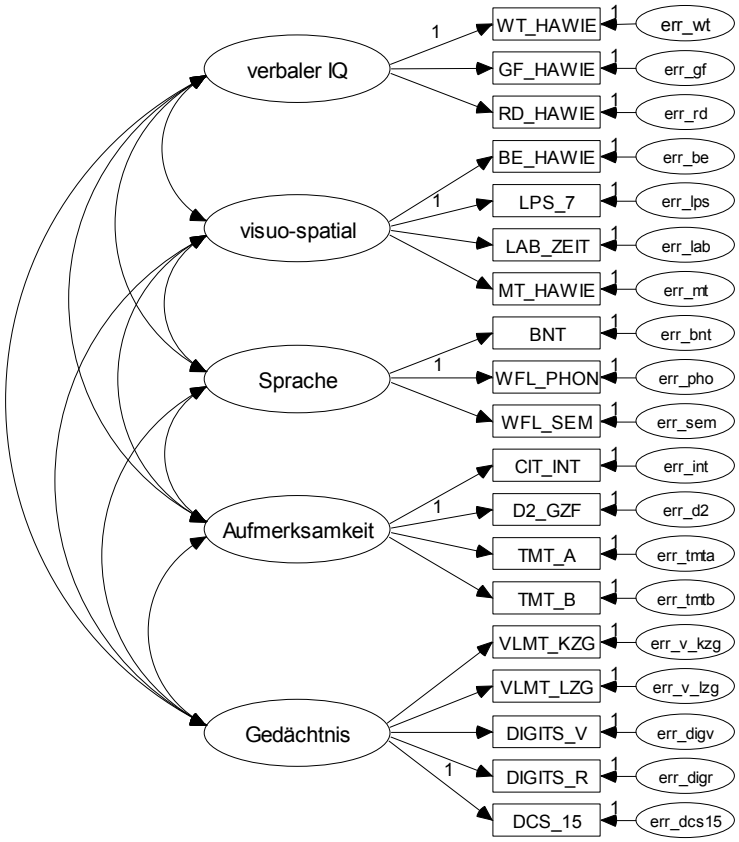


Abbildung 7-H: Operationalisierung des neuropsychologischen Intelligenzmodells.

7.3.2 Post-hoc-Modifikationen

Psychologische Theorien sind selten soweit elaboriert, dass sie eine eindeutige Modellspezifikation erzwingen. Insbesondere gilt dies für die vorliegende Fragestellung, in der eine mit einer komplett anderen Zielsetzung zusammengestellte Testbatterie an Theorien zur Intelligenzstruktur nachträglich angepasst wird. Anders wäre es, wenn ausgehend von einem theoretischen Ausgangspunkt eine Testbatterie extra zur Überprüfung der Theorien zusammengestellt würde. In dieser Situation wären Post-hoc-Modifikationen schlecht zu rechtfertigen.

Um durch inflationäre Aufnahme nachträglicher Modellmodifikationen eine Überanpassung der Modelle an die Daten zu vermeiden, soll ein enger Rahmen möglicher Post-hoc-Modifikationen vorgegeben werden:

- Nicht-signifikante Pfade (Ladungen) können zur Erhöhung der Parsimonität eliminiert werden.

- Hinweise auf fehlende Pfade ergeben sich aus den Modifikationsindizes der Ladungen. Eine Einführung zusätzlicher Pfade erfolgt, wenn die Modifikationsindizes größer als 6.6 und somit signifikant sind (χ^2 mit einem Freiheitsgrad, p <.01) und wenn die Einführung eines zusätzlichen Pfades theoretisch plausibel ist.
- Fehlerkovarianzen können nachträglich nur dann modelliert werden, wenn relevante Modifikationsindizes der Fehlerkovarianzen vorliegen und wenn deren Aufnahme auch theoretisch aufgrund gemeinsamer Methodenvarianz begründbar ist. Letzteres erscheint für die Messfehlerpaare folgender Tests gegeben zu sein: Trail Making A und B, VLMT Langzeit- und Kurzzeitgedächtnis, semantische und phonematische Wortflüssigkeit, Zahlennachsprechen vorwärts und rückwärts.

7.4 Überprüfung der Invarianz im Zwei-Gruppen-Fall

Für die Überprüfung der Fragestellungen B, C und D ist der Vergleich zweier Stichproben (Zwei-Gruppen-Fall) mithilfe unabhängiger konfirmatorischer Faktorenanalysen notwendig. Alle Modelle werden mittels Maximum-Likelihood-Schätzung mit dem AMOS-Programm (Version 4.0 Arbuckle, 1997; Arbuckle & Wothke, 1999) untersucht. Während die Fragestellungen nach konfiguraler und schwacher metrischer Invarianz auf Basis der Kovarianzstruktur der Stichproben beantwortet werden können, erfordern die Fragestellungen nach starker und strenger Invarianz als Datenbasis zusätzlich die Mittelwerte der Indikatoren (Byrne, 2001).

Vandenberg und Lance (2000) stellen in ihrer Übersichtsarbeit dar, dass es keine vereinheitlichte Methodik der Testung auf Invarianz mittels konfirmatorischer Faktorenanalysen gibt. Die Autoren empfehlen, die genaue Spezifizierung der Methode von der Fragestellung, d.h. vom Einsatzziel der Testbatterie und von den zugrunde liegenden Vergleichen, abhängig zu machen. Die vorliegende Arbeit betrachtet die Normierung einer Testbatterie für Patienten mit Epilepsie an einer nicht-klinischen Kontrollgruppe. Übergeordnetes Ziel für die klinische Anwendung ist die Untersuchung der Konstruktvaliditäten. Für Forschungsfragen steht zudem die Vergleichbarkeit von Mittelwerten im Vordergrund. Hierzu ist eine Vergleichbarkeit der beobachteten und der latenten Mittelwerte notwendig, so dass hohe Kriterien an die Invarianz zu stellen sind.

Im Zwei-Gruppen-Fall der konfirmatorischen Faktorenanalyse wird simultan in zwei Stichproben untersucht, wie sich die Anpassungsgüte durch die Gleichheitsbeschränkungen der Strukturgleichungen verändert. Untersuchungen auf

Invarianz implizieren jeweils einen Vergleich zwischen zwei Modellen, einem stärker und einem weniger stark restringierten Modell. Ist die Güte der Anpassung des stärker beschränkten Modells signifikant schlechter als die des freieren Modells, ist die eingeführte Parameterbeschränkung unzulässig, da das Modell mit frei schätzbaren, also gruppenunabhängigen Parametern die Daten signifikant besser beschreibt. Die vom strengeren Modell implizierte Invarianzhypothese ist also abzulehnen. Getestet wird die Nullhypothese, dass das stärker beschränkte Modell richtig ist, vorausgesetzt das weniger beschränkte Modell ist richtig. Das Modell mit weniger Beschränkungen hat weniger Freiheitsgrade, da es mehr zu schätzende Parameter hat. Modelle können immer dann paarweise miteinander verglichenen werden, wenn sie in einer genesteten Beziehung zueinander stehen: Modelle sind genestet, wenn sich eines durch Parameterbeschränkungen aus dem anderen erzeugen lässt.

Die Untersuchung der zunehmend stringenten Hypothesen unterliegt folgendem Schema (in Anlehnung an Arbuckle, 2003)[9]:

- Im ersten Schritt wird die Nullhypothese der Gleichheit der Form betrachtet. Die Parameterbeschränkungen beziehen sich lediglich auf das Muster aus Ladungen und Null-Ladungen, weitere Beschränkungen der Parameter des Mess- oder des Strukturmodells werden nicht vorgenommen. Dieses Modell bildet die Grundlage für die folgenden Vergleiche (*baseline model, unconstrained model*).
- Der zweite Schritt überprüft die Nullhypothese, dass alle Ladungen zwischen den Gruppen gleich sind [H_0: Λ(g) gleich]. Als Parameterbeschränkungen werden gleiche Ladungen in beiden Gruppen gefordert.
- Der dritte Schritt dient der Überprüfung der Gleichheit der Höhenlagen [H_0: τ(g) gleich].
- Der vierte Schritt überprüft die Haltbarkeit der zusätzlichen Annahme gleicher struktureller Mittelwerte [H_0: κ(g) gleich].
- Der fünfte Schritt überprüft die Haltbarkeit der Annahme gleicher struktureller Varianzen und Kovarianzen [H_0: Φ(g) gleich].
- Der letzte Schritt überprüft die Haltbarkeit der Annahme gleicher Varianz-Kovarianzmatrizen der Messfehler [H_0: Θ(g) gleich].

Siehe hierzu Tabelle 7-L und Abbildung 7-I.

[9] Zur Vermeidung von Identifizierbarkeitsproblemen sind spezifische Modellannahmen zwingend. Vor allem benötigen Schätzung auf Basis der Mittelwertstruktur die Definition von gleichen Ladungen und Höhenlagen in den Gruppen und die Festlegung der Faktorenmittelwerte einer Gruppe (möglichst der Normgruppe) auf Null. Letzteres bedeutet, dass lediglich die relative Unterschiedlichkeit zwischen den Gruppen in den Faktorenmittelwerten erfasst werden kann, aber nicht simultan für beide Gruppen deren tatsächlicher Wert. Des Weiteren sind die Fehlermittelwerte auf Null zu setzen, die Fehlervarianzen sind frei schätzbar.

Tabelle 7-L: Abfolge der Invarianzhypothesen.

Parameter			**Modell**				
	baseline[a]	**1**	**2**	**3**	**4**	**5**	**6**
Ladungen		x	x	x	x	x	x
Höhenlagen			x	x	x	x	x
Latente Mittelwerte					x	x	x
Latente (Ko-)Varianzen						x	x
Residuen				x			x
	konfigurale Invarianz	schwache Mess-invarianz	starke Mess-invarianz	strenge Mess-invarianz	strukturelle Invarianz		

Anmerkungen: Die Freiheitsgrade für die Modelle steigen mit zunehmender Modellnummer. [a] In den statistischen Tests mittels konfirmatorischer Faktorenanalysen fungiert das konfigural-invariante Modell als Vergleichsgrundlage (*baseline-model*).

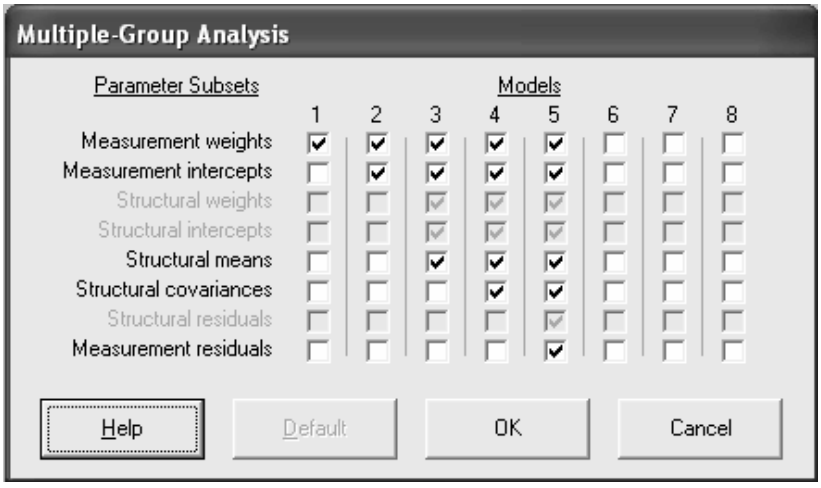


Abbildung 7-I: Screenshot der von Amos 5.0 vorgeschlagenen Abfolge zur Überprüfung der Invarianzhypothesen. Da sich die Mehr-Gruppen-Analyse auf konfirmatorische Faktorenanalysen bezieht, sind einige Strukturkomponenten ausgeblendet; sie sind nur für den Vergleich von pfadanalytischen Modellen relevant.

8 Ergebnisse

Der Ergebnisteil dieser Arbeit beschäftigt sich zunächst mit der deskriptiven Statistik auf Einzeltestebene und beantwortet anschließend die in Kapitel 5 umrissenen Fragestellungen. Fragestellung A überprüft, welches Intelligenzstrukturmodell sich am besten zur Beschreibung der Stichprobendaten eignet. Die Fragestellungen B, C und D wenden sich den verschiedenen Aspekten der Invarianz zu und Fragestellung E schließlich befasst sich mit der Konstruktvalidität des Faktorenmodells.

8.1 Deskriptive Statistik

Grundsätzlich basieren konfirmatorische Faktorenanalysen auf der Varianz-Kovarianzmatrix und auf den Mittelwerten der Tests. Da aber Korrelationen durch Standardisierungen der Kovarianzen ermittelt werden können ($r = \frac{\mathrm{cov}(x, y)}{s_x \cdot s_y}$), eignen sich als Rohdaten ebenso die Korrelationen, Mittelwerte und Standardabeichungen der Tests. Tabelle 8-A gibt die Mittelwerte und Standardabweichungen sowie die Ergebnisse eines Mittelwertvergleichs wieder. Die Testleistungen der Patientenstichprobe lagen zwischen 0.34 und 1.45 Standardabweichungen unter den Leistungen der Normierungsstichprobe. Die Korrelationen für beide Stichproben können Tabelle 8-B entnommen werden.

Tabelle 8-A: Deskriptive Statistik der Einzeltestmittelwerte und Standardabweichungen der Normstichprobe und der Patientenstichprobe; n = jeweils 190.

	Normstichprobe		**Patienten**			
	MW	***SD***	***MW***	***SD***	***F*** [a]	***Diff*** [b]
BE_HAWIE	14.63	1.85	13.36	2.50	31.673	-0.69
WT_HAWIE	23.44	4.13	18.85	4.80	99.392	-1.11
MT_HAWIE	33.67	8.43	29.46	8.15	24.427	-0.50
GF_HAWIE	27.43	3.26	23.35	4.74	95.442	-1.25
RD_HAWIE	15.19	2.49	11.88	3.59	109.077	-1.33
DIGITS_V	6.44	1.21	5.94	1.06	18.397	-0.41
DIGITS_R	5.19	1.25	4.57	1.16	24.864	-0.49
VLMT_KZG	13.93	3.46	12.18	3.37	24.836	-0.50
VLMT_LZG	11.11	4.24	7.81	5.23	45.817	-0.78
DCS_15	28.18	9.00	21.21	9.66	53.058	-0.78
TMT_A	29.85	10.58	33.42	13.46	8.231	-0.34
TMT_B	63.43	25.35	87.81	41.98	46.957	-0.96
D2_GZF	439.89	83.59	372.02	79.42	65.834	-0.81
CIT_INT	18.53	4.09	22.46	6.04	55.217	-0.96
LAB_ZEIT	275.22	114.78	346.17	192.48	19.045	-0.62
LPS_7	20.88	6.81	16.83	6.05	37.523	-0.59
WFL_PHON	41.04	9.15	27.75	8.30	219.961	-1.45
WFL_SEM	25.91	5.88	19.26	6.04	118.428	-1.13
BNT	54.44	3.44	51.54	5.86	34.546	-0.84

Anmerkungen: F-Werte einer multivariaten Varianzanalyse, alle *p*'s < .001, außer TMT_A: $p < .01$; [b] *Diff* drückt aus, wie viele Standardabweichungen die Leistungen der Patienten unter den Leistungen der Normprobanden liegen.

Tabelle 8-B: Korrelationen zwischen den 19 Subtests in der Normierungsstichprobe (grau) und in der klinischen Stichprobe (weiß); n = jeweils 190.

	BE_HAWIE	WT_HAWIE	MT_HAWIE	GF_HAWIE	RD_HAWIE	DIGITS_V	DIGITS_R	VLMT_KZG	VLMT_LZG	DCS_15	TMT_A	TMT_B	D2_GZF	CIT_INT	LAB_ZEIT	LPS_7	WFL_PHON	WFL_SEM	BNT
BNT	.218	.361	.132	.313	.229	.121	.081	.166	.142	.256	.110	.224	.150	.120	.092	.208	.235	.223	1
WFL_SEM	.202	.223	.217	.265	.073	.147	.150	.303	.274	.244	.267	.300	.324	.233	.110	.269	.453	1	.282
WFL_PHON	.294	.566	.202	.546	.124	.277	.318	.296	.326	.288	.236	.343	.399	.385	.077	.322	1	.543	.227
LPS_/	.286	.202	.456	.221	.180	.127	.174	.288	.134	.408	.330	.422	.416	.350	.357	1	.315	.345	.176
LAB_ZEIT	.208	-.001	.441	.095	-.079	.174	.023	.295	.208	.446	.344	.476	.364	.222	1	.440	.324	.320	.101
CIT_INT	.283	.244	.361	.315	.114	.242	.289	.386	.322	.419	.333	.448	.449	1	.446	.311	.453	.371	.170
D2_GZF	.303	.353	.421	.316	.164	.136	.170	.417	.345	.428	.385	.513	1	.484	.406	.324	.408	.328	.026
TMT_B	.348	.184	.468	.188	.112	.225	.244	.445	.337	.541	.620	1	.497	.514	.491	.371	.356	.373	.170
TMT_A	.293	.078	.399	.122	.035	.051	.074	.325	.308	.374	1	.521	.459	.451	.438	.213	.321	.235	.126
DCS_15	.393	.227	.580	.224	.147	.187	.260	.535	.440	1	.330	.373	.374	.289	.380	.345	.196	.162	.158
VLMT_LZG	.316	.218	.334	.229	-.029	.137	.244	.425	1	.329	.164	.230	.350	.239	.146	.121	.236	.221	-.034
VLMT_KZG	.294	.227	.319	.292	.021	.176	.227	1	.465	.237	.207	.346	.287	.353	.153	.106	.342	.277	.059
DIGITS_R	.143	.206	.219	.259	.242	.542	1	.375	.206	.182	.211	.361	.266	.411	.200	.128	.410	.276	.098
DIGITS_V	.088	.266	.181	.245	.186	1	.491	.325	.138	.128	.132	.279	.256	.289	.162	.140	.346	.214	.168
RD_HAWIE	.018	.235	.185	.175	1	.391	.347	.204	.102	.178	.251	.353	.223	.266	.182	.198	.452	.196	.204
GF_HAWIE	.152	.607	.275	1	.305	.171	.299	.252	.100	.220	.265	.224	.273	.306	.187	.255	.380	.339	.434
MT_HAWIE	.326	.108	1	.282	.259	.238	.255	.277	.263	.504	.334	.440	.500	.419	.450	.414	.270	.223	.120
WT_HAWIE	.189	1	.221	.608	.403	.226	.228	.162	.065	.187	.194	.107	.166	.164	.022	.163	.302	.235	.475
BE_HAWIE	1	.205	.350	.257	.130	.072	.103	.110	-.007	.246	.151	.255	.234	.157	.264	.331	.065	.153	.213

Wie in Kapitel 7.1.2 beschrieben, unterscheiden sich beide Stichproben in der Geschlechts- und Altersstruktur. Eine Kontrolle der Unterschiedlichkeiten anhand der jeweiligen Testnormen ist nicht zulässig, da die Normierungen der Tests auf unterschiedlichen Normierungsstichproben beruhen und die Standardwerte nur begrenzt vergleichbar sind. Die Alters- und Geschlechtseffekte werden anhand der Rohwerte der gesamten Normierungsstichprobe (n = 245) korrigiert. Tabelle 8-C gibt die mittels multifaktorieller Varianzanalyse ermittelten Abhängigkeiten der Testvariablen von Alter und Geschlecht wieder.

Tabelle 8-C: Geschlechts- und Altersunterschiede der Rohdaten.

	Geschlecht (Faktor)		**Alter** (Kovariate)	
	p	**F[1]**	***p***	**F[1]**
BE_HAWIE	n.s.		$p < .001$	18.857
WT_HAWIE	n.s.		n.s.	
MT_HAWIE	$p < .01$; m > w	7.933	$p < .001$	78.224
GF_HAWIE	n.s.		n.s.	
RD_HAWIE	$p <. 001$; m > w	19.701	$p < .01$	10.737
DIGITS_V	$p <. 01$; m > w	10.045	n.s.	
DIGITS_R	n.s.		n.s.	
VLMT_KZG	$p < .05$; w > m	5.464	$p < .001$	66.827
VLMT_LZG	$p < .001$; w > m	21.465	$p < .001$	30.489
DCS_15	n.s.		$p < .001$	66.594
TMT_A	n.s.		$p < .001$	48.163
TMT_B	n.s.		$p < .001$	88.218
D2_GZF	n.s.		$p < .001$	50.589
CIT_INT	n.s.		$p < .001$	41.234
LAB_ZEIT	n.s.		$p < .001$	93.968
LPS_7	$p < .01$; m > w	10.370	$p < .001$	31.727
WFL_PHON	n.s.		n.s.	
WFL_SEM	n.s.		$p < .05$	8.553
BNT	$p < .05$; m > w	9.526	n.s.	

Zur Ermittlung alters- und geschlechtskorrigierter Werte wurden anhand der Daten der Normierungsstichprobe (Ausgangsstichprobe; n = 245) mit den Prädiktoren Alter und Geschlecht für jede Testvariable schrittweise lineare Regressionen berechnet. Die erhaltenen Regressionsgleichungen wurden für beide Gruppen zur Vorhersage unstandardisierter Residuen genutzt, die schließlich in die Analysen eingingen. Tabelle 8-D gibt die standardisierten Regressionsgewichte wieder, die in der

Normierungsstichprobe geschätzt wurden. Eine Alters- und Geschlechtskorrektur war nicht für alle Untertests notwendig.

Tabelle 8-D: Standardisierte Regressionsgewichte für die Variablen Geschlecht und Alter.

n = 245	Standardisierte Regressions-gewichte	
	β [Geschlecht]	**β [Alter]**
BE_HAWIE		-.272
WT_HAWIE		
MT_HAWIE	-.155	-.487
GF_HAWIE		
RD_HAWIE	-.270	.199
DIGITS_V	-.203	
DIGITS_R		
VLMT_KZG	.132	-.463
VLMT_LZG	.271	-.324
DCS_15		-.465
TMT_A		-.409
TMT_B		-.518
D2_GZF		-.417
CIT_INT		-.383
LAB_ZEIT		-.530
LPS_7	-.191	-.334
WFL_PHON		
WFL_SEM		-.183
BNT	-.190	

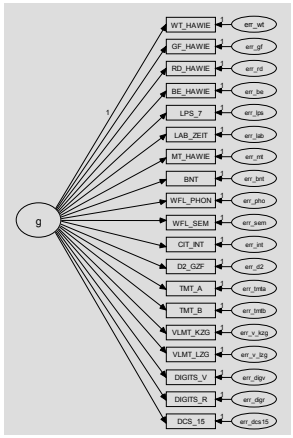
Anmerkungen: Ein positives Gewicht des Prädiktors Geschlecht verweist auf bessere Leistungen der Frauen, positives Gewicht des Prädiktors Alter verweist auf bessere Leistungen der älteren Probanden. Kriterium für Aufnahme eines Prädiktors (schrittweise): F ≤ .05.

8.2 Fragestellung A: Konfirmatorische Faktorenanalyse (Ein-Gruppen-Fall)

Im Folgenden wird die Frage beantwortet, welches Intelligenzstrukturmodell am besten mit den Daten der Normstichprobe vereinbar ist. Dazu sollen die Anpassungsgüten der oben beschriebenen fünf Modelle mittels konfirmatorischer Faktorenanalysen ermittelt und verglichen werden. Zur Erhöhung der Lesbarkeit wird in der folgenden Ergebnisdarstellung jedes Modell durch eine stark verkleinerte Abbildung in Erinnerung gerufen. Für die theoretische Darstellung der Modelle wird auf Kapitel 2.3 verwiesen, die konkrete Operationalisierung wurde in Kapitel 7.3.1 beschrieben. Die gewählte Datenbasis zur Überprüfung der Modelle ist die Normstichprobe mit n = 190 Probanden. Zunächst wird die Güte der Anpassung der Ad-hoc-Modelle ermittelt. Sollte sie nicht

hinreichend sein, werden Modellspezifikationen anhand der von AMOS 4.0 ausgegeben Modifikationsindizes durchgeführt (vergleiche Kapitel 7.3.2).

8.2.1 Das Generalfaktormodell



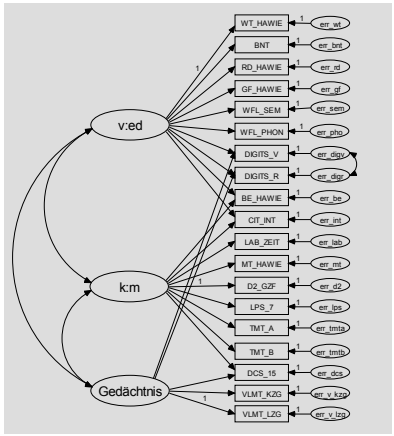
Das Generalfaktormodell impliziert, dass jeder Test signifikant auf dem Generalfaktor lädt. In der Ad-hoc-Version weist dieses Modell keine hinreichende Güte der Anpassung auf (siehe Tabelle 8-E). Die Einführung zusätzlicher Ladungen ist nicht möglich, auch die Elimination bestehender Ladungen würde dem Modell grundsätzlich widersprechen. Aufgrund der Modifikationsindizes der Fehlerkovarianzen werden korrelierte Messfehler zwischen den Indikatoren *Zahlennachsprechen vorwärts* und *rückwärts*, *Trail Making Test A* und *B* sowie *Wortflüssigkeit semantisch* und *phonematisch* eingeführt. Die Anpassungsgüte des Modells bleibt auch nach diesen Modifikationen unbefriedigend, da keiner der zentralen Indizes der Anpassungsgüte (GFI, NNFI, CFI) den Grenzwert von 0.9 überschreitet (Tabelle 8-E; Abkürzungen gültig für alle weiteren Tabellen).

Tabelle 8-E: Güte der Anpassung des Generalfaktormodells.

Modell	Status	χ^2 [FG]	GFI	NNFI	CFI	RMSEA [90%-KI]	ECVI [gesättigt]
g-Faktor	ad hoc	339.99 [152]	0.838	0.754	0.781	0.081 [0.069-0.092]	2.201 [2.011]
	post hoc	246.14 [149]	0.876	0.870	0.887	0.059 [0.045-0.072]	1.736 [2.011]

Abkürzungen: FG: Freiheitsgrade; GFI: goodness of fit index; NNFI: non-normed fit index; CFI: comparative fit index; RMSEA: root mean squared error of approximation; KI: Konfidenzintervall; ECVI: expected cross-validation index

8.2.2 Das hierarchische Intelligenzmodell von Vernon



Das hierarchische Intelligenzmodell von Vernon wurde als dreifaktorielles Modell mit den Faktoren *v:ed (verbal-edukational)*, *k:m (*kinesthetisch-mechanisch) und *Gedächtnis* operationalisiert. Die konfirmatorische Faktorenanalyse erbrachte keine eindeutige Lösung, da das Ad-hoc-Modell offenbar deutliche Fehlspezifikationen enthält. Zum Erreichen einer identifizierbaren Lösung wurden Fehlerkovarianzen bezüglich der beiden

Zahlennachsprechaufgaben (vorwärts und rückwärts) eingefügt. Tabelle 8-F gibt die insgesamt unbefriedigenden Indizes der Anpassungsgüte des modifizierten Ad-hoc-Modells wieder. Die drei zentralen Indizes der Anpassungsgüte (GFI, NNFI, CFI) liegen unterhalb des Grenzwerts von 0.9.

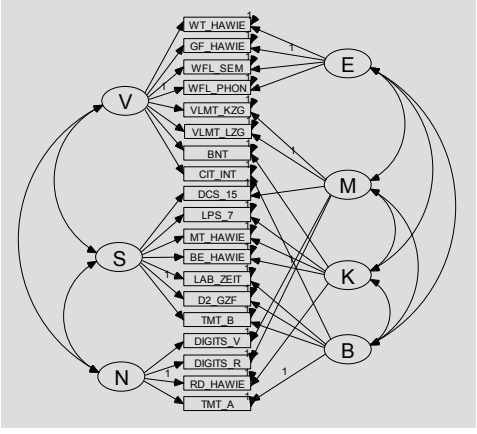
Für die Post-hoc-Modifikationen wurden zunächst die Ladungen der Indikatoren betrachtet, die bei der Ad-hoc-Operationalisierung anhand der Faktorendefinitionen nicht klar zugeordnet werden konnten (*Interferenztest*, *Bilderergänzen*). Es zeigt sich, dass beide Indikatoren starke und hoch signifikante Ladungen insbesondere auf dem kinesthetisch-mechanischen Faktor aufweisen, während die Ladungen auf dem verbal-edukativen Faktor zwar ebenfalls signifikant sind, aber deutlich schwächer ($p < .05$). Daher wird eine Entscheidung zugunsten des kinesthetisch-mechanischen Faktors getroffen. Zudem erweist sich die Aufgabe *Zahlennachsprechen vorwärts* als nicht signifikant für den Gedächtnisfaktor. Daher wird diese eliminiert, so dass diese Aufgabe nur noch auf dem verbal-edukativen Faktor lädt. Der Untertest *Zahlennachsprechen rückwärts* und der bildhafte Gedächtnistest *DCS* wurden nach Analyse der Ladungen alleinig dem Gedächtnisfaktor zugeordnet. Die Betrachtung der Fehlerkovarianzen zeigt hohe Modifikationsindizes für den *Trail Making Test*. Dies entspricht den vorformulierten Änderungsmöglichkeiten, so dass die entsprechende Kovarianz freigesetzt wurde. Die Anpassungsgüte des so modifizierten Modells kann nun als ausreichend bewertet werden (siehe Tabelle 8-F).

Tabelle 8-F: Güte der Anpassung des hierarchischen Intelligenzmodells von Vernon.

Modell	**Status**	**χ^2 [FG]**	**GFI**	**NNFI**	**CFI**	**RMSEA [90%-KI]**	**ECVI [gesättigt]**
Vernon	ad hoc (modifiziert)	229.72 [143]	0.891	0.879	0.899	0.057 [0.043-0.070]	1.713 [2.011]
	post hoc	212.78 [147]	0.896	0.911	0.923	0.049 [0.033-0.063]	1.581 [2.011]

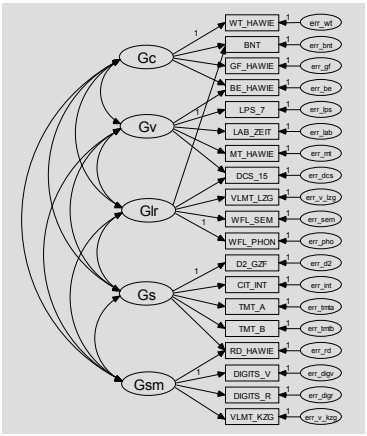
8.2.3 Das Berliner Intelligenzstrukturmodell

Die Operationalisierung des Berliner Intelligenzstrukturmodells beinhaltet dem Originalmodell entsprechend vier Operations- und drei Inhaltsfaktoren. Die Parameterschätzung war für dieses Modell nicht erfolgreich. Die ausgegebene Fehlermeldung verweist darauf, dass



entweder das spezifizierte Modell die Daten schlecht anpasst oder dass die Stichprobengröße zu klein ist. Die aktuelle Stichprobengröße von n = 190 liefert genau die von oben genannter Faustregel geforderten 10 Probanden pro Indikator. Die zweite Faustregel besagt zudem, dass das Verhältnis zwischen Probanden und frei schätzbaren Parametern zwischen 5:1 und 10:1 liegen sollte. Im diesem komplexen Modell hat jeder Indikator Doppelladungen, daher scheint die Stichprobengröße von 190 nicht ausreichend zu sein. Insgesamt enthält das Modell 66 frei zu schätzende Indikatoren (31 Ladungen, 19 Fehlervarianzen, 7 Faktorvarianzen und 9 Faktorenkovarianzen). Das Verhältnis zwischen Probanden und Parametern liegt bei 2.87:1 und unterschreitet die untere Grenze von 5:1. Diese würde eine minimale Stichprobengröße von 5*66 = 330 erfordern. Zu betonen ist, dass die erfolglose Modellanpassung aufgrund der zu geringen Stichprobengröße nicht bedeutet, dass das Modell *falsch* ist.

8.2.4 Das Cattell-Horn-Carroll-Modell



Zur Operationalisierung des CHC-Modells entsprechend der Cattell-Horn-Carroll-Theorie (McGrew, 1997) wurden die 19 Indikatoren fünf breiten Fähigkeitsfaktoren (*stratum II*) zugeordnet (Gc, Gv, Gs, Gsm, Glr). Die Anpassungsgüte der Ad-hoc-Spezifikation des CHC-Modells verweist laut konfirmatorischer Faktorenanalyse auf Spezifikationsfehler (Tabelle 8-G), so dass Post-hoc-Modifikationen notwendig sind. Anhand der Signifikanzniveaus der Regressionsgewichte sowie der Modifikationsindizes der Ladungen und Fehlerkovarianzen wurden hierzu folgende Post-hoc-Respezifikationen vorgenommen:

- Die Indikatoren der bildhaften Lernleistung (DCS_15) und des Tests *Bilderergänzen* (BE_HAWIE) laden nur auf dem visuell-räumlichen Faktor Gv.
- Der Benenntest BNT lädt nur auf dem Faktor der kristallinen Intelligenz Gc.
- Die Aufgabe zum rechnerischen Denken (RD_HAWIE) lädt nur auf dem Geschwindigkeitsfaktor Gs.
- Korrelierte Messfehler wurden für den *Trail-Making-Test* eingeführt.

Somit konnte empirisch eine Entscheidung über die Zuordnung auch der Indikatoren herbeigeführt werden, die theoretisch nicht eindeutig zuzuordnen waren. Alle verbliebenen Pfade sind nun modellkonform signifikant und positiv. Die ad hoc vorgenommene Zuordnung der Aufgabe zur phonematischen Wortflüssigkeit zum

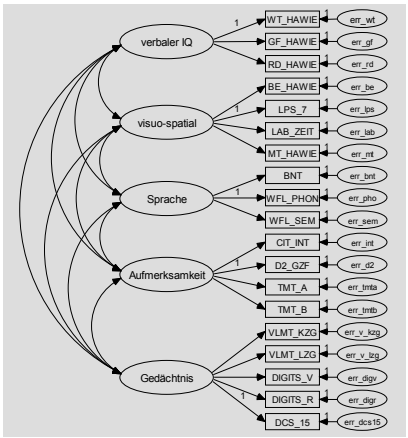
Langzeitgedächtnisfaktor Glr wurde bestätigt. Tabelle 8-G stellt die Güte der Anpassung des Post-hoc-Modells dar. Es zeigt sich, dass die Indikatoren NNFI und CFI über dem Grenzwert von 0.9 liegen. Insgesamt kann somit eine hinreichende – wenn auch nicht exzellente – Passung zwischen Modell und Daten angenommen werden.

Tabelle 8-G: Güte der Anpassung des Cattell-Horn-Carroll-Modells.

Modell	**Status**	**χ^2 [FG]**	**GFI**	**NNFI**	**CFI**	**RMSEA [90%-KI]**	**ECVI [gesättigt]**
CHC	ad hoc	229.60 [138]	0.886	0.868	0.893	0.059 [0.045-0.073]	1.765 [2.011]
	post hoc	210.28 [140]	0.895	0.902	0.919	0.051 [0.035-0.065]	1.631 [2.011]

8.2.5 Das neuropsychologische Modell

Bei der Operationalisierung des neuropsychologischen Modells wurden die 19 Indikatoren den Faktoren Gedächtnis, Aufmerksamkeit, visuell-räumliche Fähigkeiten, verbale Intelligenz und Sprache zugeordnet. Die Beurteilung der Modellgüte zeigt wieder, dass im hypothetischen Modell noch klare Fehler der Anpassung vorliegen (Tabelle 8-H). Daher sollen im nächsten Schritt wiederum die Modifikationsindizes betrachtet werden und auf deren Basis gegebenenfalls Post-hoc-Modifikationen vorgenommen werden.



Post-hoc-Modifikationen der Ladungen sind nicht notwendig: Alle ad hoc operationalisierten Ladungen sind hoch signifikant, so dass keine Notwendigkeit zur Entfernung überschüssiger Ladungen besteht; die Sichtung der Modifikationsindizes der Ladungen legt keine Doppelladungen nahe, so dass auch keine weiteren Pfade aufgenommen werden müssen. Bezüglich der Fehlerkovarianzen scheint eine Steigerung der Anpassungsgüte durch Modellierung korrelierter Fehlervarianzen zwischen den Tests *Zahlennachsprechen vorwärts* und *rückwärts* und den beiden Untertests des *Trail-Making-Test*s erreichbar zu sein. Beide korrelierte Fehlervarianzen sind als Operationalisierung gemeinsamer Methodenvarianz zu verstehen. Die Bewertung des modifizierten Modells zeigt, dass alle Ladungen weiterhin hoch signifikant sind, ebenso alle Kovarianzen einschließlich der Fehlerkovarianzen. Tabelle

8-H gibt die globale Güte der Modellpassung wieder, wobei alle relevanten Indikatoren eine ausreichende Güte der Anpassung anzeigen.

Tabelle 8-H: Güte der Anpassung des neuropsychologischen Modells (NP).

Modell	Status	χ^2 [FG]	GFI	NNFI	CFI	RMSEA [90%-KI]	ECVI [gesättigt]
NP	ad hoc	272.74 [142]	0.870	0.817	0.848	0.070 [0.057-0.082]	1.951 [2.011]
	post hoc	206.94 [140]	0.901	0.905	0.922	0.050 [0.035-0.064]	1.624 [2.011]

8.2.6 Zusammenschau der Ergebnisse

Welches der vier Modelle ist nun am besten zur Beschreibung der Daten der Normierungsstichprobe geeignet? Zur Beantwortung dieser Frage gibt Tabelle 8-I die Anpassungsgüte der Modelle in deren Ad-hoc-Modellspezifikation sowie nach den erfolgten Modellmodifikationen wieder:

Tabelle 8-I: Vergleichende Zusammenstellung der Indizes der Anpassungsgüte.

Modell	Status	χ^2 [FG]	GFI	NNFI	CFI	RMSEA [90%-KI]	ECVI [gesättigt]
g-Faktor	ad hoc	339.99 [152]	0.838	0.754	0.781	0.081 [0.069-0.092]	2.201 [2.011]
	post hoc	246.14 [149]	0.876	0.870	0.887	0.059 [0.045-0.072]	1.736 [2.011]
Vernon	ad hoc	229.72 [143]	0.891	0.879	0.899	0.057 [0.043-0.070]	1.713 [2.011]
	post hoc	212.78 [147]	0.896	0.911	0.923	0.049 [0.033-0.063]	1.581 [2.011]
BIS	ad hoc	- Lösung nicht identifizierbar -					
CHC	ad hoc	229.60 [138]	0.886	0.868	0.893	0.059 [0.045-0.073]	1.765 [2.011]
	post hoc	210.28 [140]	0.895	0.902	0.919	0.051 [0.035-0.065]	1.631 [2.011]
NP	ad hoc	272.74 [142]	0.870	0.817	0.848	0.070 [0.057-0.082]	1.951 [2.011]
	post hoc	206.94 [140]	0.901	0.905	0.922	0.050 [0.035-0.064]	1.624 [2.011]

Eine Beschreibung der Stichprobendaten mit dem Berliner Intelligenzstrukturmodell war nicht möglich. Es ist anzunehmen, dass die Ursache hierfür in der Stichprobengröße

liegt, die für die Komplexität des Modells zu klein ist. Die hohe Komplexität des Berliner Intelligenzstrukturmodells in Vergleich zu den anderen Modellen ist durch die Doppelladungen aller Indikatoren und hohe Anzahl von Faktoren bedingt. So besitzt beispielsweise bei gleicher Indikatorenanzahl das CHC-Modell 14 Freiheitsgrade mehr und ist demnach deutlich sparsamer. Für das Berliner Intelligenzstrukturmodell kann lediglich geschlussfolgert werden, dass nach dem Gütekriterium der Parsimonität die einfachen Faktorenmodelle günstiger abschneiden.

Von den vier Modellen, bei denen die Modellanpassung erfolgreich war, hatte das Generalfaktormodell eine so niedrige Anpassungsgüte, dass es für die Beschreibung der Daten als nicht gültig betrachtet werden kann.

Des Weiteren ist aus Tabelle 8-I ersichtlich, dass kein Modell in der Ad-hoc-Spezifikation ohne Spezifikationsfehler ist. Nach den empirisch basierten Post-hoc-Modifikationen weisen jedoch das hierarchische Modell von Vernon, das Cattell-Horn-Carroll-Modell sowie das neuropsychologische Modelle eine hinreichende Güte der Anpassung auf. Dabei ist die Güte der Anpassung bei allen drei Modellen sehr ähnlich und lässt keine eindeutige und empirisch fundierte Entscheidung für oder gegen eines der Modelle zu. Daher wird in einem nächsten Schritt die Eignung der modifizierten Modelle zur Beschreibung der *Patientenstichprobe* überprüft. Tabelle 8-J gibt die Ergebnisse dieser konfirmatorischen Faktorenanalysen, weiterhin im Ein-Gruppen-Fall, wieder:

Tabelle 8-J: Güte der Anpassung der Post-hoc-Modelle zur Beschreibung der Patientendaten.

Modell	**Status**	**χ^2 [FG]**	**GFI**	**NNFI**	**CFI**	**RMSEA [90%-KI]**	**ECVI [gesättigt]**
CHC	post hoc	220.29 [139]	0.891	0.904	0.921	0.055 [0.040-0.068]	1.684 [2.011]
NP	post hoc	281.76 [140]	0.857	0.827	0.858	0.073 [0.061-0.086]	2.020 [2.011]
Vernon	post hoc	302.81 [147]	0.845	0.819	0.844	0.075 [0.063-0.087]	2.057 [2.011]

In der Patientenstichprobe unterscheiden sich die drei Modelle durchaus in ihrer Güte der Anpassung. Das neuropsychologische Modell als auch das Modell nach Vernon bleiben in allen relevanten Indikatoren der Anpassungsgüte unbefriedigend, so dass hier wenig Berechtigung vorliegt, weitere Post-hoc-Modifikationen zu erproben[10]. Beim CHC-

[10] Tatsächlich bieten sich bei dem Modell nach Vernon auch keine an; beim neuropsychologischen Modell verweisen die Modifikationsindizes auf eine fehlende Fehlerkovarianz für die beiden Indikatoren des verbalen Lern- und Gedächtnistests. Die Aufnahme dieser Fehlerkovarianz führt jedoch zu keiner substanziellen Verbesserung der Anpassungsgüte.

Modell liegt lediglich der GFI-Index noch unter den Grenzwert von 0.9. Die Möglichkeit weiterer Post-hoc-Modifikationen kann durchaus überprüft werden. Für die Ladungen liegen keine Modifikationen nahe. Für die Fehlerkovarianzen zeigen die Modifikationsindizes, dass sich mithilfe der korrelierten Messfehler der beiden Indikatoren des verbalen Lern- und Gedächtnistests das Modell signifikant verbessern lässt. Nach Aufnahme der Fehlerkovarianz kann die Anpassungsgüte des CHC-Modells insgesamt als befriedigend gewertet werden (Tabelle 8-K).

Eine erneute konfirmatorische Faktorenanalyse in der Normierungsstichprobe zeigt, dass die zusätzlich eingefügte Fehlerkovarianz auch hier signifikant ist (r = .152; p = .045). Die Güte der Modellanpassung hat sich erwartungsgemäß leicht verbessert (Tabelle 8-K). Das CHC-Modell eignet sich also als Basis für die Untersuchungen zur Invarianz.

Tabelle 8-K: Güte der Anpassung des modifizierten CHC-Modells.

Modell	**Stichprobe**	**χ^2 [FG]**	**GFI**	**NNFI**	**CFI**	**RMSEA [KI]**	**ECVI [gesättigt]**
CHC	Patienten	200.42 [140]	0.901	0.926	0.940	0.048 [0.032-0.062]	1.590 [2.011]
CHC	Norm	206.48 [140]	0.898	0.905	0.923	0.050 [0.035-0.064]	1.622 [2.011]

Zur näheren Charakterisierung des Modells gibt Tabelle 8-L die R^2-Werte für jeden Indikator wieder. Diese Werte drücken aus, wie gut die latenten Variablen die Varianzen der beobachteten Variablen erklären. Es zeigt sich, dass die Streuung der R^2-Werte sehr hoch ist und somit die Erklärungskraft des Modells für die verschiedenen Variablen sehr unterschiedlich ausfällt. Abbildung 8-A schließlich gibt die Parameterschätzungen für die Ladungen und Korrelationen in der Normierungs- und in der Patientenstichprobe wieder. Diese Parameterschätzungen resultieren aus Modelltests, die für jede Gruppe *unabhängig* durchgeführt wurden (Ein-Gruppen-Fall der konfirmatorischen Faktorenanalyse). Separate Modelltests bieten zur Invarianz lediglich einen groben Überblick über die Konsistenz der Modelle, beantworten aber nicht die Frage, ob die Parameterschätzungen zwischen den Gruppen gleich und somit invariant sind. Diese Frage wird in den folgenden Kapiteln durch konfirmatorische Faktorenanalysen im Zwei-Gruppen-Fall beantwortet.

Tabelle 8-L: Anteil der durch die Faktoren aufgeklärten Varianz der Indikatoren (R^2).

	Normstichprobe	**Patienten**
BE_HAWIE	0.158	0.108
BNT	0.223	0.318
CIT_INT	0.310	0.451
D2_GZF	0.395	0.408
DCS_15	0.446	0.265
DIGITS_R	0.545	0.500
DIGITS_V	0.500	0.411
GF_HAWIE	0.580	0.639
LAB_ZEIT	0.097	0.332
LPS_7	0.336	0.305
MT_HAWIE	0.367	0.447
RD_HAWIE	0.214	0.378
TMT_A	0.163	0.290
TMT_B	0.371	0.475
VLMT_KZG	0.107	0.271
VLMT_LZG	0.131	0.090
WFL_PHON	0.736	0.656
WFL_SEM	0.254	0.419
WT_HAWIE	0.630	0.582

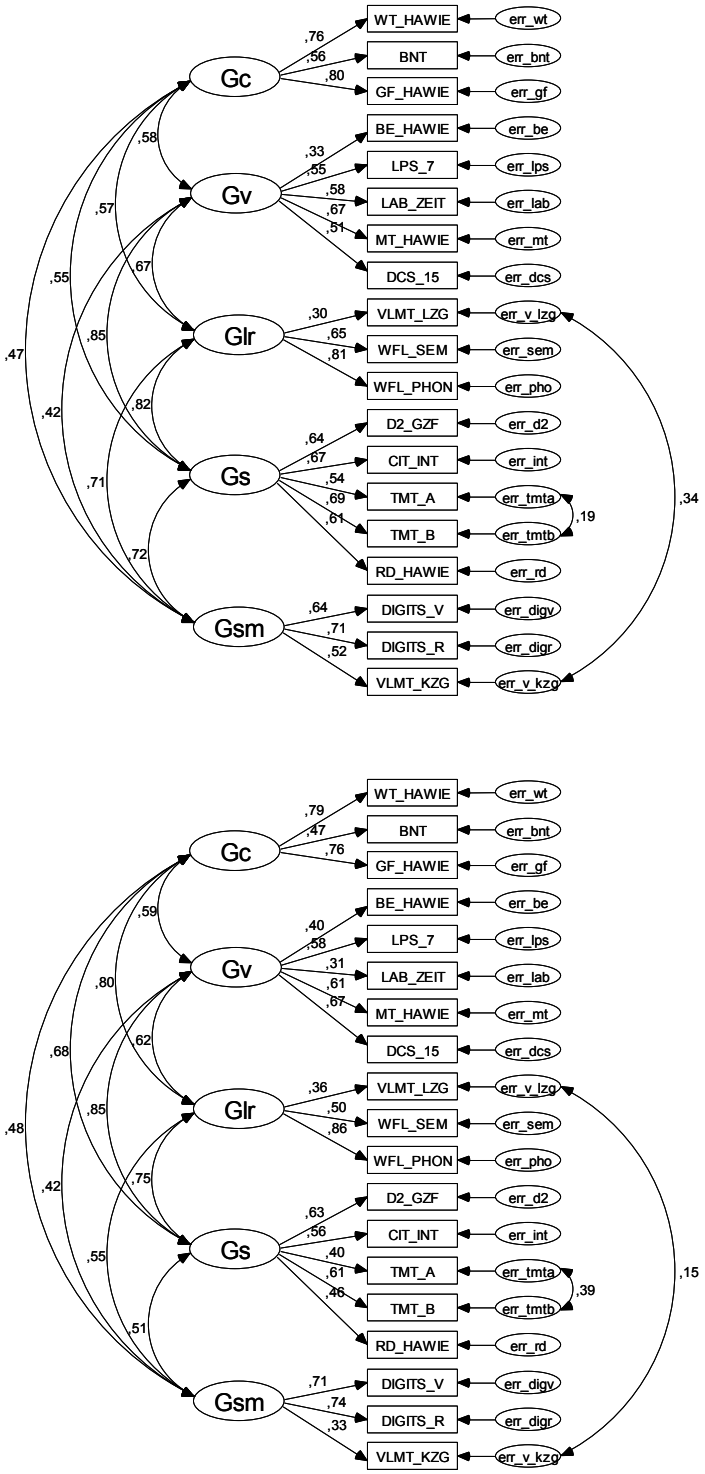


Abbildung 8-A: Das CHC-Modell mit gruppenspezifischen Ladungen und Korrelationen bei den Patienten (oben) und bei der Normstichprobe (unten).

8.3 Fragestellung B: Untersuchung auf konfigurale Invarianz

Im vorherigen Kapitel wurde gezeigt, dass das CHC-Modell für jede der beiden Stichproben Gültigkeit besitzt. Daraus ergibt sich ein erster Hinweis auf das Vorliegen konfiguraler Invarianz. Methodisch stringent ist konfigurale Invarianz jedoch nur im Zwei-Gruppen-Fall der konfirmatorischen Faktorenanalyse nachzuweisen. Dieses Vorgehen erfordert die simultane Parametrisierung in beiden Gruppen und ermöglicht gleichzeitig Testung auf gruppenspezifische Unterschiede der Parameter. Weil nicht alle Indizes der Anpassungsgüte einen direkten Modellvergleich zulassen, wird für diese und die folgenden Analysen ein reduzierter Satz von Indizes zur vergleichenden Bewertung der Modelle herangezogen (vergleiche Kapitel 4.3).

Die Indizes bestätigen das Ergebnis der vorherigen Analyse: Das CHC-Modell hat auch bei simultaner Analyse in beiden Gruppen eine hohe Anpassungsgüte und eignet sich zur Beschreibung der Daten der Patientenstichprobe ebenso gut wie für die Daten der Normstichprobe (Tabelle 8-M, erste Zeile).

An dieser Stelle soll das Modell zusätzlich in einer leichten Modifikation mit einem Generalfaktor auf die Anpassungsgüte für beide Stichproben überprüft werden (Abbildung 8-B). Die Anpassungsindizes in Tabelle 8-M (zweite Zeile) zeigen zunächst, dass auch dieses Modell gut mit den Daten vereinbar ist. Verglichen mit dem vorherigen Modell mit obliquen Gruppenfaktoren aber ohne Generalfaktor ist die Güte der Anpassung allerdings geringfügig schlechter. Das Modell ohne Generalfaktor wird daher Gegenstand der weiteren Analysen sein. Der Vorteil des Modells mit Generalfaktor wäre seine Parsimonität gewesen. Mit 290 Freiheitsgraden ist es sparsamer als das Modell mit den obliquen Gruppenfaktoren mit 280 Freiheitsgraden.

Tabelle 8-M: Test auf konfigurale Invarianz des CHC-Modells ohne oder mit g-Faktor.

CHC-Modell	**χ^2 [FG]**	**CFI**	**RMSEA [KI]**	**ECVI [KI]**
baseline (oblique)	406.902 [280]	0.932	0.035 [0.027-0.042]	1.606 [1.474-1.758]
mit g-Faktor	438.925 [290]	0.920	0.037 [0.030-0.044]	1.637 [1.499-1.797]

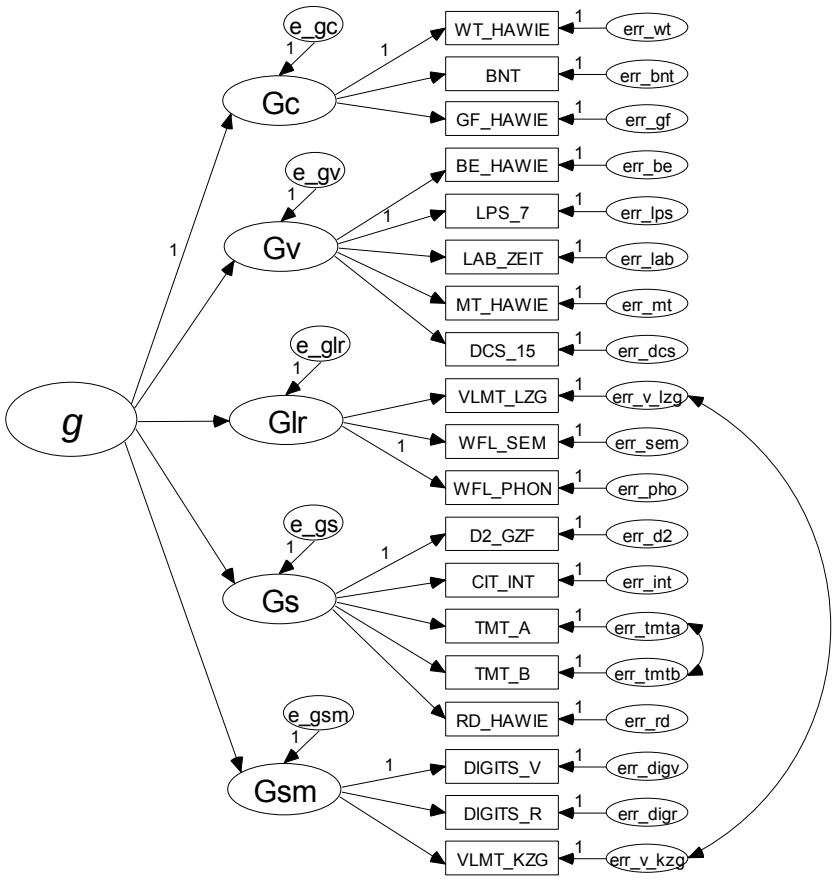


Abbildung 8-B: Hierarchisches CHC-Modell mit Generalfaktor.

8.4 Fragestellung C: Untersuchungen auf metrische Invarianz

Durch weitere konfirmatorische Faktorenanalysen soll, bezogen auf das CHC-Modell, die Invarianz des Messmodells in der Normierungs- und in der Patientenstichprobe untersucht werden. Zunächst wird die Gleichheit der Faktorladungen in beiden Gruppen überprüft und die Frage nach *schwacher metrischer Invarianz* beantwortet. Dazu wird simultan für beide Gruppen ein Modelltest mit der zusätzlichen Annahme identischer Ladungsparameter durchgeführt. Tabelle 8-N gibt die Güte der Anpassung des restringierten Modells im Vergleich zum unrestringierten Modell (baseline) wieder:

Tabelle 8-N: Test auf schwache metrische Invarianz des CHC-Modells.

Status	**χ^2 [FG]**	**CFI**	**RMSEA [KI]**	**ECVI [KI]**
baseline	406.902 280	0.932	0.035 [0.027-0.042]	1.606 [1.474-1.758]
Ladungen invariant	462.990 294	0.909	0.039 [0.032-0.046]	1.680 [1.536-1.845]

Die Differenz des χ^2-Wertes beträgt 56.088. Der χ^2-Differenztest macht die Nullhypothese der Gleichheit der Faktorladungsmatrizen unhaltbar: Die obere Grenze zur Beibehaltung der Nullhypothese (p = .01) ist bei 14 Freiheitsgraden der tabellarische χ^2-Wert von 29.14. Auch die Differenz des CFI-Index der Anpassungsgüte liegt mit 0.023 über dem bei Invarianz maximal zulässigen Differenzwert von 0.02. Die Indizes RMSEA und ECVI des begrenzteren Modells liegen dagegen innerhalb der Konfidenzintervalle des *baseline*-Modells und wären mit der Invarianzannahme vereinbar. Insgesamt ist aber die Hypothese der Gleichheit der Faktorladungsmatrizen zu verwerfen.

Möglicherweise können besonders invariante Parameter identifiziert werden, um sie zu eliminieren oder freizusetzen. Hierzu werden, ausgehend vom unbeschränkten Modell (*baseline-Modell, mit* χ^2 = 406.902), Schritt für Schritt die Ladungen aller Indikatoren einzeln als invariant modelliert. Tabelle 8-O gibt den jeweiligen Verlust der Anpassungsgüte als χ^2–Differenzwert ($\Delta\ \chi^2$, 1 Freiheitsgrad) wieder.

Tabelle 8-O: Abfall der Anpassungsgüte des *baseline*-Modells durch Festsetzen einzelner Ladungen.

Faktor	**Indikator**	χ^2	$\Delta\ \chi^2$
Gc	BNT	414.102	7.200**
	GF_HAWIE	410.686	3.784
Gv	BE_HAWIE	407.413	0.511
	LAB_Z	426.459	19.557**
	MT_HAWIE	408.395	1.493
	DCS_15	406.920	0.018
Glr	WFL_SEM	412.135	5.233*
	VLMT_LZG	407.312	0.410
Gs	CIT_INT	416.042	9.140**
	TMT_A	412.559	5.657*
	TMT_B	420.425	13.523**
	RD_HAWIE	415.394	8.492**
Gsm	VLMT_KZG	412.157	5.255*
	DIGITS_R	407.035	0.133

Anmerkung: *: p < .05; **: p < .01, 1 Freiheitsgrad

Die χ^2–Differenzen zeigen, dass insbesondere bei der Ladung des *Labyrinthtests* und des *Trail-Making-Tests* große Gruppenunterschiede bestehen. Deshalb sollen diese beiden Indikatoren eliminiert werden. Im Hinblick auf eine sinnvolle Operationalisierung des CHC-Modells ist das machbar, da die zugehörigen Faktoren (Gv bzw. Gs) jeweils

über fünf Indikatoren operationalisiert sind, während die anderen Faktoren über je drei Indikatoren operationalisiert sind. Durch den Ausschluss des *Labyrinthtests* und des *Trail-Making-Tests* entsteht als positiver Nebeneffekt eine ausgewogenere Operationalisierung der Faktoren. Die Güte der Anpassung des reduzierten Modells für die unbeschränkte Konfiguration und für die schwache Invarianzhypothese gibt Tabelle 8-P wieder.

Tabelle 8-P: Test auf schwache metrische Invarianz des reduzierten CHC-Modells (ohne Lab_Z und TMT_B).

Modell	**χ^2 [FG]**	**CFI**	**RMSEA [KI]**	**ECVI [KI]**
baseline	314.782 [216]	0.937	0.035 [0.026-0.043]	1.309 [1.194-1.445]
Ladungen invariant	344.976 [228]	0.925	0.037 [0.029-0.045]	1.325 [1.204-1.468]

Die Differenz des χ^2-Wertes zwischen den Modellen mit und ohne Gleichheitsbeschränkung für die Ladung beträgt nun 30.194. Der χ^2-Differenztest zeigt, dass der empirische χ^2-Wert über dem tabellarischen Wert von 26.22 liegt (12 Freiheitsgrade, p = 0.01). Ein anderes Bild ergeben die anderen Indikatoren: Die Differenz des CFI-Indexes liegt mit 0.012 deutlich unter dem bei Invarianz maximal zulässigen Differenzwert von 0.02. Auch die Indizes RMSEA und ECVI des Invarianzmodells liegen innerhalb der Konfidenzintervalle des *baseline*-Modells und sind mit der Invarianzannahme vereinbar. Vor dem Hintergrund, dass der χ^2-Differenztest als sehr streng gilt, ist die Gesamtkonstellation der Veränderungen der Anpassungsgüte mit der Hypothese der Gleichheit der Faktorladungsmatrizen vereinbar. Das – nun auf 17 Indikatoren reduzierte – CHC-Modell weist also schwache metrische Invarianz auf. Entsprechend erübrigt sich die weitere Suche nach nicht-invarianten Indikatoren und die nächste Stufe der Invarianz, die starke metrische Invarianz, kann überprüft werden.

Zunächst wird untersucht, ob die Höhenlagen der Indikatoren gruppenunabhängig sind. Dabei fließt das ganze faktorenanalytische Modell mit sämtlichen Modellparametern ein, so dass nicht nur die Varianz-Kovarianzmatrix als Datengrundlage dient, sondern auch die Mittelwertvektoren der Indikatoren. Die Invarianzanforderungen beziehen sich auf die Faktorladungen und Höhenlagen. Die Strukturparameter (Faktorenkovarianzen, Faktorvarianzen und Faktorenmittelwerte) können unterschiedliche Werte in den Gruppen annehmen, ebenso die Fehlervarianzen und -kovarianzen. Zur Identifizierbarkeit der Modelle sind die Faktorenmittelwerte der Normgruppe auf 0 zu setzen und die Faktorenmittelwerte der Patienten sind frei

schätzbar. Zusätzlich werden die Höhenlagen der standardisierenden Indikatoren (Ladung = 1) auf Null gesetzt.

Ein Vergleich des Modells mit gleichgesetzten Höhenlagen mit einem unbeschränkten Modell ist erübrigt sich, weil bei Modellierung der latenten Mittelwertstruktur unter Berücksichtigung aller Parameter schon das unbeschränkte vollständige Messmodell eine ungenügende Anpassungsgüte aufweist (Tabelle 8-Q).

Tabelle 8-Q: Indizes der Anpassungsgüte zur Untersuchung der latenten Mittelwertsstruktur.

Modell	**χ^2 [FG]**	**CFI**	**RMSEA [KI]**	**ECVI [KI]**
baseline	1138.159 [233]	0.883	0.101 [0.096-0.107]	3.577 [3.309-3.865]

Zusammenfassend konnte für das Cattell-Horn-Carroll-Modell die Annahme schwacher metrischer Invarianz bestätigt werden. Das heißt die Daten der Normierungsgruppe und der Patientengruppe sind mit dem gleichen Intelligenzstrukturmodell beschreibbar und die Faktorladungen entsprechen sich in beiden Gruppen. Dagegen konnten die strengeren Invarianzannahmen des Messmodells (starke und strenge metrische Invarianz) nicht bestätigt werden.

8.5 Fragestellung D: Untersuchungen auf strukturelle Invarianz

Als Bedingung für die Untersuchung der strukturellen Invarianz der Faktorenvarianzen und -kovarianzen reicht der erbrachte Nachweis schwacher metrischer Invarianz. Die Überprüfung der Invarianz dieser beiden Strukturparameter kann anhand der empirischen Kovarianzmatrix durchgeführt werden. Zur Überprüfung der Faktorenmittelwerte auf Gleichheit sind zusätzlich die empirischen Mittelwertvektoren heranzuziehen. Im nächsten Schritt wird, ausgehend vom schwachen Invarianzmodell (*baseline-Modell*), die Gleichheit der Faktorkovarianzen untersucht.

Tabelle 8-R: Test auf strukturelle Invarianz der Faktorenkovarianzen und Faktorenvarianzen.

Modell	**χ^2 [FG]**	**CFI**	**RMSEA [KI]**	**ECVI [KI]**
Schwache metrische Invarianz	344.976 [228]	0.925	0.037 [0.029-0.045]	1.325 [1.204-1.468]
Faktorenkovarianzen	369.223 [238]	0.916	0.038 [0.030-0.046]	1.337 [1.209-1.485]
Faktorvarianzen	391.900 [243]	0.905	0.040 [0.033-0.047]	1.370 [1.237-1.524]

Die Differenz des χ^2-Wertes der Modelle mit und ohne Gleichheitsbeschränkung der Faktorenkovarianzen beträgt nun 24.257. Dieser Wert überschreitet den tabellarischen Grenzwerter von 23.21 geringfügig (p = .01; 10 Freiheitsgrade). Die Differenz des CFI-Index der Anpassungsgüte liegt mit 0.009 deutlich unter dem bei Invarianz maximal zulässigen Differenzwert von .02. Auch die Indizes RMSEA und ECVI haben sich durch die Invarianzrestriktionen nur sehr geringfügig und innerhalb der Grenzen des Konfidenzintervalls des unbegrenzten Modells verändert. Die Veränderungen der Indizes der Anpassungsgüte insgesamt sind mit der Hypothese der Gleichheit der Matrizen der Faktorenkovarianzen vereinbar.

Für die Faktorvarianzen stellt sich die Lage ähnlich dar: Die Differenz des χ^2-Wertes der Modelle mit und ohne Gleichheitsbeschränkung der Faktorvarianzen beträgt bei 16 Freiheitsgraden 46.924 und überschreitet den tabellarischen Grenzwert von 32.00 (p = .01). Die Differenz des CFI-Index der Anpassungsgüte entspricht mit 0.02 dem bei Invarianz maximal zulässigen Differenzwert von 0.02. Die Indizes RMSEA und ECVI haben sich durch die Invarianzrestriktionen verschlechtert, liegen aber weiterhin innerhalb der Grenzen des Konfidenzintervalls des unbegrenzten Modells. In der Gesamtkonstellation sind die Veränderungen der Indizes der Anpassungsgüte mit der Hypothese der Gleichheit der Matrizen der Faktorenvarianzen vereinbar.

Bei gleichen Faktorenkovarianzen und Faktorvarianzen sind die Faktorkorrelationen ebenfalls als gleich anzusehen. Tabelle 8-S gibt die Faktorinterkorrelationen wieder. Sie streuen zwischen.827 und .444, die durchschnittliche Faktorinterkorrelation beträgt .685.

Tabelle 8-S: Faktorinterkorrelationen für beide Stichproben.

	Gsm	**Gs**	**Glr**	**Gv**
Gsm				
Gs	.621			
Glr	.660	.806		
Gv	.444	.827	.635	
Gc	.496	.643	.678	.616

Anmerkungen: 20 bis 68 Prozent gemeinsame Varianz, .658 als durchschnittliche Interkorrelation. Gsm: Kurzzeitgedächtnis, Gs: kognitive Verarbeitungsgeschwindigkeit, Glr: Langzeitspeicherung und Abruf, Gv: visuell-räumliche Fähigkeiten, Gc: kristalline Intelligenz

Zusammenfassend kann angenommen werden, dass die beiden Strukturparameter Faktorvarianz und Faktorenkovarianz in beiden Stichproben gleich sind. Wenn die Parameter als invariant gelten können, sind die Unterschiede der Faktorkorrelationen als statistisch nicht relevant anzusehen. Die Frage nach der Gleichheit der Faktorenmittelwerte konnte nicht beantwortet werden, da für eine sinnvolle

Beantwortung dieser Frage starke metrische Invarianz vorliegen muss. Für die Faktorenmittelwerte wäre jedoch ohnehin keine Invarianz anzunehmen gewesen.

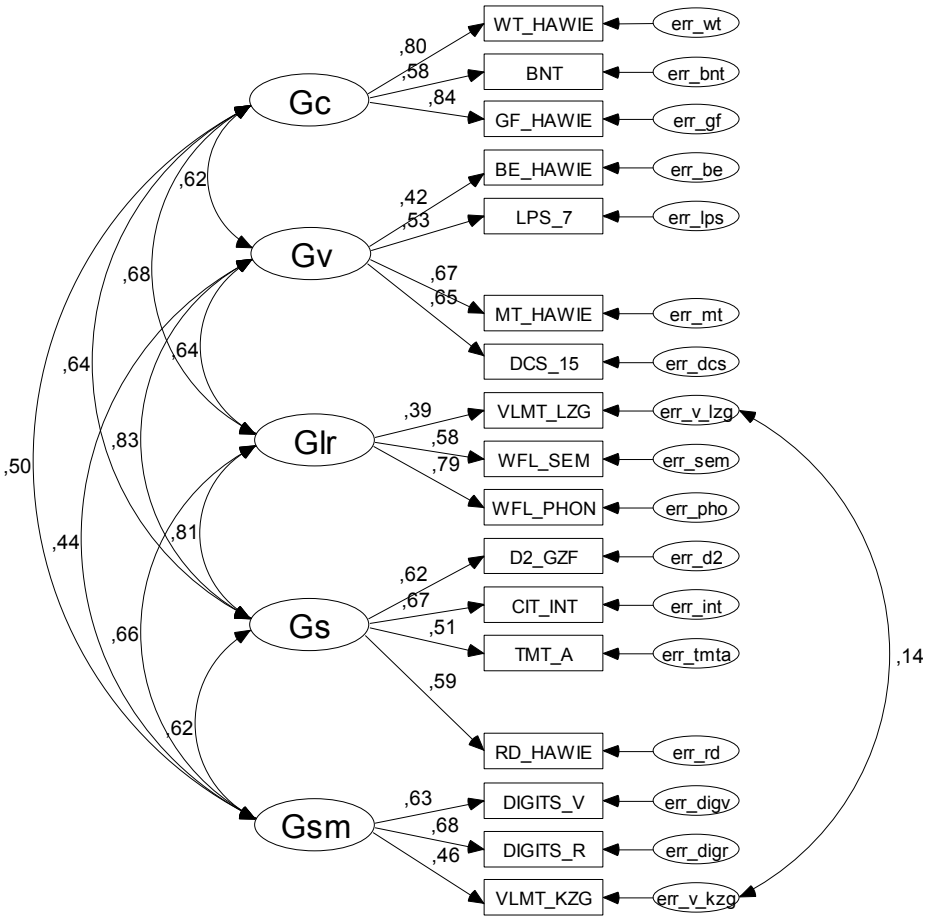


Abbildung 8-C: Standardisierte Parameter (Korrelationen und Ladungen) des endgültigen CHC-Invarianzmodells (für beide Stichproben).

8.6 Fragestellung E: Untersuchungen zur Konstruktvalidität

Die folgenden Analysen dienen der Konstruktvalidierung der fünf Faktoren des CHC-Modells. Zunächst wurden anhand der Regressionsgewichte, die im Rahmen der konfirmatorischen Faktorenanalyse geschätzt wurden, individuelle Faktorwerte auf Basis der Testrohwerte ermittelt. Es sei an dieser Stelle angemerkt, dass es sich bei den Faktorwerten um alterskorrigierte Werte handelt (siehe 8.1).

Abhängig von den verfügbaren Kriteriumsvariablen wurden zur Konstruktvalidierung verschiedene Zugänge gewählt. Die Ergebnisse der Invarianztests determinieren jedoch die Entscheidung, welche weiteren Untersuchungen zur Konstruktvalidierung des faktoriellen Modells möglich sind. Gezeigt wurde konfigurale und schwache metrische Invarianz sowie strukturelle Invarianz der Faktorvarianzen und Faktorenkovarianzen. Die

Formen der Invarianz, die auf die Kovarianzstruktur bezogen sind, wurden bestätigt; die darüber hinausgehenden strengeren Invarianzannahmen, die zusätzlich die Mittelwertstruktur mit einbeziehen, konnten nicht bestätigt werden. Folglich können ausschließlich korrelative Analysen zur Validierung der latenten Konstrukte herangezogen werden. Die Betrachtung latenter Mittelwertdifferenzen zur Überprüfung vorhergesagter Gruppenunterschiede ist nicht möglich.

Tabelle 8-T gibt die Kriterienkorrelationen der fünf Faktoren mit relevanten Hintergrund- und klinischen Variablen wieder. Um die Alphafehler-Inflation aufgrund multipler Vergleiche zu reduzieren, wurden die Signifikanzniveaus auf $p < .01$ beziehungsweise $p < .001$ abgesenkt. Einen negativen Einfluss auf die Faktoren, insbesondere auf die Langzeitgedächtniskomponente, hat die Anzahl der antiepileptischen Medikamente. Die Schulbildung hat einen hoch signifikanten positiven Einfluss auf alle fünf Faktoren; die höchste Interkorrelation besteht mit dem Faktor der kristallinen Intelligenz. Der Einfluss einer negativen Stimmungslage ist insgesamt in der Tendenz negativ auf die kognitive Leistung, wobei sich lediglich für den kristallinen Intelligenzfaktor eine (hoch) signifikante Korrelation zeigt.

Tabelle 8-T gibt auch Auskunft darüber, welchen Einfluss der Beginn und die Dauer der Epilepsie auf die kognitive Leistung haben. Während es keinen signifikanten Zusammenhang zwischen Dauer der Krankheit und Intelligenz gibt, besteht zwischen dem Alter bei Beginn der Erkrankung und den Intelligenzfaktoren insgesamt ein positiver Zusammenhang: Je später der Krankheitsbeginn, desto besser die aktuelle kognitive Leistung. Der Faktor kristalline Intelligenz hat dabei die höchste Interkorrelation mit dem Krankheitsbeginn. Für die Leistung in diesem Konstrukt ist also ein früher Krankheitsbeginn insbesondere ungünstig.

Tabelle 8-T: Kriterienkorrelationen zwischen den Faktorwerten und wichtigen Hintergrundvariablen.

	n	Gsm	Gs	Glr	Gv	Gc
Anzahl Medikamente [a]	189	-.186	-.239*	-.260**	-.204*	-.192*
Schulbildung [a]	190	.330**	.399**	.360**	.384**	.442**
BDI-Wert [b]	177	-.125	-.135	-.128	-.127	-.214**
Dauer der Erkrankung	190	-.127	-.141	-.164	-.095	-.122
Beginn der Erkrankung	190	.117*	.192*	.201*	.195*	.269**

Anmerkungen: *: $p < .01$, ** $p < .001$; [a] Rangkorrelation; [b] Beck-Depressions-Inventar (Hautzinger et al., 1995). Gsm: Kurzzeitgedächtnis, Gs: kognitive Verarbeitungsgeschwindigkeit, Glr: Langzeitspeicherung und Abruf, Gv: visuell-räumliche Fähigkeiten, Gc: kristalline Intelligenz.

Der Einfluss der Variablen Beginn und Dauer kann auch in einem Modell mit latenten Variablen modelliert werden (siehe Abbildung 8-D). So können die Effekte der Hintergrundvariablen auf die latenten Variablen direkt erfasst werden. Die in Tabelle 8-U

dargestellten Ergebnisse zeigen, dass der Krankheitsbeginn lediglich einen hochsignifikanten Einfluss auf die kristalline Intelligenz (Gc) hat und einen signifikanten Einfluss auf die visuell-räumlichen Fähigkeiten (Gv), nicht aber auf die anderen Intelligenzfaktoren. Die Krankheitsdauer hat auch bei dieser Analyseart keinen Einfluss auf die kognitiven Leistungsfaktoren.

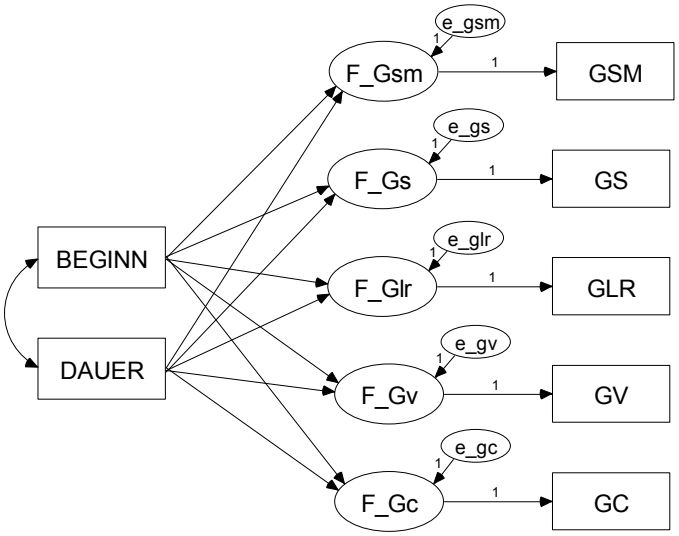


Abbildung 8-D: Pfaddiagramm zum Einfluss von Beginn und Dauer der Krankheit.
Abkürzungen: Gsm: Kurzzeitgedächtnis, Gs: kognitive Verarbeitungsgeschwindigkeit, Glr: Langzeitspeicherung und Abruf, Gv: visuell-räumliche Fähigkeiten, Gc: kristalline Intelligenz.

Tabelle 8-U: Ergebnis der Pfadanalyse zum Einfluss der Variablen Beginn und Dauer auf die Intelligenzfaktoren.

	Beginn	**Dauer**
Gc	.289, *p* = .001	.036
Gv	.204, *p* = .016	.017
Glr	.159	-.077
Gs	.163	-.052
Gsm	.068	-.090

Anmerkung: Korrelation zwischen Beginn und Dauer: *r* = - 0.545, *p* < . 001. Gsm: Kurzzeitgedächtnis, Gs: kognitive Verarbeitungsgeschwindigkeit, Glr: Langzeitspeicherung und Abruf, Gv: visuell-räumliche Fähigkeiten, Gc: kristalline Intelligenz.

9 Diskussion

Viele der veröffentlichen Standardtestbatterien, so beispielsweise die Wechsler-Gedächtnistestbatterie oder der Wechsler-Intelligenztest, wurden speziell zur Erfassung breiter Fähigkeitsfaktoren entworfen und sind deshalb für eine Ergebnisinterpretation auf Einzeltestebene nicht unbedingt geeignet (McDennott, Fanttizzo & Glutting, 1990). Die hier untersuchte neuropsychologische Testbatterie setzt sich im Unterschied dazu aus etablierten Einzelverfahren zusammen. Sie wurden anhand der erhofften oder belegten Validität der Einzelverfahren ausgewählt. Die Validität der Einzelverfahren wurde für jeden Test separat bestätigt, wenn auch in jeweils unterschiedlichen Untersuchungskontexten und anhand von unterschiedlichen Stichproben. Es liegt also nahe, die Interpretation der Testergebnisse auf Ebene der Einzelverfahren vorzunehmen. Bezogen auf Patienten mit Epilepsie konnte beispielsweise die klinische Validität für den verbalen Gedächtnistest VLMT (Helmstaedter, Grunwald, Lehnertz, Gleissner & Elger, 1997), den bildhaften Gedächtnistest DCS (Gleissner et al., 1998), die Wortflüssigkeit (Gleissner & Elger, 2001) oder verschiedene exekutive Testaufgaben (Helmstaedter, Kemper & Elger, 1996) belegt werden.

Die vorliegende Arbeit fokussiert jedoch nicht auf die Validität der Einzelverfahren, sondern auf die Validität breiter Leistungsfaktoren. Auf psychometrischer Ebene besteht der Vorteil der faktoriellen Zusammenfügung bestimmter Einzeltests darin, dass eine reliablere Erfassung des Zielkonstruktes erreicht wird. Inhaltlich ermöglicht die Faktoreninterpretation mit Bezug auf etablierte faktorielle Strukturmodelle die Übernahme des empirischen Fundaments und des theoretischen Hintergrundes dieser Konstrukte. Dieses Vorgehen verbessert also nicht nur die Reliabilität, sondern auch die Konstruktvalidität der Testbatterie.

Die Aussagekraft faktorieller Lösungen ist jedoch nicht unumstritten (siehe z. B. Bortz, 1993; Delis et al., 2003; Dodrill, 1999; Larrabee, 2003): Kritikpunkte sind unter anderem die Abhängigkeit faktorieller Lösungen von den Tests und den Probanden und von der gewählten Methodik der Faktorenanalyse. Diese methodischen Probleme beziehen sich jedoch insbesondere auf die explorative Faktorenanalyse. Konfirmatorische Faktorenanalysen können durch ein stringentes und theoriegeleitetes Vorgehen diese Probleme teilweise gut lösen. Davon unberührt bleibt aber, dass für ganz spezifische Fragestellungen durch Akkumulierung der Einzelverfahren zu Faktoren die Sensitivität der Tests verwässert werden kann. Sowohl die Interpretation von Faktoren, ebenso wie die Interpretation von Einzelverfahren, hat jeweils Vor- und Nachteile. Die beiden Ansätze sollten deshalb nicht gegeneinander ausgespielt werden.

Vielmehr sollte zur umfassenden Interpretation der Ergebnisse einer neuropsychologischen Untersuchung aus jedem Ansatz der maximale Nutzen gezogen werden.

Gegenstand der vorliegenden Untersuchung ist die neuropsychologische Testbatterie der Epileptologie der Universität Bonn (Helmstaedter, 2000). Zur Bestimmung der Konstruktvalidität dieser Testbatterie wurde untersucht, welche Intelligenztheorie die Daten am besten beschreibt. Von besonderem Interesse war dabei die Frage, ob es ein Intelligenzstrukturmodell gibt, das eine Stichprobe von gesunden Probanden ebenso gut beschreibt wie eine Stichprobe von Epilepsiepatienten. Um die Fragen sinnvoll und methodisch schlüssig zu beantworten, bieten sich konfirmatorische Faktorenanalysen an. Im Vergleich zur explorativen Faktorenanalyse bilden die Intelligenztheorien dabei nicht nur die theoretische Matrix zur abschließenden Faktoreninterpretation, sondern sind der Ausgangspunkt der Untersuchungen. Ein Vorteil der konfirmatorischen Faktorenanalyse, der für den aktuellen Untersuchungszusammenhang besonders betont werden soll, ist die Möglichkeit, die verschiedenen Intelligenztheorien parallel in mehreren Stichproben zu überprüfen.

Die bisher veröffentlichten Arbeiten mit ähnlicher Zielsetzungen untersuchen zum einen fast nur Standardtestbatterien, zum anderen sind die betrachteten faktoriellen Modelle wenig theoriebasiert und beruhen zumeist auf explorativen Vorstudien mit ähnlichen Daten (vergleiche Kapitel 5). Es werden also keine Modelle zur Struktur der Intelligenz betrachtet, sondern lediglich Modelle zur Struktur der Intelligenztests. Solche Studien sind dennoch von praktischer Bedeutung, weil sie die Konstruktvalidität der Standardtestbatterien für verschiedene klinische Stichproben aufklären. Die vorliegende Arbeit unterscheidet sich davon in zwei wichtigen Punkten: Zum einen wird eine weitgehend klinikinterne Testbatterie und keine Standardtestbatterie untersucht, zum anderen bezieht sich die Arbeit theoriebasiert auf etablierte Intelligenzstrukturmodelle. Die Ergebnisse können so in einem übergeordneten theoretischen Rahmen eingeordnet werden und sind damit testbatterieunabhängig und klinikübergreifend vergleichbar. Eine Arbeit, die ein ähnliches methodisches Vorgehen wählt und die Ergebnisse der Intelligenztestung explizit in Bezug auf aktuelle Strukturmodelle einordnet und somit deutlich über den theoretischen Rahmen der eingesetzten Einzeltests hinausgeht, ist dem Autor nicht bekannt.

Zunächst wurde untersucht, welches Intelligenzstrukturmodell sich zur Beschreibung der empirischen Kovarianzstruktur der gesunden Kontrollprobanden am besten eignet

(Fragestellung A). Dazu konnte keine klare Entscheidung zwischen dem Modell nach Cattell, Horn und Carroll, einem klassischen neuropsychologischen Modell und dem Intelligenzmodell von Vernon getroffen werden: Alle drei Modelle wiesen eine annähernd gleiche Güte der Anpassung auf. Erst die Patientenstichprobe lieferte eine Entscheidung: Einzig das Cattell-Horn-Carroll-Modell (CHC-Modell, McGrew, 2003) besaß eine hinreichende Anpassungsgüte in beiden Stichproben. Die separate Testung der Anpassungsgüte in jeder Gruppe einzeln zeigt aber nur überblicksartig die Gültigkeit der Modelle in beiden Gruppen. Ein methodisch schlüssiger Nachweis, ob die Testbatterie in beiden Gruppen die gleichen Konstrukte erfasst, kann nur durch parallele konfirmatorische Faktorenanalysen in beiden Gruppen erbracht werden (Mehr-Gruppen-Fall). Bei der simultanen Analyse wies das CHC-Modell in beiden Gruppen eine gute Anpassungsgüte auf, so dass die Hypothese konfiguraler Invarianz als bestätigt gelten kann (Fragestellung B). Bei Vorliegen konfiguraler Invarianz können die resultierenden Faktoren als ähnlich interpretiert werden. Doch damit die Faktoren als gleich interpretiert werden können, müssen sowohl die Indikatoren, die den Faktor konstituieren, gleich sein, als auch die Faktorladungen. Ordinale Unterschiede in den Faktorladungen zwischen den Gruppen führen dazu, dass die Indikatoren für die Konstruktdefinition unterschiedliche Bedeutung haben und die Konstrukte nicht direkt vergleichbar sind. Für die Faktoren der hier untersuchten Testbatterie konnte jedoch Gleichheit der Faktorladungen nachgewiesen werden (schwache metrische Invarianz, Fragestellung C). Hierfür mussten jedoch die nicht-invarianten Indikatoren *Labyrinthtest* und *Trail-Making-Test B* ausgeschlossen werden.

Das Cattell-Horn-Carroll-Modell (McGrew, 2005) wurde entwickelt, um durch die Definition breiter und enger Fähigkeitskonstrukte eine umfassende Taxonomie kognitiver Leistungen aufzustellen. Inzwischen wird dieses Modell als das aktuellste, umfassendste und einflussreichste Modell betrachtet: Flanagan und Kaufman (2004, S. 14, zitiert nach Floyd, Bergeron, McCormack, Anderson & Hargrove-Owens, 2005) schreiben: "*Never before in the history of intelligence testing has a single theory (indeed any theory) played so prominent a role in test development and interpretation*". Tatsächlich beziehen sich die meisten der relevanten Neuentwicklungen oder Neuauflagen von Intelligenztests auf dieses Modell[11].

Das CHC-Modell ist als hierarchisches Intelligenzmodell mit ca. 16 breiten Fähigkeitskonstrukten (*broad abilities, Stratum II*) konzipiert. Diesen Faktoren sind

[11] Unter anderem*: Woodcock-Johnson III Tests of Cognitive Abilities (WJ III)*, *Stanford-Binet Intelligence Scales, Fifth Edition* (SB5), *Kaufman Assessment Battery for Children. Second Edition* (KABC-U), und, wenn auch nicht explizit, die *Wechsler Intelligence Scale for Children, Fourth Edition* (siehe Floyd et al., 2005).

ungefähr 70 Faktoren geringerer Breite untergeordnet (*narrow abilities, Stratum I*), denen schließlich einzelne kognitive Testaufgaben zugeordnet werden. Auf oberster Ebene (*Stratum III*) wird „*in the minds of some*“ (Floyd et al., 2005, S. 330) der allgemeine Intelligenzfaktor g angenommen. Anhand der 19 ausgewählten Testindikatoren der Bonner Testbatterie konnten fünf der 16 breiten Faktoren der zweiten Schicht abgebildet werden, namentlich kristalline Intelligenz (Gc), visuell-räumliche Fähigkeiten (Gv), kognitive Verarbeitungsgeschwindigkeit (Gs), Kurzzeitgedächtnis (Gsm) sowie Langzeitspeicherung und Abruf (Glr).

Es wurde vorhergesagt (Kapitel 6), dass das CHC-Modell sich am besten zur Beschreibung der Daten eignet. Die Prognose wurde aufgrund der vielfältigen empirischen Bestätigungen des Modells, dessen Vollständigkeit, der klaren Faktorendefinitionen und der belegten Unabhängigkeit der Faktorenstruktur von den eingesetzten Testverfahren getroffen. Auch ergaben sich aus einer klinischen Vorstudie Hinweise auf die Eignung dieses Modells (Lutz & Helmstaedter, 2004). Ein zentraler Unterschied der Vorstudie zur aktuellen Studie ist, dass in der Vorstudie mittels explorativer Faktorenanalyse untersucht wurde, wie viele und welche Faktoren sich aus dem Kovarianzmuster der Testbatterie, ergeben, wenn sie bei Patienten mit Epilepsie eingesetzt wird. Das CHC-Modell diente dabei als Rahmen zur Benennung der Faktoren. Wie sich gezeigt hat, waren die Faktorendefinitionen des CHC-Modells zur Faktorenbenennung sehr gut geeignet. In der hier durchgeführten konfirmatorischen Analyse steht dagegen die vergleichende Eignung mehrerer theoretischer Modelle im Mittelpunkt.

Die Frage, ob in der obersten Hierarchieebene des CHC-Modells ein Generalfaktor konzipiert werden sollte, ist noch nicht entschieden (McGrew, 1997). Der Generalfaktor drückt aus, was allen Test gemeinsam ist und wird als die einzige kognitive Leistung angesehen, die alle Tests beeinflusst (Carroll, 1993). Andere Forscher betonen dagegen die Bedeutung obliquer Gruppenfaktoren und betrachten den Generalfaktor als eine relativ bedeutungslose Zusammenfassung spezieller kognitiver Fähigkeiten (z. B. Horn, 1991). Unabhängig davon, ob ein Generalfaktor oder oblique Gruppenfaktoren angenommen werden, impliziert das CHC-Modell eine gewisse Abhängigkeit aller Tests untereinander und trägt somit der positiven Mannigfaltigkeit Rechnung.

Zur Frage nach einem Generalfaktor ist anhand der Ergebnisse dieser Studie zunächst zu sagen, dass das oblique Faktorenmodell entsprechend der CHC-Theorie eine deutlich günstigere Anpassungsgüte aufwies als das klassische Generalfaktormodell nach Spearman (1927). Dies bestätigt einmal mehr, dass die alleinige Annahme eines Generalfaktors nicht haltbar ist und stattdessen entweder

Faktorenmodelle mit obliquen Gruppenfaktoren anzunehmen sind oder hierarchische Modelle mit einem g-Faktor an der Spitze, über den die Beziehungen zwischen untergeordneten Gruppenfaktoren erklärt werden. In dieser Arbeit wurde ein hierarchisches Generalfaktormodell mit einem obliquen Faktormodell verglichen. Dazu wurde für die Fünf-Faktorenlösung des obliquen CHC-Modells untersucht, ob die Güte der Modellanpassung durch Annahme eines Generalfaktors auf oberster Ebene verbessert werden kann. Dies war nicht der Fall – zumindest nicht bei Betrachtung der reinen Indizes der Anpassungsgüte. Ein methodisches Problem besteht allerdings darin, dass sich nach Byrne (2001) bei gleicher Anzahl zu schätzender Parameter ein Modell mit korrelierten Gruppenfaktoren bezüglich der Anpassungsgüte nicht von einem Modell zweiter Ordnung mit einem Generalfaktor an der Spitze unterscheidet. Nach Byrne (2001) sollte daher die Annahme eines Generalfaktors von der zugrunde liegenden Theorie abhängig gemacht werden. Die Theorie lässt jedoch keine eindeutige Schlussfolgerung zu.

Indem fünf Gruppenfaktoren operationalisiert wurden, konnte die Mehrdimensionalität der Testbatterie klar belegt werden. Trotzdem bleibt die faktorielle Breite der Batterie weit hinter der des vollständigen CHC-Modells zurück: Das Modell enthält über die oben aufgelisteten fünf Faktoren hinaus die Fähigkeitsdimensionen *fluid intelligence/reasoning* (Gf), *general (domain-specific) knowledge* (Gkn), *auditory processing* (Ga), *decision/reaction time or speed* (Gt), *psychomotor speed* (Gps), *quantitative knowledge* (Gq), *reading/writing* (Grw), *psychomotor abilities* (Gp), *olfactory abilities* (Go), *tactile abilities* (Gh) und *kinesthetic abilities* (Gk). Somit kann die Testbatterie zwar als mit dem Modell vereinbar angesehen werden, stellt aber bei weitem keine umfassende und adäquate Repräsentation des Konstruktes Intelligenz im Sinne der CHC-Theorie dar. Gerade in einem Modell, das die Gruppenfaktoren stärker betont als den Generalfaktor, ist das problematisch. Insbesondere fehlt der fluide Intelligenzfaktor. Die fluide Intelligenz wird häufig als Kern der Intelligenz betrachtet oder teilweise ganz mit ihr gleichgesetzt (Gustafsson, 1988). In der Vorgängerstudie zu dieser Studie (Lutz & Helmstaedter, 2004) wurde ein Faktor aus den Untertests *Blockspanne* (*vorwärts*), *Labyrinthtest*, *Mosaiktest*, *semantische Wortflüssigkeit* und *motorische Sequenzierung nach Luria* als Faktor für die fluide Intelligenz interpretiert. Diese Tests werden klassisch überwiegend dem Konstrukt exekutiver Funktionen zugeordnet, so dass aufgrund der in neuerer Zeit erfolgten weitgehenden Gleichsetzung von exekutiven Funktionen und fluider Intelligenz (Gray et al., 2003) diese Einordnung gerechtfertigt ist. Im Rahmen des CHC-Modells greift diese Gleichsetzung aber zu kurz:

Dort wird fluide Intelligenz definiert als „*the use of deliberate and controlled mental operations to solve novel ‚on the spot' problems (i.e. tasks that cannot be performed automatically)*" (Internetquelle, McGrew, 2003). Diese Definition ist noch gut vereinbar mit dem Konstrukt der exekutiven Funktionen. Die nähere Bestimmung der mentalen Operationen, die der fluiden Intelligenz zugrunde liegen, entfernt sich aber vom Konstrukt der exekutiven Funktionen: „*Mental operations often include drawing inferences, concept formation, classification, generating and testing hypothesis, identifying relations, comprehending implications, problem solving, extrapolating, and transforming information. Inductive (inference of a generalized conclusion from particular instances) and deductive reasoning (the deriving of a conclusion by reasoning; specifically: inference in which the conclusion about particulars follows necessarily from general or universal premises) are generally considered the hallmark indicators of Gf*." (McGrew, 2003). In der neuropsychologischen Testbatterie sind Tests, die diese Operationen hinreichend erfassen können, nicht enthalten. Grundsätzlich sind solche Tests im Kanon neuropsychologischer Standardverfahren selten (vergleiche Hartje & Poeck, 2002), die verfügbaren Tests sollten aber benutzt werden, um die Testbatterie um dieses Konstrukt zu erweitern. Hierzu würde sich beispielsweise der in die aktuelle Version der Wechsler-Intelligenztestbatterie neu aufgenommene Matrizentest eignen (Tewes et al., 2006).

Können die fünf Faktoren der Bonner Testbatterie tatsächlich als gültige Operationalisierungen der Konstrukte des CHC-Modells angesehen werden? Diese wichtige Frage kann in die Aspekte Inhaltsvalidität, Faktorenrepräsentation, konvergente Validität und Konstruktvalidität untergliedert werden: Zunächst kann von Inhaltsvalidität der Faktoren ausgegangen werden, da die Faktoren den operationalen Definitionen des CHC-Modells entsprechen (Amelang & Zielinski, 1997), was dank der sorgfältigen Faktorendefinitionen des Modells gut festzustellen ist. In diesen Definitionen ist das Konstrukt des breiten Faktors genau beschrieben, die zugehörigen Faktoren geringerer Breite sind erklärt und es ist empirisch basiert bestimmt, welche Testverfahren starke oder schwache Indikatoren für die Faktoren beider Ebenen sind (z. B. McGrew, 1997). Die hier getroffene Zuordnung der Testindikatoren zu den breiten Faktoren nutzte diese Informationen eingehend. Die Möglichkeit, die Faktoren durch eine individuell zusammengestellte Auswahl von gut normierten Einzeltests zu konstituieren, ist ein zentrales Element im Rahmen der CHC-Theorie (*cross-battery approach*, Flanagan & McGrew, 1997).

Eine neue Studie von Keith et al. (2006) zur Faktorstruktur der neuesten Version der Wechsler-Intelligenzskala für Kinder (WISC-IV, Wechsler, 2003) zeigt, dass sich auch

zur Beschreibung des WISC-IV das CHC-Modell am besten eignet[12]. Als finales Modell wurde ein Modell mit einem Generalfaktor an der Spitze und den fünf Gruppenfaktoren kristalline Intelligenz (Gc), visuell-räumliche Fähigkeiten (Gv), kognitive Verarbeitungsgeschwindigkeit (Gs), Kurzzeitgedächtnis (Gsm) und fluide Intelligenz (Gf) bestätigt. Der Faktor Gv (visuell-räumliche Fähigkeiten) wurde durch die Einzeltests *block design, matrix reasoning, picture completion, symbol search* operationalisiert, der Faktor Gs (kognitive Verarbeitungsgeschwindigkeit) durch die Untertests *coding, symbol search, cancellation*. Auch in der Bonner Testbatterie konnten die beiden Faktoren identifiziert werden. Dem Faktor Gv wurden die Tests *Bilderergänzen, LPS-7, Labyrinth-Test, Mosaiktest* und *DCS* zugeordnet, dem Faktor Gs die Tests *TMT A und B, d2, c.I.-Test (Interferenztest)* und *Rechnerisches Denken*. Die oberflächliche Betrachtung lässt vermuten, dass der anhand der Untertests des WISC-IV operationalisierte Faktor Gv dem Faktor Gv der Bonner Testbatterie ähnlich ist; Gleiches gilt für den Faktor Gs. Zur empirischen Untersuchung dieser Ähnlichkeit wären jedoch beide Testbatterien in einer Stichprobe gemeinsam zu faktorisieren. Erst dann könnte die psychometrische Voraussetzung, die hinter dem „*cross-battery approach*“ steht, nämlich die konvergente Validität, abgeschätzt werden. Die konvergente Validität gibt an, ob unterschiedliche Methoden, die das gleiche Konstrukt erfassen sollen, dies auch tun.

Weitere wichtige Hinweise zur Beantwortung der Frage, ob die Faktoren valide operationalisiert werden, können durch Untersuchung der kriterienbezogenen Validität der Faktoren erlangt werden. Hierzu wurden im empirischen Teil dieser Arbeit verschiedene Analysen durchgeführt. Dieser Abschnitt zur Konstruktvalidität bildete den Abschluss des Ergebnisteils, da die darin enthaltenen Analysen über die Kernfrage der Invarianz hinausgehen. Trotzdem sollen diese Ergebnisse der weiteren Diskussion der Invarianzannahmen vorangestellt werden.

Die Daten wurden nicht mit dem Ziel erhoben, die Konstruktvalidität der Faktoren zu überprüfen. Daher war nur eine eingeschränkte Menge an sinnvollen Kriterien zur Bestimmung der Kriterienkorrelationen verfügbar. Die Anzahl der antiepileptischen Medikamente hatte erwartungsgemäß auf die meisten Faktoren einen negativen Einfluss (Kwan & Brodie, 2001; Lutz & Helmstaedter, 2005; Meador, Gilliam, Kanner & Pellock, 2001). Eine Ausnahme stellt der Kurzzeitgedächtnisfaktor dar (Gsm). Auf diesen Faktor hat sich die Anzahl der Medikamente nicht negativ ausgewirkt, auch die weiteren krankheitsbezogenen Variablen hatten entweder einen sehr niedrigen Einfluss (Beginn der Erkrankung, siehe unten) oder keinen Einfluss (Depressivität, Dauer der

[12] Dieses Modell eignete sich unter anderem besser als das von Wechsler implizierte hierarchische Modell. Wechslers Modell geht von einem Generalfaktor aus und auf niedrigerer Ebene von den Faktoren verbales Verständnis, Arbeitsgedächtnis, wahrnehmungsbezogenes Problemlösen und Verarbeitungsgeschwindigkeit.

Erkrankung) auf die Kurzzeitgedächtnis-Komponente. Diese Ergebnisse stehen in Einklang mit der klinischen Literatur, wonach die Kurzzeitgedächtnisleistung eine sehr pathologieresistente Leistung ist, und beispielsweise selbst bei amnestischen Syndromen kaum beeinträchtigt ist (Hartje & Sturm, 2002). Den stärksten Einfluss hatte die Anzahl der antiepileptischen Medikamente auf die Langzeitgedächtnis-Komponente. Diese Dissoziation zwischen Kurzzeit- und Langzeitgedächtnis rechtfertigt die etablierte Unterteilung der Testparameter der Testbatterie entlang der Zeitachse (Atkinson & Shiffrin, 1968). Die Schulbildung hat einen hoch signifikanten positiven Einfluss auf alle fünf Faktoren, dabei besteht die höchste Interkorrelation mit dem Faktor kristalline Intelligenz (Gc). Die starke Bildungsabhängigkeit des Faktors Gc drückt die Definition im Rahmen der CHC-Theorie aus (McGrew, 2003): „*Gc is primarily a store of verbal or language-based declarative (knowing ‚what') and procedural (knowing ‚how') knowledge acquired through the ‚investment' of other abilities during formal and informal educational and general life experiences*". Auch die Tatsache, dass alle anderen Faktoren hoch signifikant mit dem Bildungsniveau korrelieren, steht nicht in Widerspruch zu den jeweiligen Faktorendefinitionen: Die Bildungsabhängigkeit ist ein allgemeines Kennzeichen von Intelligenz (Amelang & Bartussek, 2001).

Ein überraschendes Interkorrelationsmuster findet sich beim Einfluss der depressiven Stimmungslage (erfasst mit dem Beck-Depressions-Inventar, BDI, Hautzinger et al., 1995) auf die kognitive Leistung: In der Tendenz war dieser Einfluss für alle Faktoren negativ, wenn auch nicht signifikant. Die einzige Ausnahme stellt hier der hoch signifikante negative Zusammenhang zwischen depressiver Stimmungslage und kristalliner Intelligenz dar. Die Interkorrelationen zwischen dem BDI-Wert und den Leistungskennwerten auf Einzeltestebene bestätigen diesen differenziellen Zusammenhang: Depressivität hat einen negativen Einfluss auf nur drei der 17 Variablen, nämlich auf zwei der drei Indikatoren des Faktors kristalline Intelligenz [Wortschatztest ($p < .05$) und Gemeinsamkeitenfinden ($p < .01$)] sowie auf die verbale Langzeitgedächtniskomponente des verbalen Lern- und Gedächtnistests ($p < .05$). Da Depressionen häufig mit Dysfunktionen temporaler (Quiske, Helmstaedter, Lux & Elger, 2000; Reuber, Andersen, Elger & Helmstaedter, 2004) und frontaler Schaltkreise (Rogers et al., 2004) einhergehen, ist zumindest der Zusammenhang mit der verbalen Langzeitgedächtniskomponente nicht überraschend. Aber auch eine Beeinflussung semantischer Netzwerke, die für die Leistungen im Wortschatztest und im Test Gemeinsamkeitenfinden wichtig sind, kann angenommen werden: Für die systematische Kovariation zwischen den verbal-semantischen Leistungen und Depression könnten

nach Helmstaedter, Sonntag-Dillender, Hoppe und Elger (2004) insbesondere links temporo-frontale Schaltkreise wichtig sein.

Interessant ist auch die Beziehung zwischen Beginn und Dauer der Erkrankung und den Intelligenzfaktoren. Die Krankheitsdauer hat keinerlei (negativen) Einfluss auf die kognitive Leistungsfähigkeit. Dies hat sich sowohl bei einfacher korrelativer Analyse gezeigt als auch bei pfadanalytischer Auswertung. Im Unterschied dazu hat der Krankheitsbeginn einen differenziellen Einfluss auf die Intelligenzfaktoren: Ein früher Krankheitsbeginn ist für alle Intelligenzaspekte ungünstig, besonders aber für die kristalline Intelligenzleistung. Das könnte eine indirekte Folge der verminderten Bildungschancen von Patienten mit Epilepsie sein. Ein früher Krankheitsbeginn könnte aber auch den Erwerb kristalliner Intelligenz als sekundäre Folge einer allgemeinen kognitiven Leistungsbeeinträchtigung im Rahmen der Epilepsieerkrankung beeinträchtigen (Lutz & Helmstaedter, 2004). Zusammenfassend kann festgestellt werden, dass die Datenlage zur Konstruktvalidität der fünf Intelligenzfaktoren zwar eingeschränkt ist, aus den oben geschilderten Analysen hat sich jedoch kein Widerspruch zu den postulierten Validitäten der Konstrukte ergeben.

Auch wenn gezeigt werden konnte, dass die Faktoren in beiden Stichproben identische Konstrukte erfassen, kann noch nicht davon ausgegangen werden, dass die Faktoren auch die gleichen psychometrischen Eigenschaften in beiden Gruppen haben. Um dies zu überprüfen, wurden die Hypothesen starker und strenger Invarianz getestet (Fragestellung C). Da allerdings das vollständige Modell der Kovarianz- *und* Mittelwertstruktur schon ohne Gleichheitsannahmen der Höhenlagen und Fehlerterme eine schlechte Anpassungsgüte aufwies, konnten starke und strenge metrische Invarianz nicht gezeigt werden. Somit ist am ehesten von unterschiedlichen psychometrischen Eigenschaften der Faktoren in beiden Gruppen auszugehen. Dies bedeutet, dass Probanden aus den beiden Stichproben bei gleichen latenten Fähigkeiten unterschiedliche Testwerte erhalten. In logischer Konsequenz verbieten sich Gruppenvergleiche auf Faktorenniveau.

Die Gründe sind schwer zu ermitteln. Floyd et al. (2005) nennen verschiedene mögliche Ursachen, von denen im Zusammenhang dieser Arbeit insbesondere zwei in Betracht kommen: Zum einen können unterschiedliche Messeigenschaften der Faktoren in den beiden Stichproben durch eine Interaktion zwischen Fähigkeitsniveau und Aufgabeneigenschaften bedingt sein. Dieses Problem tritt auf, wenn die Spanne der Schwierigkeiten der Items einzelner Tests das Fähigkeitsniveau beider Stichproben nicht homogen abdeckt (Decken- oder Bodeneffekte). Folglich spiegelt der Testwert

eines Probanden nicht in jedem Falle das wirkliche Fähigkeitsniveau wider. Als kritischer Test käme hier beispielsweise der *Boston Naming Test* in Betracht. Dieser Test besitzt für den hohen Leistungsbereich nur eine sehr geringe Anzahl von schwierigen Items, die eine Differenzierung der latenten Fähigkeitsunterschiede gewähren können. Im unteren Fähigkeitsbereich hat dieser Test dagegen ein sehr hohes Auflösungsvermögen. Diese Eigenschaft des Tests ist für sich genommen schon problematisch, für die untersuchten Invarianzannahmen aber umso problematischer, da sich die Stichproben deutlich im Fähigkeitsniveau unterscheiden. Durch diese Unterschiedlichkeit im allgemeinen Leistungsniveau beider Stichproben können selbst Tests, die weniger massive Decken- oder Bodeneffekte aufweisen, zu Verzerrungen der psychometrischen Eigenschaften der Faktoren führen, da die Auflösung jedes Tests am Rande seines Messbereichs ungenauer wird und gleichzeitig die beiden Gruppen an entgegen gesetzten Enden des Messbereichs liegen. Ein zweiter Punkt, der für die gruppenabhängigen Messeigenschaften der Faktoren verantwortlich sein könnte, ist die mögliche Interaktion zwischen den Probandeneigenschaften und Aufgabenanforderungen: Wie genau die Tests, die einen Faktor konstituieren, tatsächlich die konstruktrelevanten Fähigkeiten erfassen, kann von Proband zu Proband unterschiedlich sein und von krankheitsbezogenen Faktoren – und somit gruppenweise unterschiedlichen Faktoren – beeinflusst werden. Das kann etwa der Fall sein, wenn der visuelle Gedächtnistest nur in der Patientenstichprobe zum großen Teil visuo-konstruktive Fähigkeiten, die für den zugehörigen Langzeitgedächtnis-Faktor irrelevant sind, erfassen würde.

Im letzten Schritt zur Untersuchung auf Invarianz wurden die Strukturparameter Faktorvarianzen und Faktorenkovarianzen auf Gleichheit in beiden Stichproben untersucht. Dabei konnte Invarianz bestätigt werden. Indem beide Parameter als invariant gelten können, sind auch Gruppenunterschiede in den Faktorenkorrelationen als statistisch nicht bedeutsam anzusehen. Im Allgemeinen werden höhere Interkorrelationen im niedrigeren Intelligenzbereich berichtet (Mackintosh, 1998, S. 213). Aber auch Bowden et al. (2004) fanden in einer neurologischen Stichprobe und in der hirngesunden Kontrollgruppe fast gleiche Faktoreninterkorrelationen.

Die Frage nach der Gleichheit der Faktorenmittelwerte konnte nicht beantwortet werden, da sie an den Nachweis starker metrischer Invarianz gebunden ist. Für die Faktorenmittelwerte ist jedoch keine Invarianz anzunehmen, da Patienten mit pharmakoresistenter Epilepsie als Gruppe grundsätzlich niedrigere kognitive Leistungen als gesunde Normprobanden aufweisen (vergleiche Tabelle 8-A).

Zusammenfassend konnte gezeigt werden, dass die Daten beider Stichproben mit dem gleichen Intelligenzstrukturmodell beschreibbar sind (konfigurale Invarianz). Anhand eines geringfügig reduzierten Modells konnte auch gezeigt werden, dass die Höhen der Faktorladungen gleich waren (schwache metrische Invarianz). Die Hypothesen starker und strenger metrischer Invarianz konnten nicht bestätigt werden. Für die Parameter des Strukturmodells hat sich Invarianz der Faktorvarianzen und -kovarianzen gezeigt. Die Gleichheit der Faktorenmittelwerte konnte nicht überprüft werden, wurde aber auch nicht erwartet. Somit können die Daten mit dem gleichen Satz von Intelligenzfaktoren beschrieben werden und die Faktoren können auch einheitlich für beide Stichproben interpretiert werden. Auch die Faktorwerte können stichprobenübergreifend korrelativ untersucht werden. Der Vergleich der Faktorwerte zwischen den Stichproben und die Normierung der Patientenleistungen durch die Normierungsstichprobe sind dagegen bei der vorliegenden Ergebniskonstellation nicht zulässig. Normative Aussagen können sich nur auf den intraindividuellen Vergleich zwischen den Faktoren beziehen, offen bleibt aber, wie die Leistung in den verschiedenen Fähigkeitsdimensionen zu werten ist.

Auch wenn der Nachweis starker und strenger Invarianz erbracht worden wäre, wäre ein Vergleich der Faktorenleistungen zwar möglich, aber nicht zulässig, da die Normierungsstichprobe nicht vollständig repräsentativ ist. Die starke Unterschiedlichkeit der Stichproben, bedingt durch die Nichtrepräsentativität der Normierungsstichprobe, ist eine wesentliche Limitation der Studie. Zur Annäherung der allgemeinen kognitiven Niveaus beider Stichproben wurde die Patientenstichprobe im unteren und die Normierungsstichprobe im oberen Leistungsbereich gekappt. Zusätzlich wurden den Analysen alters- und geschlechtskorrigierte Werte zugrunde gelegt. Das war das günstigste Vorgehen, um die Unterschiedlichkeit zwischen den Gruppen zu verringern, ohne allzu viele Probanden oder Patienten ausschließen zu müssen. Für die Ergebnisse ist aufgrund der starken Verschiedenheit der Stichproben, die auch nach der Korrektor weiterhin bestand, anzunehmen, dass die Invarianz eher unter- als überschätzt wird. Letztlich bleiben die Ergebnisse vorläufig, bis eine Validierung anhand einer repräsentativen Normstichprobe durchgeführt wurde.

Die Patientenstichprobe war eine heterogene Gruppe von Patienten mit medikamentös schwer einstellbaren Epilepsien. Es gab keine Selektion bezüglich Ätiologie oder Lokalisation. Für die Wahl einer unselektierten Gruppe spricht folgende Überlegung: Falls die Faktorenmuster, die aus einer Testbatterie abgeleitet werden, sensitiv auf eine veränderte Zusammensetzung der Stichprobe reagieren würden, dürften folglich nur sehr homogene Stichproben mit einer Kontrollgruppe verglichen werden. Beispielsweise könnte vermutet werden, dass sich die Faktorenstrukturen bei

links- und rechts-temporalen Epilepsiepatienten aufgrund der unterschiedlichen kognitiven Profile unterscheiden. Eine pathologieabhängige Homogenisierung der Stichprobe hätte aber verschiedene Nachteile: Zum einen wird die Stichprobengröße deutlich reduziert und die Verlässlichkeit der faktoriellen Ergebnisse nimmt ab. Zum anderen wären, sollten sich tatsächlich für jede Substichprobe unterschiedliche Faktorenmuster ergeben, jeweils andere Interpretationsrichtlinien und -heuristiken anzunehmen, so dass eine Testprofilinterpretation auf Basis bedingter Wahrscheinlichkeiten notwendig würde. Der klinische Nutzen wäre somit begrenzt, zumal für die unterschiedlichen Interpretationen Ad-hoc-Hypothesen über die Gruppenzugehörigkeit vorliegen müssten – gerade diese Gruppenzugehörigkeit zu bestimmen ist jedoch eine Teilaufgabe der neuropsychologischen Testung. Die Tatsache, dass keine Selektion bezüglich Ätiologie, Pathologie oder Lokalisation stattfand, soll nicht darüber hinwegtäuschen, dass es sich bei der gewählten Stichprobe um keine repräsentative Stichprobe für die Gesamtgruppe von Patienten mit Epilepsie handelt. Die untersuchten Patienten litten unter pharmako-therapeutisch schwer einstellbaren Epilepsien. Die einer solchen Gruppe zugrunde liegenden Syndrome unterschieden sich von der Gesamtgruppe beispielsweise durch eine längere Erkrankungsdauer, durch einen höheren Schweregrad der Erkrankung oder durch einen höheren Prozentsatz von Temporallappenepilepsien.

Die zentrale Methode der Studie war die konfirmatorischen Faktorenanalyse. Diese Art der Faktorenanalyse ermöglicht es, verschiedene Modelle oder Varianten eines Modells direkt miteinander zu vergleichen. Grundsätzlich können mittels konfirmatorischer Faktorenanalyse Modelle nicht bestätigt, sondern nur widerlegt werden. French und Tait (2004) merken an, dass konfirmatorische Faktorenanalysen bestätigen können, welches Strukturmodell am besten mit den Daten vereinbar ist – dabei könnte aber ein anderes, ungetestetes Modell eine ebenso gute oder noch bessere Güte der Anpassung aufweisen. Der Index der Anpassungsgüte gibt zudem keine Auskunft über die Plausibilität oder Nützlichkeit eines Modells. So folgern die Autoren: „*model selection is ultimately a subjective process to identify a meaningful and parsimonious model*“ (S. 5). Daher ist es für die sinnvolle Anwendung konfirmatorischer Faktorenanalysen wichtig, von vorneherein eine enge und gut begründete Auswahl theoretischer Modelle zu untersuchen. Für die Interpretation dieser Studie muss beachtet werden, dass die Ergebnisse zwar darüber Auskunft geben, welche Intelligenztheorie sich zur Beschreibung der empirischen Daten eignet, aber keinen direkten Rückschluss darauf zulassen, ob die jeweilige Theorie richtig oder falsch ist.

Trotzdem sollte eine Intelligenztheorie, sofern sie einen generellen Anspruch erhebt, auch zur Beschreibung neuropsychologischer Testbatterien geeignet sein.

Es wurde gezeigt, dass die Testbatterie zwar gut mit dem CHC-Modell vereinbar ist, aber das Gesamtmodell nur sehr rudimentär operationalisiert. Für die Weiterentwicklung der Testbatterie kann nun abgeleitet werden, welche zentralen Fähigkeitskonstrukte bisher nicht erfasst werden und welche Maße zu deren Messung noch aufgenommen werden sollten. Dieses Vorgehen ist freilich nur interessant, wenn der Erweiterung der Testbatterie ein zusätzlicher klinischer Nutzen gegenüber steht. Hierzu ist allerdings die Kluft zwischen der Intelligenzforschung und klinisch-neuropsychologischer Forschung zu überwinden. Aus Sicht der Intelligenzforschung ist für die kognitive Testung zu fordern, dass diese das gegenwärtige theoretische und empirische Wissen bezüglich kognitiver Fähigkeiten berücksichtigt und alle relevanten Intelligenzdimensionen erfasst. Eine Aufstockung der aktuellen Testbatterie mit dem Ziel der adäquaten Operationalisierung des CHC-Modells wäre effizienter als beispielsweise die Erweiterung der neuropsychologischen Testbatterie durch die aktuelle Version des Wechsler-Intelligenztests, da auch dieser die aktuellen Intelligenzstrukturmodelle nicht adäquat umsetzt. Aus klinischer Perspektive ist der Nachweis des klinischen Mehrwerts der Ergebnisinterpretation auf Faktorenebene einzufordern.

Für die neuropsychologische Diagnostik bei Patienten mit Epilepsie erscheint es als ungünstig, dass das CHC-Modell zwar explizit Gedächtnisleistungen berücksichtigt, diese aber lediglich entlang der Zeitachse und nicht entlang vermeintlich klinisch relevanter inhaltlicher Kategorien einteilt. Aber auch der klinischen Forschung gelingt der Nachweis einer klaren Dissoziation zwischen verbalen und bildhaften Gedächtnisleistungen nicht vollkommen widerspruchsfrei (Bowden et al., 2004; Lee, Yip & Jones-Gotman, 2002; Wilde et al., 2003). Beispielsweise konnten Bowden et al. (2004) bei gemeinsamer Faktorisierung der Gedächtnistestbatterie nach Wechsler und des Intelligenztests nach Wechsler keinen klaren Vorteil einer verbal-nonverbalen Dichotomie der Faktoren im Vergleich zu einem Modell mit getrennten Langzeit- und Kurzzeitgedächtnisfaktoren finden. Perspektivisch könnte als Alternative eine bimodale faktorielle Struktur der Gedächtnisleistungen in Anlehnung an das Berliner Intelligenzstrukturmodell (Jäger, 1984) aufgestellt und überprüft werden. In einem solchen Modell könnte jeder Indikator sowohl auf einem materialspezifischen Faktor als auch auf einem zeitlichen Faktor laden.

Die psychometrisch orientierte Neuropsychologie kann sich nur weiterentwickeln, wenn die diagnostische Qualität und die Testgüte der Verfahren ständig kontrolliert und

verbessert werden. Dazu ist es einerseits notwendig, durch Weiterentwicklung des Testinventars die verfügbare Datenlage zu verbessern, andererseits gilt es auch, die aufgrund empirischer Daten gezogenen klinischen Schlüsse durch verbesserte Ergebnisinterpretationsschemata zu verfeinern. Hierzu ist die Wahl des theoretischen Rahmens, in dem die Ergebnisse interpretiert werden, wichtig. Von diesen drei Punkten (Weiterentwicklung von Testverfahren, Erweiterung der Interpretationsschemata der Ergebnisse und theoretische Einordnung der Ergebnisse) wurden in dieser Arbeit die zwei letzteren bearbeitet. Es konnte durch Untersuchungen der Validität der neuropsychologischen Testbatterie mit neuen analytischen Methoden und zeitgenössischen Theorien gezeigt werden, dass ein deutlicher interpretatorischer Zugewinn auch für traditionelle psychometrische Verfahren, deren Entstehung teilweise viele Jahrzehnte zurückliegt, erreicht werden kann.

10 Literaturverzeichnis

Amelang, M. & Bartussek, D. (2001). Intelligenz. In *Differentielle Psychologie und Persönlichkeitsforschung* (pp. 190-265). Stuttgart: Kohlhammer.

Amelang, M. & Zielinski, W. (1997). *Psychologische Diagnostik und Intervention*. Berlin: Springer-Verlag.

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

Anderson, M. (2005). Cortex Forum on the concept of general intelligence in neuropsychology. *Cortex, 41*, 99-100.

Arbuckle, J. L. (1997). Amos 4.0 (Version 4.01 Build 344) [Computer Software]. Chicago, IL: SmallWaters Corp.

Arbuckle, J. L. (2003). Amos 5.0 (Version 5.0 Build 5134) [Computer software]. Chicago, IL: SmallWaters Corp.

Arbuckle, J. L. & Wothke, W. (1999). *Amos 4.0 User's Guide*. Chicago, IL: SmallWaters Corp.

Ardila, A. (1999). A neuropsychological approach to intelligence. *Neuropsychology Review, 9*(3), 117-136.

Atkinson, R. C. & Shiffrin, R. M. (1968). Human memory: A proposed system and its control process. In K. W. Spence & S. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 85-195). New York: Academic Press.

Bachetzky, N. & Jahn, T. (2005). Faktorielle Validität des deutschsprachigen CVLT in der neuropsychologischen Diagnostik von Gedächtnisstörungen. *Zeitschrift für Neuropsychologie, 16*(2), 63-75.

Baer, J. C., Prince, J. D. & Velez, J. (2004). Fusion or familialism: A construct problem in studies of Mexican American adolescents. *Hispanic Journal of Behavioral Sciences, 26*(3), 263-273.

Banos, J. H., Roth, D. L., Palmer, C., Morawetz, R., Knowlton, R., Faught, E., Kuzniecky, R. I., Bilir, E. & Martin, R. C. (2004). Confirmatory factor analysis of the California Verbal Learning Test in patients with epilepsy: Relationship to clinical and neuropathological markers of temporal lobe epilepsy. *Neuropsychology, 18*(1), 60-68.

Beauducel, A. & Kersting, M. (2002). Fluid and crystallized intelligence and the Berlin model of intelligence structure (BIS). *European Journal of Psychological Assessment, 18*(2), 97-112.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality.* Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Boring, E. G. (1923). Intelligence as the test tests it. *The New Republic, 6*, 35–37.

Bortz, J. (1993). *Statistik für Sozialwissenschaftler*. Berlin: Springer-Verlag.

Bowden, S. C., Cook, M. J., Bardenhagen, F. J., Shores, E. A. & Carstairs, J. R. (2004). Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence, 32*(4), 363-389.

Bowden, S. C., Ritter, A. J., Carstairs, J. R., Shores, E. A., Pead, J., Greeley, J. D., Whelan, G., Long, C. M. & Clifford, C. C. (2001). Factorial invariance for combined Wechsler Adult Intelligence Scale-Revised and Wechsler Memory Scale-Revised scores in a sample of clients with alcohol dependency. *The Clinical Neuropsychologist, 15*(1), 69-80.

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

Bucik, V. & Neubauer, A. C. (1996). Bimodality in the Berlin model of intelligence structure (BIS): A replication study. *Personality and individual Differences, 21*(6), 987-1005.

Burt, C. (1909). Experimental tests of general intelligence. *British Journal of Psychology, 3*, 94-177.

Burton, D. B., Ryan, J. J., Axelrod, B. N. & Schellenberger, T. (2002). A confirmatory factor analysis of the WAIS-III in a clinical sample with crossvalidation in the standardization sample. *Archives of Clinical Neuropsychology, 17*(4), 371-387.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement ent invariance. *Psychological Bulletin, 105*, 456-466.

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

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

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* Boston: Houghton Mifflin.

Chapuis, F. (1992). *Labyrinthtest*. Göttingen: Hogrefe.

Cheung, G. W. & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27.

Cheung, G. W. & Rensvold, R. B. (2000). Testing measurement invariance using critical values of fit indices: a Monte Carlo study. *Research Methods Forum.* Retrieved 06-02-05, from http://www.aom.pace.edu/rmd/cheung_files/cheung.htm

Daniel, M. H. (2000). Interpretation of intelligence test scores. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 477-491). Cambridge: Cambridge University Press.

Das, J. P., Naglieri, J. A. & Kirby, J. R. (1994). *Assessment of cognitive processes: The PASS theory of intelligence.* Needham Heights, MA: Allyn & Bacon.

Delis, D. C., Jacobson, M., Bondi, M. W., Hamilton, J. M. & Salmon, D. P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: lessons from memory assessment. *Journal of the International Neuropsychological Society, 9*(6), 936-946.

Delis, D. C., Kramer, J. H., Kaplan, E. & Ober, B. A. (1987). *The California Verbal Learning Test: Research edition.* San Antonio: The Psychological Corporation.

Dodrill, C. B. (1999). Myths of neuropsychology: further considerations. *The Clinical Neuropsychologist, 13*(4), 562-572.

Duncan, J. (2003). Intelligence tests predict brain response to demanding task events. *Nature Neuroscience, 3*(3), 207-208.

Finney, S. & Davis, S. (2003). *Examining the invariance of the achievement goal questionnaire across gender.* Paper presented at the American educational research association AERA, Chicago, IL.

Flanagan, D. P. & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 314-325). New York: The Guilford Press.

Flanagan, D. P. & Ortiz, S. O. (2006). CHC Cross-Battery online. Official site of the CHC Cross-Battery Approach. Retrieved 03-11-06, from http://facpub.stjohns.edu/~ortizs/cross-battery/index.html

Floyd, R. G., Bergeron, R., McCormack, A. C., Anderson, J. L. & Hargrove-Owens, G. L. (2005). Are Cattell-Horn-Carroll broad ability composite scores exchangeable across batteries? *School Psychology Review, 34*(5), 329-357.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171-191.

French, D. J. & Tait, R. J. (2004). Measurement invariance in the General Health Questionnaire-12 in young Australian adolescents. *European Child & Adolescent Psychiatry, 13*(1), 1-7.

Fujii, D. E., Lloyd, H. A. & Miyamoto, K. (2000). The salience of visuospatial and organizational skills in reproducing the Rey-Osterrieth Complex Figure in subjects with high and low IQs. *The Clinical Neuropsychologist, 4*(4), 551-554.

Gardner, H. (1993). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gleissner, U. & Elger, C. E. (2001). The hippocampal contribution to verbal fluency in patients with temporal lobe epilepsy. *Cortex, 37*(1), 55-63.

Gleissner, U., Helmstaedter, C. & Elger, C. E. (1998). Right hippocampal contribution to visual memory: a presurgical and postsurgical study in patients with temporal lobe epilepsy. *Journal of Neurology, Neurosurgery, and Psychiatry, 65*, 665–669.

Gould, J. S. (1997). *The mismeasure of man*. London: Penguin Books.

Gray, J. R., Chabris, C. F. & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*(3), 316-322.

Gresch, M. (2005). *Normierung einer neuropsychologischen Testbatterie für die prächirurgische Epilepsiediagnostik bei Erwachsenen.* Unveröffentlichte Diplomarbeit, Universität Bonn.

Gustafsson, J.-E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence.* (Vol. 4, pp. 35-71). Hillsdale, NJ: Erlbaum.

Härting, C., Markowitsch, H. J., Neufeld, H., Calabrese, P. & Deisinger, K. (Eds.). (2000). *WMS-R - Wechsler Gedächtnis Test - Revidierte Fassung. Deutsche Adaptation der revidierten Fassung der Wechsler-Memory-Scale.* Göttingen: Hogrefe.

Hartje, W. (2002). Funktionelle Asymmetrie der Großhirnhemisphären. In W. Hartje & K. Poeck (Eds.), *Klinische Neuropsychologie* (pp. 67-92). Stuttgart: Thieme.

Hartje, W. & Poeck, K. (Eds.). (2002). *Klinische Neuropsychologie*. Stuttgart: Thieme.

Hartje, W. & Sturm, W. (2002). Amnesie. In W. Hartje & K. Poeck (Eds.), *Klinische Neuropsychologie* (pp. 248-295). Stuttgart: Thieme.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling.* Cary, NC: SAS Institute.

Hautzinger, M., Worall, H. & Keller, F. (1995). *BDI. Beck-Depressions-Inventar von A.T. Beck, Dt. Bearbeitung.* Göttingen: Hogrefe.

Heijden, P. v. d. & Donders, J. (2003). A confirmatory factor analysis of the WAIS—III in patients with traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 25* (1), 59-65.

Helmstaedter, C. (2000). Neuropsychologie bei Epilepsie. In W. Sturm, M. Herrmann & C.-W. Wallesch (Eds.), *Lehrbuch der Klinischen Neuropsychologie* (pp. 571-580). Lisse, NL: Swets & Zeitlinger.

Helmstaedter, C., Brosch, T., Kurthen, M. & Elger, C. E. (2004). The impact of sex and language dominance on material-specific memory before and after left temporal lobe surgery. *Brain, 127*, 1518-1525.

Helmstaedter, C., Gleissner, U., Zentner, J. & Elger, C. E. (1998). Neuropsychological consequences of epilepsy surgery in frontal lobe epilepsy. *Neuropsychologia, 36*(7), 681-689.

Helmstaedter, C., Grunwald, T., Lehnertz, K., Gleissner, U. & Elger, C. E. (1997). Differential involvement of left temporolateral and temporomesial structures in verbal declarative learning and memory: Evidence from temporal lobe epilepsy. *Brain and Cognition, 35*, 110–131.

Helmstaedter, C., Kemper, B. & Elger, C. E. (1996). Neuropsychological aspects of frontal lobe epilepsy. *Neurpsychologia, 34*(5), 399-406.

Helmstaedter, C. & Kurthen, M. (2001). Memory and epilepsy: characteristics, course, and influence of drugs and surgery. *Current Opinion in Neurology, 14*(2), 211-216.

Helmstaedter, C., Kurthen, M., Lux, S., Reuber, M. & Elger, C. E. (2003). Chronic epilepsy and cognition: A longitudinal study in temporal lobe epilepsy. *Annals of Neurology, 54*, 425–432.

Helmstaedter, C., Lendt, M. & Lux, S. (2001). *VLMT Verbaler Lern- und Merkfähigkeitstest*. Göttingen: Beltz Test.

Helmstaedter, C., Pohl, C., Hufnagel, A. & Elger, C. E. (1991). Visual learning deficits in nonresected patients with right temporal lobe epilepsy. *Cortex, 27*(4), 547-555.

Helmstaedter, C., Sonntag-Dillender, M., Hoppe, C. & Elger, C. E. (2004). Depressed mood and memory impairment in temporal lobe epilepsy as a function of focus lateralization and localization. *Epilepsy & Behavior, 5*, 696-701.

Hoelter, J. W. (1983). The analysis of covariance-structures - goodness-of-fit indexes. *Sociological Methods & Research, 11*, 325-344.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder & R. W. Woodcock (Eds.), *WJ-R technical manual*. Chicago: Riverside.

Horn, W. (1983). *LPS. Leistungsprüfsystem.* Göttingen: Hogrefe.

Institute for Applied Psychometrics. (2002). Supporting Evidence for CHC-Theory. Retrieved 02-19-06, from http://www.iapsych.com/chcevidence.htm

Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica, 28*, 195-225.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau, 35*(21-35).

Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997). *Der Berliner Intelligenzstruktur-Test (BIS-Test; Form 4)*. Göttingen: Hogrefe.

Jöreskog, K. G. & Sörbom, D. (1993). *LISREL8: The SIMPLIS command language.* Chicago: Scientific Software.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In E. Dennis, M. Posner, D. Stein & K. Thomson (Eds.), *Clinical neuropsychology and brain functioning: Research, measurement, and practice.* (pp. 129-166). Washington DC: American Psychological Association.

Kaplan, E., Goodglass, H. & Weintraub, S. (1976). *Boston Naming Test.* Boston: Aphasia Research Center, Boston University.

Karnath, H.-O. & Sturm, W. (2002). Störungen von Planungs- und Kontrollfunktion. In W. Hartje & K. Poeck (Eds.), *Klinische Neuropsychologie* (pp. 393-411). Stuttgart: Thieme.

Kaufman, A. S. (2000). Tests of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 445-476). Cambridge: Cambridge University Press.

Kaufman, A. S., Kaufman, N. L., Melchers, P. & Preuß, U. (2001). *K-ABC Kaufman Assessment Battery for Children, Deutsche Version*. Göttingen: Hogrefe.

Keith, T. Z. (1997). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 373-402). New York: The Guilford Press.

Keith, T. Z., Fine, J. G., Taub, G. E., Reynolds, M. R. & Kranzler, J. H. (2006). Higher order, multisample, confirmatory factor analysis of the Wechsler Intelligence Scale for Children—fourth edition: What does it measure? *School Psychology Review, 35*(1), 108-127.

Kelloway, E. A. (1998). *Using LISREL for structural equation modeling. A researcher's guide.* Thousand Oaks, CA: Sage Publications.

Kline, R. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

Kwan, P. & Brodie, M. J. (2001). Neuropsychological effects of epilepsy and antiepileptic drugs. *Lancet, 357*, 216–222.

Larrabee, G. J. (2003). Lessons on measuring construct validity: a commentary on Delis, Jacobson, Bondi, Hamilton, and Salmon. *Journal of the International Neuropsychological Society, 9*(6), 947-953.

Lee, T. M. C., Yip, J. T. H. & Jones-Gotman, M. (2002). Memory deficits after resection from left or right anterior temporal lobe in humans: A meta-analytic review. *Epilepsia, 43*(3), 283-291.

Lehrl, S. (1999). *MWT-B. Mehrfachwahl-Wortschatz-Intelligenztest.* Göttingen: Hogrefe.

Lehrl, S. & Fischer, B. (1997). *c.I.-Test. Kurztest für cerebrale Insuffizienz.* Göttingen: Hogrefe.

Levine, D. W., Kaplan, R. M., Kripke, D. F., Bowen, D. J., Naughton, M. J. & Shumaker, S. A. (2003). Factor structure and measurement invariance of the Women's Health Initiative Insomnia Rating Scale. *Psychological Assessment, 15*(2), 123-136.

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

Lubke, G. H., Dolan, C. V., Kelderman, H. & Mellenbergh, G. J. (2003a). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*(6), 543-566.

Lubke, G. H., Dolan, C. V., Kelderman, H. & Mellenbergh, G. J. (2003b). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56*(2), 231-248.

Luria, A. R. (1973). *The Working Brain*. London: Penguin Press.

Lutz, M. T. & Helmstaedter, C. (2004). Effects of age at onset and duration of epilepsy on cognition in the framework of Cattell's theory of fluid and crystallized abilities. *Epilepsia, 45*(Supp 7), 346.

Lutz, M. T. & Helmstaedter, C. (2005). EpiTrack: Tracking cognitive side effects of medication on attention and executive functions in patients with epilepsy. *Epilepsy & Behavior, 7*(4), 708-714.

MacCallum, R. C. (2003). Studying Measurement Invariance Using Confirmatory Factor Analysis. Retrieved 02-26-06, from http://www.unc.edu/~rcm/psy236/measinv.pdf

Mackintosh, N. J. (1998). *IQ and Human Intelligence*. Oxford: Oxford University Press.

McDennott, P. A., Fanttizzo, J. W. & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment., 8*, 289-302.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 151-180). New York: The Guilford Press.

McGrew, K. S. (2003). Cattell-Horn-Carroll (CHC) Definition Project. Retrieved 11-28-03, from http://www.iapsych.com/chcdef.htm

McGrew, K. S. (2005). The Cattell-Hom-Carroll theory of cognitive abilities: Past, present, and future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues (2nd ed.).* New York: Guilford.

Meador, K. J., Gilliam, F. G., Kanner, A. M. & Pellock, J. M. (2001). Cognitive and behavioural effects of antiepileptic drugs. *Epilepsy & Behavior, 2 (Suppl.)*, 1–17.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*(4), 525-543.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex ‘‘frontal lobe’’ tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100.

Müller, H., Hasse-Sander, I., Horn, R., Helmstaedter, C. & Elger, C. E. (1997). Rey Auditory-Verbal Learning Test: structure of a modified German version. *Journal of Clinical Psychology, 53*(7), 663-671.

Naglieri, J. A. (1997). Planning, attention, simultaneous, and successive theory and the cognitive assessment system: A new theory-based measure of intelligence. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 247-267). New York: The Guilford Press.

Naglieri, J. A. & Das, J. P. (1996). *Das Naglieri cognitive assessment system.* Chicago: Riverside.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J. & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Nettelnstroth, W. (2003). *Intelligenz im Rahmen der beruflichen Tätigkeit - Zum Einfluss von Intelligenzfacetten, Personenmerkmalen und Organisationsstrukturen -.* Unpublished Dissertation, Freie Universität, Berlin.

Orgass, B., De Renzi, E. & Vignolo, L. A. (1982). *Token Test*. Göttingen: Hogrefe.

Ortiz, S. O. & Flanagan, D. P. (2002). Cross-Battery assessment revisited: Some cautions concerning "some cautions" (Part I). *Newspaper of the National Association of School Psychologists, 30*(7), 32-34.

Osterreith, R. (1944). Le test de copie d'une figure complex: Contribution a l'étude de la perception et de la memoire. *Archives de Psychologie, 30*(206-356).

Pubmed. (2006). U.S. National Library of Medicine [data base].

Quiske, A., Helmstaedter, C., Lux, S. & Elger, C. E. (2000). Depression in patients with temporal lobe epilepsy is related to mesial temporal sclerosis. *Epilepsy Research, 39*, 121-125.

Raven, J. C. (1998). *SPM - Standard Progressive Matrices*. Göttingen: Hogrefe.

Raykov, T. & Marcoulides, G. A. (2000). *A first course in Structural Equation Modeling.* Mahwah: Lawrence Erlbaum Associates, Inc.

Reitan, R. M. (1979). *Trail Making Test (TMT)*. Göttingen: Hogrefe.

Rensvold, R. B. (2002). Metric equivalence / invariance across multiple groups: Comparing apples with apples with apples etc. [Conference paper], *Academy of Management*. Denver, CO.

Reuber, M., Andersen, B., Elger, C. E. & Helmstaedter, C. (2004). Depression and anxiety before and after temporal lobe epilepsy surgery. *Seizure, 13*, 129–135.

Rietz, C. (1996). *Faktorielle Invarianz - Die inferenzstatistische Absicherung von Faktorstrukturvergleichen.* Bonn: PACE.

Rogers, M. A., Kasai, K., Koji, M., Fukuda, R., Iwanami, A., Nakagome, K., Fukuda, M. & Kato, N. (2004). Executive and prefrontal dysfunction in unipolar depression: a review of neuropsychological and imaging evidence. *Neuroscience Research, 50*(1), 1-11.

Russ, M. O. (2003). Klinische Diagnostik mit dem Frankfurter Neuropsychologischen Testprofil (FNTP). Retrieved 10-06-05, from http://www.kgu.de/znn/neurologie

Satzger, W., Fessmann, H. & Engel, R. R. (2002). Liefern HAWIE-R, WST und MWT-B vergleichbare IQ-Werte? *Zeitschrift für Differentielle und Diagnostische Psychologie, 23*(2), 159-170.

Schmid, J. & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*(1), 53-61.

Schmidt, F. L. & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence, 27*(3), 183-198.

Schmidt, J. U. (1984). Simultane Überprüfung der Zweimodalität im Berliner Intelligenzstrukturmodell. *Diagnostica, 30* (2), 93-103.

Schmidt, J. U. (1993). Thurstones primary mental abilities und das Berliner Intelligenzstrukturmodell – Eine empirische Gegenüberstellung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 87-100.

Schmitt, M., Maes, J. & Seiler, U. (1999). Theoretische Überlegungen und empirische Befunde zur Meßäquivalenz und strukturellen Invarianz von Indikatoren der seelischen Gesundheit bei Ost- und Westdeutschen. *Magdeburger Arbeiten zur Psychologie, 1*(1).

Schonemann, P. H. (1997). Famous artefacts: Spearman's hypothesis. *Cahiers de Psychologie Cognitive - Current Psychology of Cognition, 16*(6), 665-694.

Schonemann, P. H. (2001). Better never than late: Peer review and the preservation of prejudice. *Ethical human sciences and services, 3*(1), 7-21.

Schwarzkopf-Streit, C. (2000). *Die Schätzung der Gesamtintelligenz aus Testkurzformen im Intelligenzkonzept nach Wechsler.* Unpublished Dissertation, Medizinische Hochschule, Hannover.

Spearman, C. (1927). *The abilities of man*. London: Macmillan.

Spreen, O. & Strauss, E. (1998). *A compendium of neuropsychological tests. Administration, norms, and commentary.* New York: Oxford University Press.

SPSS. (2003). SPSS 12.0G for Windows (Version 12.0.1, 11 Nov 2003) [Computer software]. Chicago, IL: SPSS Inc.

Steenkamp, J. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research., 25*, 78-90.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (Ed.). (2000). *Handbook of intelligence*. Cambridge: Cambridge University Press.

Stöcker, C. (2005). Intelligenzmessung - Rückkehr der Rassenlehre*. Spiegel online*.

Sturm, W. (2000). Aufgaben und Strategien neuropsychologischer Diagnostik. In W. Sturm, M. Herrmann & C.-W. Wallesch (Eds.), *Lehrbuch der Klinischen Neuropsychologie* (pp. 265-276). Lisse, NL: Swets & Zeitlinger.

Sturm, W., Herrmann, M. & Wallesch, C.-W. (Eds.). (2000). *Lehrbuch der Klinischen Neuropsychologie*. Lisse, NL: Swets & Zeitlinger.

Taub, G. E., McGrew, K. S. & Witta, E. L. (2004). A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale-Third Edition. *Psychological Assessment, 16*(1), 85-89.

Tewes, U. (1991). *Hamburg-Wechsler-Intelligenztest für Erwachsene – Revision 1991*. Bern: Huber.

Tewes, U., Neubauer, A. & Aster, M. v. (2006). *Wechsler Intelligenztest für Erwachsene (WIE). Deutschsprachige Bearbeitung und Adaptation des WAIS III von David Wechsler.* Frankfurt: Harcourt Test Services.

Thompson, P. J. & Duncan, J. S. (2005). Cognitive decline in severe intractable epilepsy. *Epilepsia, 46*(11), 1780-1787.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Tiffin, J. (1968). *Perdue Pegboard: Examiner Manual.* Chicago: Science Research Associates.

Vandenberg, R. L. & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4-70.

Vernon, P. E. (1950). *The structure of human abilities*. London: Methuen.

Vernon, P. E. (1971). *The structure of human abilities.* London: Methuen.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale: Revised*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale - Revised*. New York: Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale - Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale - Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children - Fourth Edition: Technical and interpretive manual.* San Anionio, TX: Psychological Corporation.

Weidlich, S., Lamberti, G. & Hartje, W. (2001). *DCS. Diagnosticum für Cerebralschädigungen. Ein visueller Lern- und Gedächtnistest nach F. Hillers.* Göttingen: Hogrefe.

Weiss, L. & Price, L. (2001). An update on the factor structure of the Wechsler Memory Scale—third edition. Retrieved 04-03-06, from http://harcourtassessment.com/hai/Images/resource/library/Wms/wmsfactor.html

Widaman, K. F. & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J.

Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research.* (pp. 281-324). Washington, DC: APA.

Wilde, N. J., Strauss, E., Chelune, G. J., Hermann, B. P., Hunter, M., Loring, D. W., Martin, R. C. & Sherman, E. M. (2003). Confirmatory factor analysis of the WMS-III in patients with temporal lobe epilepsy. *Psychological Assessment, 15*(1), 56-63.

Woodcock, R. W. & Johnson, M. B. (1989). *Woodcock-Johnson - revised, Tests of cognitive ability: Standard and supplemental batteries.* Chicago: Riverside.

Eidesstattliche Erklärung

Hiermit erkläre ich, Martin Lutz, Dipl.-Psych., geboren am 18. November 1973 in Baden-Baden, an Eides statt, dass ich die hier vorgelegte wissenschaftliche Arbeit eigenständig und ausschließlich unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel erstellt habe.

_____

Dresden, im Juni 2006