Validierung von Fähigkeitstests auf Grundlage stochastischer Testmodelle

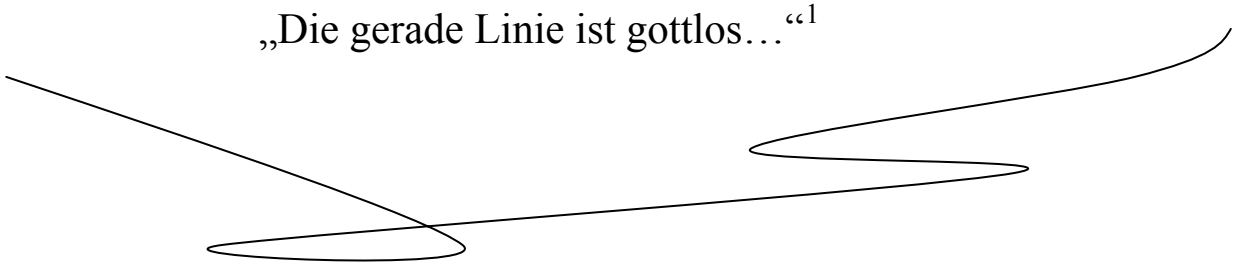
Dissertation zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften (Dr.rer.nat.)
im Fachbereich Psychologie der Universität Konstanz

gewissenhaft angefertigt, schriftlich vorgelegt und
menschlich verantwortet von
Franz Kindermann

Datum der mündlichen Prüfung: 20.6.2007
Referenten:
Prof. Dr. Wilhelm Kempf & Prof. Dr. Peter Steck

Konstanzer Online-Publikations-System (KOPS)
URL: http://www.ub.uni-konstanz.de/kops/volltexte/2008/4525/
URN: http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-45253

„Die gerade Linie ist gottlos…“[1]



[1] Friedensreich Hundertwasser, *Verschimmelungs – Manifest*, 1958 und Zusätze 1959 & 1964

Jenen Einzelnen

und

allen anderen Märchengestalten

An der vagen Grenze naturwissenschaftlich möglicher Erkenntnis
des vernunftbegabten Lebewesens Mensch
gilt mein besonderer Dank ihnen allen und jedem im Einzelnen:

Prof. Dr. Wilhelm Kempf,
Prof. Dr. Peter Steck,
Dr. Willi Nagl,
Hans-Joachim Niermeyer,
Jan Rasmus Böhnke,
Christoph Stadler
und Frau Ulrike Buttgereit

Diese Arbeit wäre unmöglich gewesen
ohne die freundliche Unterstützung sämtlicher 615 Testteilnehmer,
insbesondere der Patienten, Mitarbeiter und Verantwortlichen folgender
Kliniken:

Herz-Kreislauf-Klinik, Radolfzell
Lurija Institut und Kliniken Schmieder, Konstanz
Psychiatrische Klinik, Münsterlingen, Schweiz
Zentrum für Psychiatrie Bad Schussenried
Zentrum für Psychiatrie Reichenau
Zentrum für Psychiatrie Weißenau

Ergebenst
Franz Kindermann
Konstanz am Bodensee, im Dezember 2006

INHALTSVERZEICHNIS:

1 EINLEITUNG 17

1.1 Prolog 17

1.2 Thema 17

1.3 Psychologie, Diagnostik und Testtheorie 20

1.4 Die Realisierung einer psychologischen Messung als Score in der klassischen Testtheorie 22

1.5 Aufbruch zu einer naturwissenschaftlichen Psychologie? 23

1.6 Die Realisierung einer psychologischen Messung als Itemantwort in der probabilistischen Testtheorie 25

1.7 Theoretische Relevanz stochastischer Testmodelle 27

1.8 Praktische Relevanz stochastischer Testmodelle 29

1.8.1 Adaptives Testen 30

1.8.2 „person-fit“ 30

1.8.3 „item-bias“ und „Differential Item Functioning“ 31

1.8.4 Identifizierung systematischer Unterschiede im Antwortverhalten 32

1.9 Mischverteilungsmodelle als Weiterentwicklungen stochastischer Testmodelle 35

1.10 Identifizierung unterschiedlicher Bearbeitungsstrategien mit Mischverteilungsmodellen 39

1.11 Kritische Anmerkungen zu Untersuchungen der Vorhersagevalidität von Testverfahren 42

1.12 Allgemeine Bemerkungen zu den verwendeten Tests 44

1.12.1 Raventest 44

1.12.2 Bentontest 45

1.12.3 Cattelltest 46

2 FRAGESTELLUNG 49

2.1 Fragestellung 1: Werden von stochastischen Testmodellen bei den einzelnen Testverfahren unterschiedliche Arten der Testbearbeitung (latente Klassen) identifiziert? ... 49
2.1.1 Graphischer Modelltest ... 50
2.1.2 Andersen-Test ... 50
2.1.3 Prüfung der Modellgeltung mittels Informationstheoretischer Maße ... 51
2.2 Fragestellung 2: Wie können identifizierte Arten der Testbearbeitung (latente Klassen) bei den drei Tests interpretiert werden? ... 51
2.3 Fragestellung 3: Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen der einzelnen Tests? ... 52
2.4 Fragestellung 4: Können die identifizierten Arten der Testbearbeitung die Vorhersagevalidität der drei Tests verbessern? ... 52
2.5 Fragestellung 5: Ergibt sich durch eine Kombination der Einzeltests zu einer Testbatterie eine Verbesserung der Vorhersagevalidität? ... 53
2.6 Fragestellung 6: Können Lösungsmuster oder Typen der Testbatterie-bearbeitung identifiziert und differentialdiagnostisch genutzt werden? ... 53
3 THEORIE PROBABILISTISCHER TESTANALYSEN ... 55
3.1 Das dichotome logistische Modell nach RASCH ... 55
3.1.1 Parameterschätzung ... 60
3.1.2 Prüfung der Modellgeltung ... 63
3.1.3 Modellvergleiche ... 64
3.1.4 Statistische Information und das Prinzip der adaptiven Testung ... 65
3.2 Probabilistische Testmodelle ... 68
3.3 Das Saturierte Modell ... 71
4 METHODEN ... 73
4.1 Beschreibung der Stichproben ... 73
4.1.1 „Gesamtstichprobe“ (N=615) ... 73
4.1.2 Teilstichprobe „GPN parallel“ (N=210) ... 76
4.1.3 Teilstichprobe „Parallelisierte Stichprobe“ (N=468) ... 76
4.1.4 Teilstichprobe „Stichprobe `parallel-cerebral`“ (N=269) ... 77

4.1.5 Teilstichprobe „Testbatterie“ (N=590) ... 77

4.2 Testmaterial ... 77

4.2.1 Begründung der Testauswahl ... 78

4.3 Testdurchführung ... 85

4.4 Auswertung der Testergebnisse ... 86

4.5 Statistische Auswertung ... 87

4.6 Mathematisches Vorgehen ... 87

4.6.1 Entropiereduktion in % ... 89

4.6.2 Adjustierte Entropiereduktion in % ... 89

4.6.3 Fläche unter der Receiver Operating Characteristic Kurve (ROC-Kurve) ... 90

4.6.4 Wald-Effekt-Tests ... 93

4.7 Statistische Evaluation der Datengrundlage ... 94

4.7.1 Vergleich der Testkennwerte ... 94

4.7.2 Scorevergleiche und Testreihenfolgen ... 96

4.7.3 Scorevergleiche der Diagnosegruppen in der Altersgruppe 0 (≤ 39. Lebensjahr) ... 97

4.7.4 Statistische Absicherung des Validitätskriteriums ... 98

4.7.4.1 Testscoreunterschiede der drei Diagnosegruppen G, P und N ... 99

4.7.4.2 Testtrennschärfen der Diagnosegruppen ... 102

4.7.4.3 Konklusion und Forschungskonsequenz ... 103

4.7.5 Korrelationen in der Gesamtstichprobe ... 104

4.7.6 Ergebnisse der Hauptkomponentenanalyse ... 106

5 ERGEBNISSE ... 109

5.1 Raventest ... 109

5.1.1 Prüfung der Homogenität ... 109

5.1.1.1 Überprüfung des Raschmodells mit graphischen Modelltestungen ... 109

5.1.1.2 Überprüfung des Raschmodells mit der „Andersen-Testmethode“ ... 111

5.1.1.2.1 Modelltestung anhand des externen Teilungskriteriums „diagnostische Zugehörigkeit“ .. 111

5.1.1.2.2 Modelltestung anhand des internen Teilungskriteriums „Testleistung“ ... 112

5.1.1.3 Geltungsprüfung unterschiedlicher probabilistischer Modelle mittels informationstheoretischer Maße und Likelihoodquotiententests ... 112

5.1.2 Interpretation latenter Klassen ... 114

5.1.2.1 Summenscoreverteilung der 3-Klassen-Lösung des Mixed-Rasch-Modells ... 114

5.1.2.2 Charakterisierung der 3-Klassen-Lösung des Mixed-Rasch-Modells ... 114

5.1.2.3 Die Profile der Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells ... 117
5.1.3 Diagnostische Relevanz der latenten Klassen ... 118
5.1.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen ... 118
5.1.3.2 Scorevergleiche von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung ... 120
5.1.3.3 Vorhersagevalidität ... 122
5.1.3.3.1 Prädiktion „klinische Auffälligkeit“ ... 122
5.1.3.3.2 Prädiktion „cerebrale Beteiligung“ ... 124

5.2 Bentontest ... 127
5.2.1 Prüfung der Homogenität ... 127
5.2.1.1 Überprüfung des Raschmodells mit graphischen Modelltestungen ... 127
5.2.1.2 Überprüfung des Raschmodells mit der „Andersen-Testmethode“ ... 129
5.2.1.2.1 Modelltestung anhand des externen Teilungskriteriums „diagnostische Zugehörigkeit“ .. 129
5.2.1.2.2 Modelltestung anhand des internen Teilungskriteriums „Testleistung“ ... 129
5.2.1.3 Geltungsprüfung unterschiedlicher probabilistischer Modelle mittels informationstheoretischer Maße und Likelihoodquotiententests ... 130
5.2.2 Interpretation latenter Klassen ... 131
5.2.2.1 Summenscoreverteilung der 3-Klassen-Lösung des Mixed-Rasch-Modells ... 132
5.2.2.2 Charakterisierung der 3-Klassen-Lösung des Mixed-Rasch-Modells ... 132
5.2.2.3 Die Profile der Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells ... 136
5.2.3 Diagnostische Relevanz der latenten Klassen ... 137
5.2.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen ... 137
5.2.3.2 Scorevergleiche von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung ... 139
5.2.3.3 Vorhersagevalidität ... 141
5.2.3.3.1 Prädiktion „klinische Auffälligkeit“ ... 141
5.2.3.3.2 Prädiktion „cerebrale Beteiligung“ ... 143
5.2.3.3.3 Prädiktion „klinische Auffälligkeit“ und Fehlermaß BTF ... 145
5.2.3.3.4 Prädiktion „cerebrale Beteiligung“ und Fehlermaß BTF ... 147

5.3 Cattelltest ... 149
5.3.1 Identifikation latenter Klassen ... 149
5.3.1.1 Geltungsprüfung unterschiedlicher probabilistischer Modelle mittels informationstheoretischer Maße und Likelihoodquotiententests ... 149
5.3.2 Interpretation latenter Klassen ... 151
5.3.2.1 Summenscoreverteilung der 4-Klassen-Lösung der Latenten-Klassen-Analyse ... 151
5.3.2.2 Charakterisierung der 4-Klassen-Lösung der Latenten-Klassen-Analyse ... 151
5.3.2.3 Die Profile der Itemparameter der 4-Klassen-Lösung der Latenten-Klassen-Analyse ... 155
5.3.2.4 Homogenität der Items ... 156

5.3.3 Diagnostische Relevanz der latenten Klassen ... **158**
5.3.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen ... ***158***
5.3.3.2 Scorevergleiche von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung ... ***160***
5.3.3.3 Vorhersagevalidität ... ***162***
5.3.3.3.1 Prädiktion „klinische Auffälligkeit“ ... 162
5.3.3.3.2 Prädiktion „cerebrale Beteiligung“ ... 164

5.4 Raven-, Benton- und Cattelltest gemeinsam ... **167**
5.4.1 Prädiktion „klinische Auffälligkeit“ ... **167**
5.4.2 Prädiktion „cerebrale Beteiligung“ ... **169**

5.5 Zusammenfassung der Kennwerte zur Prädiktion „klinische Auffälligkeit“ und „cerebrale Beteiligung“ ... **173**

5.6 Latente-Klassen-Analyse der gesamten Testbatterie ... **177**
5.6.1 Klassischer Auswertungsmodus der Testbatterie ... **178**
5.6.1.1 Identifikation latenter Klassen ... ***178***
5.6.1.1.1 Inferenzstatistische Prüfung ... 179
5.6.1.2 Interpretation der indizierten Klassen ... ***180***
5.6.1.2.1 Charakterisierung der indizierten Klassen ... 180
5.6.1.2.1.1 Merkmalsprofile der indizierten Klassen ... *183*
a) Die Klasse 3 „Hohes Leistungsniveau“ ... 185
b) Die Klasse 1 „Mittleres Leistungsniveau“ ... 186
c) Die Klasse 2 „Niedriges Leistungsniveau“ ... 187
5.6.1.3 Diagnostische Relevanz der latenten Klassen ... ***188***
5.6.1.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen ... 188
5.6.1.3.2 Vergleich der Klassenzusammensetzungen ... 192
5.6.1.3.2.1 Diagnostische Gruppen ... *192*
5.6.1.3.2.2 Krankheitsbilder ... *193*
5.6.1.3.3 Validierung der 3-Klassen-Lösung ... 194
5.6.2 Probabilistischer Auswertungsmodus der Testbatterie ... **196**
5.6.2.1 Identifikation latenter Klassen ... ***196***
5.6.2.1.1 Inferenzstatistische Prüfung ... 198
5.6.2.2 Interpretation der indizierten Klassen ... ***199***
5.6.2.2.1 Charakterisierung der indizierten Klassen ... 199
5.6.2.2.1.1 Merkmalsprofile der indizierten Klassen ... *205*
a) Die Klasse 1 der „jüngeren Personen mit hohem Leistungsniveau“ ... 207
b) Die Klasse 4 der „jüngeren Personen mit leicht reduziertem Leistungsniveau“ ... 208
c) Die Klasse 3 der „jüngeren Personen mit reduziertem Leistungsniveau“ ... 209

d) Die Klasse 2 der „älteren Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“ ... 210

e) Die Klasse 6 der „älteren Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“ ... 211

f) Die Klasse 5 der „älteren Personen mit Leistungseinbrüchen in allen drei Tests“ ... 212

5.6.2.3 Diagnostische Relevanz der latenten Klassen ... 213

5.6.2.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen ... 213

5.6.2.3.2 Vergleich der Klassenzusammensetzungen ... 218

5.6.2.3.2.1 Diagnostische Gruppen ... 218

5.6.2.3.2.2 Krankheitsbilder ... 220

5.6.2.3.3 Validierung der 6-Klassen-Lösung ... 222

5.6.3 Zusammenfassung der Prädiktionskennwerte der Klassen der Testbatterie beider Auswertungsmodi ... 225

6 DISKUSSION ... 227

6.1 Fragestellung 1: Werden von stochastischen Testmodellen bei den einzelnen Testverfahren unterschiedliche Arten der Testbearbeitung (latente Klassen) identifiziert? ... 228

6.2 Fragestellung 2: Wie können die von stochastischen Testmodellen identifizierten Arten der Testbearbeitung bei den drei Tests interpretiert werden? ... 228

6.3 Fragestellung 3: Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen der einzelnen Tests? ... 230

6.4 Fragestellung 4: Können die identifizierten Arten der Testbearbeitung die Vorhersagevalidität der drei Tests verbessern? ... 230

6.5 Fragestellung 5: Ergibt sich durch eine Kombination der Einzeltests zu einer Testbatterie eine Verbesserung der Vorhersagevalidität? ... 232

6.6 Fragestellung 6: Können Lösungsmuster oder Typen der Testbatteriebearbeitung identifiziert und differentialdiagnostisch genutzt werden? ... 232

6.6.1 Fragestellung 6.1: Werden unterschiedliche Bearbeitungstypen (latente Klassen) der gesamten Testbatterie identifiziert? ... 233

6.6.2 Fragestellung 6.2: Wie können die von der LCA identifizierten latenten Klassen interpretiert werden? ... 233

6.6.3 Fragestellung 6.3: Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen in den einzelnen Testergebnissen? ... 234

6.6.4 Fragestellung 6.4: Bieten die identifizierten latenten Klassen der Testbatterie diagnostisch verwertbaren Informationsgewinn? ... 235

6.7 Allgemeine Diskussion ... 239

6.8 Verschimmelungsmanifest für das Fach Psychologie als Wissenschaft vom Menschen ... 251

6.9 Epilog ... 253

7 A) ZUSAMMENFASSUNG ... 255

7 B) ABSTRACT ... 259

8 LITERATURVERZEICHNIS ... 263

9 ANHANG ... 281

9.1 Anhang A: Tabellenverzeichnis ... 281

9.2 Anhang B: Abbildungsverzeichnis ... 285

9.3 Anhang C: Ergänzende Tabellen ... 287

1 EINLEITUNG

1.1 Prolog

„…was wird aus dem Gedanken selbst werden, der unter den *Druck* der Krankheit gebracht wird? Dies ist die Frage, die den Psychologen angeht: und hier ist das Experiment möglich“.

(Friedrich Nietzsche, *Vorrede zur fröhlichen Wissenschaft*, 1886)

1.2 Thema

In den sechziger Jahren entwickelten sich stochastische Testmodelle (LAZARSFELD, 1950) als neuer grundlegender methodischer Ansatz für die Konstruktion von Testverfahren. Der Begriff fasst formale Modelle zusammen, die das Antwortverhalten von Versuchspersonen auf Testitems unter einer Wahrscheinlichkeitsbeziehung modellieren und damit eine Abkehr von älteren klassischen testdiagnostischen Ansätzen bedeuten. Diese waren aus wissenschafts- und methodentheoretischer Sicht wegen formaler Mängel stark in die Kritik geraten (vgl. LUMSDEN, 1976).

Entwicklung und Anwendung neuartiger Verfahren stecken im klinisch-diagnostischen Bereich, trotz denkbarer Einsatzmöglichkeiten, bis heute in den Kinderschuhen. Selbst der grundlegendste Anwendungsfall für stochastische Testmodelle, bleibt in der Mehrzahl der Fälle ungeprüft oder zweifelhaft. So gilt es bei Testkonstrukteuren als seltene Auszeichnung, wenn der Annahme der Eindimensionalität der gemessenen psychologischen Fähigkeit mit „dichotomem Raschmodell“ entsprochen werden kann. Nur dieser Fall sichert die Gültigkeit und Fairness der Scorebildung theoretisch ab[2]. Die daraus resultierenden äußerst praktischen Konsequenzen für den summenscoreverwöhnten Testanwender sind rechentechnischer Natur: Die Anzahl gelöster Items darf dann[3] als Skalierung einer Fähigkeit betrachtet werden.

Stochastische Ansätze haben jedoch weit mehr als diesen, unter Anwendungsgesichtspunkten vielleicht „langweiligsten Spezialfall“ einer probabilistischen Messtheorie (ROST, 1999, S.141) zu bieten. Die grundsätzlich humanistische Gesinnung der probabilistischen Test- oder

---

[2] damit ist gemeint, dass latente Fähigkeitsausprägungen durch den numerischen Wert des Scores abgebildet werden dürfen.

[3] und nur dann!

Item Response Theorie schließt nämlich nicht aus, dass Menschen äußerst unterschiedliche Strategien verwenden, um vorgelegte Testitems zu bearbeiten. Augenscheinlichkeiten und Instruktionen der Testpsychologen ändern nichts daran, dass Testpersonen gedanklich nicht zwangsläufig dem postulierten und vermeintlich einleuchtenden Lösungsweg folgen. Die Gefahr, dass Menschen deshalb eher ***keiner*** als ***nur einer*** Metrik gerecht werden, weil heterogene Personenpopulationen als gleich angesehen werden, muss von Geisteswissenschaftlern angemahnt, jedoch von Testtheoretikern erkannt werden. Die Auswirkungen derartiger Fehleinschätzungen können zahlreiche Erkenntnisse der Psychologie in Schieflage bringen (LAZARSFELD, 1959; BEM & ALLEN, 1974; SIXTL, 1993).

Um solche, bislang kaum zugängliche Sachverhalte aufzudecken, werden große Versuchspersonenpools benötigt. Dieser vorab zu leistende, immense Aufwand in der Testentwicklungsphase mag dazu beitragen, dass sich der Wille zur Konstruktion neuartiger, anwendungsorientierter Testverfahren mit probabilistischen Methoden eher bescheiden gibt. Es sind jedoch auch andere hinderliche Gründe zu nennen: An exakten Wissenschaften angelehnte Testhistorie und die daraus resultierende Dominanz quantitativer Methodik (GRÜNWALD, 1980; KEMPF, 2003), mathematische Parameterschätzprobleme, Konsumorientierung der Testverlage, Bequemlichkeit, Ideen-, Konzept- und Mutlosigkeit der Testautoren sowie mangelndes Interesse der Praxis. Die häufig wenig publikumsträchtig zu nennende Vorstellung der hinter einer Fülle mathematischer Formeln verborgenen testtheoretischen Ideen, Möglichkeiten und Belange tut ihr Übriges (THISSEN et al., 1986).

Längst gehören Computer zur Standardausrüstung psychologischer Praxen (STOLL, 1992 & 1998). Trotzdem markiert bislang nur das als testökonomischer Sonderfall quantitativer Messung zu bezeichnende „Adaptive Testen“ (KUBINGER, 1986 & 1996) eine für die Testpraxis geeignete und als moderner Ansatz erkennbare Anwendung (vgl. *Abschnitt 1.8.1*).

Mit neueren Testmodellentwicklungen z.B. Mixed-Rasch-Modell (ROST, 1990) kann ein theoriegeleiteter, tiefer, beinahe mikroskopischer Blick in individuelle Testbearbeitungscharakteristiken gelingen, der dem Anwender mit den weniger genauen Auflösungsmodi klassischer Theorie oder faktorenanalytischer Methode bisher verschlossen war. Derartige Entwicklungen könnten OZER & REISE (1994) zu der euphorischen Aussage bewogen haben, dass die interessantesten diagnostischen Forschungen in den nächsten zehn Jahren im Bereich der Analyse von strukturellen Gruppenunterschieden stattfinden werden.

Dass der 50-jährigen (!), ausgiebigen und fruchtbaren Testmodellforschung und der damit scheinbar verbundenen Fortschritte der Testkonstruktion keine grundsätzlich neuen

verwertbaren Ideen für den praktischen diagnostischen Alltag folgen, ist von einem testtheoretischen Standpunkt in jeglicher Hinsicht zu bedauern.

Unter Betrachtung ihrer damaligen und - nicht merklich veränderten - derzeitigen Lage weist STECK (1981) darauf hin, dass klinisch-psychologische Diagnostik ohne Revision der klassischen Testtheorie ihrer absehbaren Entwicklung kaum gerecht werden kann. JÄGER & KRIEGER (1994) bezeichnen es als unausweichlich, dass eine angewandte Disziplin wie die Psychologische Diagnostik aus Sackgassen herausgeleitet und neuen Perspektiven eröffnet werden muss. Zumindest angedacht wurde die grundsätzliche Neuorientierung (JÄGER, 2003) bis hin zum Paradigmawechsel (SCHULZ, 1997).

Die vorliegende Arbeit will neue Wege aufzeigen und Anstöße für praxisorientierte neuartige Entwicklungen geben. Neben üblichen quantitativen, soll besonders die Möglichkeit der Aufdeckung qualitativer Testbearbeitungsunterschiede ins psychodiagnostische Blickfeld gerückt werden. Im Mittelpunkt der Betrachtungen steht die grundsätzliche Frage, ob von stochastischen Testmodellen identifizierte Arten der Testbearbeitung bei drei untersuchten klassischen psychometrischen Verfahren (Benton-, Raven- und Cattelltest)[4] zu einer Verbesserung der Vorhersagevalidität klinischer Kriterien beitragen.

Als Anstoß zu diesem Thema dienten Ergebnisse der probabilistischen Analyse einer neuen Form des Benton-Tests (KINDERMANN, 2001). Dabei wurde festgestellt, dass in einer von zwei identifizierten Klassen des Mixed-Rasch-Modells 93 Prozent aller klinisch auffälligen Probanden enthalten waren. In diesem Ergebnis spiegelt sich die grundsätzliche psychologische Bedeutung der Idee von Mischverteilungsmodellen (ROST & ERDFELDER, 1996) wieder. Diese berücksichtigen, dass für die Lösung derselben Testitems, unter Umständen mehrere Bearbeitungsstrategien eingesetzt werden. Eine Aufdeckung unterschiedlicher Strategien bedeutet, dass Fähigkeiten nicht verallgemeinerbar skalierbar sind und eine Unterscheidung der Bearbeitungsart im vorliegenden Fall differentialdiagnostisches Interesse verdient, weil Patienten mit hohem Sicherheitsgrad nur einer der beiden Klassen zugeordnet werden.

Die zwei möglichen Konsequenzen dieser bemerkenswerten Sachlage aus KINDERMANN (2001) waren klar verteilt:

Einerseits war es wegen unklarer Datenlage unbedingt notwendig, mittels einer Kreuzvalidierung der „sensationellen“ Befunde, die Gefahr eines Artefakts auszuschließen.

---

[4] Es handelt sich beim Bentontest um die von STECK (2005) verlängerte 20-Item-Form, beim Raventest um eine aus 20-Items der Advanced Progressive Matrices extra für diese Untersuchung zusammengestellte Form, beim Cattelltest um die 13 regulären Items des Untertests „Matrizen“ im CFT-3, 2. Teil .

Andererseits war die Möglichkeit gegeben, dass die testdiagnostische Zukunft (z.B. BOOTH, 1990; SCHALLBERGER, 1992; JÄGER & KRIEGER, 1994; GUTHKE & WIEDL, 1996; BECKMANN & HERZBERG, 2001) durch Mischverteilungsmodelle Impulse erfährt, die einen revolutionären Strategiewechsel testdiagnostischer Prinzipien implizieren und neue Anwendungsfelder erschließen.
Auf die Resultate von KINDERMANN (2001) wird im Verlauf der vorliegenden Arbeit nochmals Bezug genommen und die darauf aufbauende Validierung anschließend präsentiert.

1.3 Psychologie, Diagnostik und Testtheorie

Jene Wissenschaft vom Menschen, die sich mit den subjektiven Lebensvorgängen beschäftigt, die gesetzmäßig mit den objektiven verknüpft sind, wird im Lexikon als **Psychologie** bezeichnet (HÄCKER & STAPF, 2004).
Die sachliche Feststellung psychologischer Charakteristika eines Individuums unter Anwendung besonderer Methoden wird **Psychodiagnostik** genannt. Im klinischen Bereich lässt sie sich in ihrer allgemeinsten Form als Suche nach Information für Behandlung verstehen (PAWLIK, 1976).
Eine häufige, wichtige aber auch umstrittene Hilfe für die Informationsfindung (REXILIUS, 1978; GRÜNWALD, 1980; SCHALLBERGER, 1992; SCHULZ, 1997; GOULD, 1999; SACKS, 1999), stellen für den in der Klinik tätigen Psychologen **Testanwendungen** dar. Mit Testverfahren wird versucht, Information über mögliche Ursachen der Beschwerden des Patienten zu erhalten, um die Diagnosefindung und Behandlung zu unterstützen und auf wissenschaftlicher Grundlage abzusichern.
Die angestrebte **Professionalität psychodiagnostischer Aussagen** bestimmt sich durch die Verwendung wissenschaftlich verankerter Konstrukte und Testverfahren. Sie zielt darauf ab, den Anteil von Mess- und Prognosefehlern zu verringern und deren Ausmaß abzuschätzen (RAUCHFLEISCH, 1989; WOTTAWA & HOSSIEP, 1997). Eine Störung wird dann aufgezeigt, wenn eine Normabweichung per Testscore als „Negation des Normalen“ (STECK, 1981; S.78) festgestellt werden kann.
Empirische Sozialwissenschaften operieren mit einer großen Menge an Begriffen, die nur gedankliche Konstruktionen, also **theoretische Konstrukte** darstellen. Diese zeichnen sich dadurch aus, dass sie - vorausgesetzt sie existieren - auch für den Experten weder direkter Beobachtung zugänglich, noch allgemeingültig klar und exakt definiert sind. Die **Entwicklung objektiver psychologischer Messverfahren** ist deshalb - im Gegensatz zu den

Messinstrumenten der Physik - mit Schwierigkeiten verbunden. Dennoch orientiert sich psychologische Forschung und Praxis an diesem Vorbild, um (natur-)wissenschaftlichen Ansprüchen[5] gerecht zu werden. Schon GULLIKSEN (1950) war klar, dass der Messfehler in der Psychologie im Vergleich zur Physik viel größer ist.
Das Für und Wider nach wissenschaftlichen Kriterien entwickelter, standardisierter Verfahren ist in entscheidendem Maße an die so genannten **„klassischen" Gütekriterien** Objektivität, Reliabilität und Validität gebunden. Diese, auch auf andere Forschungsdaten anwendbaren, besonders jedoch bei psychologischen Tests gebräuchlichen (Un-) Sicherheitsmaße waren aus folgenden Gründen erforderlich geworden (FISCHER 1974):

- Psychologen können auch bei gewissenhaftester Anwendung eines Testverfahrens stets nur annähernde Angaben über Begabungsstruktur und Persönlichkeitsdimensionen eines Probanden machen (≙ **Messfehler**),
- verschiedene Psychologen müssen durchaus nicht in ihrem Urteil über dieselbe Person übereinstimmen (≙ **Objektivität**),
- derselbe Psychologe erhält beim selben Probanden zu verschiedenen Zeitpunkten unter gleichen Bedingungen oft recht verschiedene Resultate (≙ **Reliabilität**),
- die Übereinstimmung zwischen Vorhersagen aufgrund von Testbefunden einerseits und Außenkriterien andererseits lässt oft zu wünschen übrig (≙ **Validität**).

Entschließt sich der praktisch tätige Psychologe aus unterschiedlichsten diagnostischen Gründen für eine Testdurchführung, befindet sich die kleinste darstellbare und mitteilbare Menge diagnostischer Information zunächst „verschlüsselt" in jeder einzelnen Beantwortung der vom Patienten oder auch anderen Personen bearbeiteten Testitems. Als zuständiger Fachbereich für diese Informationskomponenten, als Wächter über die Grenzen der Naturwissenschaftlichkeit psychodiagnostischer Aussagen mit Tests und als Vermittler zwischen Psychologie (Mensch und Natur) und Mathematik (Zahl und Abstraktion) entwickelte sich das Fach **Testtheorie**. Als **zentrales Problem** dieser Fachrichtung gilt die fragliche Beziehung zwischen **latenter** psychologischer und **manifester** numerischer Variable (GULLIKSEN, 1961). Die theoretische Absicherung der Zuordnungsregeln zwischen Zahlen und empirischen Elementen wird Messtheorie, Testtheorie oder Skalierung genannt.

[5] Über die wissenschaftstheoretischen Probleme dieser Annäherung der Psychologie an die Naturwissenschaften vgl. HARTMANN (1993)

1.4 Die Realisierung einer psychologischen Messung als Score in der klassischen Testtheorie

Die von GULLIKSEN (1950) mathematisch formulierten und von NOVICK (1966) zum anschaulichen psychologischen Modell ausgearbeiteten Axiome der klassischen Testtheorie markieren einen wichtigen Schritt aus der vorwissenschaftlichen und intuitiven Deutung testpsychologischer Untersuchungsbefunde. Sie stellen den bis heute am weitesten verbreiteten testtheoretischen Ansatz dar (STUMPF, 1996).
Die **klassische Testtheorie** basiert auf folgendem Theorem:

> *Der individuelle Wert x bei einem Test ist eine Zufallsvariable mit einer meist unbekannten Häufigkeitsverteilung. Der Erwartungswert dieser Verteilung wird als wahrer individueller Wert τ bezeichnet.*

Den Ausgangspunkt des klassischen Modells bilden **drei Voraussetzungen** (NOVICK, 1966):

1.) Jeder Testung t eines Probanden v entspricht eine zufällige Scorevariable möglicher Testergebnisse X_{vt} mit endlichem Erwartungswert E (X_{vt}) und endlicher Varianz $\sigma^2(X_{vt})$.
2.) Das Testergebnis x_{vt} der Testperson, ist eine unabhängige Realisation dieser Scorevariable.
3.) Per definitionem wird festgelegt, dass der wahre Wert der Testperson τ_{vt} dem Erwartungswert der Scorevariable entspricht: τ_{vt}= Erw (X_{vt}).

Als Indikator für die **diagnostisch relevante Information** in den beantworteten Testitems, wird von der klassischen Testtheorie der **Summenscore x** (Anzahl richtig gelöster Items) einer Testanwendung verwendet. Dieser Testscore ist eine vollkommene Analogie zum fehlerbehafteten Messwert im physikalisch-technischen Bereich. Nur weil in der **Physik** exakt definierte Größen existieren, kann ein physikalischer Sachverhalt dort mit Messgeräten[6] beobachtet werden. „Dabei begnügt sie[7] sich nicht mit qualitativen Angaben, sondern sucht immer quantitative Aussagen zu gewinnen, indem sie die gleichartigen Eigenschaften zweier

[6] und normalverteilter Streuung um den wahren Wert

[7] die Physik (Anmerkung d.V.)

Objekte (ihre gleichen „Qualitäten“) vergleicht, oder indem sie gleichartige Ereignisse (Zustandsänderungen) zählt“ (WALCHER, 1989, S. 16). Im Gegensatz dazu, verkörpert x in der **Psychologie** die fehlerbehaftete Abbildung einer nicht direkt beobachtbaren und nur über die Testitems operationalisierten latenten psychologischen Eigenschaftsausprägung einer mit menschlichen Fähigkeiten ausgestatteten Testperson. Die Messung erfolgt in der klassischen Testtheorie auf der psychologisch heiklen „Skala“ möglicher erreichbarer Summenscorewerte (FISCHER, 1974; ROST, 2004).

Weder wahrer Wert τ, noch Messfehler, sondern nur der Summenscore x ist bei Testungen direkter Beobachtung zugänglich. Als einzelner Datenpunkt einer hypothetischen Wahrscheinlichkeitsverteilung weiterer Summenscorewerte streut er bei standardisierten Verfahren um die **„wirkliche“ Eigenschaftsausprägung τ**, weil ein zufälliger Einfluss von Situationsfaktoren bei Testungen angenommen wird (HILKE, 1980). „Darüber hinausgehende Fehlerquellen (z.B. Störfaktoren wie Müdigkeit oder die Bevorzugung hochbegabter Probanden durch den Test) sind schlichtweg nicht Gegenstand der klassischen Testtheorie, und die Analyse von Tests mittels der klassischen Testtheorie sagt über sie nichts aus“ (KEMPF, 2003, S. 80). Weil sie **keinen empirischen Gehalt** aufweist und daher universell anwendbar ist, wird das axiomatische System der klassischen Testtheorie von WALTER (1991) als „nachträglicher theoretischer Überbau“ und von ROST (1996) als „Messfehlertheorie“ bezeichnet.

1.5 Aufbruch zu einer naturwissenschaftlichen Psychologie?

Zur Überwindung der vorwissenschaftlichen Phase der Psychologie, brachte die **klassische Testtheorie** wichtige Impulse, sie erlaubt jedoch keine tiefer gehenden Erkenntnisse. Der klassische Theorieansatz ist auf wohl definierte und beobachtbare physikalisch-technische Größen anwendbar und dort zur Beschreibung von Messvorgängen möglich. Er reicht aber für die Zuschreibung psychologischer Eigenschaften nach komplexem menschlichem Testverhalten nicht aus: „...there can be no solution without a structural theory” (GUTTMAN, 1953, S. 129).

Der Fortschritt sozialwissenschaftlicher Theoriebildung hängt für FISCHER (1974) davon ab, ob es gelingt **realistische überprüfbare Modelle** zu konstruieren und sie dann auch anzuwenden. Der erste falsifizierbare Modellansatz mit diesem Ziel (GUTTMAN, 1950), hat den entscheidenden Nachteil, dass er sehr schnell durch Daten widerlegt werden kann, weil ein **deterministischer Zusammenhang** zwischen Itembeantwortungen und Fähigkeiten

vorausgesetzt wird. Dagegen lassen stochastische Modelle größeren Spielraum zu. Sie berücksichtigen, dass menschliches Verhalten in jeder Situation von vielen zufälligen Faktoren abhängig ist und niemals mit völliger Sicherheit deterministisch vorhergesagt werden kann (GUTHKE, BÖTTCHER & SPRUNGE, 1990). Charakterisierend bezeichnet es SIXTL (1993) als **„probabilistische Wendung“** des neuzeitlichen Denkens, wenn wissenschaftlicher Erkenntnisgewinn ohne deterministische Wenn-Dann-Beziehungen zustande kommt.

In Abkehr von der nicht falsifizierbaren Messfehlertheorie klassischer Art und den zu strengen Voraussetzungen jedes deterministischen Ansatzes, wurden **probabilistische Konzepte** als Theorie der Psychologie menschlichen Verhaltens und Erlebens im deutschsprachigen Raum erstmals von FISCHER (1968) aufgegriffen, diskutiert und später lehrbuchmäßig propagiert. FISCHER (1974) setzt große Hoffnungen in die Entwicklung und Anwendung probabilistischer Testmodelle. Er will aufzeigen, dass sich testtheoretische Belange - unter neuem realistischerem probabilistischem Ansatz - nicht auf die Konstruktion diagnostischer Instrumente beschränken lassen, sondern Grundfragen der empirischen Psychologie schlechthin betreffen.

Die von Fischer propagierte Psychologie steht auf folgenden **Grundpfeilern**:

(1) erkenntniskritische Psychologie kann nicht ohne Auseinandersetzung mit der Test- und Messtheorie erfolgen;
(2) psychologische Theorie ist nur dann wissenschaftlich, wenn aus ihr in exakter Weise empirisch prüfbare Aussagen deduzierbar sind.

Der **mathematische Beweis**, dass nur ein bestimmtes Testmodell „spezifisch objektive“ Vergleiche latenter psychologischer Variablen zwischen Personen und auch Items zulässt (RASCH, 1961), markiert eine bahnbrechende, fundamentale Mess- und Testerkenntnis. Diese wurde noch von früheren Fachvertretern der Psychologie für unmöglich gehalten (FERGUSON et al., 1940). Die Unabhängigkeit der Differenz zweier Personenparameter von der Verteilung der Itemparameter[8] in einem Test – also die **spezifische Objektivität** - ist von entscheidender Bedeutung für die Fundierung einer naturwissenschaftlich ausgerichteten Psychologie: „Wenn zwei Personen, für welche die Aufgaben des Tests gleiche Schwierigkeit

[8] und umgekehrt, die Unabhängigkeit der Differenz zweier Itemparameter von der Verteilung der Personenparameter.

haben, verschieden viele Lösungen erzielen, dann ist erwiesen, dass die Fähigkeit der beiden Personen unterschiedlich hoch entwickelt ist. Alle Fragen nach Unterschieden zwischen den Geschlechtern, sozialen Schichten, Rassen werden mit Tests objektiv beantwortet, deren Aufgaben die Personen spezifisch objektiv klassifizieren“ (SIXTL, 1996, S. 64).

Das resultierende probabilistische Testmodell wird als **Rasch-Modell** bezeichnet. Dass es zunächst nur am Rande Beachtung fand, ist auf die mathematische Komplexität, die teilweise unübersichtliche Darstellung des Verfahrens und den rechentechnischen Aufwand zurückzuführen. Viele weitere Modellentwicklungen, die in ihren vielfältigen Ausgestaltungen üblicherweise das Raschmodell zur Grundlage haben, markieren einen anhaltenden Boom „moderner Testtheorie“ (KUBINGER, 1988). Der Weg aus den Forschungslabors in das Bewusstsein des Anwenders muss aber noch gefunden werden (STOLL, 1998; JÄGER, 2003).

1.6 Die Realisierung einer psychologischen Messung als Itemantwort in der probabilistischen Testtheorie

James McKeen Cattell der erste Professor für Psychologie, gilt als Erfinder des Begriffes „mental tests“ und Vorreiter testtheoretischer Prinzipien nach dem strengen **Vorbild der Physik**. In seiner testpsychologischen Maxime bringt er Gedankengut zum Ausdruck, für das erst die sieben Jahrzehnte später entstandene probabilistische Testtheorie den nötigen theoretischen Rahmen bietet: „Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement. (...) The scientific and practical value of such tests would be much increased should a uniform system be adopted, so that determinations made at different times and places could be compared and combined” (CATTELL, 1890, S. 373).

Die als modern bezeichnete **probabilistische Test- oder Item Response Theorie (IRT)** verwendet die einzelnen direkt beobachtbaren (manifesten) Itemantworten eines Tests als Symptome einer zugrunde liegenden vermuteten (latenten) psychologischen Eigenschaft oder Fähigkeit (im Überblick bei HAMBLETON, SWAMINATHAN & ROGERS, 1991; ROST, 2004). Dieser Vermutung wird durch einen **stochastischen Zusammenhang zwischen jeder einzelnen Itemantwort und der latenten Personeneigenschaft** Ausdruck verliehen. Jede einzelne Testaufgabe wird so zu einer winzigen experimentellen Versuchsanordnung: Testitems dienen als Auslöser latenter (innerpsychischer) Prozesse bei Testpersonen. Diese

münden in eine vorgegebene und bei Fähigkeitstests häufig dichotome[9] Menge manifester Antwortmöglichkeiten ein (KEMPF, 1972).
Grundsätzlich werden **Latent Trait** von **Latent Class Modellen** unterschieden. Weil sich Quantifizieren mathematisch als Spezialfall von Klassifizieren darstellen lässt (vgl. *Abschnitt 3.2)*, ergibt sich eine enge theoretische Verwandtschaft der beiden Ansätze. Während ersterer die Homogenität von Tests prüft und als Rechtfertigung für die Gültigkeit der Scorebildung bei Fähigkeitstests herangezogen wird, funktioniert letzterer auch, wenn sich Inhomogenitäten ergeben und qualitative Unterschiede festgestellt werden sollen (ROST, 1996).

Zwei Annahmen bestimmen die Struktur stochastischer Testmodelle mit dichotomen Itemantwortmöglichkeiten:

(1) Die Wahl der **Form einer Itemcharakteristik** (ICC[10]) oder Itemfunktion f, die die Wahrscheinlichkeit einer richtigen Itemantwort $p(X_{vi} = 1)$ in Abhängigkeit von der Personenvariable ξ_v beschreibt.
(2) **Lokale stochastische Unabhängigkeit** der Itemantworten, d.h. die Unabhängigkeit der Antwortwahrscheinlichkeit einer Testperson auf ein Item A, von der Antwortwahrscheinlichkeit derselben Testperson auf ein anderes Item B.

Die Annahme (2) schränkt zwar das Anwendungsgebiet aufgrund folgender Voraussetzungen ein:

a) keine erfolgsabhängigen Lernprozesse während der Testung;
b) keine logisch abhängigen Itemlösungen, d.h. für die Lösung des einen darf nicht die Lösung des anderen Items nötig sein.

Sie rechtfertigt aber die Anwendung des Multiplikationstheorems der Wahrscheinlichkeitsrechnung für unabhängige Ereignisse:

$$p(A \cap B) = p(A) * p(B) \qquad (1.1)$$

[9] richtig/falsch

[10] Item Characteristic Curve

Ihren **Grundstein** legt die probabilistische Testtheorie in überprüfbarer mathematischer Form:

Die Wahrscheinlichkeit p das Item i von Person υ gelöst wird, ist eine Funktion f der Personenfähigkeit ξ_υ und Itemschwierigkeit σ_i

$$(1.2) \quad p(X_{\upsilon i} = 1) = f(\xi_\upsilon - \sigma_i)$$

Für Fähigkeitstests (Latent Trait Modelle) ergibt sich folgende probabilistische Konzeption menschlichen Testverhaltens:

Es ist wahrscheinlich, dass eine Person mit größerer Fähigkeit auch schwierigere Testitems lösen kann.

Dieses einleuchtende menschliche Reaktionsgesetz, bestehend aus den **drei Verhaltenseinflüssen Item, Person und Zufall** ist grundlegend, um die interne Struktur von Antwortmustern[11] für eine Analyse zugänglich zu machen. Die grundsätzliche Frage, ob die zahlreichen in einem Test vorliegenden manifesten Itemantworten zu modellgültigen Messwerten latenter Variablen von Testpersonen umgewandelt werden dürfen, kann im Gegensatz zur klassischen Testtheorie erst dann beantwortet werden, wenn das **Ergebnis einer Modelltestung** vorliegt. Nur wenn die Modellgültigkeit statistisch belegt werden kann, darf **jedes einzelne Item als winziges Messinstrument** mit eigenem Messgenauigkeitswert (Informationswert) betrachtet werden. Sämtliche verfügbare Information in der Anzahl beantworteter Items repräsentiert dann eine metrisch bestimmbare Testleistung.

1.7 Theoretische Relevanz stochastischer Testmodelle

Die angebliche Messbarkeit psychologischer Konstrukte kann mit den Mitteln klassischer Testtheorie nicht falsifikatorisch abgesichert werden. Für klare wissenschaftliche Aussagen ist es jedoch notwendig, dass definierte theoretische und empirische Grundbegriffe vorliegen (ROSKAM, 1996). Die **kritische Frage** lautet: Vorausgesetzt latente psychologische

[11] Ein Antwortmuster - auch Antwortpattern genannt - beinhaltet die vollständige Beschreibung der Testleistung einer Versuchsperson. x_υ=($x_{\upsilon 1}$, $x_{\upsilon 2}$,... $x_{\upsilon k}$) wobei: υ=Versuchsperson; K=Itemanzahl; im dichotomen Auswertungsmodus x=1 (gelöst) oder 0 (ungelöst)

Eigenschaften und Fähigkeiten entziehen sich einer eindeutigen Definition und direkten Beobachtung, gibt es dann Möglichkeiten, um einen der wenigen wissenschaftlichen Zugänge zu diesen Konstrukten - nämlich mit Tests - auf eine **tragfähige Grundlage** zu stellen?

POPPER (1982; S. 75) warnt, dass die **empirische Basis objektiver Wissenschaft nichts „Absolutes“** ist, da Wissenschaft nicht auf Felsengrund gebaut ist. „Es ist eher ein Sumpfland, über dem sich die kühne Konstruktion ihrer Theorien erhebt; sie ist ein Pfeilerbau, dessen Pfeiler sich von oben her in den Sumpf senken – aber nicht bis zu einem natürlichen, „gegebenen“ Grund. Denn nicht deshalb hört man auf die Pfeiler tiefer hineinzutreiben, weil man auf eine feste Schicht gestoßen ist: wenn man hofft, dass sie das Gebäude tragen werden, beschließt man, sich vorläufig mit der Festigkeit der Pfeiler zu begnügen.“

Unter Bezugnahme auf dieses Zitat, kann die Relevanz stochastischer Testmodelle folgendermaßen bestimmt werden: Stochastische Testmodelle bieten definitive Möglichkeiten, die Festigkeit, der mit klassischer Theorie per definitionem gesetzten Pfeiler, zu erschüttern, sprich zu falsifizieren. **Der probabilistische Ansatz bietet leistungsstarke mathematische Prozeduren um standfeste von maroden Pfeilern zu unterscheiden**.

Äußerst spannend ist ein weiterer Verwendungszweck stochastischer Testmodelle, der einen Schwerpunkt dieser Arbeit darstellt und auch künftige diagnostisch-klinische Forschung begleiten kann: So genannte **Mischverteilungsmodelle** (ROST & ERDFELDER, 1996) eignen sich zur **selbständigen „Pfeilersuche im Datensumpf“** (hier: Identifizierung latenter Klassen). Mischverteilungsmodelle sind in der Lage heterogene Personenpopulationen auf unterschiedliche Weise in jeweils homogene Teilpopulationen (latente Klassen) aufzuteilen. Im Verlauf dieser Arbeit wird dargestellt, welche Vorteile eine Analyse mit derartigen probabilistischen Modellen bringt.

Hält man die **Unterschiede** fest, begnügt sich die klassische Theorie mit einem Modell das nur **„schwache“ Anforderungen an die Daten** stellt, während probabilistische Modelle mit **„strengen“ Bedingungen** verknüpft sind. Beide Ansätze entsprechen sehr verschiedenen wissenschaftstheoretischen Positionen: Der klassische Ansatz gibt sich mit der **Beschreibung von Empirie** zufrieden und sucht möglichst flexible Modelle, wobei die Quantifizierung ein definitorischer Akt bleibt. Der probabilistische Ansatz versucht den Messvorgang - auch mit der klassischen Methode - durch **empirische Gesetzmäßigkeiten** auf ein Fundament zu stellen, dessen relative Tragfähigkeit kontrolliert und abgesichert werden kann (LORD & NOVICK, 1968; FISCHER, 1974; BECHGER et al., 2003).

1.8 Praktische Relevanz stochastischer Testmodelle

Die **Anforderungen an ein modernes Testkonzept** sind von WAGNER-MENGHIN (2002a) folgendermaßen zusammengefasst worden:

- raschhomogener Aufgabenpool
- Minimierung der Ratewahrscheinlichkeit und Verfälschbarkeit
- Testökonomie
- Auswertung mehrerer Testkennwerte zur Leistungsbeschreibung

Bei der Konstruktion psychometrischer Verfahren ist das **Ziel einer eindimensionalen, folglich raschhomogenen Messung** aus mehreren Gründen notwendig (VAN ABSWOUDE et al., 2004):

(1) Testergebnisse können jeder Person praktikabel in Form eines einzelnen Testscores zugeordnet und fair interpretiert werden.
(2) Eindimensionale Testmodelle beinhalten weniger Parameter als multidimensionale, was zu genaueren Parameterschätzungen aus den Stichproben führt.
(3) Multidimensionale Konstrukte - so wird angenommen - lassen sich prinzipiell in mehrere eindimensionale Subtests aufspalten und können einzeln dann weniger problematisch vermessen werden.

Eine Auflistung einiger als raschhomogen ausgewiesener, im Handel erhältlicher Verfahren zeigt die *Tabelle A-47* im Anhang. Weitere, über die Ansprüche klassischer Testtheorie hinausgehende, **klinisch-diagnostische Anwendungsbereiche für stochastische Testmodelle** sind:

1) Adaptives Testen
2) Identifizierung von Personen mit „verdächtigen“ Antwortmustern („person-fit“)
3) Identifizierung problematischer Items („item-Bias“ und „DIF“)
4) Identifizierung systematischer Unterschiede im Antwortverhalten

1.8.1 Adaptives Testen

Mit **„Adaptivem Testen“** (z.B. KUBINGER, 1986; KUBINGER, 1996) wird eine häufig computergestützte Form quantitativer psychologischer Messung bezeichnet, die für den Praktiker besonders ökonomisch und für die Testperson wenig frustrierend also optimal herausfordernd sein soll. Die Testidee wurde schon von LORD (1977) angeregt und ist nur mit probabilistischen Methoden und raschhomogenem Itempool realisierbar. Genutzt wird eine von anderen Items unabhängige Bestimmung des Informationsbeitrags, den jedes Item für verschiedene Fähigkeitsausprägungen leistet. Nach jeweils neu berechnetem vorläufigem Leistungsniveau werden an die Testperson angepasste Aufgaben vorgelegt, um die Ausprägung einer Fähigkeit weiter zu differenzieren. Da nicht mehr jeder Versuchsperson sämtliche Testitems vorgelegt werden müssen, kann eine Reduktion der Testzeit bei Konstanthaltung der Messgenauigkeit erreicht werden (vgl. auch *Abschnitt 3.1.4*).

Beim **„branched testing“** handelt es sich um eine suboptimale Form adaptiver Testadministration. Die Anpassung an das jeweilige Fähigkeitsniveau einer Testperson erfolgt nicht im Anschluss an die Vorgabe jedes einzelnen Items, sondern immer erst nach einer Reihe von wenigen Items.

Computereinsatz ermöglicht beim **„tailored testing“** eine optimale oder „maßgeschneiderte“ Auswahl jedes einzelnen Items, angepasst an das jeweils neu berechnete Fähigkeitsniveau jeder Testperson. Die wegen computergestützter Vorgabe zusätzlich mögliche, **exakte Zeitregistrierung** bei der Bearbeitung einzelner Aufgaben gewinnt hinsichtlich ihrer diagnostischen Relevanz zunehmendes wissenschaftliches Interesse (z.B. BECKMANN et al, 1997; HORNKE, 1997; RAMMSAYER & BRANDLER, 2003).

Eine Auflistung einiger im Handel erhältlicher adaptiver Verfahren zeigt *Tabelle A-48* im Anhang.

1.8.2 „person-fit“

Die Item Response Theorie bietet Berechnungsmöglichkeiten für den **„person-fit“**. Derartige Kennwerte sind von diagnostischem Interesse, um auffällige Testbearbeitungen durch Vergleich mit dem der Testanalyse zugrunde gelegten und für die große Personenmehrheit gültigen Testmodell herausfiltern zu können (DRASGOW et al., 1987; LI & OLEJNIK, 1997; MEIJER & SIJTSMA, 2001; ROST, 2004). Gerade aus diesem Grund stellen sie wichtige und aufschlussreiche Testergebnisse dar. Nicht modellkonform werden solche Antwortmuster von

Personen bezeichnet, die eine zu geringe oder möglicherweise auch eine zu hohe Modellgeltungswahrscheinlichkeit haben.
Anwendungsbereiche für „untypische“ oder „verdächtige“ Antwortvektoren mit fraglicher Mess- und Summenscoregültigkeit sind vielfältig. „Unskalierbare Personengruppen“ spielen eine Rolle bei: Rateverhalten (KÖLLER, 1994), Simulation und Dissimulation (FERRANDO & CHICO, 2001; PONOCNY & KLAUER, 2002), „gestohlenem“ Vorwissen (Mc LEOD et al., 2003), Schummeln (MADSEN, 1987), sozialer Erwünschtheit (SCHMITT et al., 1993), Antwortinkonsistenz (REISE & WALLER, 1993), besonderer Kreativität (LEVINE & DRASGOW, 1982), kulturellen Unterschieden (FLIER, 1982), ungleicher Testinstruktion (HARNISCH, 1983), Missverständnissen (TATSUOKA, 1985), test-bias (FRARY, 1982).

1.8.3 „item-bias“ und „Differential Item Functioning“

Konzepte der Item Response Theorie eignen sich zur Identifizierung von Items, die konstante oder systematische Fehler im Testscore verursachen (THISSEN, STEINBERG & GERRARD, 1986).
Die **„item-bias“**-Forschung (BERK, 1982) befasst sich mit dem Problem, dass Testaufgaben heterogene Fähigkeiten erfassen, oder andere Effekte auftreten, aus denen ungültige Scorevergleiche resultieren. Um die angestrebte Skalierbarkeit der fraglichen Fähigkeit und somit (Rasch-) Modellpassung zu erreichen, ist eine Aussonderung problematischer, bias-belasteter Items bei Testentwicklungen derzeit gängige Praxis (ROST, 2002).
Mit größerer Betonung der subgruppenspezifisch fairen Auslegung von Testergebnissen, gehört ins Umfeld der „item-bias“-Forschung auch der Begriff des **„Differential Item Functioning“ (DIF)** (HOLLAND & WAINER, 1993). DIF geht davon aus, dass Personen gleicher Fähigkeit, aber aus verschiedenen Subpopulationen stammend, unterschiedliche Lösungswahrscheinlichkeiten bei einem gegebenen Item besitzen. Dies hat z.B. in Schulvergleichtests zur Konsequenz, dass bestimmte Bevölkerungsgruppen wegen einzelner Items systematisch benachteiligt werden. Im Mittelpunkt des Interesses steht bei DIF die Frage, warum ein bestimmtes Item keine Modellpassung zeigt. Unterschiedliche äußere Variablen wie sozioökonomischer Status, kulturelle Herkunft oder Geschlecht können für DIF verantwortlich sein. Besonders in der Bildungsforschung wurde bemerkt, dass DIF-Parameter eine Gefährdung der Testfairness bedeuten (ARTELT & BAUMERT, 2004). Gerade der umgekehrte Zugang zu diesem Thema kann jedoch für die testdiagnostische Zukunft wegweisend werden: DIF-Items können wegen systematischer und spezifischer

Varianzanteile im allgemeinen Messfehler wichtige Hilfsmittel bei der Identifizierung qualitativ unterschiedlicher Personengruppen sein (KLIEME & BAUMERT, 2001). KUBINGER (1998) untersuchte Testverfahren mit Aufgaben, die leistungsschwachen Probanden leichter fallen als leistungsstarken und folglich Auswirkungen auf diagnostische Entscheidungen haben.
Methoden der DIF-Erkennung werden bei LINDEN & HAMBLETON (1997) näher beschrieben.

1.8.4 Identifizierung systematischer Unterschiede im Antwortverhalten

Mischverteilungsmodelle[12] bieten die Möglichkeit **Testpersonen mit systematisch anderem Antwortverhalten** zu erkennen und erlauben dadurch eine neuartige Verbindung kognitionspsychologischer und psychometrischer Testinteressen. Während es sich bei DIF eher um die einzelne Betrachtung problematischer Items handelt, die weitestgehend auf einer Dimension messen, werden mit Mischverteilungsmodellen auch Subpopulationen erkannt, die bei Testitems systematisch andere Lösungswege verwenden.
Kritisch hinterfragt wurde im **Bildungsbereich** (FISKE & BUTLER, 1963; MESSICK, 1984), wie sich herkömmliche Testdurchführungen und die Validität ihrer Interpretationen durch den Nutzer vertragen, wenn mittels klassischer Konstruktion nur der momentane Leistungsgesichtspunkt als Summenscore registriert und interpretiert wird, während strukturelle Gesichtspunkte, die jeder Fähigkeit zur Itemlösung von Testpersonen vorangehen ohne testdiagnostische Abklärung bleiben müssen. BRUDER (1973, S. 123) bringt das problematische Vorgehen auf den Punkt: „Der objektive Maßstab wird in die subjektive Dimension getragen, dadurch, dass man diese selber dem objektiven Maßstab der Leistungsanforderungen unterwirft: nicht unter die Frage, welches Verhalten äußert das Individuum, sondern äußert das Individuum das *geforderte* Verhalten oder nicht“.
FISCHER (1974) kommt über die klassisch unabdingbare Voraussetzung eines homogenen Itempools auf die testdiagnostischen Konsequenzen der „Differenzierungshypothese der Intelligenzentwicklung“ zu sprechen. Diese besagt, dass sich **Intelligenzentwicklung** vom älteren Kind zum Erwachsenen weniger quantitativ durch Zunahme als vielmehr qualitativ durch Differenzierung vollzieht. Einerseits bedeutet dies, „dass man verschiedene Personen (…) überhaupt nicht quantitativ in Bezug auf ihre Leistung in einem Intelligenztest vergleichen kann, und andererseits, dass derselbe Test bei verschiedenen Personen auch

[12] vgl. *Abschnitte 1.9*, *1.10* und *3.2*

Verschiedenes misst. Damit hätten Intelligenztests überhaupt ihren Sinn verloren!“ (FISCHER, 1974, S. 311). Analog wird auch in Längsschnittstudien angezweifelt, ob mit einem bestimmten Itemsatz über einen längeren Zeitraum konstant dieselbe Fähigkeit erfasst wird, weil Entwicklungsfortschritte innerhalb einer Kohorte, oder Unterschiede der betrachteten Kohorten mit der Art der Itembearbeitung interferieren (SPIEL et al., 2001).

Wenn Itemparameter laut FISCHER (1973) die Komplexität oder Anzahl nötiger Denkoperationen zur Itemlösung reflektieren, dann ergeben sich vor einem diagnostischen Hintergrund folgende **Fragestellungen**:

1) Führen Bildungsunterschiede, pathologischer Leistungsabbau (z.B. demenzielle Prozesse), mentale Verfassungen (z.B. depressive Zustände), chronische Erkrankungen (z.B. Epilepsie) oder Hochbegabungen zu einer Beeinflussung oder Veränderung der latenten Bearbeitungsform von Testitems?

2a) Besteht die Möglichkeit Testitems zu konstruieren mit denen derartige Phänomene aufgedeckt werden?

2b) Können Tests oder Testbatterien erstellt werden, die verschiedenartige differentialdiagnostisch nutzbringende Profilcluster ergeben?

THORNDIKE (1947)[13] teilt mögliche **Varianz- oder Informationsanteile** die sich negativ auf die Reliabilität und Validität einer Fähigkeitsmessung per Summenscore auswirken, in fünf Kategorien ein:

(i) **Stabile** und **allgemeine** Varianzanteile (z.B. allgemeine Leistungsfähigkeit in einem bestimmten Test, das „Ideal des Summenscores“)

(ii) **Stabile** und **spezifische** Varianzanteile (z.B. bestimmtes Spezialwissen oder chronische Erkrankungsprozesse, welche die Itemlösungen beeinflussen)

(iii) **Temporäre** und **allgemeine** Varianzanteile (z.B. Müdigkeitsfaktoren)

(iv) **Temporäre** und **spezifische** Varianzanteile (z.B. vorübergehende mentale Zustände die Itemlösungen beeinflussen)

(v) **Andere** Varianzanteile (z.B. Ratefaktoren)

[13] zitiert nach CRONBACH et al. (1972)

Das klassische Maß „Summenscore“ setzt sich aus wahrem Testleistungsanteil (i) und weiteren Varianzanteilen (ii)-(v) zusammen. Die klassische Perspektive - ergebnisorientiert an der allgemeinen Leistungsfähigkeit des Probanden - kommt im klinischen Bereich dann an ihre Grenzen, wenn sich gemessene Eigenschaften etwa bei psychischen Störungen qualitativ andersartig als bei Menschen ohne psychische Auffälligkeiten darstellen (STECK, 1981). Denkbar ist, dass zur Kompensation neuropsychologischer, kognitiver und sonstiger chronischer Veränderungen (ii) oder vorübergehender mentaler Verfassungen (iv) eine differentialdiagnostisch aufschlussreiche **Umstellung auf qualitativ andere latente Verarbeitungsmuster von Testitems** erfolgt.

Solche Phänomene sind für den Testanwender mit den Mitteln **klassischer Testauswertung** nicht erkennbar. Die klassische Testtheorie ebenso wie die Faktorenanalyse bieten keine „Werkzeuge“, um psychologisch heterogene Testbearbeitungen auseinanderzuhalten. Eine Differenzierung der latenten Testbearbeitungsform scheitert hier, weil diese auf Information innerhalb der Daten beruht, die zusätzlich im psychologisch ungeklärten Wert des klassischen Summenscores enthalten ist (ii)-(v). So sollen die gering geschätzten Zusatzinformationen (ii)-(v) innerhalb der Testdaten mit klassischen Testgütekontrollen zwar minimiert werden (GRÜNWALD, 1980; ANASTASI, 1976; SIXTL, 1996), eine klare Abgrenzung erwünschter Testleistungsinformation von dem komplexen Messfehlerbündel unterschiedlichster Störinformationen ist jedoch unmöglich. Mangels Theoriegrundlage können somit auch krankheitsspezifische Störeinflüsse[14] nicht separat als systematische Varianzanteile identifiziert werden.

Es sind diese „nebensächlichen“ Informationen (ii)-(v) die zunehmend das Interesse der **modernen Testtheorie** gewinnen. Das erweiterte Anwendungsfeld hat sich als Nebenprodukt[15] ihrer ursprünglichen Aufgabe - einer Sicherung der Homogenität (i) mit dem Raschmodell - ergeben (LINDEN & HAMBLETON, 1997). Weil zur Lösung von Testitems häufig mehr als nur eine Fähigkeit oder Strategie eingesetzt werden (ANSLEY & FORSYTH, 1985; RECKASE & MC KINLEY, 1991; KÖLLER et al., 1994), oder andere klinisch relevante systematische Störeinflüsse (z.B. Stress- und Ermüdungsfaktoren) existieren, können sich aus der Identifizierung solcher Zusatzinformationen Vorteile für den Diagnostiker ergeben (KINDERMANN, 2001; KAHRAMAN & KAMATA, 2004).

Nur mit Weiterentwicklungen von Testmodellen - insbesondere Mischverteilungsmodellen - eröffnen sich über die Itemantworten einer Testperson testökonomisch elegante

[14] stabile und spezifische (ii) oder temporäre und spezifische Varianzanteile (iv)

[15] und selbstbewusstes Zeichen der Ablösung von ungelösten Problemen der klassischen Testtheorie

Möglichkeiten[16], um **vermeintliche Störinformation** innerhalb der Daten - im Sinne unterschiedlicher Strategien, Denkprozesse, Eigenschaften, Personenmerkmale oder Fähigkeiten - erkennbar, interpretierbar und spezifischer differentialdiagnostischer Nutzung zugänglich zu machen. Sicher ausgeschlossen sind derartige Phänomene testtheoretisch nur dann, wenn die Gültigkeit des einfachen Raschmodells in klinisch auffälligen und unauffälligen Populationen nachgewiesen wurde.

Neben der Frage nach latenten Bearbeitungsstrategien von Testitems ergeben sich im klinischen Einsatz **weitere Forschungsfragen und praxisorientierte Anwendungsfelder für Mischverteilungsmodelle** bei der Ergebnisanalyse herkömmlicher Testbatterien[17]:
Im Gegensatz zu den starr vorgegebenen, von reger Forschungstätigkeit begleiteten und diagnostisch weitgehend erfolglos gebliebenen Testprofilvorgaben[18] (ANASTASI, 1976), bieten Mischverteilungsmodelle die Möglichkeit latente Personengruppen nachzuweisen, die sich bei der Bearbeitung einer Testbatterie unterscheiden. Qualitative Unterschiede der erbrachten Testleistungen in sämtlichen Einzeltests können so als gruppenspezifische Leistungs- oder Testmerkmalsprofile dargestellt werden. Erst nach erfolgreicher Aufdeckung solcher Profile wird geprüft, ob für die Klassifizierung differentialdiagnostisch sinnvolle psychologische, psychopathologische oder andere Kriterien gefunden werden können. Einerseits ist anzunehmen, dass klinisch unauffällige Personengruppen wegen differentieller Persönlichkeitsunterschiede verschiedenartige Testprofile zeigen (BUKASA et al., 1990). Andererseits könnten latente Klassenbildungen von Testmerkmalen wichtige Informationen zur diagnostischen Abklärung klinischer Auffälligkeiten liefern (KUBINGER et al., 1980).

1.9 Mischverteilungsmodelle als Weiterentwicklungen stochastischer Testmodelle

In Anlehnung an das Raschmodell wurde eine **ganze Reihe von Testmodellen** entwickelt, die sich mit weniger strengen Voraussetzungen und Anforderungen für unterschiedlichste Antwortformate und Anwendungen eignen. Computerunterstützt eröffnen sich vielfältige

[16] im Gegensatz zu den wenig praxisrelevanten und umständlichen Methoden der Analyse von Antwortzeiten (BLOUGH & SLAVIN, 1987), der Selbstbeobachtung (COCHRAN & WHEATLEY, 1989), des lauten Denkens oder der Blickbewegungsanalyse etc. (PUTZ-OSTERLOH, 1981; CARPENTER et al., 1990)

[17] als Zusammenschluss verschiedener eindimensionaler Einzeltests

[18] z.B. bestimmter Patientengruppen beim Wechsler-Intelligenz-Test

Möglichkeiten, um den Umgang mit multidimensionalen Konstrukten oder polytomen item- und geschwindigkeitsabhängigen Antwortformaten zu ermöglichen. Auch den zur Itemlösung notwendigen kognitiven Operationen oder Lerneffekten bei Testanwendungen kann mit entsprechenden Modellen auf die Spur gekommen werden.
Deutliche **Einschränkungen** der Anwendungsmöglichkeiten ergeben sich, weil realistischere, komplexere Modelle zusätzliche Parameter erfordern. Diese sind in annähernd realisierbaren Personenstichproben - auch mit modernen Algorithmen - häufig kaum identifizierbar oder bleiben nicht replizierbar. Als noch entscheidenderer Einwand, gehen bei fast allen Modellen die äußerst vorteilhaften Eigenschaften „erschöpfender" oder „suffizienter Statistiken" verloren. Die Anzahl gelöster Aufgaben repräsentiert dann nicht mehr die gesamte Information über die in Frage stehende Testleistung einer Versuchsperson.
Einen Überblick bisheriger Modellentwicklungen geben LINDEN & HAMBLETON (1997).

Klassische Testkonstruktion ist insbesondere bei folgender Frage entscheidend auf die messtheoretische Grundlage probabilistischer Testtheorie angewiesen:

Darf die mit einem Test per Summenscore bestimmte Anzahl richtiger Itemlösungen als Maßzahl einer allgemeingültigen latenten Fähigkeit der Population betrachtet werden?

Die klassische Testtheorie selbst stellt sich nicht dem Problem, dass mit Itembeantwortungen nicht in jedem Fall ein empirisch abgegrenztes und nur dann skalierbares Merkmal der Gesamtpopulation vorliegt. Es ist nicht auszuschließen, dass durch diese Vereinfachung, an einem **für die psychologische Diagnostik entscheidenden Punkt**, wertvolle Zusatzinformation verloren geht, weil Subpopulationen unerkannt bleiben:

„Die Daten der Psychologie stammen von Personen mit Individualität, d.h. sie erfüllen die Bedingung identischer Verteilung nicht" (SIXTL, 1999, S. 519).

Nur **Testmodelle mit „erschöpfenden Statistiken"** (z.B. Raschmodell) unterscheiden **Zusatzinformationen in den Daten** von sämtlicher im Summenscore enthaltener quantitativer Testleistungsinformation. Fehlende Raschmodellpassung drückt somit immer aus, dass zusätzliche qualitative Information vorhanden ist. Diese kann aus den verwendeten Items, den getesteten Personen, den Beantwortungsmöglichkeiten oder der latenten Variable resultieren (ROST, 2002). Zusatzinformation ist unabhängig von der Anzahl gelöster Items

(Summenscore) und abhängig davon, welche Items von Versuchspersonen gelöst werden. Der üblicherweise vernachlässigte Informationsanteil stört jede quantitative Aussage über die Testleistung einer Person als ungeklärter Messfehleranteil.

Für die Aufdeckung qualitativer Informationen oder latenter Strukturen in den Daten sind **Mischverteilungsmodelle** (z.B. Latent Class Analyse, Mixed-Rasch-Modell) (ROST & ERDFELDER, 1996) von besonderem Interesse. Mit diesen hochentwickelten statistischen Modellen können latente Subpopulationen von Versuchspersonen identifiziert werden, in denen verschiedene statistische Modelleigenschaften gelten. Mischverteilungsmodelle eignen sich zur Erfassung qualitativer Personenunterschiede und stehen probabilistischen Methoden zur Messung quantitativer Unterschiede in nichts nach. Sie werden „ebenso wie diese den Erfordernissen einer wissenschaftstheoretischen Position des kritischen Rationalismus gerecht“ (ROST, 1995, S. 393). Der große Nutzen besteht darin, dass eine **Verschiedenartigkeit menschlicher Reaktionsweisen auf Testitems** im zugrunde gelegten Mischverteilungsmodell Berücksichtigung findet. Deshalb müssen keine restriktiven Anforderungen an die verwendbaren Testitems bezüglich Homogenität oder Eindimensionalität gestellt werden, wie sie zur Feststellung quantitativer Unterschiede in der Gesamtpopulation mit dem Raschmodell notwendigerweise erforderlich sind. Als Bonus kann zusätzliche, klassisch unbeachtete und vernachlässigte, eventuell entscheidungsrelevante qualitative Information aufgespürt werden. Diese existiert, weil in den Daten der Gesamtstichprobe heterogene Personenpopulationen enthalten sind, die - aufgedeckt mit Mischverteilungsmodellen - als latente Klassen bezeichnet werden.

Eine **Berücksichtigung qualitativer Aspekte** einer Testbearbeitung zur Differenzierung von Testpersonen anhand der Klassenparameter eignet sich prinzipiell auch für den praktischen diagnostischen Alltag, weil sich für jede Testperson qualitative Messgrößen aus dem persönlichen Antwortmuster ergeben. Aus diesem lassen sich computerunterstützt individuelle Zugehörigkeitswahrscheinlichkeiten zu jeder einzelnen latenten Klasse berechnen[19]. Die maximale Zuordnungssicherheit einer Testperson zu einer latenten Klasse bestimmt einen nominal skalierten personenspezifischen Klassenparameter, der die Zugehörigkeit zu einer Klasse festlegt.

Die **testdiagnostischen Möglichkeiten und die historischen Vorbehalte** einer Messung qualitativer Unterschiede mittels Typenkonzepten wurden bei GIEGLER & ROST (1993) sowie ROST (1995) behandelt.

[19] die zugehörige Berechnungsformel findet sich in *Tabelle A-49* im Anhang.

Zwei Arten von Mischverteilungsmodellen werden in dieser Arbeit verwendet:

i) Die **„Latente-Klassen-Analyse“** (LAZARSFELD & HENRY, 1968) ist eine Möglichkeit, um unter Vermeidung einschränkender Voraussetzungen bezüglich der Dimension und Metrik eines Merkmalsraums und der einzelnen manifesten Variablen (Items) qualitative Testbearbeitungsunterschiede aufzudecken.
Jede einzelne identifizierte latente Klasse zeichnet sich durch größtmögliche Homogenität, bei gleichzeitig maximaler Heterogenität zwischen den Klassen aus (vgl. auch *Abschnitt 3.2*).
Die Latente-Klassen-Analyse wurde bei folgenden Untersuchungen im Testbereich angewendet:
KUBINGER, FORMANN & SCHUBERT (1980) verwendeten die Latente-Klassen-Analyse zur Identifizierung von Lösungsprofilen einer aus Subtests des Hamburg-Wechsler-Intelligenz-Tests für Kinder (HAWIK) zusammengestellten Testbatterie (vgl. auch KUBINGER, 1983b).
GIEGLER & ROST (1993) untersuchten mit Latenter-Klassen-Analyse Typenbildungen des Persönlichkeitsinventars „Gießen-Test“.
TARNAI & DOTTERWEICH (1996) verwendeten die Latente-Klassen-Analyse beim Allgemeinen-Interessen-Struktur-Test (A-I-S-T) und prüften, wie unterschiedliche Referenzstichproben die Klassifizierung von Personen beeinflussen.

ii) Speziell das **„Mixed-Rasch-Modell“** (ROST, 1990) wurde unter Verzicht auf die äußerst strenge Raschmodell-Voraussetzung einer personenhomogenen Gesamtpopulation für die Identifizierung latenter raschhomogener Populationsanteile entwickelt.
Wird eine Population von diesem Modell wegen qualitativer Zusatzinformationen in mehrere Klassen aufgeteilt, so unterscheiden sich sämtliche Klassen in den Itemschwierigkeiten. Dieselben Items, die demnach für Personen in einer Klasse relativ schwierig sind, können für Personen einer anderen Klasse sehr einfach zu lösen sein. Ein solcher Befund kann auch mit qualitativ unterschiedlichen Itembearbeitungsstrategien interpretiert werden (vgl. auch die *Abschnitte 1.10* und *3.2*). Innerhalb der Klassen besitzt das Raschmodell für die quantitative Bestimmung der Testleistung per Summenscore Gültigkeit.

1.10 Identifizierung unterschiedlicher Bearbeitungsstrategien mit Mischverteilungsmodellen

KÖLLER et al. (1994) konnten unter Anwendung des Mixed-Rasch-Modells und der Latenten-Klassen-Analyse (LCA) **unterschiedliche Lösungsstrategien bei den Würfelaufgaben des IST-70 Intelligenzstrukturtests** von AMTHAUER (1973) aufzeigen. Die Autoren wählten die zwei verschiedenen Analysemöglichkeiten aus, „um an einer Fragestellung beispielhaft zeigen zu können, wie sie mit alternativen probabilistischen Modellen adäquat beantwortet werden kann“ (KÖLLER et al., 1994, S. 71).

Während die Lösungsstrategie für **„Raumwürfel“** eine mentale Rotation im dreidimensionalen Raum darstellt, kann zur Lösung von **„Flächenwürfeln“** auch alternativ die als platte Ebene eingeprägte Würfeloberfläche verwendet werden. Dabei wird eine gedanklich zweidimensionale Rotation um den Umkreismittelpunkt ausgeführt.

Diese zwei latenten Bearbeitungsstrategietypen konnten in einer Stichprobe von N=264 Personen mit dem Mixed-Rasch-Modell identifiziert werden:

Die so genannten **„Holistiker“** lösten „Raumwürfel“ mit vergleichsweise höherer und „Flächenwürfel“ mit niedrigerer Wahrscheinlichkeit.

Die **„Analytiker“** im Gegensatz dazu, lösten „Flächenwürfel“ mit hoher und „Raumwürfel“ mit relativ niedriger Wahrscheinlichkeit.

In einem zweiten Durchgang, diesmal mit N=2558 Schülern, zeigten sich analoge Ergebnisse unter Verwendung einer Latenten-Klassen-Analyse. Ein Vergleich mit den Resultaten aus dem Mixed-Rasch-Modell verdeutlichte, dass „Holistikern“ und „Strategieflexiblen“ sowie „Analytikern“ und „Flächenstrategen“ ähnliche Lösungsprofile zuzuordnen waren. Weitere mit Latenter-Klassen-Analyse identifizierte Profile konnten auf motivationale Faktoren in der Testbearbeitung zurückgeführt werden.

Relevanz für die Praxis der Testdiagnostik zeigt sich in diesen Ergebnissen, weil Antwortmuster verwendet werden können, um Personengruppen zu identifizieren, die gleiche Testitems mit unterschiedlichen Lösungsstrategien bearbeiten. Dieser Gesichtspunkt wurde bei HOSENFELD et al. (1997) zum Thema einer Untersuchung, die sich mit den Geschlechtsunterschieden von „Holistikern“ oder „Analytikern“ befasst hat.

Bezogen auf die einzelne Testperson lässt sich im diagnostischen Alltag mit Hilfe einer **computergestützten Umsetzung** der Resultate von KÖLLER et al. nach jeder Testdurchführung eine Wahrscheinlichkeitsaussage darüber machen, welche Bearbeitungsstrategie im Test verwendet wurde. Eine Studie von GLÜCK et al. (2002)

befasste sich mit den Auswirkungen eines mehrtägigen, individuellen Lerntrainings mentaler Rotation auf die Strategieverwendung bei Würfeltests.

KINDERMANN (2001) musste nach Analyse einer neuen - auf 20 Items verlängerten - Form des Bentontests, die Annahme der Raschhomogenität verwerfen. In einer Gesamtstichprobe von N=1124 klinisch auffälligen und unauffälligen Personen sowie in einer Teilstichprobe ausschließlich zusammengesetzt aus klinisch unauffälligen Personen (N=885) konnte er jeweils eine **2-Klassen-Lösung** des Mixed-Rasch-Modells identifizieren.
Bei der genaueren Betrachtung der 2-Klassen-Lösung der Gesamtstichprobe überraschte, dass **93 Prozent der Patienten mit unterschiedlichsten psychiatrischen und neurologischen Diagnosen in nur <u>einer</u> der beiden raschhomogenen Klassen** enthalten waren. In den gemittelten Itemlösungswahrscheinlichkeiten schnitt diese - Patienten beinhaltende - Klasse außerdem weniger erfolgreich ab, als eine vergleichbare raschhomogene Klasse, die sich nur aus klinisch unauffälligen Personen (N=885) zusammensetzte (vgl. *Tabelle 1-1*).

Tabelle 1-1 mittlere klassenspezifische Itemlösungswahrscheinlichkeiten (KINDERMANN, 2001)

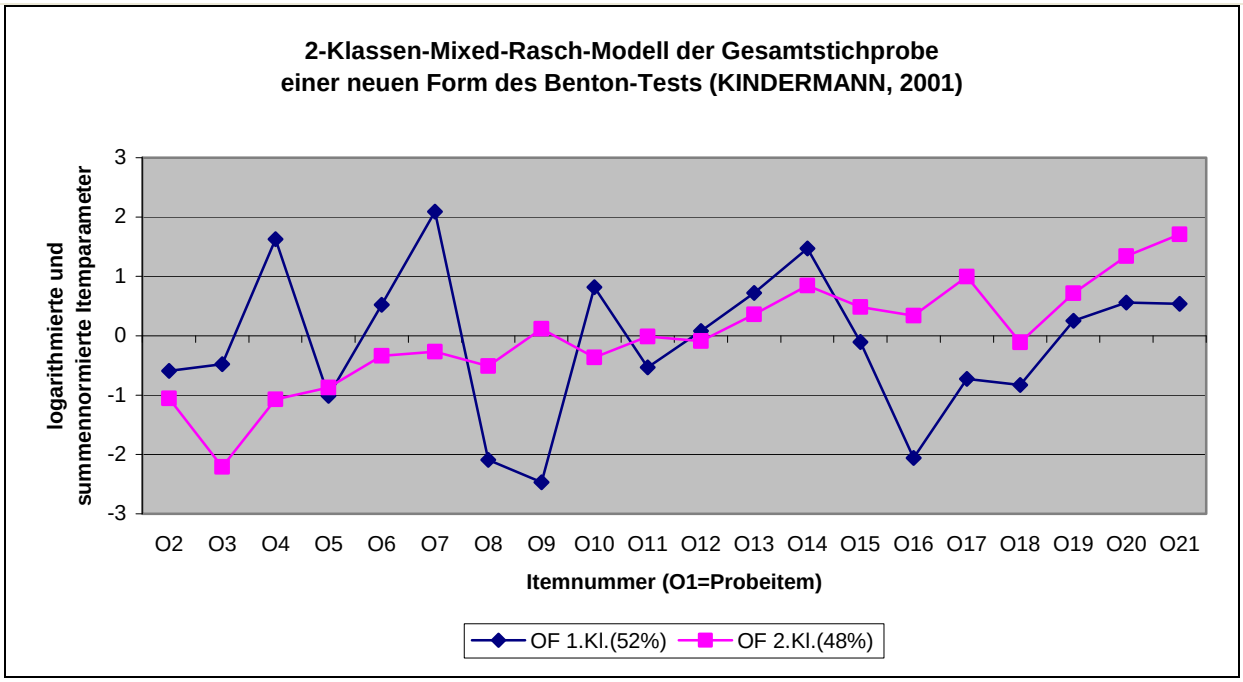
Zwei-Klassen Mixed-Rasch-Modelle einer neuen Version des Bentontests	**Gesamtstichprobe mit Patienten (N=1124) (vgl. auch *Abb. 1-1*)**	Unter-schied Ilwk	**Teilstichprobe ohne Patienten (N=885) (ohne Abb.)**
mittlere Ilwk sämtlicher Items in Klasse 1	0,533 (inkl. 93% des Patientenanteils)	<	0,623 (ohne Patientenanteil)
mittlere Ilwk sämtlicher Items in Klasse 2	0,663 (inkl. 7% des Patientenanteils)	≅	0,666 (ohne Patientenanteil)

Ilwk=Itemlösungswahrscheinlichkeit

Der vorliegende Befund lässt vermuten, dass Testpersonen zur Lösung von Benton-Items zwei **unterschiedliche Bearbeitungsstrategien** verwenden. Alternativ ist auch an klassenspezifische Aufmerksamkeitsverläufe zu denken.
In *Abbildung 1-1* sind die auffallend stark variierenden Profilverläufe der Itemparameter der zwei Klassen des Mixed-Rasch-Modells der Gesamtstichprobe dargestellt. Die 93 Prozent der klinisch auffälligen Personen befinden sich in Klasse 1[20].

[20] zusammen mit immerhin auch 41 Prozent aller klinisch unauffälligen Personen

Abbildung 1-1 Itemschwierigkeitsprofile (KINDERMANN, 2001)



Die Ergebnisse von KINDERMANN (2001) beim Benton-Test - prinzipiell ein Analogon zu den Würfelaufgaben des IST-70 von KÖLLER et al. (1994) - legen praktische Konsequenzen für den Testanwender nahe:

Die Vergleichbarkeit der Testleistung per Summenscore hängt davon ab, ob Versuchspersonen der einen oder anderen Klasse zugeordnet werden.

Die von KINDERMANN (2001) identifizierten Klassen implizieren, dass sich die Aufdeckung qualitativer Testbearbeitungsunterschiede mit Mischverteilungsmodellen durch Berechnung von Klassenzugehörigkeiten auch **für diagnostisch relevante Entscheidungen** eignen kann. Im Wert des Summenscores scheint zusätzlich zur Testleistungsinformation ein spezifischer Varianzanteil innerhalb der Antwortmuster von Testpersonen enthalten zu sein, der - hier erstmalig - aus pathologischen Prozessen oder klinisch relevanten Zusatzinformationen resultiert. Gelingt es also in einem ersten Schritt, anhand der Klassenzugehörigkeit mit hinreichender Sicherheit Unterscheidungsmerkmale dahingehend zu identifizieren, welche Lösungsstrategie angewendet wurde, gibt die erreichte Leistungsmenge in Form des dann klassenspezifischen Summenscores[21] erst in einem zweiten Schritt den entscheidenden prognostischen Ausschlag dafür, ob die Testergebnisse eines

[21] oder entsprechend Personenparameters

beliebigen Probanden klinische Auffälligkeiten vermuten lassen.
Ausgangspunkt der vorliegenden Untersuchung ist die Annahme, dass bei der Replizierung derartiger Ergebnisse beim Bentontest in der Kreuzvalidierung, **validere Prädiktionen pathologischer Prozesse** gelingen können, wenn Klassenzugehörigkeiten, Testscores und deren Interaktionen gemeinsam betrachtet werden.

1.11 Kritische Anmerkungen zu Untersuchungen der Vorhersagevalidität von Testverfahren

Als **Validität** eines Tests wird das Ausmaß dessen bezeichnet, mit dem der Test das misst, was er zu messen vorgibt. Die Konzepte der Validitätsbestimmung lassen sich allgemein in zwei große Gruppen unterteilen (RAUCHFLEISCH, 1989):

1) Korrelation mit einem Kriterium: Innere und äußere Validierung
2) Konstruktvalidierung

Die Korrelationsinterpretation setzt einen linearen Zusammenhang zwischen einem Testmaß und einem Kriterium voraus. Korreliert ein Testmaß hoch mit anderen als valide anerkannten Tests, die dasselbe Persönlichkeitsmerkmal erfassen, wird von der **inneren Validität** des Testverfahrens gesprochen.
ROST (2004) verwendet den Begriff **interner Validierung** auch dann, wenn die Gültigkeit eines Testmodells anhand der Testdaten überprüft wird. Im Mittelpunkt des Interesses steht dabei, ob Versuchspersonen auf vorgelegte Testitems in einer Weise antworten, wie es die zugrunde gelegte Theorie über die zu bestimmende Personeneigenschaft erwarten lässt.
Man spricht von **äußerer** oder **externer Validität**, wenn der Zusammenhang eines Testergebnisses mit äußeren Variablen geprüft wird, die der Test erfassen soll. Externe Validität setzt voraus, dass eine **andere Messung** dessen, was gemessen werden soll, verfügbar ist. Diese liegt bei klinischen Tests dann vor, wenn Messungen mit dem Testverfahren und der Diagnosezuschreibung hoch korrelieren. Über die Diagnosezuschreibung gibt WITTLING (1983, S. 203) zu bedenken, dass häufig als Außenkriterium verwendete Untersuchungsmethoden, wie etwa psychiatrisches Urteil oder neurologische Standarduntersuchungen selbst „nachgewiesenermaßen invalide Außenkriterien“ darstellen. Zudem beinhaltet eine Patientenstichprobe häufig Personen mit heterogenen Diagnosen, welche zugleich im Querschnitt ohne Beachtung spezieller

Krankheitsverlaufscharakteristika untersucht wurden. Externe Validitätsbestimmungen[22] von Testverfahren greifen im klinischen Bereich oftmals aus zeitökonomischen Gründen auf dieses - wegen des verwendeten unsicheren Außenkriteriums - problematisch zu bezeichnende Vorgehen (STECK, 1991) zurück. Die Validierungsstudien von RAUCHFLEISCH (1983), WYTEK et al. (1984) beim Raventest und STECK (1990), SCHEURER et al. (1994) beim Benton-Test sind in diesem Sinne angelegt.
Auch die zur Korrelation erforderliche **zweite Messung** in Form eines Testergebnisses, muss kritisch gesehen werden. Die Aufdeckung hirnorganischer Störungen[23] mittels kurzer testdiagnostischer Ausleseverfahren, gilt als veraltetes Ziel der klinischen Neuropsychologie. Unter Verwendung eines vormals eindimensionalen Hirnschädigungskonzeptes, wurde angenommen, dass sich Leistungseinschränkungen in bestimmten Tests ungeachtet ihrer Ätiologie und Lokalisation zeigen. Mit dem Einsatz moderner apparativer Diagnoseverfahren, wie Computer- (CT), Kernspin- (MRT) und Positronen-Emissions-Tomographie (PET), sollte ein derartiger Ansatz der Vergangenheit angehören (STURM & HARTJE, 1989; BENTON SIVAN & SPREEN, 1996; MERTEN, 1999).

Bei Validitätsstudien entziehen sich die Einflüsse vieler Störvariablen weitgehend der Kontrolle. Kennwerte der Validität von klassischen Testverfahren im klinischen Bereich sind deshalb nur äußerst vage Abschätzungen, die von folgenden Faktoren störend beeinflusst werden (vgl. WESTMEYER, 1972):

- veraltetes Testkonzept (eindimensionales Konzept von Beeinträchtigungen)
- fragliche Diagnosezuschreibung und inadäquate nosologische Beschreibung
- heterogene und eventuell multiple Diagnosen
- freiwillige Testteilnahme (deshalb eventuell bestimmte Patientensubgruppen)
- Querschnittsdesign klinischer Validitätsstudien
- Versuchsleitereffekte
- Umgebungsbedingungen

[22] prädiktive Validität oder Vorhersagevalidität als eine Form der externen Validität

[23] Hirnorganisch Erkrankte und Hirnverletzte (z.B. Hirntumore, Schädelhirnverletzungen und Hirngefäßstörungen) werden dabei mit charakteristischen Störungen des Gedächtnisses, des Denkens und der Affektivität in Verbindung gebracht (FRÜHAUF, 1975).

Weil Tests keine direkt beobachtbaren Merkmale erfassen[24], sondern nur bestimmte z.B. mit Krankheiten assoziierte Teilfunktionen objektivieren, schufen CRONBACH & MEEHL (1955) das Konzept der **Konstruktvalidierung**. Der Begriff umschreibt Analysevorgänge der Zusammenhänge zwischen abgeleiteten psychologischen Konstrukten und Testergebnissen. Die erforderlichen Validitätsbestimmungen können unter diesen Prämissen nicht metrisch erfolgen, weil Teilvarianzen in der Testleistung von Versuchspersonen aus unterschiedlichsten Störfaktoren resultieren und letztlich ungeklärt bleiben müssen. Konstruktvalidierung bleibt daher eine permanente Aufgabe, die es immer wieder notwendig machen wird, theoretische Ansätze zu modifizieren. Sie beinhaltet äußerst unterschiedliche Validierungsmethoden, deren Prüfung sich mangels operationaler Fassbarkeit der Konstrukte häufig schwierig gestaltet.
Die Kontrolle der inneren und äußeren Validität eines Testverfahrens bildet einen wesentlichen Gesichtspunkt des Vorgangs zur Ermittlung der Konstruktvalidität.

1.12 Allgemeine Bemerkungen zu den verwendeten Tests

Ausgangspunkt der vorliegenden Arbeit sind drei klassische psychometrische Fähigkeitstests, die in ursprünglicher (Cattelltest), abgewandelter (Raventest) oder neuer Testform (Bentontest[25]) verwendet worden sind. Zur Vereinfachung werden sie hier als Raven-, Benton-, und Cattelltest bezeichnet. Dieser *Abschnitt* erläutert ihre Anwendungsbereiche und die Ergebnisse früherer Raschanalysen. Eine ausführliche Darstellung des verwendeten Untersuchungsmaterials erfolgt in den *Abschnitten 4-2* bis *4-4*.

1.12.1 Raventest

Der Raventest nimmt laut CARPENTER et al. (1990) eine zentrale Stellung unter den allgemeinen Fähigkeitstests ein.
In den verschiedenen Ausführungen des Raventests[26] sollen unvollständige graphische Darstellungen ergänzt werden, wobei aus acht Lösungsmöglichkeiten[27] die richtige

[24] z.B. klinische Tests keine Krankheiten

[25] „neue Version des Bentontests“, STECK (2005)

[26] Raven`s Progressive Matrices (RPM) als Überbegriff folgender Raventest-Versionen: Standard Progressive Matrices (SPM), Advanced Progressive Matrices (APM) und Coloured Progressive Matrices (CPM))

[27] bei CPM sechs Lösungsmöglichkeiten

auszuwählen ist. „Da Wahrnehmung grundsätzlich ein Prozess des Begreifens ist, der die Fähigkeit beinhaltet, den Sinn im heilen Durcheinander zu erkennen, kann von Raven's Progressive Matrices (RPM) behauptet werden, dass sie die Fähigkeit des logischen Denkens und Erkennens messen. Spearman und andere Psychologen stimmen überein, dass der Generalfaktor (g-Faktor), gemessen mit den meisten Intelligenztests, aus zwei Hauptkomponenten besteht: der eduktiven und der reproduktiven Fähigkeit. Aus diesem Sichtwinkel messen die RPM eine der grundlegendsten menschlichen Fähigkeiten“ (SCHUHFRIED GmbH, 1995).

Die Advanced Progressive Matrices (APM) wurden entwickelt, weil sich aus der breiten Anwendung der Standard Progressive Matrices (SPM) zusätzlicher Bedarf für einen Test ergeben hatte, der auch leistungsstarke Testpersonen differenzieren kann (vgl. RAVEN, 2000).

PUTZ-OSTERLOH (1981) setzte sich kritisch mit einer Vielzahl faktorenanalytischer und korrelationsstatistischer Studien zu den Raven Matrizen Tests auseinander. Die APM verwendete sie als Untersuchungsmaterial für denkpsychologische Ansätze. Mit den Methoden des lauten Denkens und der Blickbewegungsanalyse konnten itemabhängige intra- und interindividuelle Problemlösestrategien und Problembearbeitungen aufgezeigt werden. „An intraindividuellen Unterschieden ist zu erkennen, inwiefern Tests wirklich konsistente Problemlösestrategien erfassen. An interindividuellen Unterschieden lässt sich die Eindeutigkeit der Interpretation von Testleistungen überprüfen. So halten wir es für wichtig, auch bei anderen Tests die Bedeutung der Lösungsalternativen zu prüfen“ (PUTZ-OSTERLOH, 1981, S. 200).

CARPENTER et al. (1990) haben sich mit ähnlichen Fragestellungen befasst und vergleichbare Resultate erhalten.

VEN & ELLIS (2000) führten eine Raschanalyse mit den Raven`s Standard Progressive Matrices (SPM) durch. Bei mehreren Untertests musste das Kriterium der Raschhomogenität verworfen werden.

Für die Advanced Progressive Matrices (APM) konnten vom Autor der vorliegenden Arbeit keine Studien zur Raschmodellgültigkeit gefunden werden.

1.12.2 Bentontest

Der Visual Retention Test nach A. L. Benton, genannt Bentontest, zählt zu den klassischen Hirnorganikertests. Er wurde „unter dem Druck der klinischen Praxis im Kriege“ (BENTON,

1996, S. 7) als so genanntes „allgemeines Ausleseverfahren“ für Hirnschädigungen konzipiert und später „auch bei anderen Erkrankungen mit Nutzen gebraucht“ (BENTON, 1996, S. 7).
Der Bentontest kann auf drei Arten angewendet werden: Er wird als Zeichenform zur Prüfung der visuellen Merkfähigkeit, als Abzeichenform zur Prüfung visuo-konstruktiver Fähigkeiten und in der Wahlform zur Prüfung der visuellen Rekognition eingesetzt.
BENTON SIVAN & SPREEN (1996, S. 12) charakterisieren das Verfahren wie folgt:
„Seit der ersten Veröffentlichung hat der Test als Instrument der klinischen und der Forschungs-Arbeit gedient, mit dem die visuelle Wahrnehmung, das visuelle Gedächtnis und visuo-konstruktive Fähigkeiten gemessen werden können. Ursprünglich wurde er als eine kurze Methode für die klinische Arbeit zur Identifizierung von offensichtlichen und manchmal auch wenig erkennbaren Schwierigkeiten des Gedächtnisses, der Raum-Orientierung und der Motorik entwickelt. Seither hat sich der Test aber auch als besser als die meisten anderen zur Differenzierung von Aufmerksamkeitsschwächen, die charakteristisch für viele psychiatrisch oder neurologisch auffällige Patienten sind, erwiesen.“
KINDERMANN (2001) und ZWESCHPER (1992) mussten die Annahme der Raschhomogenität bei den untersuchten Formen des Bentontests verwerfen.
Insbesondere bei KINDERMANN (2001) wies der Analysebefund darauf hin, dass in einer beliebigen Population zumindest zwei unterschiedliche Strategien oder latente Fähigkeiten bei der Lösung der Benton-Items Verwendung finden (vgl. *Abschnitt 1.10*).

1.12.3 Cattelltest

In den sechziger Jahren entwarf Raymond Bernard Cattell (1905-1998) das Musterbild klassisch quantitativer Testkonstruktion indem er „culture free tests“ entwickelte.
Das Ziel bei der Entwicklung derartiger Testverfahren bestand darin, dass diese unabhängig vom Kulturkreis reliable und somit vergleichbare Testresultate ergeben können.
Nachdem das ursprüngliche Vorhaben, kulturfreie Tests zu konstruieren, immer mehr relativiert werden musste, hatte sich eine Namensänderung in „culture fair tests“ (CFT) durchgesetzt (ANASTASI, 1985).
Mit der deutschen Adaption des als „Grundintelligenztest“ veröffentlichten CFT-3 (WEIß, 1971), soll ein spezifischer Intelligenzfaktor Cattell`scher Terminologie - die „flüssige Intelligenz“ - erfasst werden. Die erfolglose Konstruktvalidierung von LANGFELDT-NAGEL (1982) stellt diese ursprüngliche Intention des Tests in Frage.

Der in der vorliegenden Arbeit verwendete, dem CFT-3[28] entnommene Subtest „Matrizen“, verlangt vom Probanden figurale Muster zu vervollständigen. Dieser Subtest unterscheidet sich vom zeitunabhängigen Raventest dadurch, dass bei begrenzter Zeit aus sechs Alternativen die richtige Lösung auszuwählen ist.

EDJIEH (1980) hat eine für Raschanalysen erforderliche Testvorgabe ohne Zeitbegrenzung durchgeführt. Unter Verwendung der damals üblichen graphischen Modellkontrolle zeigte sich, dass die 13 Items des „Matrizen“-Subtests nicht den Anforderungen des Raschmodells genügen.

[28] CFT-3, 2.Teil, Test 3 „Matrizen“

2 FRAGESTELLUNG

In dieser Arbeit werden drei interkorrelierte „Klassiker“ der testpsychologischen Diagnostik - Raven-, Benton- und Cattelltest - jeweils separat und als Testbatterie eingehenden testtheoretischen Prüfungen unterzogen. Am Beispiel der Vorhersagevalidität klinischer Kriterien wird der klassische mit dem probabilistischen Theorieansatz verglichen. Dabei lautet die zentrale Ausgangsfrage:

Liefert eine mittels stochastischer Testmodelle identifizierte Art der Testbearbeitung einen klinisch-prognostisch verwertbaren Informationsgewinn?

Dieses zentrale Problem gliedert sich in die folgenden sechs Fragestellungen.

2.1 Fragestellung 1: Werden von stochastischen Testmodellen bei den einzelnen Testverfahren unterschiedliche Arten der Testbearbeitung (latente Klassen) identifiziert?

Bei den beiden Power-Tests[29], dem Raven- und Bentontest, soll die Gültigkeit des Raschmodells in Bezug auf seine zentrale Modellannahme der Personenhomogenität geprüft werden. Diese besagt, dass alle getesteten Personen den jeweiligen Test mit derselben Eigenschaft oder Fähigkeit bearbeiten.

Für die Prüfung der Power-Tests wird ein externes und internes Teilungskriterium verwendet.

Das *externe Teilungskriterium* leitet sich aus der „diagnostischen Zuordnung“ ab. Die Gesamtstichprobe wird in klinisch auffällige und unauffällige Personen geteilt.

Für das *interne Teilungskriterium* wird die „Testleistung“ herangezogen. Die Gruppe der Hoch- und Niedrigscorer wird über einen Mediansplit ermittelt.

Zu beachten ist, dass eine Prüfung der Raschhomogenität beim Cattelltest nicht möglich ist, weil Speed-Power-Tests[30] so konzipiert sind, dass Versuchspersonen unter Zeitdruck nicht zur Bearbeitung aller Aufgaben kommen.

---

[29] Als Power-Tests werden Tests bezeichnet, die sich aus Items unterschiedlicher Schwierigkeit zusammensetzen und ohne Zeitdruck bearbeitet werden.

[30] Als Speed-Power-Tests werden Tests bezeichnet, die sich aus Items unterschiedlicher Schwierigkeit zusammensetzen und außerdem unter Zeitdruck bearbeitet werden.

Alle im folgenden genannten Modelltests, die im Theorieteil ausführlicher dargestellt werden, überprüfen die Personenhomogenität als die zentrale Annahme des Raschmodells.
Das leitende Kriterium bleibt bei allen Prüfungen unverändert: Sind die geschätzten Itemparameter unabhängig von den jeweils verwendeten Teilstichproben?

2.1.1 Graphischer Modelltest

Der Graphische Modelltest wurde schon von RASCH (1960) angewendet. Er ist kein statistischer Modelltest. Raschmodellgültigkeit bedeutet, dass sich für sämtliche Teilstichproben gleiche Itemparameterschätzungen ergeben. Abweichungen von der 45^0-Linie im Schaubild zeigen graphisch Modellverletzungen an, da sich die geschätzten Itemparameter für verschiedene Teilstichproben unterscheiden.

2.1.2 Andersen-Test

Beim Andersen-Test handelt es sich um einen statistischen Modellgeltungstest. Anhand der manifesten Teilungskriterien „diagnostische Zuordnung“ und „Testleistung“ wird die vom Raschmodell geforderte Personenhomogenität geprüft. Das von ANDERSEN (1973) entwickelte und für die bedingte Likelihood bewiesene Testverfahren, beruht auf dem Vergleich von Personengruppen mit gleich vielen richtigen Antworten. Weil das verwendete Programm WINMIRA 2001 1.36 (DAVIER, 2001) die logarithmierte marginale Likelihood berechnet, wird die Prüfung der Raschhomogenität mit der marginalen Likelihood durchgeführt. Diese verläuft rechnerisch äquivalent zum bedingten Likelihoodquotiententest. Hierbei werden vom WINMIRA-Programm normalverteilte Scores oder Personenparameter angenommen, um die Anzahl der zu schätzenden Modellparameter gering zu halten.
Für FISCHER (1974, S.300) ist dieser Test „das exakteste der derzeit bekannten Verfahren zur globalen Modellkontrolle“.
Getestet werden die Hypothesen:

H_0 : Der Raven/Bentontest ist modellkonform, d.h. raschhomogen.

H_1 : Die Raven/Bentontest ist nicht modellkonform.

Signifikanzniveau α=5%

2.1.3 Prüfung der Modellgeltung mittels Informationstheoretischer Maße

Informationstheoretische Maße wählen das beste Modell durch einen optimalen Kompromiss zwischen der Genauigkeit der Datenbeschreibung und der Sparsamkeit der dazu nötigen Modellparameter aus. Sie setzen den Wert der Likelihoodfunktion mit der Parameteranzahl in Beziehung und erlauben so eine Beurteilung der Modellgeltung.

Je kleiner ein informationstheoretisches Maß (z.B. AIC, BIC) ist, desto besser passt ein Modell.

Der Vorteil ist, dass auch Modelle miteinander verglichen werden können, die in keiner hierarchischen Beziehung stehen (Latent-Class ⇔ Latent-Trait).

Der Nachteil ist, dass es sich um keinen statistischen Signifikanztest handelt.

In dieser Arbeit wird zur Abschätzung der Modellpassung aller drei verwendeten Tests aus pragmatischen Gründen der AIC-Wert verwendet. Dieser ist weniger konservativ als der ebenfalls aufgeführte BIC-Wert und indiziert gegebenenfalls Modelle mit größeren Klassenanzahlen, die differenziertere diagnostische Aussagen ermöglichen.

Die vom informationstheoretischen Maß angezeigten Klassen-Lösungen aller Tests werden mit Likelihoodquotiententests abgesichert.

2.2 Fragestellung 2: Wie können identifizierte Arten der Testbearbeitung (latente Klassen) bei den drei Tests interpretiert werden?

Abhängig von den Ergebnissen aus Fragestellung 1 sollen die identifizierten Klassenmodelle vorgestellt werden.

Klassenspezifisch betrachtet werden:

a) die Verteilung der Summenscores
b) die zugeordneten Personen hinsichtlich Testleistung, Alter, Bildung und klinischer Auffälligkeit[31]
c) die geschätzten Itemparameterprofile der einzelnen Tests.

[31] Klassenunterschiede sollen außerdem statistisch beurteilt werden: Signifikanzniveau α=5%, t-Tests von Testleistung und Alter mit zweiseitiger Testung

2.3 Fragestellung 3: Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen der einzelnen Tests?

Die diagnostische Relevanz der einzelnen latenten Klassen soll durch den Vergleich des klassenspezifischen, klinisch auffälligen und unauffälligen Personenanteils differenziert betrachtet und statistisch geprüft werden. Dabei ist abzuklären, inwiefern sich Testleistungsunterschiede innerhalb der einzelnen Klassen ergeben.
Verglichen werden die klassenspezifischen Testleistungen klinisch unauffälliger Personen mit

i) klinisch auffälligen Personen und
ii) Patienten mit cerebraler Beteiligung

Signifikanzniveau α=5%, t-Tests mit zweiseitiger Testung

2.4 Fragestellung 4: Können die identifizierten Arten der Testbearbeitung die Vorhersagevalidität der drei Tests verbessern?

Unter Verwendung von Kennwerten aus nominal logistischen Regressionen soll die Vorhersagevalidität der Tests bezüglich klinischer Kriterien beurteilt werden.
Mit den Prädiktoren

a) Klassischer Prädiktor (Summenscore)
b) Probabilistischer Prädiktor (Klassenzugehörigkeit) und
c) Kombinierte Prädiktoren (Summenscore, Klassenzugehörigkeit und deren Interaktion)

wird die Prädiktion der klinischen Kriterien

i) klinische Auffälligkeit und
ii) cerebrale Beteiligung

geprüft, beurteilt und verglichen.
Die Hypothesen der logistischen Regressionen lauten:

H_0 : Der Einfluss der im logistischen Modell verwendeten Prädiktoren auf das Kriterium ist Null (H_0: β_i=0)

H_1 : Der Einfluss der im logistischen Modell verwendeten Prädiktoren auf das Kriterium ist von Null verschieden (H_1: $\beta_i \neq 0$)

Signifikanzniveau α=5%

2.5 Fragestellung 5: Ergibt sich durch eine Kombination der Einzeltests zu einer Testbatterie eine Verbesserung der Vorhersagevalidität?

Die Vorhersagevalidität der gesamten Testbatterie wird analog zur Fragestellung 4 geprüft.

2.6 Fragestellung 6: Können Lösungsmuster oder Typen der Testbatteriebearbeitung identifiziert und differentialdiagnostisch genutzt werden?

Es soll untersucht werden, ob mit einer Latenten-Klassen-Analyse (LCA) der gesamten Testbatterie, klinisch relevante Lösungstypen identifiziert werden können.
Mit den nachfolgend genauer zu erläuternden Auswertungsmodi „klassisch“ (LCA der Summenscorequartile) und „probabilistisch“ (LCA der Klassenzugehörigkeiten und klassenspezifischen Summenscorequartile), werden jeweils separat die oben genannten Fragestellungen 1) bis 4) bearbeitet:

1) **Werden unterschiedliche Bearbeitungstypen (latente Klassen) der gesamten Testbatterie identifiziert?**

   a) Prüfung der Gültigkeit von Latent Class Modellen mittels informationstheoretischer Maße

   b) Absicherung der Ergebnisse von a) mit Likelihoodquotiententests

2) **Wie können die von der LCA identifizierten latenten Klassen interpretiert werden?**

3) **Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen in den einzelnen Testergebnissen?**

4) **Bieten die identifizierten Bearbeitungstypen (latenten Klassen) der Testbatterie diagnostisch verwertbaren Informationsgewinn?**

3 THEORIE PROBABILISTISCHER TESTANALYSEN

3.1 Das dichotome logistische Modell nach RASCH

Ausgehend von seinen Überlegungen zur Poisson-Funktion entdeckt RASCH (1960) eine probabilistische Funktionsklasse, deren fundamentale Eigenschaften sich auf die Separierbarkeit von Modellparametern beziehen:

a) Durch den *Personenparameter* ξ_v lässt sich jede Person *v* hinsichtlich ihrer Fähigkeit, ein bestimmtes Testitem zu lösen, auf einer eindimensionalen Skala charakterisieren.
b) Durch den *Itemparameter* σ_i lässt sich jedes Item *i* hinsichtlich seiner Schwierigkeit auf einer eindimensionalen Skala charakterisieren.
c) Beide Parameter lassen sich gemeinsam auf derselben Skala abbilden.
d) Es existiert ein probabilistischer Zusammenhang zwischen der Lösung eines Items und den beiden Parametern.

Aus der notwendigen und hinreichenden Bedingung der erschöpfenden oder suffizienten Statistik folgt eindeutig die „Logistische Funktion", das sog. Rasch-Modell:

$$(3.1) \qquad p_{vi} = p(x_{vi} = 1 | \xi_v, \sigma_i) = f_i(\xi_v) = \frac{e^{(\xi_v - \sigma_i)}}{1 + e^{(\xi_v - \sigma_i)}}$$

(x_{vi}=1 bedeutet, dass die *v*-te Versuchsperson die *i*-te Aufgabe gelöst hat, x_{vi}=0 bedeutet, dass dieselbe Aufgabe nicht gelöst wurde)

oder nach antilogarithmischer Transformation:

$$(3.2) \qquad p(x_{vi} = 1 | \theta_v, \varepsilon_i) = \frac{\theta_v \varepsilon_i}{1 + \theta_v \varepsilon_i}$$

wobei: $\theta_v = e^{\xi_v}$: Personenparameter einer Person *v*,

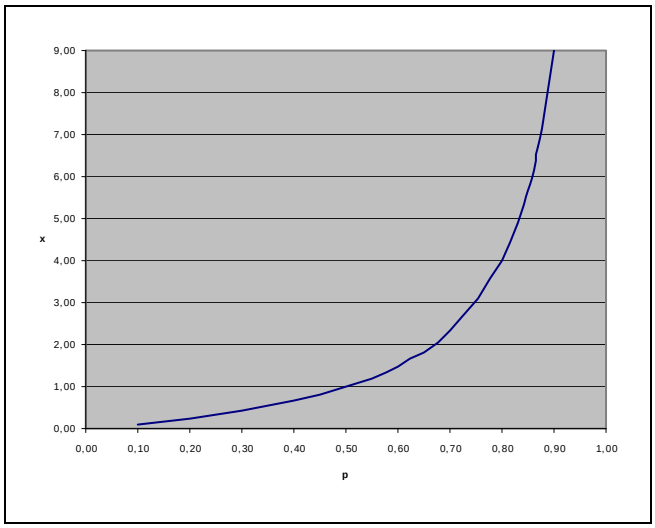
$\varepsilon_i = e^{-\sigma_i}$: Itemparameter des Items i.

Nur in diesem Modell ist die Zahl x_{v0} der von einer Versuchsperson gelösten Aufgaben eine erschöpfende Statistik für den Personenparameter ξ_v. Identische Rohwerte indizieren nur dann unabhängig vom Lösungsmuster gleiche Merkmalsausprägungen oder latente Fähigkeitswerte. In jedem anderen Testmodell ist die Zahl der richtigen Antworten einer Versuchsperson - der Summenscore - als Rohscore ihrer Leistung „ungerechtfertigt", da er mit Informationsverlust verbunden ist.
Diese erschöpfende Statistik gilt jedoch nur unter der Voraussetzung, dass alle Items des Tests bearbeitet wurden, weil es nur dann zufallsabhängig sein kann, ob eine Testperson eher leichte oder auch schwierige Items löst. Das Raschmodell eignet sich aus diesem Grund nicht zur Analyse von Tests, die unter Zeitdruck bearbeitet werden müssen.

Mögliche Herleitung der logistischen Funktion in drei Schritten (ROST, 1996; NAGL, 2005):
1.) Als Ausgangspunkt dient die Darstellung der Lösungswahrscheinlichkeit p als Wettquotient x oder **odds-ratio** (*Abbildung 3-1*):

$$(3.3) \qquad x = \frac{p}{1-p};$$

Abbildung 3-1 Wahrscheinlichkeit als odds-ratio

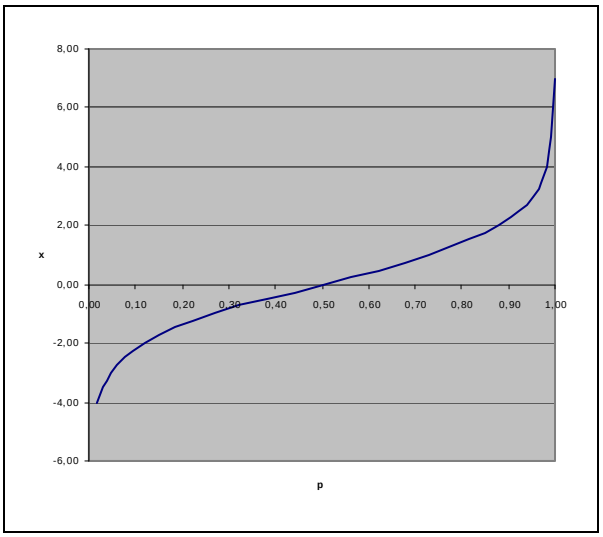


2.) Aus der Logarithmierung des Wettquotienten, ergibt sich eine symmetrische Intervallerweiterung, (wobei x=0 bei p=0,5). Das Intervall [0,1] wird mit dieser, als **logit-Transformation** bezeichneten Umwandlung, auf den gesamten Zahlenstrahl abgebildet. So

wird jeder Wahrscheinlichkeit p eine reelle Zahl zugeordnet, die zwischen $-\infty$ und $+\infty$ liegt (dargestellt in *Abb. 3-2*):

$$(3.4) \qquad x = \ln \frac{p}{1-p} = \text{logit (p)};$$

Abbildung 3-2 Logit der Lösungswahrscheinlichkeit



3.) Um die Itemcharakteristiken (IC) zu erhalten wird die **Umkehrfunktion** durch Umformung gebildet. (Berechnung vgl. *Infokasten* unten) (*Abbildung 3-3*):

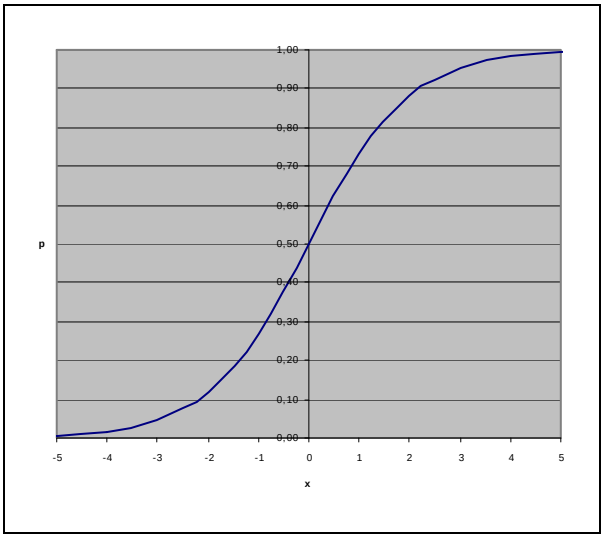
$$(3.5) \qquad p = \frac{e^x}{1+e^x};$$

Infokasten: Berechnungsschritte der Umformung zur Herleitung der logistischen Funktion

$$x = \ln \frac{p}{1-p} \Leftrightarrow e^x = \frac{p}{1-p} \Leftrightarrow (1-p)e^x = p \Leftrightarrow e^x - e^x p = p \Leftrightarrow$$

$$p + e^x p = e^x \Leftrightarrow p(1+e^x) = e^x \Leftrightarrow p = \frac{e^x}{1+e^x}$$

Abbildung 3-3 Logistische Funktion

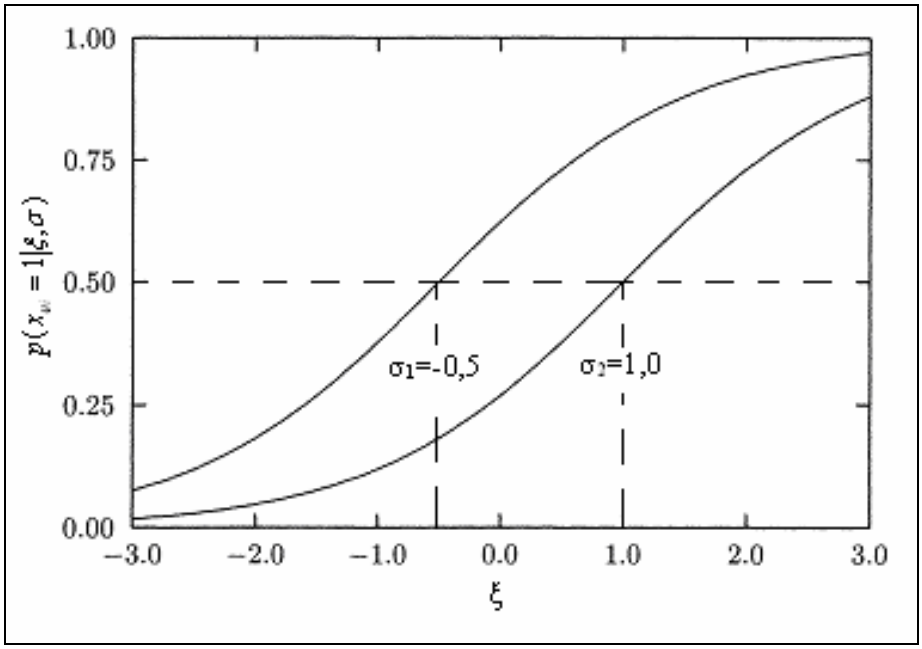


Im Raschmodell gilt: x=(ξ - σ)

Weil der Personenparameter ξ und der Itemparameter σ im Raschmodell auf derselben Skala gemessen werden, ergibt sich eine lineare Funktion der beiden Parameter. Damit der Itemparameter die Schwierigkeit und nicht die Leichtigkeit ausdrückt, sind beide Parameter mit einem Minuszeichen verknüpft.

Im Raschmodell sind die Graphen der logistischen Funktion, die so genannten Itemcharakteristiken (IC), notwendigerweise parallel weil nur ein Itemparameter, nämlich der Schwierigkeitsparameter, im Modell enthalten ist. Für zwei unterschiedlich schwierige Items *i* (hier: $\sigma_1 = -0{,}5$ und $\sigma_2 = 1{,}0$) ergibt sich *Abbildung 3-4*. Der Wert jedes Itemparameters definiert sich über den Abszissenwert der 50%-Lösungswahrscheinlichkeit.

Abbildung 3-4 Aufgabenkennlinien (ICC) zweier Testaufgaben nach dem Rasch-Modell



Jede Itemcharakteristik besitzt im Raschmodell die folgenden Eigenschaften:

- streng monotone Steigung der Lösungswahrscheinlichkeiten als Funktion der latenten Eigenschaft ξ
- $\xi_v, \sigma_i \in \mathbf{R}$
- $0 < f_i(\xi) < 1$
- $\lim_{\xi \to -\infty} f_i(\xi) = 0$
- $\lim_{\xi \to +\infty} f_i(\xi) = 1$
- $p(x_{vi} = 1 | \xi_v = \sigma_i) = \frac{1}{2}$

Die wichtigste Eigenschaft des Modells - und das Erkenntnisziel einer Raschanalyse - ist die sog. „spezifische Objektivität“ des Vergleichs zweier Personen oder Items. Sie bedeutet (KEMPF, 1972):

a) dass Aussagen über die Relation zwischen Versuchspersonen nicht davon abhängen, welche Items aus einer Itemmenge ausgewählt wurden

b) dass Vergleiche transitiv sein müssen, d.h. zu dem gleichen Ergebnis führen, wenn zwei Versuchspersonen direkt verglichen werden, wie wenn der Vergleich über andere Versuchspersonen als Zwischenstufe erfolgt. Dieses Prinzip verletzen psychologische

Daten wegen großer Fehlerbehaftetheit der Messinstrumente oder Mehrdimensionalität der beurteilten Objekteigenschaften häufig (WALTER, 1991, S. 106).

Analog gilt dies auch für den „stichprobenunabhängigen“ quantitativen Vergleich der Leichtigkeit oder Schwierigkeit mehrerer Items.

3.1.1 Parameterschätzung

Für Parameterschätzungen wird die sog. Likelihoodfunktion verwendet. Eine Likelihoodfunktion beschreibt die Wahrscheinlichkeit der Daten unter der Annahme, dass das der Analyse zugrunde gelegte Modell gilt. Die Prozedur der Parameterschätzung[32] besteht darin, das Maximum der Likelihoodfunktion zu ermitteln.
Die Idee der Maximum-Likelihood-Methode (z.B. NAGL, 2005) besagt, dass alle Modellparameter auf die Werte festgelegt werden „die das Auftreten der empirisch angetroffenen Daten am besten erklären“ (BORTZ, 1989, S. 127).
Bei Latent-Trait-Modellen gibt es verschiedene Likelihoodfunktionen:

a) <u>unbedingte oder unconditional Maximum Likelihood Methode (zur Schätzung der Item- und Personenparameter)</u>

Die Likelihoodfunktion der gesamten Datenmatrix ergibt sich unter der Voraussetzung der lokalen stochastischen Unabhängigkeit der Itemantworten als das Produkt aller Patternwahrscheinlichkeiten. Diese simultane Schätzung der Item- und Personenparameter führt aber wegen des Neyman-Scott-Problems (NEYMAN & SCOTT, 1948) zu systematischen Abweichungen des Parameterschätzwertes vom wahren Parameterwert, „deren Bias noch nicht bekannt ist“ (MÜLLER, 1999, S. 123).
Eine Schätzung der Personenparameter für die beiden Extremscores - alle Items oder kein Item gelöst - ist mit der unbedingten Maximum Likelihood Methode nicht möglich, da diese im ersteren Fall gegen plus und im letzteren Fall gegen minus unendlich divergieren.

$$(3.6) \quad uL = p(\underline{\underline{x}}) = \prod_{v=1}^{N} p(\underline{x}_v)$$

[32] Im Falle des Raschmodells die Schätzung der Item- und Personenparameter.

b) <u>bedingte oder conditional Maximum Likelihood Methode (zur Schätzung der Itemparameter)</u>

Diese Methode nutzt die Parameterseparierbarkeit des Raschmodells aus, um das Neyman-Scott-Problem potentiell unendlich vieler inzidenteller Parameter (Personenparameter) zu umgehen. Während der Schätzung der strukturellen Parameter (Itemparameter) werden die inzidentellen Parameter aus der zu maximierenden Likelihoodfunktion eliminiert.

RASCH (1977) hat gezeigt, dass die bedingte Wahrscheinlichkeit aller Itemrandsummen x_{0i} unter der Voraussetzung, dass die Personenrandsummen x_{v0} (Anzahl richtiger Lösungen) gegeben sind, eine nur von den Itemparametern abhängige Funktion ist. Umgekehrt ist auch die bedingte Wahrscheinlichkeit aller Personenrandsummen x_{v0} unter der Voraussetzung, dass die Itemrandsummen x_{0i} gegeben sind, nur von den Personenparametern abhängig. Die Maximum-Likelihood-Methode maximiert das Produkt über alle Patternwahrscheinlichkeiten unter der Bedingung des zum jeweiligen Pattern gehörenden Summenwertes durch geeignete Wahl der Itemparameter.

$$(3.7) \qquad cL = \prod_{v=1}^{N} p(\underline{x}_v / x_{v0})$$

c) <u>marginal Maximum Likelihood Methode (zur Schätzung der Itemparameter)</u>

Die marginale Maximum Likelihood Methode heißt so, weil die Likelihood der „Randsummen“, sog. marginals, maximiert wird. Die Methode geht davon aus, dass die Personenparameter in der Population eine bestimmte Verteilungsdichte haben. Diese Wahrscheinlichkeiten der Summenscores x_{v0} stellen zwar neue unbekannte Größen dar, die jedoch im Raschmodell (und nur dort) - wegen suffizienter Schätzfunktion - unproblematisch anhand der Daten durch die Häufigkeiten der x_{v0} geschätzt werden können, ohne dass Annahmen über die Verteilung der Personenparameter getroffen werden müssen. Die in der vorliegenden Studie zur Schätzung verwendete Einstellung des WINMIRA-Programms nimmt die Normalverteilung der Personenparameter an, um die Parameteranzahl zu reduzieren. PFANZAGL (1993) hat gezeigt, dass die Schätzung der Itemparameter mit der marginal Maximum Likelihood Methode beim Raschmodell unabhängig von Annahmen über Personenparameterverteilungen und identisch mit den Schätzungen der conditional Maximum Likelihood Methode ist.

Die Gleichung der marginal Likelihood beinhaltet die bedingte Likelihood und erhält bei Maximierung exakt dieselben Schätzgleichungen und Itemparameterschätzungen.

$$(3.8) \quad mL = \prod_{v=1}^{N} p(x_{v0}) * p(\underline{x}_v / x_{v0})$$

d) <u>WARM`s gewichtete Maximum Likelihood Methode (zur Schätzung der Personenparameter)</u>

Um Personenparameter auch für Extremscores angeben zu können, wenn also alle (r=k) oder keine Items (r=0) gelöst wurden und um den Bias aus dem Neyman-Scott-Problem zu korrigieren, gibt es nach WARM (1989) folgenden Ausweg, welcher der Personenparameterschätzung mit der unbedingten Maximum Likelihood Methode überlegen ist:

Die Wahrscheinlichkeit der Personenparameter wird unter der Bedingung der Daten und der Itemparameter maximiert. Die resultierende Wahrscheinlichkeit ist proportional zu einem Ausdruck, in dem die Dichtefunktion der Personenvariablen $f(\xi_v)$ vorkommt. WARM ersetzt $f(\xi_v)$ durch die Wurzel aus der Informationsfunktion, bei der extrem kleine oder große Werte für ξ sehr unwahrscheinlich sind:

$$(3.9) \quad p(\xi_v | \underline{x}_v, \sigma_i) \approx \frac{e^{(r_v \xi_v)}}{\prod_{i=1}^{k} (1 + e^{(\xi_v - \sigma_i)})} * f(\xi_v)$$

wobei:

$$f(\xi_v) = \sqrt{I(\xi_v)} = \sqrt{\sum_{i=1}^{k} p_{vi}(1 - p_{vi})}$$

Nach dem logarithmieren und differenzieren ergibt sich:

$$(3.10) \quad \sum_{i=1}^{k} p_{vi} = x_{v0} + \frac{\sum_{i=1}^{k} p_{vi}(1 - p_{vi})(1 - 2p_{vi})}{2\sum_{i=1}^{k} p_{vi}(1 - p_{vi})}$$

Dieser Term bewirkt, dass relativ zur unbedingten Maximum-Likelihood-Methode, große ξ kleiner und kleine ξ größer werden (vgl. auch ROST, 2004) und führt so zu einer Korrektur der Neyman-Scott-Verzerrung der unbedingten Maximum-Likelihood-Methode.

Die modifizierte Methode nach WARM ist der früher üblichen unbedingten Maximum Likelihood-Methode überlegen „und hat diese als `Standardverfahren` ersetzt“ (ROST, 2004, S. 314).

3.1.2 Prüfung der Modellgeltung

Zentrale Annahmen eines Modells wie Personenhomogenität können mit graphischem Modelltest oder „Andersen-Test“ überprüft werden. Anhand manifester Teilungskriterien (in dieser Arbeit: „diagnostische Zuordnung“ und „Testleistung“) wird eine Gesamtstichprobe in möglichst heterogene Teilstichproben aufgeteilt. Die vom Raschmodell geforderte Personenhomogenität kann so durch Vergleich der geschätzten Itemparameter in den Teilstichproben überprüft werden.

a) Mit dem von RASCH (1960) entwickelten **graphischen Modelltest** lassen sich Modellabweichungen zwar erkennen aber nicht statistisch prüfen.
Die geschätzten Itemparameter beliebiger Teilstichproben werden gemeinsam mit denen aus der Gesamtstichprobe in einem Schaubild aufgetragen. Weil sich bei (Rasch-) modellkonformen Items in jeder der Gesamtstichprobe entnommenen Personenteilstichproben dieselben Schwierigkeiten ergeben sollten, liegen diese im Koordinatensystem idealerweise auf einer 45°-Linie. Wenn Modellabweichungen durch unterschiedliche Itemparameter in den verschiedenen Stichproben angezeigt werden, ist die vom Raschmodell geforderte Personenhomogenität nicht gegeben.

b) Ein statistisches Maß für die Modellabweichung bietet der nach seinem Erfinder benannte **„Andersen-Test“** (ANDERSEN, 1973). Auch hier wird - wie im graphischen Modelltest - davon ausgegangen, dass die Itemschwierigkeiten über alle Stichproben gleich sind. Unter Geltung der Annahme der Personenhomogenität sollte das üblicherweise aus bedingten Likelihoods gebildete Verhältnis der Gesamtstichprobe mit dem Produkt der Teilstichproben 1 werden.

$$(3.11) \quad LR = \frac{L_0}{\prod_{g=1}^{h} L_g}$$

mit L_0 = Likelihood der Gesamtstichprobe

L_g = Likelihood der Teilstichproben g=1,..,h

k= Itemanzahl

LR lässt sich in eine χ^2-verteilte Prüfstatistik umwandeln: $-2*ln(LR) \rightarrow \chi^2$

Wird der Test wie von Andersen vorgesehen, über die bedingte Likelihood berechnet, so beträgt die Anzahl der Freiheitsgrade FG=(k-1)*(h-1). Wird er über die marginal Likelihood berechnet, so beträgt die Anzahl der Freiheitsgrade $\{(k-1)+n_p(\xi)\}*(h-1)$ worin $n_p(\xi)$ die Anzahl der Parameter bezeichnet, die in der Wahrscheinlichkeitsdichte von ξ vorkommen. Wird angenommen, dass die Personenparameter normalverteilt sind, so ist $n_p(\xi)$ =2 und die Anzahl der Freiheitsgrade ergibt sich zu FG=(k+1)(h-1). Das Programm WINMIRA (DAVIER, 2001) verwendet diese Annahme für die Berechnungen.

3.1.3 Modellvergleiche

Ob ein bei der Auswertung herangezogenes Testmodell angemessener ist als ein anderes Modell, kann mit Likelihoodquotiententests oder Informationskriterien beantwortet werden.
ROST (1996, S. 327) weist darauf hin, dass Vergleiche mit anderen Testmodellen allein anhand der marginal Likelihood (mL) möglich sind.
Da es generell so ist, dass die logarithmierte Likelihood größer wird, je mehr zusätzliche Parameter eingeführt werden, muss zur Beurteilung eines Modells anhand der Loglikelihood die Anzahl der Modellparameter berücksichtigt werden.

a) Ein **Likelihoodquotient** ist der Quotient zweier Likelihoodwerte derselben Datenmatrix in zwei unterschiedlichen Modellen:

$$(3.12) \quad LR = \frac{L_0}{L_1}$$

Unter der Bedingung, dass sich das Modell im Zähler durch eine Restriktion der Parameter des nachgewiesenermaßen gültigen Nennermodells darstellen lässt, lässt sich *LR* in eine χ^2-verteilte Prüfstatistik umwandeln: $-2*ln(LR) \rightarrow \chi^2$
Mit den Freiheitsgraden FG= $n_p(L_1)-n_p(L_0)$.

Inferenzstatistische Prüfungen mittels Likelihoodquotiententests sind zwar theoretisch möglich, erfüllen jedoch wegen der zu großen Anzahl möglicher Antwortmuster nur selten die Voraussetzungen für eine Signifikanzaussage (DAVIER, 1996). So sind bei einem Test mit nur 10 Items und dichotomen Antwortmöglichkeiten kombinatorisch 2^{10} = 1024 unterschiedliche Lösungsmuster denkbar. Informationstheoretische Maße werden als mögliche Lösung dieses Problems angesehen.

b) **Informationstheoretische Maße** erlauben den Vergleich unterschiedlichster Modelle für den gleichen Datensatz. Sie setzen den Wert der Likelihoodfunktion mit der Parameteranzahl in Beziehung. Jedes informationstheoretische Maß stellt eine Funktion aus der Loglikelihood und der Anzahl der zur Beschreibung der Daten herangezogenen Parameter dar. Der in dieser Arbeit zur Bestimmung der Klassenanzahl verwendete AIC-Wert ist der historisch erste seiner Art (Akaikes Information Criterion; (AKAIKE, 1973)). Gegenüber späteren informationstheoretischen Maßen (z.B. BIC (SCHWARZ, 1978), CAIC (BOZDOGAN & RAMIREZ, 1988)) bevorzugt er, weniger konservativ als diese, komplexere Modelle mit höherer Klassenanzahl, da auf eine Berücksichtigung der Stichprobengröße *N* verzichtet wird.

$$(3.13)\quad AIC=-2*lnL+2*n_p$$

mit n_p= Anzahl unabhängiger Modellparameter

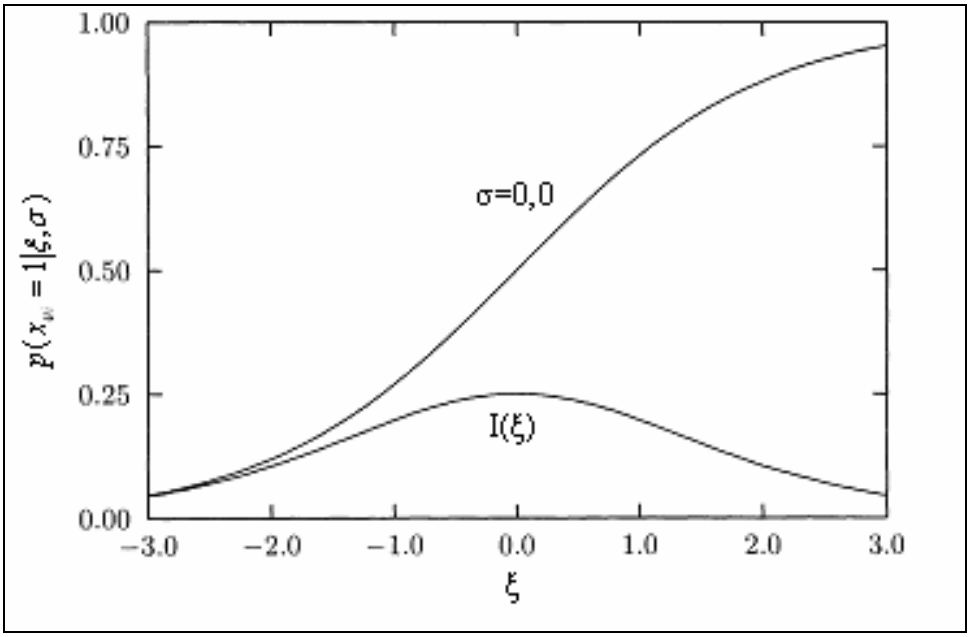
In seiner Funktion als Korrekturfaktor „bestraft“ der Term `$2*n_p$` beim AIC jeden zusätzlich eingeführten Parameter.

3.1.4 Statistische Information und das Prinzip der adaptiven Testung

Das von FISHER (1921) erstmals erwähnte Prinzip der „statistischen Information“ lässt sich beim Raschmodell als Unsicherheitsmaß der Itemlösung bei gegebener Personenfähigkeit interpretieren. Es besagt, dass ein Maximum an Information bezüglich des Personen- oder Itemparameters dort vorliegt, wo die größte Unsicherheit in Bezug auf Lösung oder Nichtlösung eines Items besteht. Im Falle des Raschmodells ergibt sich eine Lösungswahrscheinlichkeit von 0,5 - d.h. max. Unsicherheit - genau dann, wenn Item- und Personenparameter gleich groß sind ($\xi - \sigma = 0$).

Abbildung 3-5 stellt diesen Zusammenhang für ein Item mit $\sigma = 0{,}0$ und Personenfähigkeiten ξ entlang der Abszisse graphisch dar. Das Maximum der Informationsfunktion befindet sich dort, wo Personen- und Itemparameter dieselben Werte annehmen ($\sigma = \xi = 0$).

Abbildung 3-5 Kennlinie einer Aufgabe bei gleich großen Parametern und zugehörige Informationsfunktion der Testaufgabe



Für einen bestimmten Parameter π beträgt die in den Daten enthaltene statistische Information:

$$(3.14)\quad I(\pi) = Erwartungswert\left\{ \left(\frac{\partial \ \ln L}{\partial \pi} \right)^2 \right\}$$

Für die logistische Funktion von RASCH ergibt sich für ein Item i:

$$(3.15)\quad I_i(\xi_v) = \frac{e^{\xi_v - \sigma_i}}{(1 + e^{\xi_v - \sigma_i})^2} = p_i(\xi_v) * (1 - p_i(\xi_v))$$

„Die Größe [$I_i(\xi_v)$, d.V.] bezeichnen wir als die Information des Items *i* (mit gegebenem Parameter σ_i) in Bezug auf den Personenparameter ξ_v. Wegen der Symmetrie zwischen ξ und σ im logistischen Testmodell [...beschreibt diese Gleichung, d.V....] zugleich auch die Information der *v*-ten Person (mit gegebenem Parameter ξ_v) in Bezug auf den Itemparameter σ_i“ (FISCHER, 1974, S. 295).

Beim adaptiven Testen sollen häufig computerunterstützt möglichst „maßgeschneiderte“ Testitems ausgewählt werden. Diese sollen der Testperson so vorgelegt werden, dass möglichst maximale Beiträge statistischer Information zur quantitativen Schätzung des Personenparameters geliefert werden (vgl. *Abschnitt 1.8.1)*.

Es lässt sich beweisen, dass das Informationsmaß additiv ist (vgl. RASCH, 1989). Der Informationsbeitrag I_T mehrerer Items i=*1.....k* für die Schätzung eines Personenparameters ξ_v ist deshalb gleich der Summe der Informationsbeiträge jedes einzelnen Items:

$$(3.16)\quad I_T(\xi_v) = \sum_{i=1}^{k} \frac{e^{\xi_v - \sigma_i}}{(1+e^{\xi_v - \sigma_i})^2} = \sum_{i=1}^{k} p_i(\xi_v)(1 - p_i(\xi_v))$$

Aufgrund der Ungleichung von Rao-Cramér und der asymptotischen Wirksamstkeit von Maximum Likelihood Schätzungen (vgl. FISZ, 1970; KEMPF, 2003) ergibt sich für die asymptotische Fehlervarianz der Personenparameterschätzungen:

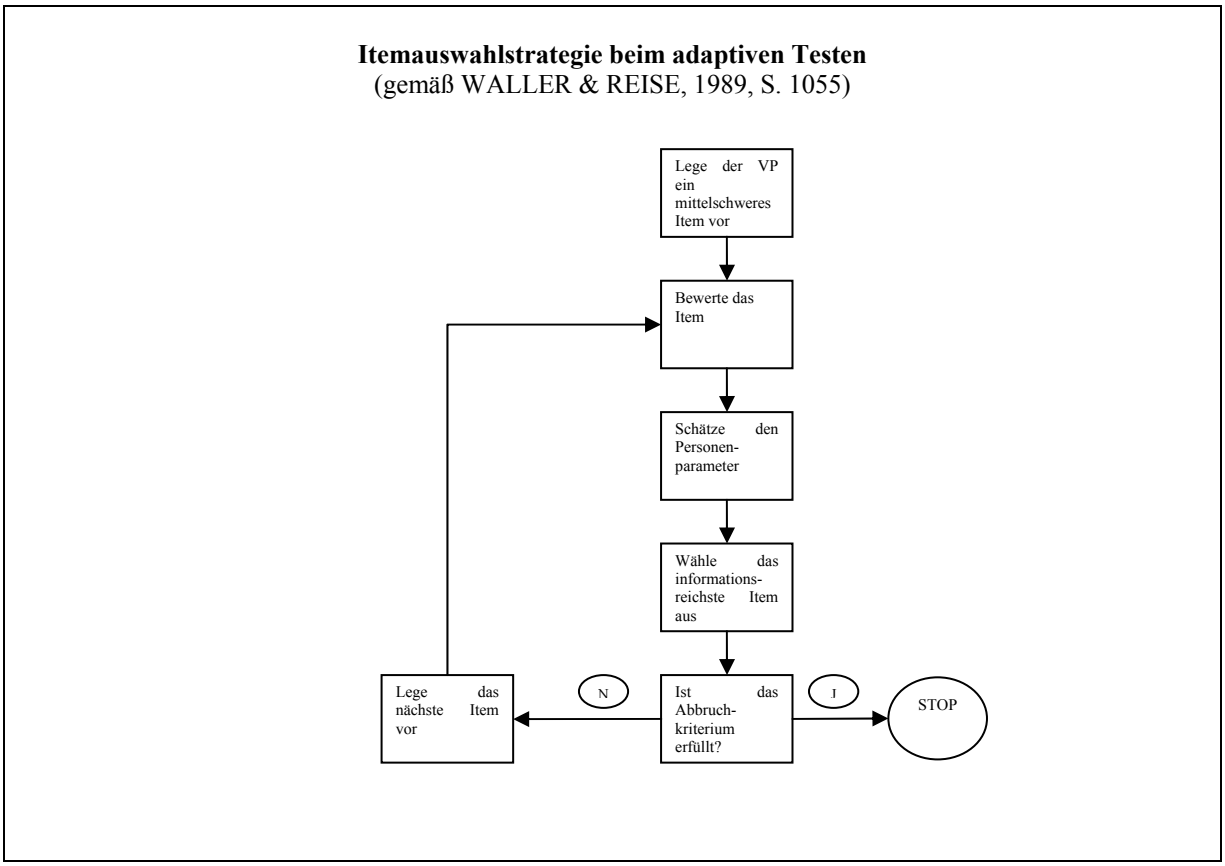
$$(3.17)\quad Var(\hat{\xi}) = \frac{1}{\sum_{i=1}^{k} p_i(\xi_v)\{1 - p_i(\xi_v)\}}$$

Diese Formel verdeutlicht:

- ➔ ein Test misst umso genauer, je mehr Items vorgelegt werden. Die Fehlervarianz des Personenparameters wird dann umso kleiner.
- ➔ ein Item mit einer Lösungswahrscheinlichkeit von 0,5 $(\xi - \sigma = 0)$ trägt zur Schätzung eines Personenparameters am meisten bei.
- ➔ um die Messgenauigkeit eines Personenparameters zu erhöhen sollten Items ausgewählt und vorgelegt werden, die der ungefähren Fähigkeitsausprägung der Testperson entsprechen und so die höchsten Informationswerte in Bezug auf den festzustellenden Personenparameter liefern.
  Dies ist das Prinzip der adaptiven Testvorgabe.

Das praktische Vorgehen der Itemauswahl beim adaptiven Testen ist in *Abbildung 3-6* anschaulich dargestellt.

Abbildung 3-6 Prinzip einer adaptiven Testung



3.2 Probabilistische Testmodelle

Das Raschmodell (RASCH, 1960) ist ein sehr simples probabilistisches Testmodell, dessen Annahme der Personenhomogenität nur sehr selten erfüllt wird (KÖLLER et al., 1994).

Weiterentwicklungen des Raschmodells in Form von **Mischverteilungsmodellen** beruhen auf der Idee, dass die beobachteten Daten eine Mischung aus verschiedenen Teilpopulationen darstellen, die durch unterschiedliche Verteilungsparameter gekennzeichnet sind.
Eine ursprünglich heterogene Personenpopulation lässt sich auf diese Weise in mehrere homogene Subpopulationen „entmischen“, in denen jeweils verschiedene statistische Modelle gelten.
Der Mittelwert der Zuordnungswahrscheinlichkeitsschätzungen aller zugeordneten Individuen[33] kann als anschauliches Maß dafür dienen, wie gut eine Stichprobe anhand ihrer Antwortmuster entmischt werden konnte. Eine Entmischung kann als umso sicherer und

[33] in dieser Arbeit mit „Zuordnungssicherheit“ benannt

gelungener bezeichnet werden, je mehr sich die gemittelten Zuordnungswahrscheinlichkeiten der einzelnen „entmischten“ Klassen dem Wert 1 annähern (ROST & ERDFELDER, 1996). Zwei Arten von Mischverteilungsmodellen werden in dieser Arbeit verwendet:

a) Bei der **Latent-Class-Analyse** (LCA) (LAZARSFELD & HENRY, 1968; FORMANN, 1984) gilt nicht das Prinzip des Rasch-Modells, dass die Leistungsmenge in Form der Anzahl gelöster Aufgaben alles über die Personenfähigkeit aussagt. Der Messwert, d.h. die Klassenzugehörigkeit einer Person, hängt bei der LCA davon ab, welche Items gelöst wurden. Anders ausgedrückt sollen ausgehend von den Daten in sich homogene Gruppen von Personen (latente Klassen) gefunden und klassifiziert werden. Die Itemparameter der einzelnen „entmischten“ Klassen einer LCA dürfen also im Gegensatz zum Rasch-Modell verschiedene Werte annehmen. Dies hat den Nachteil, dass für alle Personen innerhalb einer Klasse dieselbe Ausprägung der latenten Variable angenommen wird.
„Praktische Erfahrungen mit der Latent-Class-Analyse haben gezeigt, dass dabei in den meisten Fällen sehr klare diagnostische Entscheidungen mit großer Zuordnungssicherheit getroffen werden können“ (KEMPF, 2003, S. 132).

Vorausgesetzt, dass sich Testscores für diagnostisch relevante Aussagen eignen, ist zu erwarten, dass die Latente-Klassen-Analyse stärker nach Scores trennt als das Mixed-Rasch-Modell. Deshalb soll darauf hingewiesen werden, dass die in dieser Arbeit mittels LCA identifizierten latenten Klassen beim Speed-Power-Test „Cattell“ in engerem Zusammenhang mit den berechneten Summenscores klassischer Testtheorie stehen als die mittels Mixed-Rasch-Modell identifizierten latenten Klassen bei den Power-Tests „Raven“ und „Benton“. Der Wertebereich der enthaltenen Scores in den jeweiligen Klassen der Mixed-Rasch-Modelle ist deshalb eher größer als in den jeweiligen Klassen der LCA.

Die herkömmliche Form einer LCA beschreibt die Wahrscheinlichkeit, dass eine Person *v* Item *i* löst, als:

$$(3.18)\quad p(x_{vi}) = \sum_{g}^{G} \pi_g \pi_{i|g} \qquad \left\{ \text{ äquivalent dazu: } p(x_{vi}) = \sum_{g}^{G} \pi_g \frac{e^{(\alpha_{ig})}}{1+e^{(\alpha_{ig})}} \right) \right\}$$

mit $\sum_{g}^{G} \pi_g = 1$

wobei: π_g: Wahrscheinlichkeitsparameter der latenten Klasse g oder deren relative Größe

und $\pi_{i|g}$: Lösungswahrscheinlichkeit des *i*-ten Items in der *g*-ten Klasse.

Unter der Annahme lokaler stochastischer Unabhängigkeit innerhalb der Klassen ergibt sich beim Vorliegen von I manifesten Variablen, die den I-dimensionalen Merkmalsvektor $\underline{x}$ konstituieren für die Wahrscheinlichkeit von $\underline{x}$:

$$(3.19) \quad p(\underline{x}) = \sum_{g}^{G} \pi_g \prod_{i}^{I} \pi_{i|g}$$

Die zu maximierende Likelihood der gesamten Datenmatrix errechnet sich über eine Multiplikation der Vektorwahrscheinlichkeiten über alle Personen als:

$$(3.20) \quad mL = \prod_{v=1}^{N} p(\underline{x}_v)$$

oder als Produkt über alle möglichen Patterns als:

$$(3.21) \quad mL = \prod_{\underline{x}} p(\underline{x})^{n(\underline{x})}$$

Werden die Itemprofile der latenten Klassen in einem Schaubild aufgetragen (Abszisse: Personenparameter; Ordinate: Lösungswahrscheinlichkeiten, vgl. ROST, 1996), kann mit einer LCA auch überprüft werden, ob einem Test ein quantitatives Testmodell mit monotonen, überschneidungsfreien Itemfunktionen (z.B. Raschmodell) zugrunde gelegt werden darf. Es müssen folgende voneinander unabhängige Voraussetzungen gelten:

1. Monotoniebedingung: geordnete, überschneidungsfreie latente Klassen, d.h. dass alle Lösungswahrscheinlichkeiten einer leistungsstärkeren Klasse größer sind als die einer leistungsschwächeren Klasse.

2. Monotoniebedingung: geordnete, monoton ansteigende Itemprofile, d.h. dass die Items in jeder Klasse dieselbe Rangordnung ihrer Antwortwahrscheinlichkeiten aufweisen.

Solcherart „lokalisierte Klassen“ erlauben den Schluss, dass die anzahlmäßig begrenzten Personenparameter einer Latenten-Klassen-Analyse den Ausschnitt einer kontinuierlichen - über das ganze Spektrum verteilten - Personenvariable eines quantitativen Testmodells (z.B. Raschmodell) darstellen.

Quantifizieren wird somit zum Spezialfall von Klassifizieren.

b) Das **Mixed-Rasch-Modell** (MRM) wurde von ROST (1990) entwickelt und erlaubt gleichzeitige Quanti- und Klassifizierung. Latente Klassen werden identifiziert, die in ihrer Klasse raschhomogen sind, sich aber durch variierende Itemparameter zwischen den Klassen unterscheiden. Im Gegensatz zur LCA lassen sich nicht nur Klassenzugehörigkeitsparameter anhand der Zuordnungswahrscheinlichkeitsschätzungen, sondern auch raschmodellgültige, klassenspezifische, vom Summenscore abhängige Personenparameter angeben.
Wie im Modell der LCA wird angenommen, dass die Klassen exhaustiv und disjunkt sind, was als Modellgleichung ergibt:

$$(3.22) \quad p(x_{vi}) = \sum_{g}^{G} \pi_g \frac{e^{(\xi_{vg} - \sigma_{ig})}}{1 + e^{(\xi_{vg} - \sigma_{ig})}}$$

Wie man erkennen kann geht das MRM formal für $\xi_{vg} = 0 \;\; und \;\; -\sigma_{ig} = \alpha_{ig}$ in die LCA, für *G=1* in das Rasch-Modell mit $\xi_{vg} = \xi_v \;\; und \;\; \sigma_{ig} = \sigma_i$ über.

3.3 Das Saturierte Modell

Das saturierte Modell stellt das höchste Kriterium dar, das es in der Datenbeschreibung mit Modellen zu erreichen gibt. Es ist „das Modell, das die beobachteten Daten perfekt erklären kann und so viele Parameter enthält wie es unabhängige Daten gibt. Die Likelihood des saturierten Modells ist allein eine Funktion der Patternhäufigkeiten:

$$(3.23) \quad L_{sat} = \prod_{\underline{x}} \left(\frac{n(\underline{x})}{N} \right)^{n(\underline{x})}$$

Die Patternwahrscheinlichkeiten $p(\underline{x})$ stellen die Modellparameter des saturierten Modells dar und werden durch die relativen Häufigkeiten der Pattern geschätzt. Somit hat dieses Modell tatsächlich so viele Modellparameter wie es Antwortpattern gibt, von denen jedoch einer abhängig ist“ (ROST, 1996, S. 333).
Das saturierte Modell liefert die bestmögliche Beschreibung der Daten, benötigt dazu aber auch die maximal mögliche Anzahl von Modellparametern.

4 METHODEN

4.1 Beschreibung der Stichproben

Die Gesamtstichprobe dieser Untersuchung beinhaltet N=615 Versuchspersonen. Sie stellt den Datenpool für vier Teilstichproben, die dieser aus methodischen Überlegungen entnommen wurden.

Es handelt sich um folgende Stichproben:

- „GPN parallel“ (N=210)
- „parallelisierte Stichprobe“ (N=468)
- „Stichprobe `parallel-cerebral`“ (N=269)
- „Testbatterie“ (N=590)

Die Verwendung der Stichproben wird nachfolgend eingehend erläutert.

4.1.1 „Gesamtstichprobe“ (N=615)

Die Durchführung einer Testanalyse nach probabilistischen Kriterien erfolgt auf der Grundlage einer möglichst großen Anzahl Testteilnehmer mit unterschiedlichsten Testresultaten und Antwortmustern. Die Gesamtstichprobe bildet hierbei die Datenbasis für die Latenten Klassen-, Mixed-Rasch- und Raschanalysen von Raven-, Benton- und Cattelltest. Aus diesen Berechnungen resultiert zusätzlich zum Testscore jeweils ein testabhängiger, personenspezifischer, nominalskalierter Klassenzugehörigkeitsparameter.

Die Gesamtstichprobe (N=615) setzt sich aus klinisch unauffälligen und auffälligen Probanden zusammen. Im Gegensatz zu den Daten der klinisch unauffälligen Personen ohne berichtete neurologische oder psychiatrische Erkrankungen, die als anfallende Stichprobe in unterschiedlichen Institutionen und Umgebungen erhoben wurden, stammen die Daten der Patienten aus den verschiedenen bereits genannten neurologischen und psychiatrischen Kliniken[34].

Sämtliche in die Untersuchung eingehenden neurologischen Patienten hatten Erkrankungen oder Verletzungen mit cerebraler Beteiligung. Patienten aus dem psychiatrischen Bereich mit

[34] vollständige Liste siehe Danksagung im vorderen Teil

akuten, floriden Wahnerkrankungen wurden wegen der fraglichen Kooperationsfähigkeit nicht zur Testung zugelassen.

Die Diagnosen beruhen auf dem internationalen Klassifikationsschema psychischer Störungen ICD-10, Kapitel V (F).

Bei den Personen mit Doppeldiagnosen, also gleichzeitig diagnostizierten neurologisch cerebralen Erkrankungen und psychiatrischen F-Diagnosen, stand die F-Diagnose bei der Zuteilung der in dieser Untersuchung verwendeten Diagnosenummer im Vordergrund.

Die detaillierte Zusammensetzung der Gesamtstichprobe ergibt sich aus den *Tabellen 4-1* bis *4-7*.

Sporadisch vorhandene Unstimmigkeiten der Zellbesetzungen resultieren aus fehlenden Angaben in den Fragebögen. Zu beachten ist, dass nicht alle Teilnehmer sämtliche Tests bearbeitet haben (vgl. *Tabelle 4-7*).

Unvollständige Angaben sind für die Gesamtstichprobe, die nur zur Bestimmung der Klassenzugehörigkeitsparameter verwendet wurde, unproblematisch. In den Teilstichproben wurden Daten von Testteilnehmern mit unvollständigen Angaben umfassend eliminiert.

Hinsichtlich der Altersgruppenkodierung der verwendeten Tests wurde der derzeit gängigen und mehrfach bestätigten Einteilung beim Bentontest entsprochen (STECK, 2005).

Tabelle 4-1 Kodierung der Diagnosen in der Gesamtstichprobe

Kodierung der Diagnosen		#VP *(davon* **mit** *cerebraler Beteiligung)*
0	Klinisch unauffällige Personen	327 *(0)*
1	F1 (psychotrope Substanzen)	65 *(3)*
2	F2 (Schizophrenie)	30 *(0)*
3	F3 & F4 (Affektive & Neurotische Störungen)	48 *(14)*
4	F5 & F6 & F8 & F9 (Persönlichkeitsstörungen etc.)	14 *(0)*
5	F0 & neurologische Patienten (Schlaganfall, SHT, Tumor)	125 *(125)*
		Σ 609 *(142)*

Tabelle 4-2 Kodierung der Bildungsgrade in der Gesamtstichprobe

Kodierung der Bildungsgrade		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	Kein Schulabschluss	1	16 *(4)*	17
1	Hauptschulabschluss	88	111(52)	199
2	Mittlere Reife	111	88*(34)*	199
3	Abitur	125	72*(56)*	197
		Σ 325	**Σ 287** *(146)*	**Σ 612**

Tabelle 4-3 Bildungsgrade der Diagnosegruppen in der Gesamtstichprobe

Bildungsgrade der Diagnosegruppen					
Diagnose-gruppe	Bildungsgrad 0 #VP	Bildungsgrad 1 #VP	Bildungsgrad 2 #VP	Bildungsgrad 3 #VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	2	27	31	4	64 *(3)*
2	1	14	9	6	30 *(0)*
3	3	17	14	14	48 *(14)*
4	5	5	4	0	14 *(0)*
5	0	47	30	48	125 *(125)*
					Σ 281 *(142)*

Tabelle 4-4 Kodierung der Altersgruppen in der Gesamtstichprobe

Kodierung der Altersgruppen		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	≤ 39. Lebensjahr	155	113*(32)*	268
1	40. – 49. Lebensjahr	64	78*(38)*	142
2	50. – 59. Lebensjahr	63	60*(45)*	123
3	60. – 70. Lebensjahr	40	31*(25)*	71
4	> 70. Lebensjahr	5	6*(6)*	11
		Σ 327	**Σ 288** *(146)*	**Σ 615**

Tabelle 4-5 Altersverteilung der Diagnosegruppen in der Gesamtstichprobe

Altersverteilung der Diagnosegruppen						
Diagnose-gruppe	Alters-gruppe 0 #VP	Alters-gruppe 1 #VP	Alters-gruppe 2 #VP	Alters-gruppe 3 #VP	Alters-gruppe 4 #VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	32	21	12	0	0	65 *(3)*
2	20	6	3	1	0	30 *(0)*
3	22	15	7	4	0	48 *(14)*
4	11	2	0	1	0	14 *(0)*
5	23	34	37	25	6	125 *(125)*
						Σ 282 *(142)*

Tabelle 4-6 Geschlechterverteilung in der Gesamtstichprobe

Geschlecht		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	♀	135	89*(51)*	224
1	♂	192	199*(95)*	391
		Σ 327	**Σ 288** *(146)*	**Σ 615**

Tabelle 4-7 Durchgeführte Tests in der Gesamtstichprobe

Durchgeführte Tests	#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
Raven	327	278*(142)*	605
Cattell	327	280*(141)*	607
Benton	327	283*(142)*	610
Alle drei	327	272*(136)*	599

4.1.2 Teilstichprobe „GPN parallel“ (N=210)

Die kleinste Stichprobe „GPN parallel“ (N=210) wurde zur statistischen Absicherung des methodisch problematischen Validitätskriteriums „klinische Auffälligkeit“ gebildet.
Es wurden drei gleich große, weit gefasste Diagnosegruppen gebildet, um hinsichtlich Alters- und Bildungsgraden möglichst umfangreiche parallelisierbare Gruppengrößen zu erreichen:

G: klinisch unauffällige Personen

P: psychiatrische Patienten **ohne** cerebrale Beteiligung

N: neurologische Patienten **mit** cerebraler Beteiligung

Die drei parallelisierten Gruppen gleicher Größe bilden gemeinsam die Stichprobe „GPN parallel“.
In dieser Stichprobe sind keine Personen ohne Schulabschluss enthalten.
Die Zusammensetzung der Stichprobe „GPN parallel“ ergibt sich aus den *Tabellen A-1* bis *A-7* des Anhangs.

4.1.3 Teilstichprobe „Parallelisierte Stichprobe“ (N=468)

Die parallelisierte Stichprobe (N=468) beruht auf der methodisch notwendigen Überlegung, dass valide Testleistungsvergleiche der klinisch auffälligen und unauffälligen Personengruppen nur dann möglich sind, wenn leistungsbeeinflussende Alters- und Bildungseffekte weitestgehend ausgeschlossen werden können. Die Personen in der parallelisierten Stichprobe wurden darum nach Altersgruppen und Bildungsgraden parallelisiert. Die Kontrolle dieser Störvariablen soll den Grad der Validität der im Verlauf der Arbeit berechneten Prädiktionskennwerte erhöhen.
In der parallelisierten Stichprobe sind Personen ohne Schulabschluss nicht enthalten.

Die Zusammensetzung der parallelisierten Stichprobe ergibt sich aus den *Tabellen A-8* bis *A-14* des Anhangs.

4.1.4 Teilstichprobe „Stichprobe `parallel-cerebral`“ (N=269)

Die erheblich kleinere und deshalb aufgrund der geringer besetzten latenten Klassen problematischere „Stichprobe `parallel-cerebral`“ (N=269) geht aus analogen Überlegungen wie die zuvor beschriebene „parallelisierte Stichprobe“ hervor. Sie bezieht sich jedoch nicht auf sämtliche klinisch auffälligen Personen, sondern auf die separate Gruppe der klinisch auffälligen Personen mit cerebraler Beteiligung.
In dieser Stichprobe sind Personen ohne Schulabschluss nicht enthalten.
Die Zusammensetzung der Stichprobe ergibt sich aus den *Tabellen A-15* bis *A-21* des Anhangs.

4.1.5 Teilstichprobe „Testbatterie“ (N=590)

In der Stichprobe „Testbatterie“ sind sämtliche Personen (N=590) mit vollständigem Datensatz aller Variablen beinhaltet, die zur Berechnung der Latenten-Klassen-Analysen der Testbatterie im klassischen und probabilistischen Auswertungsmodus benötigt werden. Diese Voraussetzung erfüllen lediglich die Untersuchungsteilnehmer, die zusammenhängend alle drei Tests der Testbatterie bearbeitet haben.
Die Stichprobe „Testbatterie“ wird herangezogen, um Typenbilder oder Leistungsprofile der gesamten Testbatterie zu identifizieren.
Die Zusammensetzung der Stichprobe ergibt sich aus den *Tabellen A-22* bis *A-28* des Anhangs.

4.2 Testmaterial

Den Teilnehmern wurden folgende Tests in variierten Reihenfolgen vorgelegt:

1.) Eine 20 Item-Auswahl der Subtests der Advanced Progressive Matrices (APM) nach RAVEN (1965) ohne vorgegebenes Zeitlimit. Die Zusammenstellung beinhaltet alle 12 Items aus Set 1 (mit Item 1 als Probe-Item), sowie die nach Zahlenfolge (X=X+4) ausgewählten Items 1, 5, 9, 13, 17, 21, 25, 29 und 33 aus Set 2.

Hierbei handelt es sich um einen Power-Test, der in der Arbeit als **Raventest** bezeichnet wird.

2.) Die von STECK (2005) auf 20 Items verlängerte neue Form des Visual Retention Tests nach Benton. Vorgelegt wurden die Items nach Standardinstruktion A, d.h. Darbietungszeit 10 Sekunden und direkt anschließende zeichnerische Reproduktion.
Hierbei handelt es sich um einen Power-Test, der in der Arbeit als **Bentontest** bezeichnet wird.

3.) Der dreiminütige Subtest 3, die „Matrizen“ aus dem zweiten Teil der Adaption des Culture Fair Intelligence Tests CFT-3 (WEIß, 1971).
Hierbei handelt es sich um einen Speed-Power-Test, der in der Arbeit als **Cattelltest** bezeichnet wird.

Alle drei Tests wurden den Probanden erst im Anschluß an die vorgeschriebenen Instruktionen zur Bearbeitung vorgelegt.

4.2.1 Begründung der Testauswahl

Raven-, Benton- und Cattelltest zählen bis heute zu den am häufigsten angewendeten Verfahren in der psychologischen Diagnostik (SCHORR, 1995; STECK, 1997). Die Geschichte dieser klassischen, psychometrischen Verfahren reicht beim Raventest bis ins Jahr 1938 und beim Bentontest ins Jahr 1945 zurück. Die Culture Fair Tests (CFT) wurden von Cattell erstmals 1960 veröffentlicht.
Alle drei Tests erfassen Fähigkeiten, die weitgehend unabhängig vom Sprachvermögen der Probanden gelöst werden können. Störende Kovariationen der Testergebnisse mit Schulbildung und sozialem Status sollen auf diese Weise minimiert werden.
Insbesondere der Bentontest scheint eine für pathologische Einbußen hochsensible Fähigkeit zu erfassen. So stehen beim Bentontest zahlreiche Untersuchungen, in denen der Nachweis signifikanter Leistungsdefizite von Personen mit Hirnschädigung verschiedenster Ätiologie, Lokalisation und Schweregrad gelungen ist, nur wenige Ausnahmen gegenüber (KASCHEL, 1994). STECK et al. (1990) konnten signifikante Leistungsdefizite zudem bei Patienten folgender Diagnosegruppen feststellen: Schizophrenien, Depressionen, Neurosen sowie Persönlichkeitsstörungen, Alkoholabhängigkeiten mit nicht geklärten eventuell

hirnpathologischen Veränderungen und stationär behandelten Drogenabhängigkeiten nach dem Abklingen der Entzugserscheinungen.

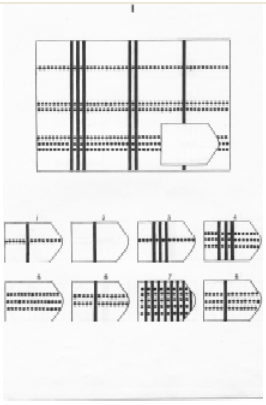
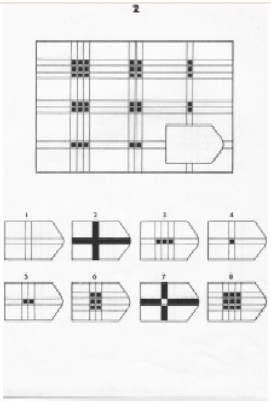
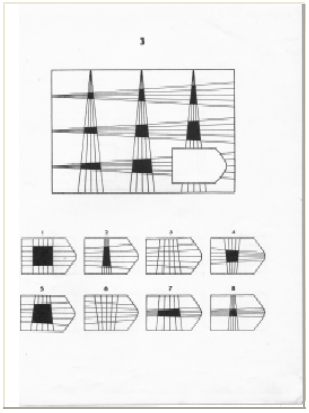
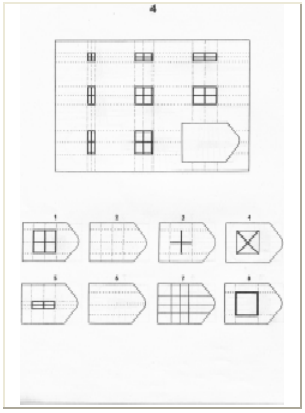
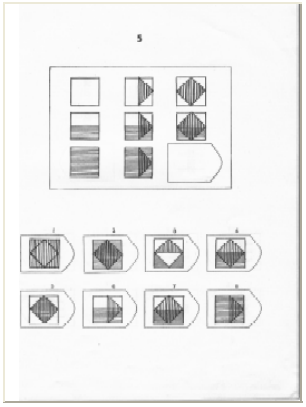
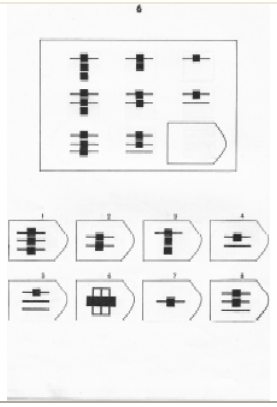
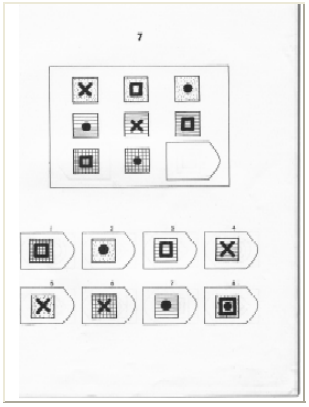
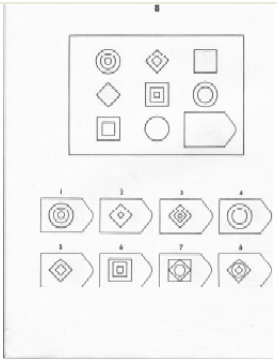
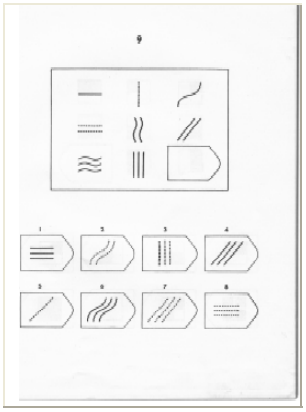
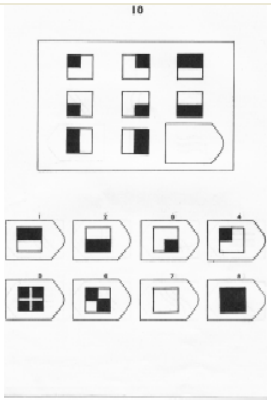
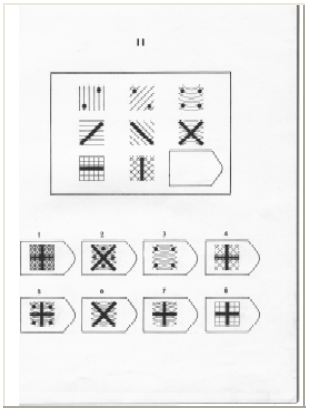
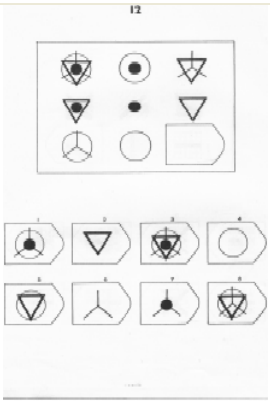
In wissenschaftlichen Untersuchungen zeigten sich hohe Korrelationen zwischen Raven- und Bentontestleistungen (KÜHNLEIN et al., 1983). Beide Tests scheinen folglich weitgehend analoge latente Fähigkeiten zu erfassen, die dem visuell räumlichen Bereich zuzuordnen sind.
Auch beim zeitabhängigen „Matrizen"-Subtest des CFT-3 weisen faktorenanalytische Untersuchungen auf einen Faktor, der als „anschauungsgebundenes Denken" bezeichnet wurde (STECK et al., 1975). Die Art der Aufgaben ähnelt denen im Raventest, deshalb wird der Vergleich der Testleistungen von Cattelltest und Raventest, mit und ohne Zeitdruck möglich.

Die Itemschwierigkeiten der Standard Progressive Matrices (SPM) von Raven gelten als niedrig (RAUCHFLEISCH, 1983). Aus diesem Grund wurde für diese Arbeit die bereits genannte Itemfolge aus den schwereren Advanced Progressive Matrices (APM) zusammengestellt, um einen Deckeneffekt zu vermeiden.
Das vollständig darin enthaltene Set 1 kann als verkürzte Ausgabe der Standard Progressive Matrices (SPM) betrachtet werden (RAVEN, 1965). Im Handbuch ist nachzulesen, aber nicht unumstritten (PUTZ-OSTERLOH, 1981), dass diese 12-Item-Version sämtliche kognitiven Vorgänge erfasst, die auch zur Lösung der SPM verwendet werden.
Die Verkürzung auf 20 Items plus einem Probe-Item soll den Vergleich der Powertests Raven und Benton ermöglichen und ebenso der Testökonomie dienen, da auf eine Zeitbegrenzung bei der Lösung des Raventests verzichtet wurde.

Die folgenden *Abbildungen 4-1* bis *4-3* zeigen die Items der drei Tests.

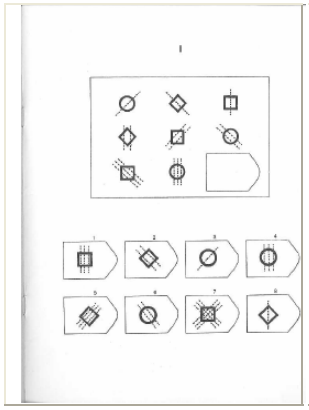
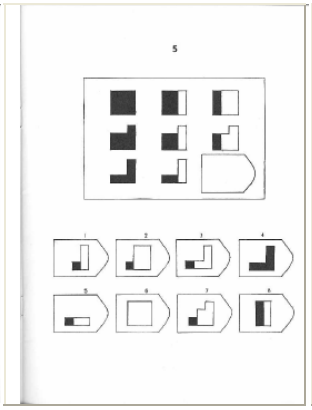
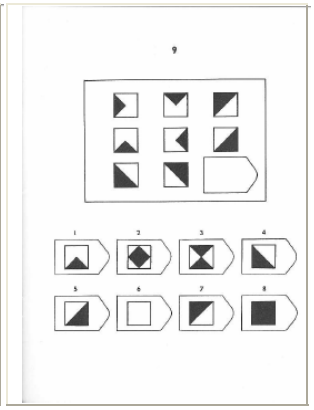
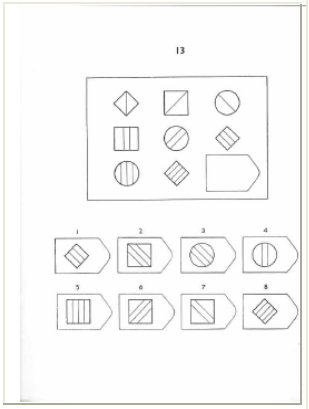
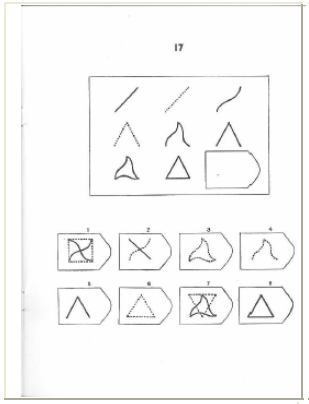
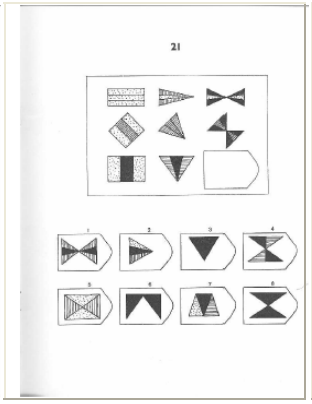
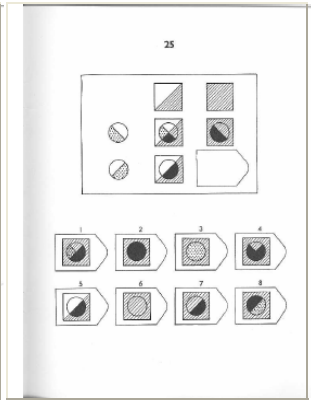
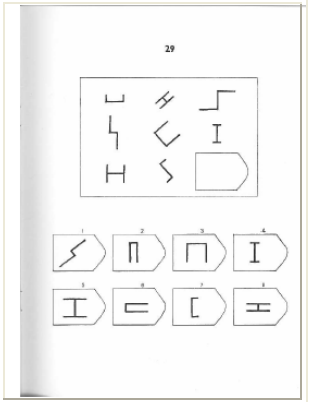
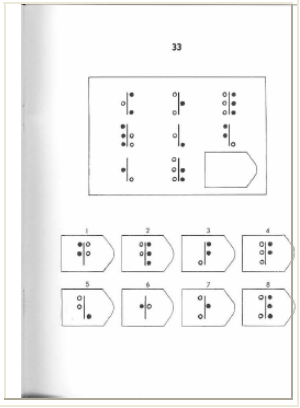
Die Items des Raventests:

Abbildung 4-1 Darstellung der Items des Raventests

Die verwendete 20-Item-Form des Raventest [APM-Quellenangabe]			
ProbeItem [APM Set1, Nr.1]	R1 [Set1, Nr.2]	R2 [Set1, Nr.3]	R3 [Set1, Nr.4]
			
R4 [Set1, Nr.5]	R5 [Set1, Nr.6]	R6 [Set1, Nr.7]	R7 [Set1, Nr.8]
			
R8 [Set1, Nr.9]	R9 [Set1, Nr.10]	R10 [Set1, Nr.11]	R11 [Set1, Nr.12]
			

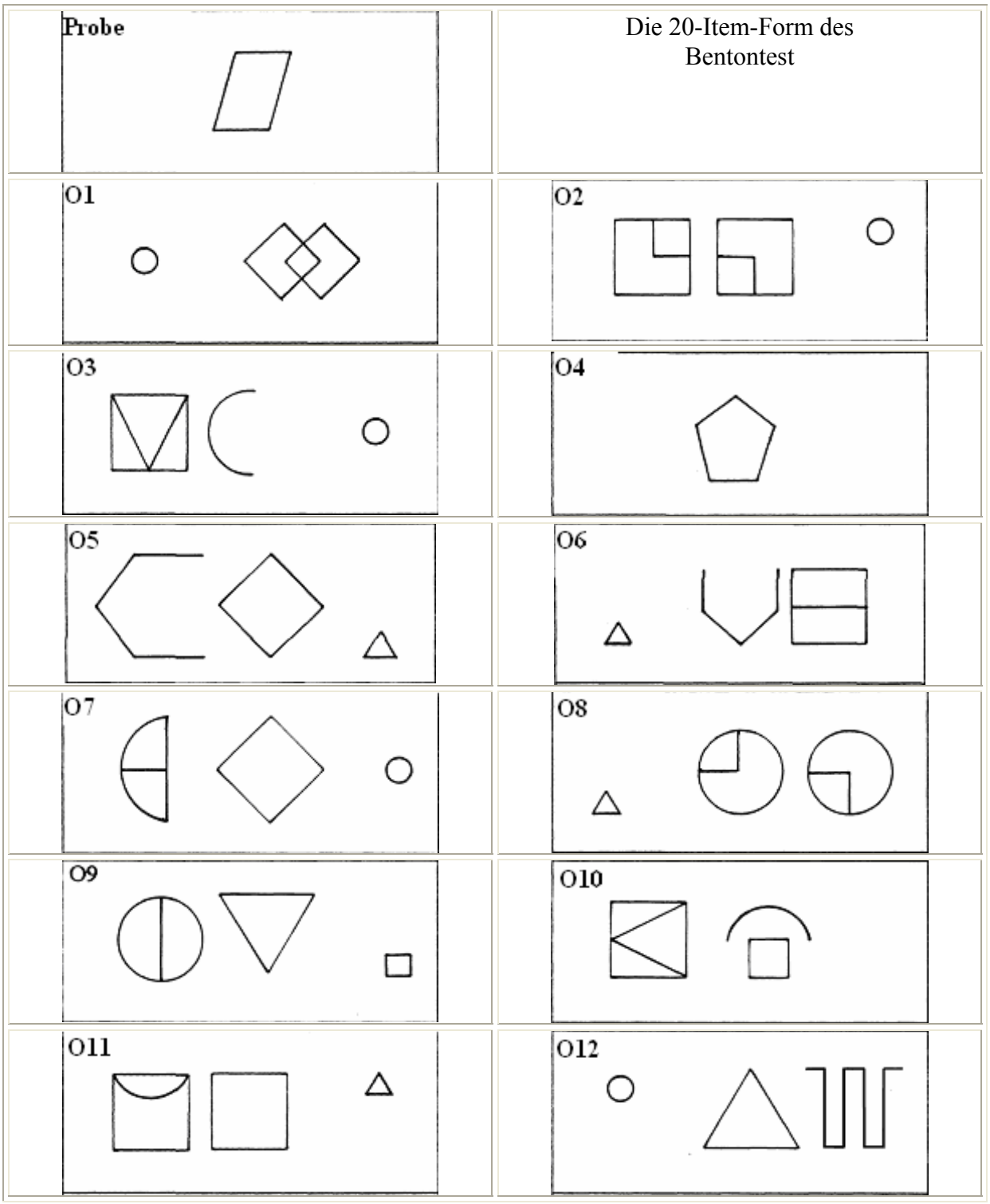
Fortsetzung nächste Seite

Abbildung 4-1 Fortsetzung

R12 [Set2, Nr.1]	R13 [Set2, Nr.5]	R14 [Set2, Nr.9]	R15 [Set2, Nr.13]
			
R16 [Set2, Nr.17]	R17 [Set2, Nr.21]	R18 [Set2, Nr.25]	R19 [Set2, Nr.29]
			
R20 [Set2, Nr.33]			
			

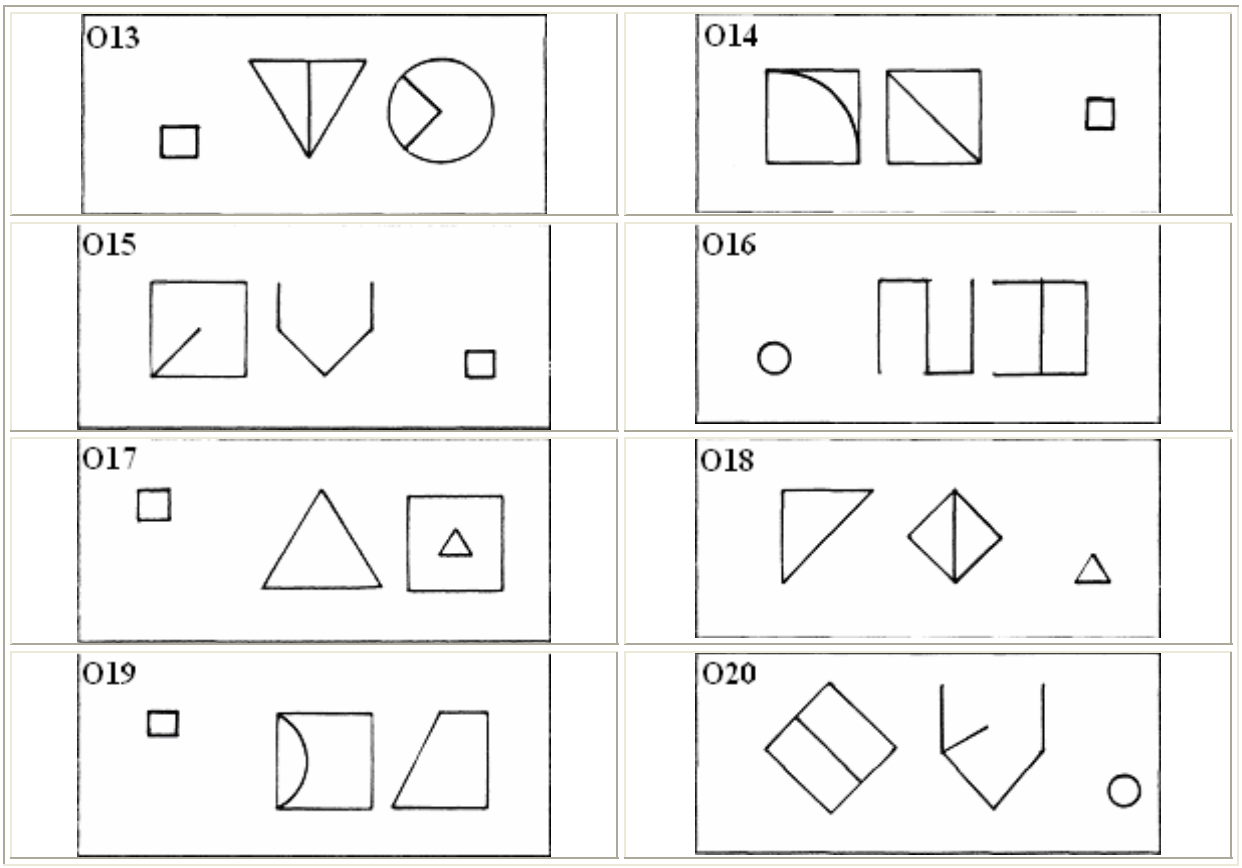
Die Items des Bentontests:

Abbildung 4-2 Darstellung der Items des Bentontests



Fortsetzung nächste Seite

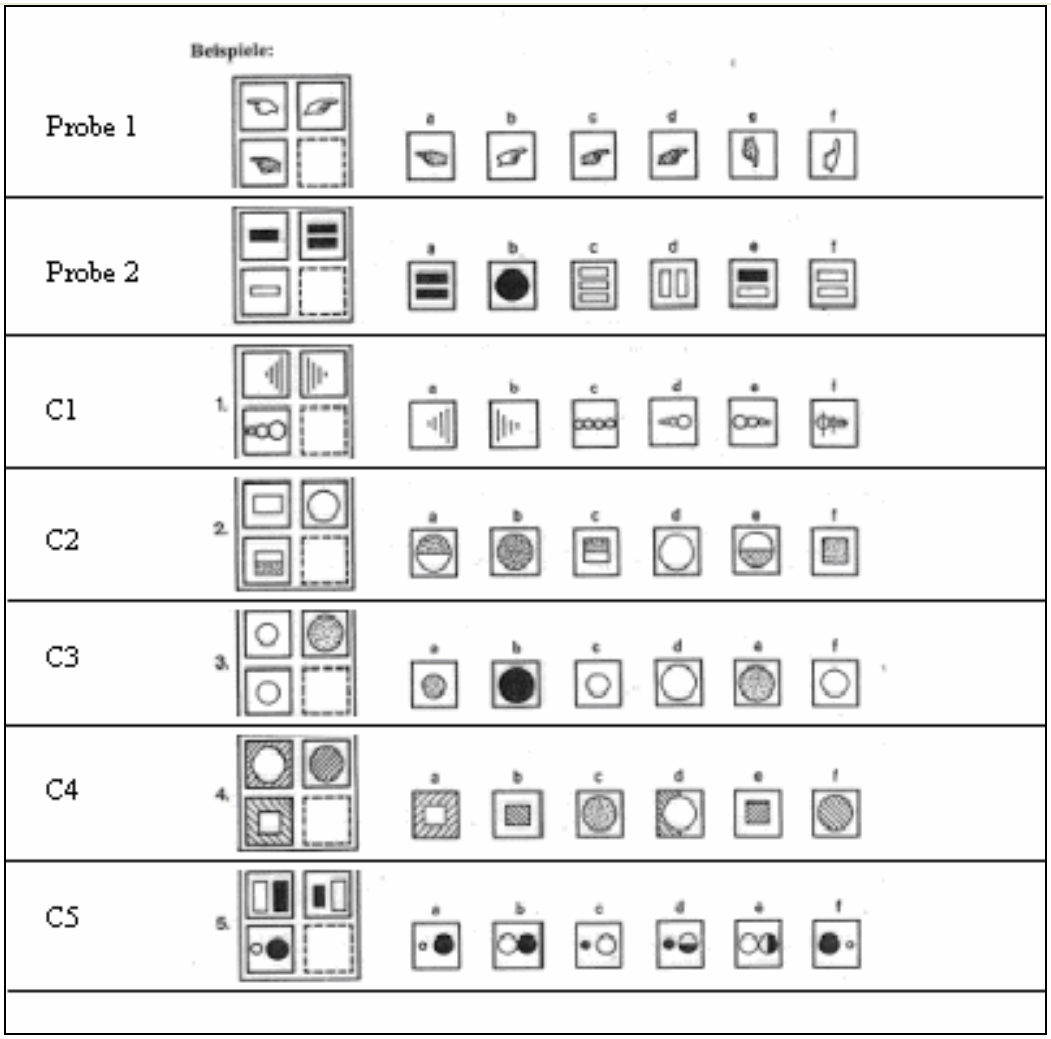
Abbildung 4-2 Fortsetzung



Die Items des Cattelltests:

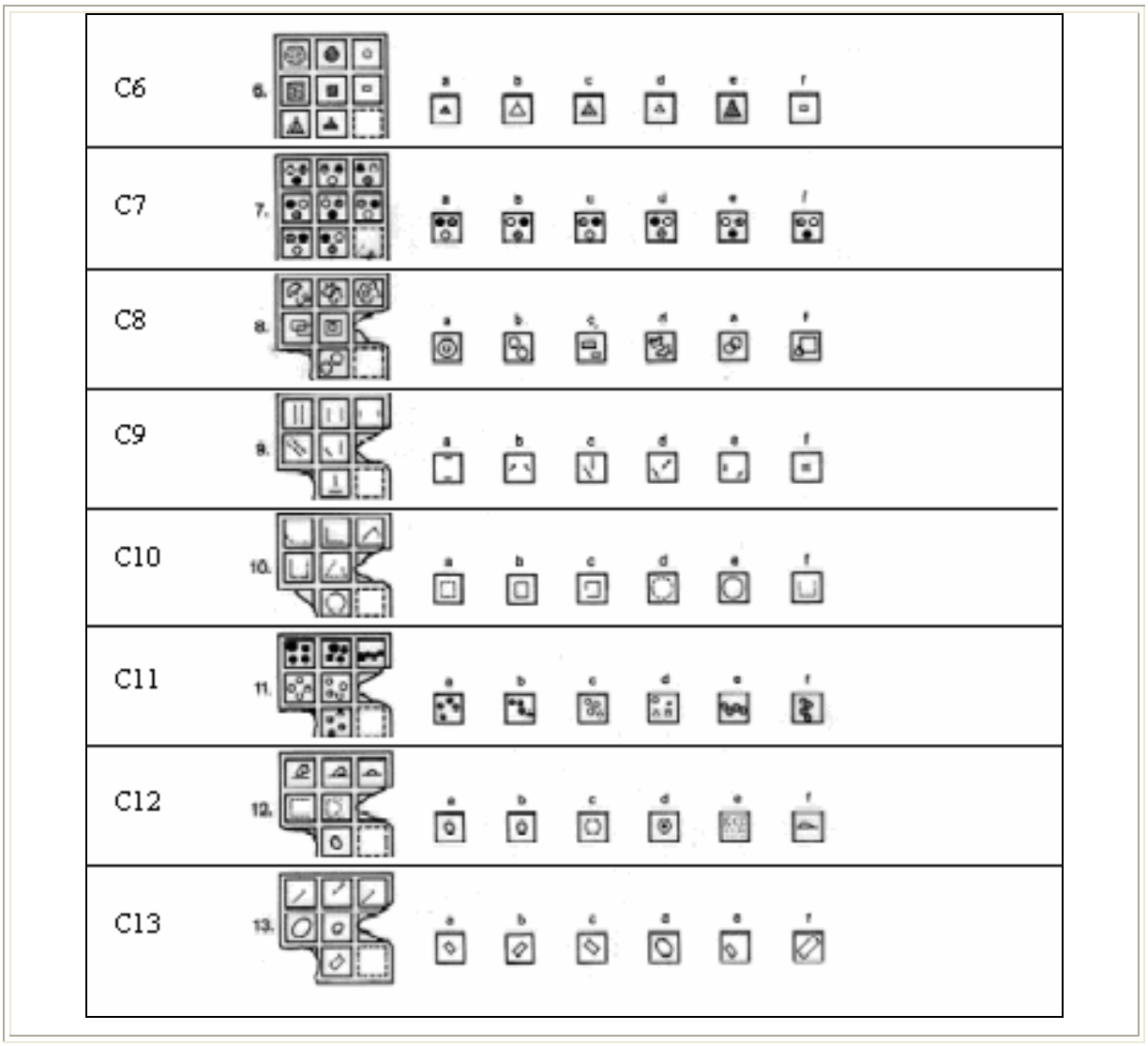
Abbildung 4-3 Darstellung der Items des Cattelltests

Die 13-Item-Form des Catteltests („Matrizen"-Subtest; CFT-3; 2.Teil)



Fortsetzung nächste Seite

Abbildung 4-3 Fortsetzung



4.3 Testdurchführung

Sämtliche Versuchspersonen wurden im Laufe des Jahres 2003/04 vom Verfasser persönlich interviewt und 97 Prozent der Teilnehmer wurden zusammenhängend mit allen drei verwendeten Testverfahren getestet.

Die Probanden wurden persönlich über Zweck und Ablauf der Tests informiert und es stand ihnen frei sich für oder gegen die Testteilnahme zu entscheiden.

Jeder Testteilnehmer, bis auf die Gruppe der neurologischen Patienten[35], erhielt einen Betrag von 10 € direkt nach der Testung.

[35] wegen klinikinterner Regelung

Der persönlich ausgehändigte anonymisierte Lösungsbogen zur Erfassung der demographischen Merkmale Lebensalter, Bildung und Geschlecht des Testteilnehmers, wurde noch vor Testbeginn im Beisein des Testleiters ausgefüllt. Mögliche neurologische oder psychiatrische Diagnosen der klinisch unauffälligen Probanden wurden durch Selbstbericht ausgeschlossen. Alle erhobenen Daten dienen dieser Arbeit zur Berechnung statistischer Zusammenhänge zwischen erbrachten Testleistungen, intellektuellen Fähigkeiten, Alterseffekten und klinischen Auffälligkeiten.

Die Teilnehmer wurden nach den gültigen Teststandards instruiert und um Positionseffekte auszuschalten in zufällig variierter Testreihenfolge getestet.
Für die Lösung des **Raventests** war kein Zeitlimit vorgegeben, was zu verlängerter Testdauer führen konnte. Der **Cattelltest** war wie im Handbuch vorgegeben, in 3 Minuten, also unter Zeitdruck zu bearbeiten. Der **Bentontest** wurde nach Standardinstruktion A, Darbietungszeit 10 Sekunden und direkt anschließender zeichnerischer Reproduktion vorgelegt.
Die Durchführung der gesamten Testbatterie dauerte im Durchschnitt etwa 45 Minuten. Kürzere Pausen wurden zwischenzeitlich, je nach Bedarf ermöglicht.

4.4 Auswertung der Testergebnisse

Die Auswertung der Tests erfolgt nach den Vorgaben der zugehörigen Testhandbücher. Probe-Items werden nicht gewertet. Bei allen drei Tests wird für jede richtige Lösung ein Punkt, für jede falsch gelöste Aufgabe kein Punkt vergeben. Auf diese Weise können bei Raven- und Bentontest maximal 20, beim Cattelltest maximal 13 Punkte erreicht werden. Wird keines der Items richtig gelöst, ergibt sich bei allen drei Tests als minimaler Summenscore 0 Punkte.
Bei allen Tests wird der dichotome Auswertungsmodus verwendet. Er ist für Raschanalysen mit kleinerem Versuchspersonenpool wegen der geringeren Anzahl zu schätzender Parameter besonders geeignet.

<u>Auswertungsalternative beim Bentontest:</u>
Der Bentontest sieht alternativ zum dichotomen Auswertungsmodus BTR, als Anzahl richtiger Lösungen, noch eine zweite Auswertungsalternative vor. Das Fehlermaß BTF gibt die Anzahl der Fehler in den Itemreproduktionen an.

In dieser Studie werden die Prädiktionen der klinischen Kriterien mit beiden Auswertungsmodi, also den Prädiktoren BTR und BTF geprüft. Allerdings wird der Prädiktor BTF nur peripher besprochen (vgl. *Abschnitt 5.2.3.3.3*). Der Auswertungsmodus BTF wird so gehandhabt, dass pro fehlerhafter Itemreproduktion bis zu vier spezifische Fehlerpunkte vergeben werden können.

In jeder Itemreproduktion können die folgenden Fehler auftreten:

a) Fehler linke große Figur
b) Fehler rechte große Figur
c) Formfehler periphere Figur
d) Positionsfehler periphere Figur

4.5 Statistische Auswertung

Für die Datenverarbeitung wurde ein AMD Athlon™ XP 3200+ verwendet.

Die Rohwerte der Datenerhebungen wurden in das Tabellenkalkulationsprogramm Microsoft EXCEL 2002 (Microsoft Corporation, 2002) eingegeben.

Für die Prüfungen der Modellgültigkeiten der untersuchten Testverfahren, die Klassengrößenberechnungen und die Zuordnungswahrscheinlichkeitsschätzungen der Testpersonen mit probabilistischen Methoden wurden drei verschiedene Programme angewendet. Diese sind:

WINMIRA 2001 1.36 (DAVIER, 2001)

DYNAMIC VERSION 2.5 (KEMPF, 2004)

LACORD 3.31 (ROST, 1989)

Die statistischen Berechnungen und die Darstellung der Ergebnisse in Abbildungen erfolgten mit den Programmen „JMP, Release 5.0.1.2“ (SAS Institute Inc., 2003) und Microsoft EXCEL 2002 (Microsoft Corporation, 2002).

4.6 Mathematisches Vorgehen

Zentraler Punkt dieser Arbeit ist die Schätzung der Prädiktion von dichotomen und multinominalen Außenkriterien[36] mit quantitativen und qualitativen Prädiktoren[37].

[36] abhängige Variablen: dichotom: klinische Auffälligkeit (Ja/Nein), cerebrale Beteiligung (Ja/Nein)
multinominal: Klassenlösung (1,2...X) vgl. *Abschnitte 5.6.1.3.3* und *5.6.2.3.3*

Weil es sich bei den Außenkriterien um qualitative Zielvariablen handelt, wird das Modell der **logistischen Regression** verwendet (vgl. HOSMER & LEMESHOW, 1989; BACKHAUS et al., 2003).

Die Herleitung der logistischen Funktion wurde am Beispiel des Raschmodells zur Prädiktion des dichotomen Außenkriteriums „Itemlösung" mit den beiden Prädiktoren Personenparameter ξ und Itemparameter σ bereits unter *Abschnitt 3.1* in den *Abbildungen 3-1* bis *3-3* dargestellt und erläutert.

Für eine dichotom[38] kodierte Kriteriumsvariable Y (1 ≅ Ereignis ist eingetreten) und die Prädiktorvariablen[39] X_1 bis X_M erklärt das logistische Regressionsmodell die bedingte Wahrscheinlichkeit für das betrachtete Ereignis (Y=1) unter der Prädiktorvariablen-kombination folgendermaßen:

$$(4.1) \quad p(Y=1|X) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} \text{ mit } \eta := \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_M X_M$$

wobei: $\beta_1, \beta_2, \ldots, \beta_M$ = Regressions- oder Logit-Koeffizienten

β_0 = Wert auf der logit-Achse an der Stelle, an der alle X-Werte 0 sind

Die Abhängigkeit der Wahrscheinlichkeit eines bestimmten dichotomen oder multinominalen Ereignisses Y von den Prädiktoren X_i wird mit dem logistischen Regressionsmodell modelliert, welches einen linearen Zusammenhang zwischen den logits der Wahrscheinlichkeit des Außenkriteriums und den Prädiktoren unterstellt. Ein Prädiktor X mit positivem Regressionskoeffizient ($\beta > 0$) bewirkt einen Anstieg der Wahrscheinlichkeit für Ereignis Y=1, entsprechend führt ein negativer Effekt auf den Logit ($\beta < 0$) zu einer Abnahme dieser Wahrscheinlichkeit.

Es werden folgende Hypothesen auf dem Signifikanzniveau $\alpha = 5\%$ getestet:

H_0: Der Einfluss der Prädiktoren auf das Kriterium ist Null
(H_0: $\beta_1 = \beta_2 = \ldots = \beta_M = 0$)

H_1: Der Einfluss der Prädiktoren auf das Kriterium ist von Null verschieden
(H_1: $\beta_1, \beta_2, \ldots, \beta_M \neq 0$)

---

[37] unabhängige Variablen: klassischer Summenscore und probabilistischer Klassenzugehörigkeitsparameter (Bildungsgrad, Altersgruppe, Diagnosegruppe vgl. *Abschnitte 5.6.1.3.3* und *5.6.2.3.3*)

[38] analog der multinominale Fall

[39] intervallskaliert oder durch „Dummysierung" von kategorialen Variablen entstanden

Nach einer Prüfung der Modellgültigkeit mit χ^2-verteilten Likelihood-Ratio-Tests werden **drei unterschiedliche Kennwerte der Prädiktion** angegeben. Wenn mehrere Prädiktoren im Modell enthalten sind, werden die einzelnen Beiträge dieser Prädiktoren mit **Wald-Effekt-Tests** statistisch geprüft.

Die berechneten Kennwerte zur Schätzung der Vorhersagevalidität werden in den nächsten *Abschnitten* genauer erläutert.
Die prozentuale Entropiereduktion in *Abschnitt 4.6.*1, die prozentuale adjustierte Entropiereduktion in *Abschnitt 4.6.2*, die Fläche unter der Receiver Operating Curve (ROC) in *Abschnitt 4.6.3* und die Wald-Effekt-Tests in *Abschnitt 4.6.4*.

4.6.1 Entropiereduktion in %

Das PRU-Maß (Proportional Reduction of Uncertainty) stellt bei qualitativen Kriterien einen Spezialfall von $r^2_{McFadden}$ (Mc FADDEN, 1973) dar. Dieses Maß beruht auf der Gegenüberstellung des vollständigen Modells L_M und des Nullmodells L_0. Dabei gilt:

$$(4.2) \quad \text{PRU} = \text{RSquare(U)} = r^2_{\text{McFadden}} = 1 - \frac{\ln L_M}{\ln L_0}$$

der Wertebereich von r^2_{McFadden} umfasst das Intervall $[0,1)$.

wobei: L_M= Likelihood des Modells mit den Prädiktoren $x_1 \ldots x_i$
L_0= Likelihood des Modells ohne Prädiktoren

Das Maß „Entropiereduktion (%)“ gibt die Reduktion der Unsicherheit der Vorhersage des Kriteriums mit den Prädiktoren gegenüber einer Zufallsentscheidung ohne die Prädiktoren an. Der aufgeführte Tabellenwert in dieser Arbeit ergibt sich durch Multiplikation von $r^2_{McFadden}$ mit Faktor 100 als Prozentwert.

4.6.2 Adjustierte Entropiereduktion in %

Da der Wert r^2_{McFadden} mit der Anzahl verwendeter Modellparameter steigt, wird er in dieser Arbeit um eine „adjustierte Version“ $\overline{r^2}_{McFadden}$ (LONG, 1997) ergänzt:

$$(4.3) \quad \text{PRU}_{\text{adjustiert}} = \overline{r^2}_{McFadden} = 1 - \frac{\ln L_M - K^*}{\ln L_0}$$

wobei K^*= Anzahl der unabhängigen Modellparameter (Freiheitsgrade).
Dieses Korrekturmaß entspricht im Wesentlichen der Logik des informationstheoretischen Maßes AIC bei Modellvergleichen, weil die Anzahl nötiger Schätzparameter verrechnet wird (vgl. *Abschnitt 3.1.3*).

4.6.3 Fläche unter der Receiver Operating Characteristic Kurve (ROC-Kurve)

ROC-Kurven (GREEN & SWETS, 1966) sind eine anschauliche Darstellung der Güte der Prädiktion eines dichotomen Kriteriums[40] mit den verwendeten Prädiktoren. Die ROC-Kurve beschreibt im Schaubild auf der Abszisse die Spezifität und auf der Ordinate die zugehörige Sensitivität eines Tests für verschiedene Diskriminanzschwellen.
Die diagnostische **Sensitivität** ist definiert als die Anzahl wahrer positiver Ergebnisse, somit auffälliger Tests bei Patienten, bei denen eine klinische Auffälligkeit mit Sicherheit besteht.
Die diagnostische **Spezifität** ist definiert als die Anzahl wahrer negativer Resultate, also unauffälliger Tests bei Gesunden, bei denen keine klinische Auffälligkeit besteht.
Die Werte von Sensitivität und Spezifität sind abhängig von der gesetzten Diskriminanzschwelle. Die korrespondierenden Werte für Spezifität und Sensitivität werden für verschiedene Referenzwerte berechnet und im Schaubild gegeneinander aufgetragen.
Die Fläche unter den ROC-Kurven ermöglicht eine Beurteilung der Prädiktionsgüte. Sie wird umso größer, je höher die Zuverlässigkeit der Prädiktion ist. Entspricht die ROC einer Diagonalen, so sind die Prädiktoren zur Vorhersage des Kriteriums für die betreffende Fragestellung unbrauchbar, da ihr Beitrag einer einfachen Zufallsentscheidung entspricht.
Eine Illustration der Zusammenhänge bei ROC-Kurven geben die *Abbildungen 4-4* bis *4-6*. Diese Graphiken zeigen Scoreverteilungen[41] bei drei fiktiven Tests, die zur Prädiktion eines dichotomen Kriteriums „A“ und „kein A“ der Testpersonen verwendet wurden.
Sensitivität[42] und Spezifität[43], sowie entsprechend WNA[44] und FPA[45] sind in den *Abbildungen* für einen ROC-Datenpunkt der jeweils gleichen Diskriminanzschwelle berechnet worden.

---

[40] Hier „klinische Auffälligkeit“ (Ja/Nein) und „cerebrale Beteiligung“ (Ja/Nein)

[41] In den *Abbildungen* normalverteilt.

[42] Testergebnisse sprechen für „A“, wenn „A“ wirklich vorliegt (WPA)

[43] Testergebnisse sprechen für „kein A“, wenn „kein A“ wirklich vorliegt (FNA)

[44] **W**ahre **N**egative Vorhersage von **A**

[45] **F**alsche **P**ositive Vorhersage von **A**

Im **ersten Fall** (*Abbildung 4-4*) gelingt keine Prädiktion des Kriteriums mit dem Testscore. Die ROC beschreibt hier eine Diagonale, weil die Scoreverteilungen der Gruppen „A“ und „kein A“ nahezu identisch sind. Die Klassifizierung einer Person nach einer Testdurchführung in eine der beiden Gruppen ist nicht besser als eine reine Zufallsentscheidung, d.h. die Fläche unter der ROC-Kurve beträgt $\approx$ 0,5.

Die Prädiktion des Kriteriums verbessert sich mit dem verwendeten Test im **zweiten Fall** (*Abbildung 4-5*) und funktioniert im **dritten Fall** am besten (*Abbildung 4-6*). Die Fläche unter der ROC-Kurve ist hier am größten, weil die Scoreverteilungen der Gruppen „A“ und „kein A“ die geringste Überlappung zeigen. Somit gelingt die trennschärfste Unterscheidung beider Gruppen mit dem dritten Test. Die Vorhersage des dichotomen Kriteriums „A“ oder „kein A“ ist infolgedessen mit diesem Test am zuverlässigsten, die Vorhersagevalidität am höchsten.

Abbildung 4-4 ROC-Kurve 1: Test 1

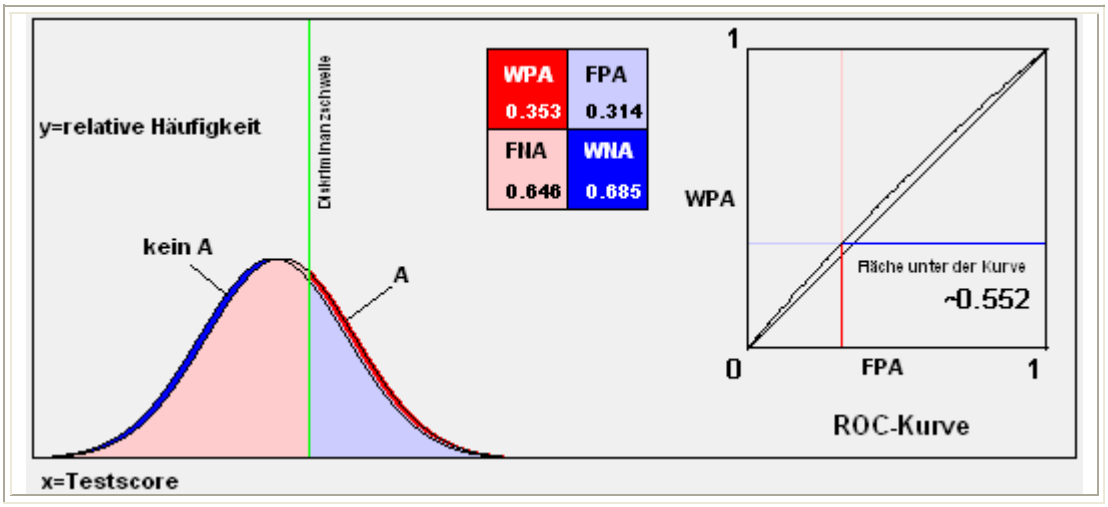


Abbildung 4-5 ROC-Kurve 2: Test 2

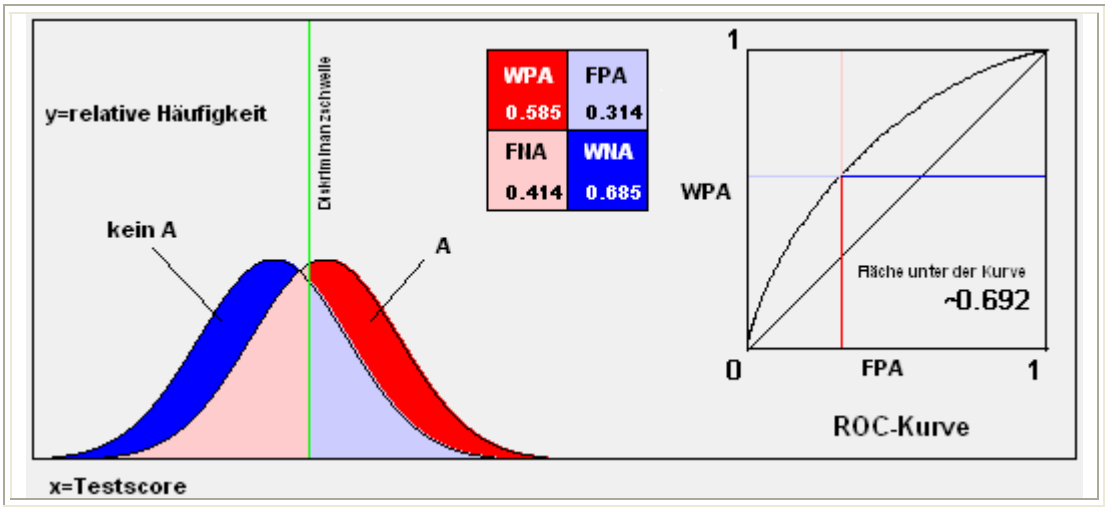
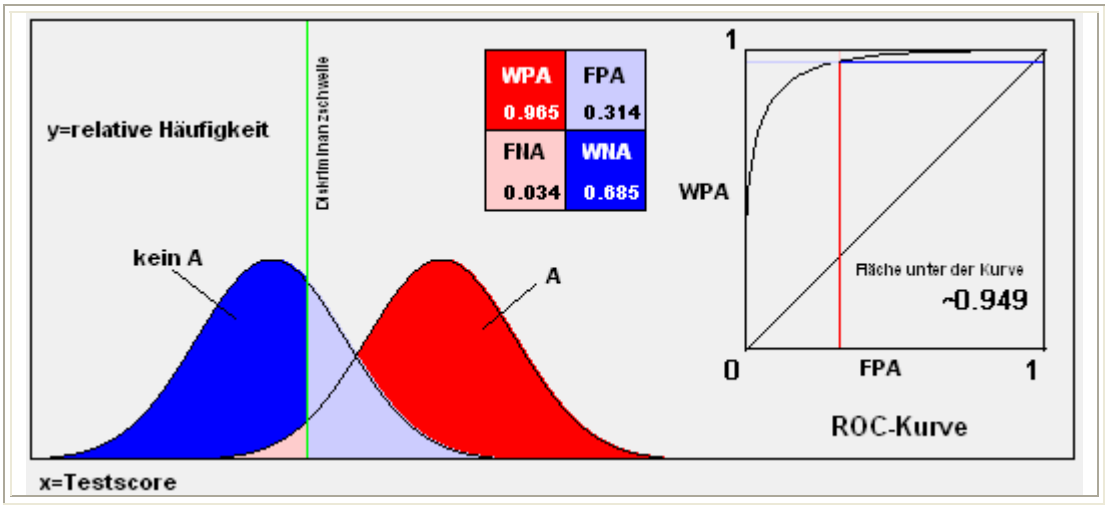


Abbildung 4-6 ROC-Kurve 3: Test 3



4.6.4 Wald-Effekt-Tests

Wald-Effekt-Tests (vgl. BACKHAUS et al., 2003) bieten eine Möglichkeit um die zur Vorhersage des Kriteriums berechneten Beiträge der einzelnen geschätzten Regressionskoeffizienten $\hat{\beta}_i$ bei kombinierten Prädiktoren einer statistischen Prüfung zu unterziehen.

Mit der in großen Stichproben annähernd χ^2 – verteilten Wald-Statistik:

$$(4.4) \quad W = \left(\frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}\right)^2,$$

werden folgende Hypothesen auf dem Signifikanzniveau α = 5% getestet:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

wobei $s_{\hat{\beta}_i}$ = Standardfehler der geschätzten Regressionskoeffizienten

Es ist zu beachten, dass der Wald-Effekt-Test ein sehr konservativer Test ist (HOSMER & LEMESHOW, 1989), der also die Annahme der Nullhypothese begünstigt.

In den aufgeführten Tabellen der Wald-Effekt-Tests bedeutet die Abkürzung FG die Anzahl unabhängiger Parameter der Regressionskoeffizienten.

4.7 Statistische Evaluation der Datengrundlage

Um die Fragestellungen im Ergebnisteil so zu bearbeiten, dass zulässige Schlussfolgerungen gezogen werden können, muss die Datengrundlage dieser Berechnungen geprüft und statistisch abgesichert werden. Um eine solide Ausgangsbasis zu schaffen werden die Testkennwerte in *Abschnitt 4.7.1*, die Auswirkungen unterschiedlicher Testreihenfolgen in *Abschnitt 4.7.2*, die Scorevergleiche der Diagnosegruppen in *Abschnitt 4.7.3*, das Validitätskriterium in *Abschnitt 4.7.4* und die Korrelationen sämtlicher Variablen in der Gesamtstichprobe in *Abschnitt 4.7.5* untersucht.

4.7.1 Vergleich der Testkennwerte

Ausgangspunkt für die Berechnungen der gebräuchlichen Kennwerte für den Vergleich von Testverfahren sind die Daten der Gesamtstichprobe (N=615).

Die gemittelten Schwierigkeits- und Trennschärfeindices sowie die Reliabilitätsmaße Cronbach-α der drei verwendeten Tests zeigt *Tabelle 4-8.*

Ergänzend zu diesen klassischen Testkennwerten wird auch ein Reliabilitätsmaß der probabilistischen Testtheorie angegeben, das von ROST (1996, S.354) behandelt wurde. Dieses Maß berücksichtigt den Anteil der Schätzfehlervarianzen $\sigma^2(E_{\xi_\upsilon})$ jedes einzelnen Personenparameters ξ_υ an der Gesamtvarianz der Personenparameter. Diese Fehlervarianzen werden im Raschmodell zu den beiden Personenparameterextremen hin größer und sind dort, wo ein Test die größte Messgenauigkeit besitzt, also im mittleren Bereich, am kleinsten.

Die zugehörige Berechnungsformel ist in der *Tabelle A-50* des Anhangs angegeben.

Tabelle 4-8 klassische Testkennwerte

Gesamtstichprobe (N=615)	Power-Test Raven	Power-Test Benton	Speed-Power-Test Cattell
Schwierigkeit Ø	0,61	0,58	0,44 (mit Inangriffnahmekorrektur)
Trennschärfe Ø	0,49	0,48	0,32 (mit Inangriffnahmekorrektur)
Cronbach-α #	0,84	0,83	0,54
Reliabilität ##	0,82	0,79	-

#=Klassische Testtheorie; ##=Probabilistische Testtheorie

Interpretation: Die klassischen Testkennwerte sind bei beiden Power-Tests ähnlich. Der Raventest ist geringfügig leichter und trennschärfer als der Bentontest. Im Gegensatz zum Cattelltest zeichnen sich Raven- und Bentontest außerdem durch relativ hohe Cronbach-α-Werte (>0,82) aus.

Ein Vergleich mit den Daumenregeln von WEISE (1975, S. 219)[46] ergibt für sämtliche Indices, bis auf das Cronbach-α des Cattelltests, Werte im mittleren Gütebereich.

Die probabilistisch berechneten Reliabilitäten der Personenparameter im einfachen Raschmodell unterscheiden sich bei den zwei analysierten Power-Tests nur unerheblich von den Cronbach-α-Werten der klassischen Testtheorie. Diese Ähnlichkeit belegt, dass die Cronbach-α-Werte bei Benton- und Raventest von großen Messfehlern verschont bleiben. Die Leistungsfähigkeiten von Raven- und Bentontest als Messinstrumente mittels Summenscores werden somit bestätigt.

Der Cattelltest unterscheidet sich von den beiden anderen Verfahren deutlich durch größere Itemschwierigkeiten und geringere Trennschärfeindices. Wird die „Inangriffnahme-Korrektur“ (LIENERT & RAATZ, 1994, S. 76) wegen der Speed-Komponente des Cattelltests berücksichtigt, führt dies zu einer geringfügigen Reduktion seiner Itemschwierigkeiten und Trennschärfeindices.

Der vergleichsweise kleine, hinsichtlich der Speed-Komponente unkorrigierte und deshalb zusätzlich „scheinerhöhte“ (LIENERT & RAATZ, 1994, S. 203) Cronbach-α-Wert (0,54) markiert die deutlich geringere und als niedrig zu bezeichnende Reliabilität des Cattelltests[47].

[46] WEISE (1975, S. 219) gibt folgende Richtgrößen für den „mittleren“ Gütebereich von Fähigkeitstests: Schwierigkeit (.80-.20), Trennschärfe (.30-.50) und Reliabilität (.80-.90).

[47] Eine „niedrige“ Reliabilität bedeutet nach WEISE (1975, S. 219) Werte ≤ .80

4.7.2 Scorevergleiche und Testreihenfolgen

Die vollständige Testbatterie wurde bei der Datenerhebung in zufällig variierten Testreihenfolgen vorgelegt, um Positionseffekte weitestgehend zu eliminieren (*Tabelle 4-9*).

Tabelle 4-9 Testreihenfolgen

	1.Test	2.Test	3.Test	# Personen
1.)	Raven	Cattell	Benton	89
2.)	Raven	Benton	Cattell	93
3.)	Cattell	Benton	Raven	108
4.)	Cattell	Raven	Benton	91
5.)	Benton	Cattell	Raven	141
6.)	Benton	Raven	Cattell	73
				Σ 595

Um Einflüsse der Testreihenfolge auf die Testleistung auszuschließen, wurde eine Varianzanalyse der erhaltenen Testscores bei unterschiedlichen Abfolgen der Tests berechnet. Hierzu wurde die Gesamtstichprobe verwendet. *Tabelle 4-10* zeigt die Ergebnisse.

Tabelle 4-10 Varianzanalysen der Testscores bei unterschiedlicher Testreihenfolge

Varianzanalysen der Testscores bei unterschiedlicher Testreihenfolge in der Gesamtstichprobe (N=599)						
	FG	QS	$\hat{\sigma}^2$	F	Prob > F	Sign.
Ravenscore						
Testreihen-folgen	5	42,137	8,4274	0,4664	0,8013	n.s.
Fehler	589	10642,09	18,0681			
Total	594	10684,23				
Cattellscore						
Testreihen-folgen	5	18,7224	3,74448	1,0810	0,3697	n.s.
Fehler	589	2040,313	3,46403			
Total	594	2059,035				
Bentonscore						
Testreihen-folgen	5	86,369	17,2737	0,8828	0,4922	n.s.
Fehler	589	11524,38	19,5660			
Total	594	11610,75				

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

In den Varianzanalysen ergeben sich keine signifikanten Unterschiede. Die verglichenen Gruppen mit unterschiedlichen Testreihenfolgen unterscheiden sich in keinem der erhobenen Leistungsmaße.

Mögliche leistungsmindernde Effekte, wie Ermüdung oder Nachlassen der Motivation im zeitlichen Verlauf der bearbeiteten Testbatterie können daher bei sämtlichen nachfolgenden Berechnungen vernachlässigt werden.

4.7.3 Scorevergleiche der Diagnosegruppen in der Altersgruppe 0 (≤ 39. Lebensjahr)

Die jüngste Altersgruppe 0 umfasst in der Gesamtstichprobe auch die größte Personenanzahl. Deshalb wurde diese Altersgruppe ausgewählt, um für sämtliche darin enthaltenen Diagnosegruppen die Scoremittelwerte der drei Tests zu berechnen[48].

Signifikante Testscoreunterschiede ergeben sich mit parameterfreien Van der Waerden-Tests nur zwischen klinisch auffälligen und unauffälligen Personen. Unterschiede der Scores innerhalb der klinischen Diagnosegruppen 1-5 lassen sich bei keinem der verwendeten Testverfahren nachweisen (*Tabellen 4-11* und *4-12*).

Wegen teilweise geringer Zellbesetzungen der Diagnosegruppen sind die Bildungsgrade nicht parallelisierbar. Kovariationen mit dem Bildungsgrad sind daher in den Ergebnissen nicht auszuschließen. Die Aussagekraft dieser Resultate bleibt deshalb fraglich.

Tabelle 4-11 Scoremittelwerte in den Diagnosegruppen der Altersgruppe 0

Scoremittelwerte in den Diagnosegruppen der Altersgruppe 0 (≤ 39.Lebensjahr)									
Diagnose-gruppe	**Ravenscore**			**Cattellscore**			**Bentonscore**		
	N	Mittelwert	σ	N	Mittelwert	σ	N	Mittelwert	σ
0	155	14,37	3,65	155	5,86	1,49	155	14,43	3,41
1	32	13,34	3,55	32	5,38	1,66	32	12,5	3,38
2	17	11,59	4,27	19	5,00	1,86	20	10,6	3,87
3	22	12,05	4,54	22	4,82	1,79	22	11,59	4,98
4	11	9,55	5,03	11	4,82	2,04	11	9,36	5,24
5	23	12,13	3,28	23	5,04	1,61	22	11,68	4,11
Σ	260			262			262		

σ = Standardabweichung

[48] Kodierung der Diagnosen in *Tabelle 4-1*

Tabelle 4-12 Van der Waerden Tests der Scores sämtlicher und klinischer Diagnosegruppen

Van der Waerden Tests sämtlicher Diagnosegruppen 0-5			**Van der Waerden Tests klinischer Diagnosegruppen 1-5**		
Ravenscore (N=260)			**Ravenscore (N=105)**		
χ^2	FG	Prob> χ^2	χ^2	FG	Prob> χ^2
25,7542	5	<,0001***	6,4960	4	0,1650
Cattellscore (N=262)			**Cattellscore (N=107)**		
χ^2	FG	Prob> χ^2	χ^2	FG	Prob> χ^2
16,1346	5	0,0065**	1,5517	4	0,8174
Bentonscore (N=262)			**Bentonscore (N=107)**		
χ^2	FG	Prob> χ^2	χ^2	FG	Prob> χ^2
38,4575	5	<,0001***	5,2871	4	0,2591

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

4.7.4 Statistische Absicherung des Validitätskriteriums

Ein methodisches Problem stellt die Wahl des Validitätskriteriums „klinische Auffälligkeit“ dar. Es soll in dieser Arbeit von einer sehr heterogenen Patientengruppe gebildet werden. Aus zeitökonomischen Gründen bei der Versuchspersonenrekrutierung beinhaltet diese Gruppe psychiatrische und neurologische Erkrankungen unterschiedlichster Ätiologie **mit** oder **ohne** cerebraler Beteiligung.

Zweifelhafte Validität des Kriteriums ergibt sich, weil die kognitiven Leistungen die zur jeweiligen Testlösung erforderlich sind, nicht bei allen vorkommenden Diagnosen aufgrund unterschiedlicher Pathologien in gleicher Weise verändert oder eingebüßt werden. Eine möglichst umfangreiche Patientenstichprobe ist aber notwendig, weil in einem nachfolgenden Teil dieser Arbeit Klassenparameter bestimmt werden, die eine weitere Aufteilung in kleinere Klassen zur Folge haben.

Um eine hinreichende statistische Absicherung dieser Problematik zu erreichen, wurden aus der Gesamtstichprobe drei anzahlmäßig gleiche, vom diagnostischen Spektrum weit gefasste, jedoch voneinander klar abgrenzbare Gruppen gebildet. Auf diese Weise werden möglichst umfangreiche, hinsichtlich Altersgruppen und Bildungsgraden parallelisierbare Gruppengrößen erreicht.

Diese Gruppen sind:

G: klinisch unauffällige Personen

P: psychiatrische Patienten **ohne** cerebrale Beteiligung

N: neurologische Patienten **mit** cerebraler Beteiligung

Zusammen bilden die drei parallelisierten Gruppen gleicher Größe die schon in *Abschnitt 4.1.2* vorgestellte Stichprobe „GPN parallel“ (N=210).

Die Stichprobe „GPN parallel“ soll folgende **Fragen** beantworten:

Frage a) Lassen sich die Patientengruppen P und N mit dem Testscore von den unauffälligen Personen G unterscheiden?

Frage b) Unterscheiden sich die Testscores der Patientengruppen P und N?

Frage c) Unterscheiden sich die Tests in der Trennschärfe der Diagnosegruppen G, P und N?

Folgende Erwägungen müssen in Hinblick auf die weiteren Analysen berücksichtigt werden:
Wenn die Frage a) mit „Ja“ beantwortet werden kann, deutet dieses Resultat auf die grundsätzliche klinische Brauchbarkeit des verwendeten Testverfahrens hin.
Wenn die Fragen b) und c) mit „Nein“ beantwortet werden können, unterscheiden sich die beiden Patientengruppen P und N statistisch nicht voneinander. Eine Zusammenfassung der Patienten in **einer** großen Gruppe „klinisch auffälliger“ Personen ist dann vertretbar.
In *Abschnitt 4.7.4.1* werden die Fragen a) und b), in *Abschnitt 4.7.4.2* wird Frage c) bearbeitet. Das weitere Vorgehen bei dieser Untersuchung wird unter Berücksichtigung der gewonnenen Resultate in *Abschnitt 4.7.4.3* klargelegt.

4.7.4.1 Testscoreunterschiede der drei Diagnosegruppen G, P und N

Zur Abklärung der Fragestellungen a) und b), werden die Testscoremittelwerte der drei Gruppen G, P und N berechnet (*Tabelle 4-13*) und Gruppenunterschiede mit einer Varianzanalyse (*Tabelle 4-14*) geprüft. Die Ergebnisse der t-Test-Einzelvergleiche im Falle signifikanter Varianzanalysen bei Cattell- und Bentontest werden in *Tabelle 4-15* zusammengefasst. Die ausführlichen Kontraste finden sich in der *Tabelle A-29* des Anhangs.

Tabelle 4-13 Testscoremittelwerte in der Stichprobe „GPN parallel“

Testscoremittelwerte in der Stichprobe: „GPN parallel“ (N=210)						
	Klinisch unauffällige Personen (G)		**Psychiatrische Erkrankungen ohne cerebrale Beteiligung (P)**		**Neurologische Erkrankungen mit cerebraler Beteiligung (N)**	
	Ø-Score	σ	Ø-Score	σ	Ø-Score	σ
Raventest	12,61	3,80	11,63	4,58	11,90	3,39
Cattelltest	5,3	1,70	4,56	1,93	4,61	1,59
Bentontest	12,57	3,70	10,84	4,30	10,87	4,22

σ= Standardabweichung

Tabelle 4-14 Varianzanalysen der Testscores in der Stichprobe „GPN parallel“

Varianzanalysen der Testscores in der Stichprobe: „GPN parallel“ (N=210)						
	FG	QS	$\hat{\sigma}^2$	F	Prob > F	Sign.
Ravenscore						
Gruppen GPN	2	35,85	17,93	1,15	0,3174	n.s.
Fehler	205	3184,57	15,54			
Total	207	3220,42				
Cattellscore						
Gruppen GPN	2	23,84	11,92	3,91	0,0216*	s.
Fehler	204	621,90	3,05			
Total	206	645,74				
Bentonscore						
Gruppen GPN	2	136,67	63,34	4,10	0,0179*	s.
Fehler	204	3396,20	16,65			
Total	206	3532,87				

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 4-15 Zusammenfassung der Kontraste in der Stichprobe „GPN parallel“

Zusammenfassung der Kontraste beim Cattelltest (*Tab.A-29*)	G	P	N
G		ns	*
P	ns		ns
N	*	ns	
*p≤ 0,0166 (Bonferoni-Korrektur)			

Zusammenfassung der Kontraste beim Bentontest (*Tab.A-29*)	G	P	N
G		*	*
P	*		ns
N	*	ns	
*p≤ 0,0166 (Bonferoni-Korrektur)			

Der empirisch ermittelte F-Wert ist beim Cattell- und Benton-, nicht jedoch beim Ravenscore auf dem 5%-Niveau signifikant.
Diese Resultate bedeuten, dass nur mit dem Cattell- und Bentontest statistisch bedeutsame Auswirkungen der Gruppenzugehörigkeiten G, P und N auf die Scores festgestellt werden konnten.
Der Raventest scheint eine resistentere Fähigkeit zu erfassen. Diese Fähigkeit zeigt keine signifikanten Einbußen durch pathologische Prozesse. Klinische Auffälligkeiten sind demnach mit dem Ravenscore nicht erkennbar.
Die Einzelvergleiche von Cattell- und Bentontest sind in *Tabelle 4-15* aufgeführt.

Beantwortung der Frage a) „Lassen sich die Patientengruppen P und N mittels Testscore von den unauffälligen Personen G unterscheiden?“
Signifikante Scoreunterschiede zeigen die t-Test-Einzelvergleiche (*Tabelle 4-15*) ausschließlich zwischen klinisch auffälligen und unauffälligen Personen bei Cattell- und Bentontest.
Nur der Cattelltest scheitert äußerst knapp beim Vergleich zwischen klinisch unauffälligen Personen (G) und psychiatrischen Patienten **ohne** cerebrale Beteiligung (P) an der Bonferoni-korrigierten Signifikanzgrenze (vgl. *Tabelle A-29* im Anhang).
Der klinische Einsatz des Benton- und Cattelltests zur Aufdeckung pathologischer Leistungseinbußen scheint grundsätzlich sinnvoll zu sein.

Beantwortung der Frage b) „Unterscheiden sich die Testscores der Patientengruppen P und N?“
Es zeigen sich in *Tabelle 4-15* keine signifikanten Scoreunterschiede der beiden Patientengruppen P und N.

Bewertung der Resultate:
Beide klinischen Gruppen unterscheiden sich in ähnlicher Weise von den klinisch unauffälligen Personen. Sie unterscheiden sich nicht untereinander. In Hinblick auf eine möglichst große Stichprobe klinisch auffälliger Personen als Validitätskriterium „klinische Auffälligkeit“, sprechen diese Ergebnisse für die Zusammenlegung der beiden untersuchten Patientengruppen P und N.

4.7.4.2 Testtrennschärfen der Diagnosegruppen

Zur Beurteilung der Fragestellung c), ob sich die Tests in der Trennschärfe der Diagnosegruppen G, P und N unterscheiden, wird die nominal logistische Regression verwendet.
Die Datengrundlage bildet die Stichprobe „GPN parallel“ (N=210).
Folgende **drei Fälle** sind im Rahmen dieser Berechnungen zu unterscheiden:

Fall 1: Der jeweilige Test unterscheidet erfolgreich Personen ohne klinische Auffälligkeit (G) von den beiden Patientengruppen (P und N);

Fall 2: Der jeweilige Test unterscheidet erfolgreich psychiatrische Erkrankungen **ohne** cerebrale Beteiligung (P) von neurologischen Erkrankungen **mit** cerebraler Beteiligung und klinischer Unauffälligkeit (N und G);

Fall 3: Der jeweilige Test unterscheidet erfolgreich neurologische Erkrankungen **mit** cerebraler Beteiligung (N) von psychiatrischen Erkrankungen **ohne** cerebrale Beteiligung und klinischer Unauffälligkeit (P und G).

Die Resultate der statistischen Prüfungen mit nominal logistischen Regressionen sind in *Tabelle 4-16* aufgeführt. Die sehr signifikanten Ergebnisse zeigen, dass nur Cattell- und Bentontest in der Lage sind, allgemein zwischen klinischer Auf- oder Unauffälligkeit, dem Fall 1 zu differenzieren.
Eine trennscharfe Unterscheidung der diagnostischen Gruppen der Fälle 2 und 3 gelingt nicht.
Mit dem Raventest ist keine signifikante Differenzierung der Diagnosegruppen möglich.

Vergleich der Trennschärfe der Diagnosegruppen G, P und N

Ausgehend von Fragestellung c) „Unterscheiden sich die Tests in der Trennschärfe der Diagnosegruppen G, P und N?“ lassen sich die statistischen Unterschiede folgendermaßen bewerten: Bei Benton- und Cattelltest zeigen sich analoge Resultate. Mit diesen beiden Tests können Personen ohne klinische Auffälligkeiten von Patientengruppen unterschieden werden. Der Raventest erlaubt keine relevante Unterscheidung der Gruppen.
In Hinblick auf das Validitätskriterium „klinische Auffälligkeit“, stützt dieses Ergebnis die Zusammenlegung beider Patientengruppen P und N zu einer großen Gruppe klinisch auffälliger Personen.

Tabelle 4-16 nominal logistische Regressionen der Stichprobe „GPN parallel“

Fall 1) Kriterium: Klinische (Un-)Auffälligkeit Stichprobe: „GPN parallel“(N=210) - nominal logistisches Modell				
Prädiktorvariable	**Modell-Test**			**Entropiereduktion**
	FG	χ^2	Prob> χ^2	%
Raven Score	1	2,181	0,1397	0,82
Cattell Score	1	7,844	0,0051**	2,96
Benton Score	1	8,311	0,0039**	3,14
Alle drei Testscores	3	12,40	0,0061**	4,71
Fall 2) Kriterium: psychiatrische Erkrankung ohne cerebrale Beteiligung Stichprobe: „GPN parallel“ - nominal logistisches Modell				
Prädiktorvariable	**Modell-Test**			**Entropiereduktion**
	FG	χ^2	Prob> χ^2	%
Raven Score	1	1,151	0,2832	0,44
Cattell Score	1	2,312	0,1283	0,88
Benton Score	1	2,130	0,1444	0,81
Alle drei Testscores	3	3,058	0,3828	1,18
Fall 3) Kriterium: neurologische Erkrankung mit cerebraler Beteiligung Stichprobe: „GPN parallel“ - nominal logistisches Modell				
Prädiktorvariable	**Modell-Test**			**Entropiereduktion**
	FG	χ^2	Prob> χ^2	%
Raven Score	1	0,159	0,6898	0,06
Cattell Score	1	1,564	0,2111	0,59
Benton Score	1	1,898	0,1683	0,72
Alle drei Testscores	3	4,138	0,2469	1,59

*p≤ 0,05;**p≤ 0,01; ***p≤ 0,001

4.7.4.3 Konklusion und Forschungskonsequenz

Die gewonnenen Resultate haben bezüglich des weiteren praktischen Vorgehens im Ergebnisteil *Kapitel 5* folgende Konsequenzen:

Den Berechnungen der Prädiktion „klinische Auffälligkeit“ wird die größere „parallelisierte Stichprobe“ (N=468), bestehend aus sämtlichen neurologischen und psychiatrischen Patientengruppen (vorgestellt in *Abschnitt 4.1.3*) zugrunde gelegt.

Als jeweilige Ergänzung und Vergleichsmöglichkeit wird für die Berechnungen der Prädiktion „cerebrale Beteiligung“ die kleinere „Stichprobe `parallel-cerebral`“ (N=269)

herangezogen. Diese umfasst nur neurologische Patienten **mit** cerebralen Krankheitsbildern (vorgestellt in *Abschnitt 4.1.4*)

4.7.5 Korrelationen in der Gesamtstichprobe

Tabelle 4-17 gibt einen Überblick über die Produkt-Moment-Korrelationen der Tests untereinander, sowie mit anderen per Fragebogen erhobenen Merkmalen der Testpersonen in der Gesamtstichprobe (N=615).

Tabelle 4-17 Korrelationen unterschiedlicher Variablen in der Gesamtstichprobe

Gesamt-stich-probe N=615	**Raven Score**	**Cattell Score**	**Benton Score**	**Klinisch. Auffällig-keit**	**Cerebr. Beteili-gung**	**Alter**	**Bildungs-grad**	**Ge-schlecht**
Raven Score		0,61*** (N=601)	0,68*** (N=601)	-0,19*** (N=605)	-0,13** (N=605)	-0,31*** (N=605)	0,44*** (N=602)	0,01 (N=605)
Cattell Score	0,61*** (N=601)		0,55*** (N=604)	-0,24*** (N=607)	-0,22*** (N=607)	-0,40*** (N=607)	0,32*** (N=604)	-0,04 (N=607)
Benton Score	0,68*** (N=601)	0,55*** (N=604)		-0,29*** (N=610)	-0,23*** (N=610)	-0,39*** (N=610)	0,41*** (N=607)	-0,06 (N=610)
Klinisch. Auffällig-keit	-0,19*** (N=605)	-0,24*** (N=607)	-0,29*** (N=610)		0,59*** (N=615)	0,12** (N=615)	-0,20*** (N=612)	0,11** (N=615)
Cerebr. Beteili-gung	-0,13** (N=605)	-0,22*** (N=607)	-0,23*** (N=610)	0,59*** (N=615)		0,29*** (N=615)	0,02 (N=612)	0,02 (N=615)
Alter	-0,31*** (N=605)	-0,40*** (N=607)	-0,39*** (N=610)	0,12** (N=615)	0,29*** (N=615)		-0,04 (N=612)	0,10* (N=615)
Bildungs-grad	0,44*** (N=602)	0,32*** (N=604)	0,41*** (N=607)	-0,20*** (N=612)	0,02 (N=612)	-0,04 (N=612)		-0,11** (N=612)
Ge-schlecht	0,01 (N=605)	-0,04 (N=607)	-0,06 (N=610)	0,11** (N=615)	0,02 (N=615)	0,10* (N=615)	-0,11** (N=612)	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Zusammenfassung der wichtigsten Ergebnisse in *Tabelle 4-17*:

● Recht hohe positive Korrelationen der Tests untereinander, weisen darauf hin, dass ähnliche Fähigkeiten bei der Testbearbeitung und Lösungssuche bei Raven-, Benton- und Cattelltest eine Rolle spielen. Der mit 0,68 höchste korrelative Zusammenhang ergibt sich dabei zwischen Benton- und Raventest.
Mit einer Hauptkomponentenanalyse werden diese Testzusammenhänge im nächsten *Abschnitt 4.7.6* differenzierter betrachtet.

● Wie schon von den jeweiligen Testautoren angenommen, bestätigt sich in den Korrelationen eine Alters- und Bildungsabhängigkeit der Testleistungen. Die erreichten Testscores werden bei allen verwendeten Tests mit zunehmendem Alter schlechter, mit höherer Bildung besser.

● Die Geschlechtsunterschiede bei den Testergebnissen sind vernachlässigbar klein, es sind jedoch mehr klinisch auffällige, ältere und weniger gebildete Männer als Frauen in der Gesamtstichprobe enthalten.

● Sämtliche Testleistungen sind in der Gesamtstichprobe negativ mit „klinische Auffälligkeit“ und „cerebrale Beteiligung“ korreliert. Beide Kriterien korrelieren ebenso mit Alter und Bildung. Zur Kontrolle dieser Störeffekte wurden die zwei Teilstichproben mit parallelisierten Altersgruppen und Bildungsgraden gebildet (vgl. *Abschnitte 4.1.3* und *4.1.4)*:
„Parallelisierte Stichprobe“ (N=468) für das Validitätskriterium „klinische Auffälligkeit“ und „Stichprobe `parallel-cerebral`“ (N=269) für das Validitätskriterium „cerebrale Beteiligung“.

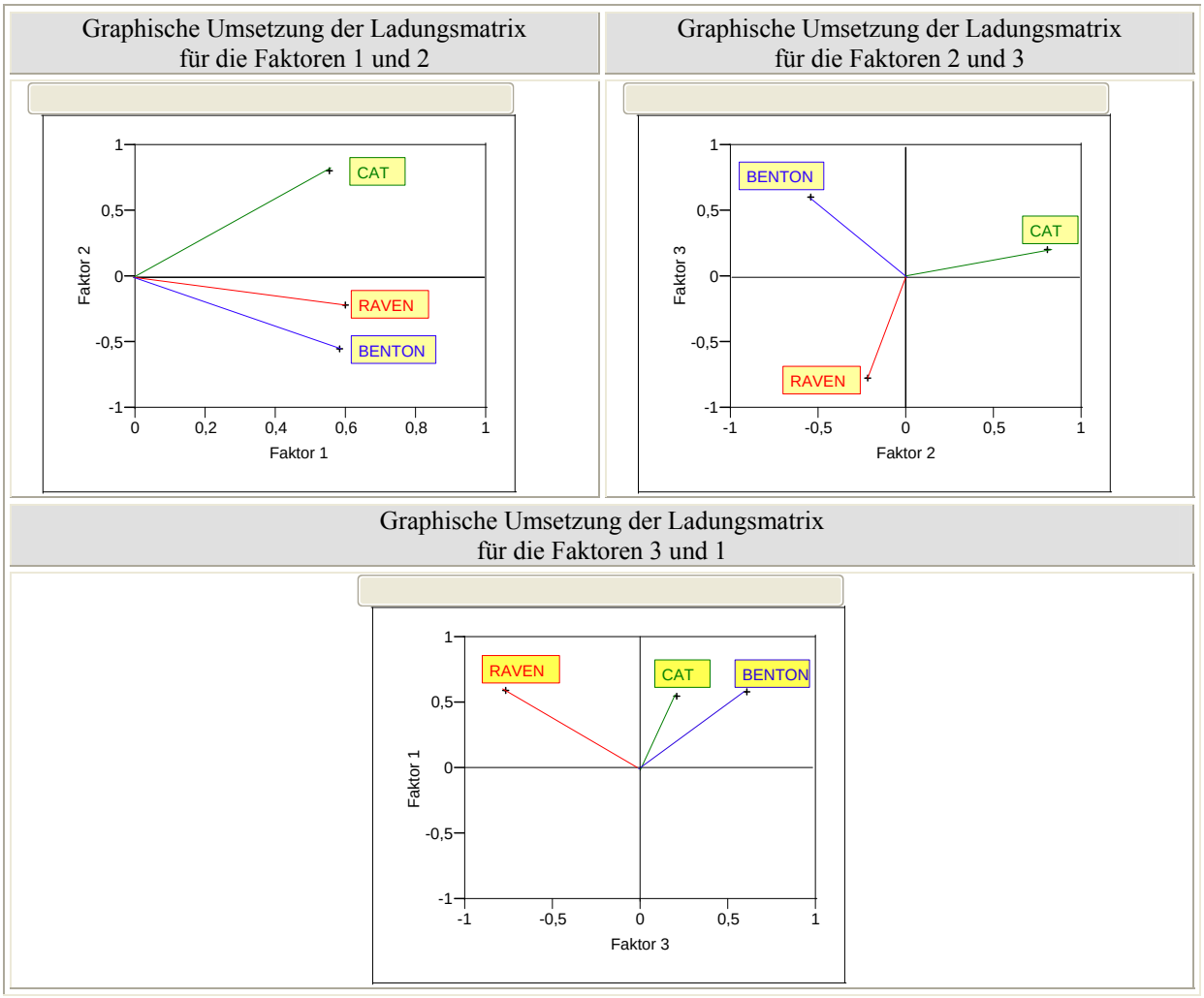
4.7.6 Ergebnisse der Hauptkomponentenanalyse

Um die korrelativen Beziehungen der Scores der drei verwendeten Tests genauer abzuklären, wurde eine Hauptkomponentenanalyse mit drei Faktoren gerechnet. Die Ergebnisse sind in *Tabelle 4-18* dargestellt. *Abbildung 4-7* verdeutlicht graphisch die Zusammenhänge.

Tabelle 4-18 Hauptkomponentenanalyse der Gesamtstichprobe

Gesamtstichprobe (N=615)	**Faktor 1**	**Faktor 2**	**Faktor 3**
Raven Score	0,598	-0,217	-0,772
Cattell Score	0,554	0,808	0,201
Benton Score	0,580	-0,548	0,603
aufgeklärte Varianz	74,0%	15,7%	10,3%

Abbildung 4-7 Graphische Darstellung der Ladungsmatrix der Hauptkomponentenanalyse



Wie die graphische Umsetzung der Beziehung zwischen Tests und Faktoren erkennen lässt, laden die drei verwendeten Tests gleichmäßig hoch auf dem ersten Faktor.

Der **erste Faktor** (erklärte Varianz 74%) scheint somit die gemeinsame Haupteigenschaft der Tests zu repräsentieren und kann als Erklärung für die hohen Testinterkorrelationen dienen. Er soll in Anlehnung an Thurstone`s Primärfaktoren (JÄGER, 1967) mit „Anschauungsgebundenem Denken (Space)“ bezeichnet werden.

Auf dem **zweiten Faktor** (erklärte Varianz 15,7%) zeigt der Cattelltest positive Ladungen, während sich Raven- und Bentontest mit umgekehrtem Vorzeichen absetzen.

Dieser zweite Faktor soll mit der Dimension „Power←→Speed“ beschrieben werden.

Der **dritte Faktor** (erklärte Varianz 10,3%) positiv für Benton- und ausgeglichen für Cattelltest, wird negativ für den Raventest. Er scheint für eine Komponente des Arbeitsgedächtnisses zu stehen. Auf der Dimension „schwache Beteiligung ←→ starke Beteiligung des Arbeitsgedächtnisses“, ist dieser dritte Faktor besonders beim Bentontest wichtig für die kurzfristige Fokussierung der Aufmerksamkeit und damit den Erfolg beim Einprägen und Wiedergeben der Bildvorlagen.

Durch die mit den Faktoren 2 und 3 beschriebenen spezifischen Komponenten, werden die drei Tests optimal differenziert. Diese Tatsache kommt in der *Abbildung 4-7 (oben rechts)* durch die 120° -Winkel der Faktorendarstellung deutlich zum Ausdruck.

5 ERGEBNISSE

In diesem *Kapitel* werden die eigentlichen Fragestellungen bearbeitet.
In *Abschnitt 5.1* zunächst für den Raventest, in *Abschnitt 5.2* für den Bentontest und in *Abschnitt 5.3* für den Cattelltest.
In *Abschnitt 5.4* werden Raven-, Benton- und Cattelltest gemeinsam als Testbatterie analysiert.
Der *Abschnitt 5.5* gibt eine Zusammenfassung dieser Ergebnisse.
In *Abschnitt 5.6* werden zwei Latente-Klassen-Analysen der gesamten Testbatterie vorgestellt. Eine Latente-Klassen-Analyse wird mit Variablen des klassischen und die andere mit Variablen des probabilistischem Auswertungsmodus durchgeführt.

5.1 Raventest

5.1.1 Prüfung der Homogenität

In den folgenden *Abschnitten* wird die Personenhomogenität als die zentrale Forderung der Raschhomogenität bei der Itemlösung des Raventests mit Modellgeltungstests geprüft. Der graphische Modelltest und der „Andersen"-Test werden mit einem externen und internen Teilungskriterium angewendet.
Die Zweiteilung der Gesamtstichprobe in „klinisch unauffällige" und „klinisch auffällige" Personen dient als externes Teilungskriterium.
Für das interne Kriterium „Testleistung" wird die Gesamtstichprobe durch einen Mediansplit in die Gruppe der Hoch- und Niedrigscorer aufgeteilt. Dabei werden die Personen mit einem Rohscorewert von ≥ 14 den „Hochscorern" und ≤ 13 den Niedrigscorern zugeordnet.

5.1.1.1 Überprüfung des Raschmodells mit graphischen Modelltestungen

Die Itemparameter σ_i der jeweiligen Teilstichproben des Raventests (*Tabelle A-30* im Anhang) werden für die graphischen Modellkontrollen verwendet. Gleiche Schwierigkeitsunterschiede bedeuten in der logarithmierten, summennormierten Form gleich große Abstände. Die Parameterschätzungen aus der Gesamtstichprobe „Raven Gesamt" sind

in den Darstellungen auf der Abszisse, die Schätzungen aus den Teilstichproben auf der Ordinate aufgetragen.

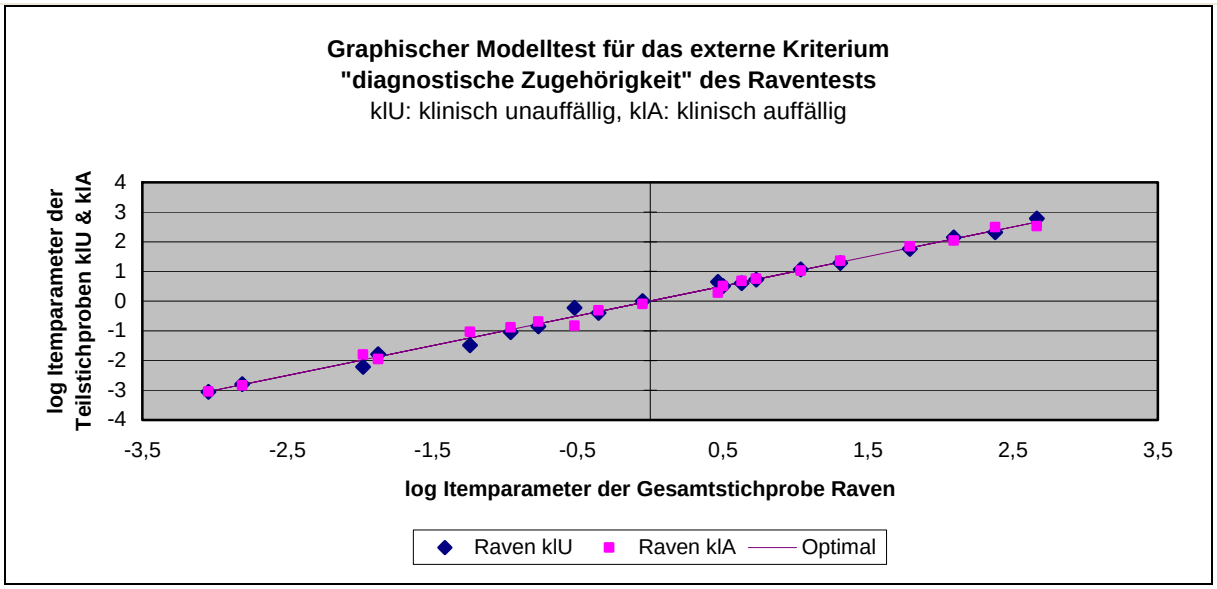
Für das externe Kriterium „diagnostische Zugehörigkeit“ wird die graphische Modellkontrolle in *Abbildung 5-1* gezeigt. *Abbildung 5-2* zeigt den graphischen Modelltest für das interne Kriterium „Testleistung“.

Bei optimaler Modellpassung befinden sich sämtliche Punkte exakt auf der durchgezogenen Geraden y=x.

In beiden *Abbildungen* sind keine groben Modellverletzungen erkennbar.

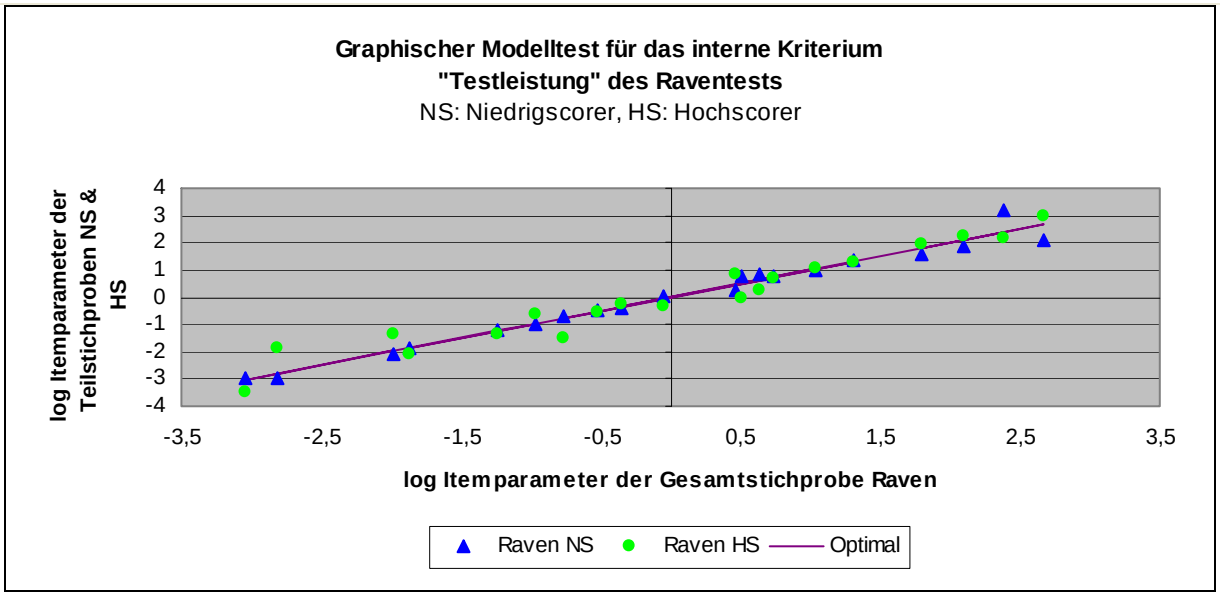
Die berechneten Produkt-Moment-Korrelationen der Itemparameter unter den *Abbildungen* bieten eine Vergleichsmöglichkeit des externen und internen Kriteriums: Die geringfügig kleinere Produkt-Moment-Korrelation beim Teilungskriterium „Testleistung“ deutet auf unterschiedlichere Itemschwierigkeiten der Hoch- und Niedrigscorer, als der klinisch auffälligen und unauffälligen Personen hin.

Abbildung 5-1 Graphischer Modelltest für das externe Kriterium beim Raventest



Die Produkt-Moment-Korrelation zwischen Raven klU und klA beträgt: 0,989

Abbildung 5-2 Graphischer Modelltest für das interne Kriterium beim Raventest



Die Produkt-Moment-Korrelation zwischen Raven NS und HS beträgt: 0,939

5.1.1.2 Überprüfung des Raschmodells mit der „Andersen-Testmethode“

Die statistische Prüfung der Raschmodellgültigkeit beim Raventest mittels „Andersen-Methode“ wird analog zum graphischen Modelltest mit einem externen und internen Kriterium durchgeführt.

5.1.1.2.1 Modelltestung anhand des externen Teilungskriteriums „diagnostische Zugehörigkeit“

Tabelle 5-1 fasst die marginalen Likelihoods beim Raventest und die daraus resultierende statistische Prüfgröße für das externe Teilungskriterium „diagnostische Zugehörigkeit“ zusammen.

Tabelle 5-1 „Andersen-Methode“ (Teilungskriterium „diagnostische Zugehörigkeit“) beim Raventest

Stichprobe	N	ln(mL)	n_p
Raven klU	327	-2983,04	21
Raven klA	278	-2711,23	21
Raven Gesamt	605	-5717,19	21
-2ln(LR)=45,84		FG=21	

Die ermittelte Prüfgröße beträgt 45,84. Sie liegt damit oberhalb der kritischen Grenze von 32,7 der χ^2-Verteilung (α =5%) bei 21 Freiheitsgraden.

Der Raventest verhält sich für das externe Teilungskriterium „diagnostische Zugehörigkeit“ nicht modellkonform. Die Null-Hypothese „Raschhomogenität“ muss verworfen werden.

5.1.1.2.2 Modelltestung anhand des internen Teilungskriteriums „Testleistung“

Die *Tabelle 5-2* fasst die marginalen Likelihoods beim Raventest und die daraus resultierende statistische Prüfgröße für das interne Teilungskriterium „Testleistung“ zusammen.

Tabelle 5-2 „Andersen-Methode“ (Teilungskriterium „Testleistung“) beim Raventest

Stichprobe	N	ln(mL)	n_p
RavenNS	343	-3434,74	21
RavenHS	262	-1900,64	21
RavenGesamt	605	-5717,19	21
-2ln(LR)= 763,62		FG=21	

Die ermittelte Prüfgröße beträgt 763,62. Sie liegt damit oberhalb der kritischen Grenze von 32,7 der χ^2-Verteilung (α =5%) bei 21 Freiheitsgraden.

Der Raventest verhält sich für das interne Teilungskriterium „Testleistung“ nicht modellkonform. Die Null-Hypothese „Raschhomogenität“ muss verworfen werden.

5.1.1.3 Geltungsprüfung unterschiedlicher probabilistischer Modelle mittels informationstheoretischer Maße und Likelihoodquotiententests

Die berechneten Kennwerte der informationstheoretischen Maße sind für die Gesamtstichprobe in *Tabelle 5-3* zusammengefasst.

Tabelle 5-3 Informationstheoretische Maße beim Raventest

N=605	loglikeli	AIC	BIC	n_p
RAVEN RM	-5717,19	11476,37	11568,88	21
RAVEN MRM2Kl	-5647,15	11380,31	11569,73	43
RAVEN MRM3Kl	-5572,65	11275,30	11561,64	65
RAVEN MRM4Kl	-5556,65	11287,30	11670,56	87
RAVEN LCA1Kl	-6567,69	13175,39	13263,49	20
RAVEN LCA2Kl	-5790,57	11663,15	11843,76	41
RAVEN LCA3Kl	-5622,13	11368,25	11641,37	62
RAVEN LCA4Kl	-5579,48	11324,96	11690,60	83
RAVEN LCA5Kl	-5540,36	11288,71	11746,86	104
RAVEN LCA6Kl	-5516,25	11282,50	11833,16	125
RAVEN LCA7Kl	-5497,58	11287,15	11930,32	146
saturiertes Modell	-3737,28	2104624,56	6723836,99	1048575
Startwert 4321 abgesichert mit 3321 und 2321				

In *Tabelle 5-3* wird ersichtlich, dass die informationstheoretischen Maße AIC (11275,3 < 11282,50) und BIC (11561,64 < 11641,37) dem Mixed-Rasch-Modell den Vorzug geben. Sowohl der BIC-Wert (11561,64), als auch der AIC-Wert (11275,30) favorisieren beim Raventest eine 3-Klassen-Lösung des Mixed-Rasch-Modells.

Dieses Ergebnis wird inferenzstatistisch mit Likelihoodquotiententests (vgl. *Tabelle 5-4*) abgesichert. Die große Parameteranzahl des saturierten Modells (n_p=1048575) und die vergleichsweise geringe Stichprobengröße (N=605) bedingen, dass die Vergleiche der Likelihoodquotienten nur begrenzt aussagekräftig sind.

Tabelle 5-4 Likelihoodquotiententests Raven

LR	-2ln(LR)	FG	Prob> χ^2
MRM3Kl/sat.Modell	3670,74	1048510	1
RM/MRM3Kl	289,08	44	<,0001***

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die Ergebnisse der beiden in *Tabelle 5-4* gebildeten Likelihoodquotienten bedeuten, dass die vom AIC und BIC-Wert indizierte 3-Klassen-Lösung des Mixed-Rasch-Modells mit den vorliegenden Daten vereinbar ist. Diese passt besser als das Raschmodell und nicht schlechter

als das saturierte Modell. Die 3-Klassen-Lösung wird deshalb als geeignetes Modell den weiteren Berechnungen zugrunde gelegt.

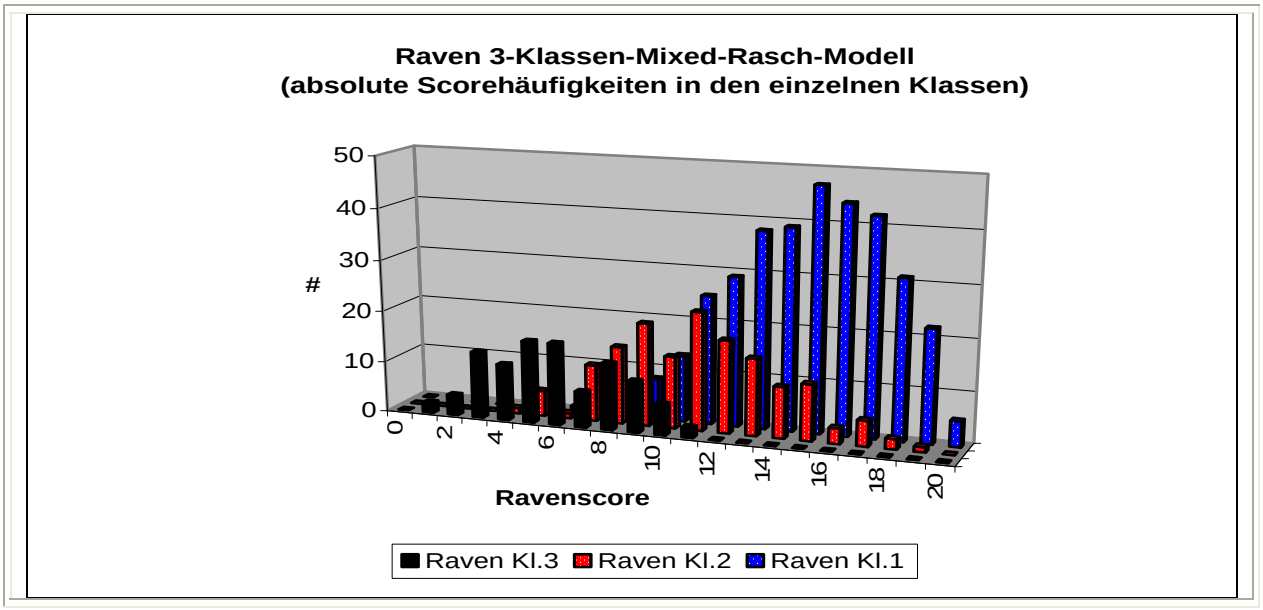
5.1.2 Interpretation latenter Klassen

Die indizierte 3-Klassen-Lösung des Mixed-Rasch-Modells wird im Detail vorgestellt. Dazu werden die einzelnen Klassen aufgrund der erhobenen Merkmale der zugeordneten Probanden voneinander abgegrenzt.

5.1.2.1 Summenscoreverteilung der 3-Klassen-Lösung des Mixed-Rasch-Modells

Abbildung 5-3 zeigt deutliche von der Testleistung abhängige klassenspezifische Zusammenhänge. Beim Raventest gibt es eine Klasse mit niedrigem, mittlerem und hohem Leistungsniveau.

Abbildung 5-3 Ravenscores in den Klassen



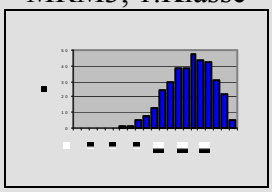
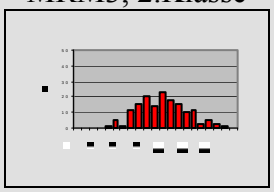
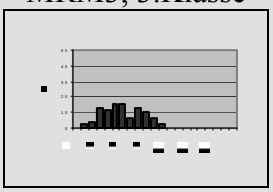
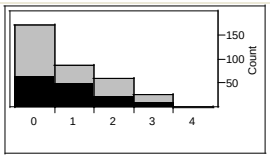
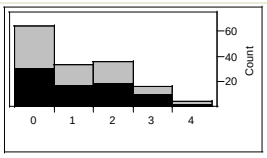
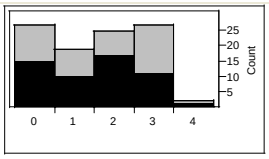
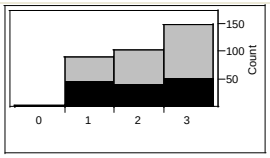
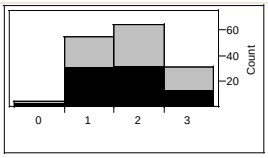
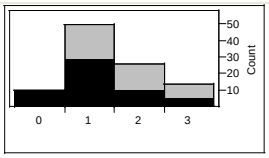
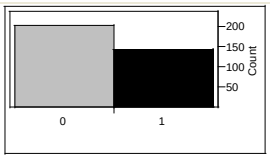
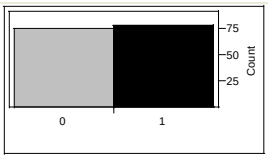
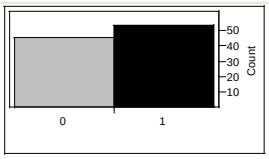
5.1.2.2 Charakterisierung der 3-Klassen-Lösung des Mixed-Rasch-Modells

Die drei Klassen werden bezüglich des Mittelwertes der Raven-Scoreleistung geordnet und mit den Bezeichnungen „Könner“ für Klasse 1, „Durchschnittliche“ für Klasse 2 und „Nicht-Könner“ für Klasse 3 versehen.

Die einzelnen Klassen werden in *Tabelle 5-5* charakterisiert.

Die Zusammenfassungen der Klassenkontraste sind in *Tabelle 5-6* aufgeführt. Die vollständigen Statistiken finden sich in der *Tabelle A-31* des Anhangs.

Tabelle 5-5 Die 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest

3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest			
	hoch ← Quantitative Testleistung → niedrig		
Raventest (N=605) Scoreverteilung Maßstab x=Score: 0-20 y=N: 0-50	MRM3, 1.Klasse  **„Könner“**	MRM3, 2.Klasse  **„Durchschnittlich“**	MRM3, 3.Klasse  **„Nicht-Könner“**
N (%-Anteil gesamt)	350 (57%)	155 (26%)	100 (17%)
mittlere klassenspezifische Zuordnungssicherheit	0,890	0,862	0,836
Cronbach-α #	0,69	0,70	0,49
Reliabilität##	0,68	0,78	0,51
Ravenscore [Ø]	14,62	11,02	5,94
Ravenscore [σ]	2,82	3,04	2,43
Alter [Ø]	39,41	43,41	49,04
Alter [σ]	14,04	14,53	14,16
Altersverteilung Hell: klinisch unauffällig Dunkel: klinisch auffällig 0=15-39LJ, 1=39-49LJ, 2=49-59LJ, 3=59-70LJ, 4= >70LJ			
Bildungsverteilung Hell: klinisch unauffällig Dunkel: klinisch auffällig 0=ohne Schulabschluss, 1=Hauptschule, 2=Realschule, 3=Abitur			
Klinische Auffälligkeit Hell=0: klinisch unauffällig Dunkel=1: klinisch auffällig			

#=Klassische Testtheorie; ##=Probabilistische Testtheorie (Reliabilität *vgl. Abschnitt 4.7.1*)

Aus den *Tabellen 5-5* und *5-6* geht hervor:

- In den Grafiken der *Tabelle 5-5* ist eine Alters- und Bildungsabhängigkeit der drei indizierten Klassen mit einer eindeutigen Tendenz erkennbar:

  In der 1. Klasse „Könner“ befinden sich junge und hoch gebildete Personen mit hohen Scores, in der 2. Klasse „Durchschnittliche“ Personen mittleren Alters und mittlerer Bildung mit durchschnittlichen Testleistungen und in der 3. Klasse „Nicht-Könner“ erreichen ältere und niedriger gebildete Personen eher niedrige Scores.

Diese Unterschiede bestätigen sich deutlich in den statistisch signifikanten Kontrasten (*Tabelle 5-6*).

- Die unterste Reihe der Grafiken zeigt eine tendenzielle Zunahme der klinisch auffälligen Personen in Richtung der Klassen niedrigerer Testleistung (aber auch höheren Alters und geringerer Bildung).
  Diese Unterschiede sind statistisch jedoch nicht signifikant (ebenfalls *Tabelle 5-6*).

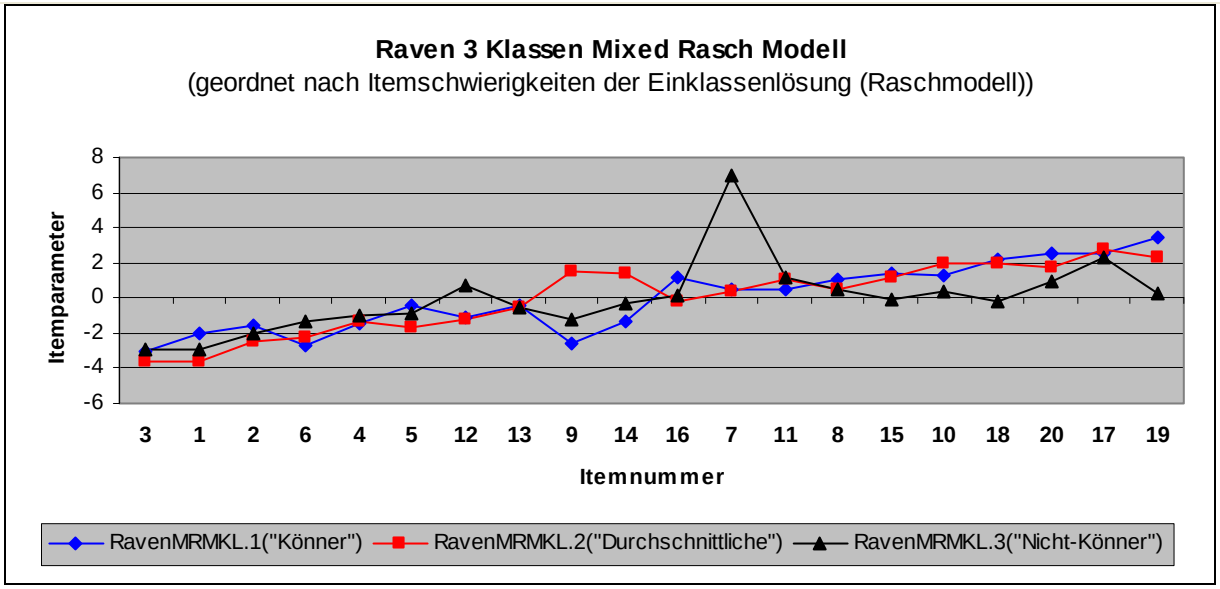
Tabelle 5-6 Zusammenfassung der Kontraste der 3-Kl.-Lösung des Mixed-Rasch-Modells beim Raventest

Kontraste der Scores beim Raventest	Könner	Durchschnitt	Nichtkönner
Könner		***	***
Durchschnittliche	***		***
Nichtkönner	***	***	
Kontraste des Lebensalters beim Raventest	Könner	Durchschnitt	Nichtkönner
Könner		*	***
Durchschnittliche	*		**
Nichtkönner	***	**	
Kontraste der Altersgruppen beim Raventest	Könner	Durchschnitt	Nichtkönner
Könner		ns	***
Durchschnittliche	ns		**
Nichtkönner	***	**	
Kontraste der Bildungsgrade beim Raventest	Könner	Durchschnitt	Nichtkönner
Könner		***	***
Durchschnittliche	***		**
Nichtkönner	***	**	
Kontraste der Kontingenztafeln (klin. Auff. ja/nein) beim Raventest	Könner	Durchschnitt	Nichtkönner
Könner		ns	ns
Durchschnittliche	ns		ns
Nichtkönner	ns	ns	
*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur) entnommen aus *Tab.A-31*			

5.1.2.3 Die Profile der Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells

In der *Abbildung 5-4* sind die Itemparameter der drei Klassen des Raventests nach den Itemschwierigkeiten des einfachen Raschmodells[49] geordnet.
Die zugehörigen Itemparameter sind in der *Tabelle A-32* des Anhangs aufgeführt.

Abbildung 5-4 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest



Bewertung der *Abbildung 5-4* und Fazit:

- Die bei der Datenerhebung verwendete Itemreihenfolge 1-20 entspricht nicht der berechneten Itemschwierigkeitsreihenfolge. Das vollständig verwendete Set 1 der APM[50] (Item 1- 12) ist diesen Ergebnissen zufolge nicht nach Itemschwierigkeit geordnet (vgl. Itemnummern der Abszisse).
- Das Item 7 ist in der *Abbildung* besonders auffällig: Es ist für die kleinste Klasse 3 „Nicht-Könner“ das mit Abstand schwerste Item. Für die Items 9, 14, 18 und 19 ergeben sich weniger ausgeprägte, aber ebenfalls erkennbare Inhomogenitäten.
- Die hochsignifikanten und positiven Spearman-Rho-Rangkorrelationen, aufgeführt in *Tabelle A-33* des Anhangs belegen einen weitgehend monotonen und ähnlichen Zusammenhang der Itemschwierigkeiten der drei Klassenprofile.

[49] von links nach rechts in ansteigender Schwierigkeit

[50] Advanced Progressive Matrices nach J.C. Raven

Unter zusätzlicher Beachtung der metrischen Informationen in den Itemparametern mittels Produkt-Moment-Korrelationen werden zwar sämtliche Klassenzusammenhänge der Itemschwierigkeiten kleiner, sie bleiben aber ebenso im signifikanten Bereich (vgl. *Tabelle A-33)*.

Die größte Produkt-Moment-Korrelation von 0,76 ist gleichzeitig hochsignifikant. Sie ergibt sich zwischen den beiden Klassen der „Könner" und „Durchschnittlichen".

Fazit: Deutliche Hinweise auf klassenspezifische Bearbeitungsstrategien, die den ganzen Test und nicht nur die Lösung einzelner Items beeinflussen, können der *Abbildung 5-4* nicht entnommen werden.

5.1.3 Diagnostische Relevanz der latenten Klassen

Der klinische Nutzen des Raventests soll genau analysiert werden. Hierzu werden die Scores innerhalb der einzelnen Klassen verglichen. Unter *Abschnitt 5.1.3.1* von klinisch auffälligen und unauffälligen Personen, zusätzlich unter *Abschnitt 5.1.3.2* von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung.

Ab *Abschnitt 5.1.3.3* wird dann die Vorhersagevalidität klinischer Kriterien mit klassischen und probabilistischen Prädiktoren des Raventests beurteilt.

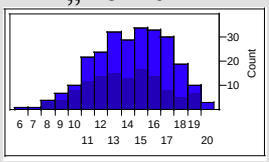
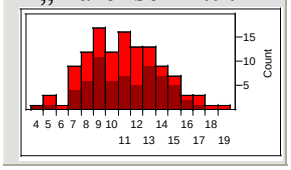
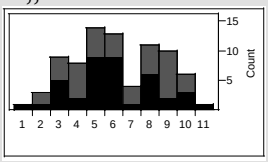
5.1.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen

Zur Beurteilung der Scoreunterschiede klinisch auffälliger und unauffälliger Personen in den einzelnen Klassen des Raventests wird die parallelisierte Stichprobe (N=460) verwendet.

Der Vergleich der Ravenscores der klinisch auffälligen und unauffälligen Personen innerhalb der Klassen ist möglich, weil in *Tabelle 5-7* keine signifikanten Unterschiede der Alters- und Bildungsgrade aufgezeigt werden.

Die Ravenscores von klinisch auffälligen Personen sind nur in der größten und zugleich leistungsstärksten Klasse 1 „Könner" sehr signifikant kleiner als die von klinisch unauffälligen Probanden. Dieser Klasse werden 57% aller Personen der parallelisierten Stichprobe zugeordnet.

Tabelle 5-7 Differenzierte Betrachtung der einzelnen Klassen beim Raventest (klinische Auffälligkeit)

Differenzierte Betrachtung der klinisch auffälligen und unauffälligen Personen in den einzelnen Klassen des Raventests						
Parallelisierte Stichprobe	hoch ← Quantitative Testleistung → niedrig					
Raventest (N=460) Scoreverteilung [ungleicher Maßstab] Hell: klinisch unauffällig (klU) Dunkel: klinisch auffällig (klA)	Klasse1 (57%) „Könner“ 		Klasse2 (26%) „Durchschnittl.“ 		Klasse3 (17%) „Nicht-Könner“ 	
	klU	klA	klU	klA	klU	klA
N (%-Anteil in Klasse)	139 (54%)	120 (46%)	56 (46%)	65 (54%)	40 (50%)	40 (50%)
t-Tests der Scores innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
Ravenscore [Ø]	14,71	13,79	10,80	11,07	6,3	5,98
Ravenscore [σ]	2,73	2,77	3,14	2,86	2,46	2,37
t-Test	2,689		-0,501		0,602	
FG	257		119		78	
Prob > |t|	0,0076**		0,6173		0,5492	
t-Tests des Lebensalters innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
Alter [Ø]	41,46	42,58	45,143	46,89	50,93	52,4
Alter [σ]	15,01	11,09	14,39	12,08	14,10	10,52
t-Test	-0,670		-0,727		-0,530	
FG	257		119		78	
Prob > |t|	0,5032		0,4687		0,5974	
Wilcoxon-Tests der Altersgruppen innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
χ^2 approx.	1,2556		0,0000		0,1467	
FG	1		1		1	
Prob> χ^2	0,2625		0,9957		0,7018	
Wilcoxon-Tests der Bildungsgrade innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
χ^2 approx.	0,1791		1,3836		1,4637	
FG	1		1		1	
Prob> χ^2	0,6721		0,2395		0,2263	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

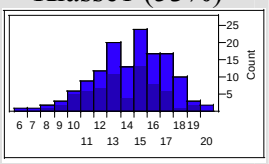
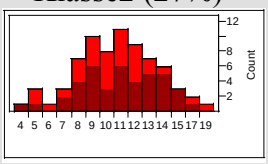
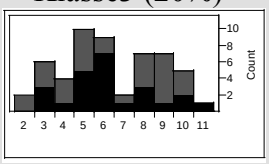
5.1.3.2 Scorevergleiche von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung

Zur Beurteilung der Scoreunterschiede von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung in den einzelnen Klassen des Raventests wird die Stichprobe `parallel-cerebral` (N=265) verwendet.

Der Vergleich der Ravenscores dieser Personengruppen innerhalb der Klassen ist möglich, weil in *Tabelle 5-8* keine signifikanten Unterschiede der Alters- und Bildungsgrade aufgezeigt werden.

Die Resultate ergeben den analogen Befund zu den klinisch auffälligen und unauffälligen Personen. Die Ravenscores von Patienten mit cerebraler Beteiligung sind nur in der größten und zugleich leistungsstärksten Klasse 1 „Könner“ signifikant kleiner als die der klinisch unauffälligen Probanden. Diese Klasse beinhaltet 53% aller Personen der Stichprobe `parallel-cerebral`.

Tabelle 5-8 Differenzierte Betrachtung der einzelnen Klassen beim Raventest (cerebrale Beteiligung)

Differenzierte Betrachtung der Patienten mit cerebraler Beteiligung und klinisch unauffälliger Personen in den einzelnen Klassen des Raventests						
Stichprobe `parallel-cerebral`	hoch ← Quantitative Testleistung → niedrig					
Raventest (N=265) Scoreverteilung [ungleicher Maßstab] Hell: klinisch unauffällig (klU) Dunkel: klinisch auffällig mit cerebraler Beteiligung (mcB)	Klasse1 (53%)  „Könner“		Klasse2 (27%)  „Durchschnittl.“		Klasse3 (20%)  „Nicht-Könner“	
	klU	mcB	klU	mcB	klU	mcB
N (%-Anteil in Klasse)	74 (53%)	66 (47%)	31 (44%)	41 (56%)	29 (55%)	24 (45%)
t-Tests der Scores innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
Ravenscore [Ø]	14,77	13,73	10,45	10,98	6,38	6,29
Ravenscore [σ]	2,84	2,54	3,02	2,89	2,61	2,22
t-Test	2,279		-0,748		0,130	
FG	138		70		51	
Prob > |t|	0,0242*		0,4571		0,8970	
t-Tests des Lebensalters innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
Alter [Ø]	49,03	46,64	49,84	50,73	54,41	55,04
Alter [σ]	10,92	9,88	13,10	12,06	11,95	10,98
t-Test	1,352		-0,300		-0,197	
FG	138		70		51	
Prob > |t|	0,179		0,765		0,8443	
Wilcoxon-Tests der Altersgruppen innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
χ^2 approx.	1,979		0,179		0,185	
FG	1		1		1	
Prob> χ^2	0,159		0,673		0,668	
Wilcoxon-Tests der Bildungsgrade innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
χ^2 approx.	0,011		0,383		0,559	
FG	1		1		1	
Prob> χ^2	0,918		0,536		0,455	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

5.1.3.3 Vorhersagevalidität

Die Vorhersagevalidität der klinischen Kriterien „klinische Auffälligkeit“ und „cerebrale Beteiligung“ wird mit Kennwerten aus nominal logistischen Regressionen beurteilt. Hierzu werden die einzelnen und die in Kombination verwendeten klassischen und probabilistischen Prädiktoren des Raventests verwendet.

5.1.3.3.1 Prädiktion „klinische Auffälligkeit“

Der Raventest wird hinsichtlich der Prädiktion des Kriteriums „klinische Auffälligkeit“ untersucht.

Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die Statistiken sind zusammen mit den zugehörigen Entropiereduktionen der *Tabelle 5-9* zu entnehmen.

Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests, sind in *Tabelle 5-10* wiedergegeben. Diese Kombination wird hier als „Raven kombiniert“ bezeichnet.

In *Tabelle 5-11* sind die ROC-Kurven der Prädiktion „klinische Auffälligkeit“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-9* bis *5-11*:

Die Modelltests werden nicht signifikant.

Die Prädiktoren des Raventests erweisen sich für die Prädiktion von „klinische Auffälligkeit“ als unbrauchbar.

Tabelle 5-9 Entropiereduktionen mit dem Raventest (Kriterium: klinische Auffälligkeit)

Kriterium: Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test (N=460)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
Raven Score	1	3,655	0,056	0,57	(0,26)
probabilistischer Prädiktor					
Raven Kl.zugehörigkeit (MRM 3Kl)	2	1,848	0,397	0,29	(0,00)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Raven kombiniert	5	9,627	0,087	1,51	(0,00)

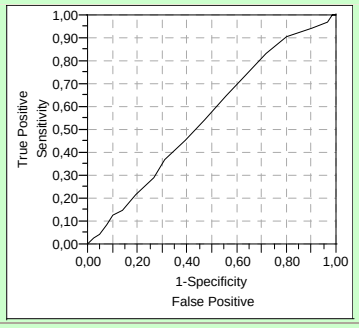
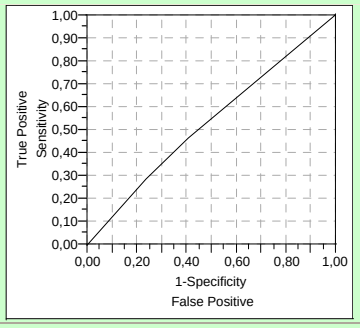
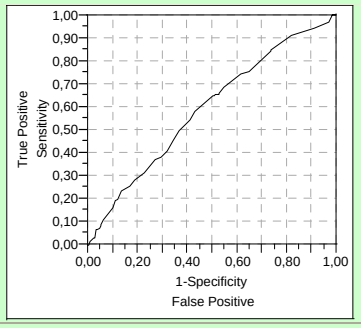
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-10 Wald Effekt Tests beim Raventest (Kriterium: klinische Auffälligkeit)

Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Raven kombiniert	Ravenscore	1	1,482	0,2235
	RavenMRM3	2	0,681	0,7115
	Interaktion	2	3,949	0,1388

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-11 ROC-Kurven beim Raventest (Kriterium: klinische Auffälligkeit)

ROC Receiver Operating Curves (klinische Auffälligkeit = positiv)		
Ravenscore	Raven-Klassenzugehörigkeit	Raven kombiniert
		
Fläche unter der Kurve: 0,556	Fläche unter der Kurve: 0,533	Fläche unter der Kurve: 0,588

5.1.3.3.2 *Prädiktion „cerebrale Beteiligung“*

Ergänzend zur Prädiktion „klinische Auffälligkeit“ wird der Raventest in diesem *Abschnitt* hinsichtlich der Prädiktion des Kriteriums „cerebrale Beteiligung“ untersucht.

Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die Statistiken sind zusammen mit den zugehörigen Entropiereduktionen der *Tabelle 5-12* zu entnehmen.

Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests, sind in *Tabelle 5-13* wiedergegeben.

In *Tabelle 5-14* sind die ROC-Kurven der Prädiktion „cerebrale Beteiligung“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-12* bis *5-14*:

Die Ergebnisse entsprechen grundsätzlich denen bei der Prädiktion des Kriteriums „klinische Auffälligkeit“.

Die Modelltests werden nicht signifikant.

Die Prädiktoren des Raventests erweisen sich für die Prädiktion von „cerebrale Beteiligung“ als unbrauchbar.

Tabelle 5-12 Entropiereduktionen mit dem Raventest (Kriterium: cerebrale Beteiligung)

Kriterium: cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test (N=265)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
Raven Score	1	0,799	0,3713	0,22	(0,00)
probabilistischer Prädiktor					
Raven Kl.zugehörigkeit (MRM 3Kl)	2	2,289	0,3183	0,62	(0,00)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Raven kombiniert	5	8,036	0,1543	2,19	(0,00)

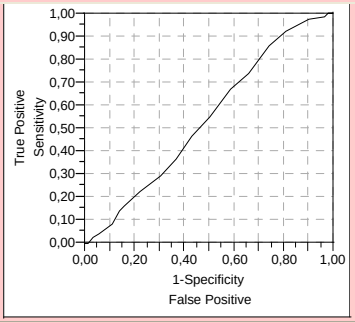
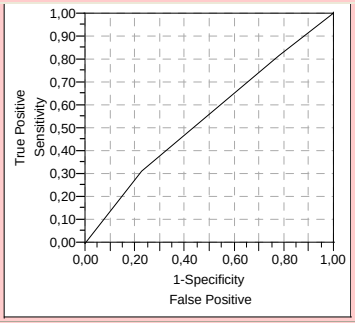
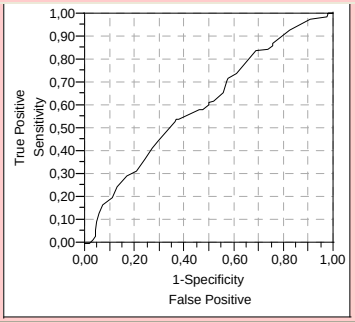
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-13 Wald Effekt Tests beim Raventest (Kriterium: cerebrale Beteiligung)

Cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Raven kombiniert	Ravenscore	1	0,386	0,534
	RavenMRM3	2	0,718	0,698
	Interaktion	2	3,996	0,136

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-14 ROC-Kurven beim Raventest (Kriterium: cerebrale Beteiligung)

ROC Receiver Operating Curves (cerebrale Beteiligung = positiv)		
Ravenscore	Raven-Klassenzugehörigkeit	Raven kombiniert
		
Fläche unter der Kurve: 0,537	Fläche unter der Kurve: 0,545	Fläche unter der Kurve: 0,603

5.2 Bentontest

5.2.1 Prüfung der Homogenität

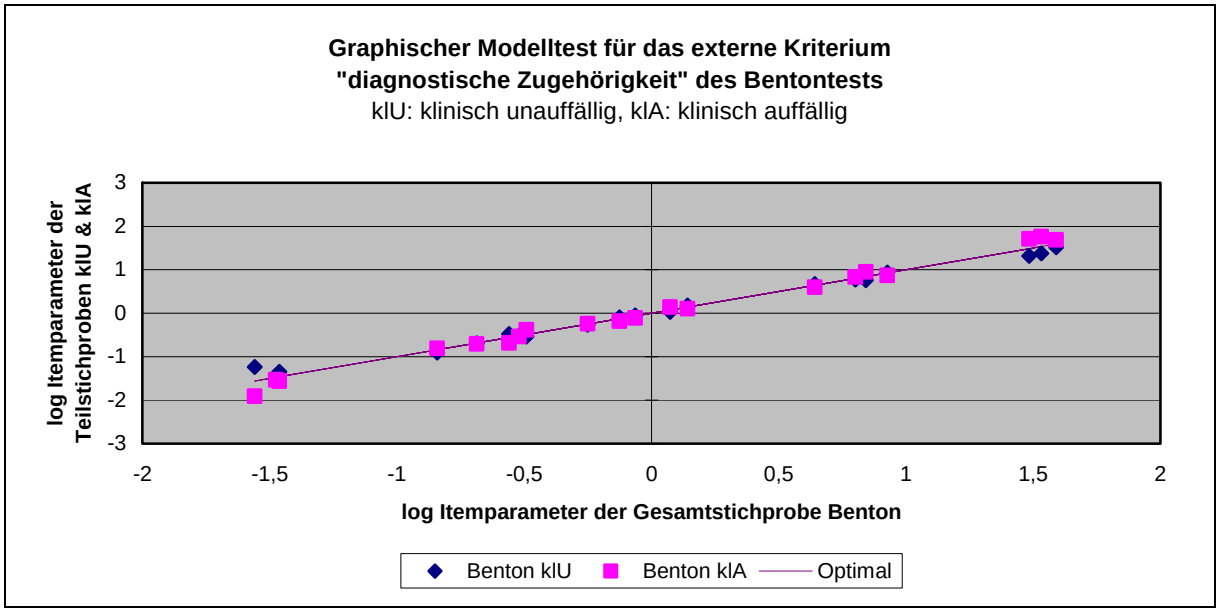
In den folgenden *Abschnitten* werden Modellgeltungtests durchgeführt. Die Personenhomogenität als die zentrale Forderung der Raschhomogenität wird bei der Itemlösung des Bentontests geprüft. Der graphische Modelltest und der „Andersen"-Test werden mit einem externen und internen Teilungskriterium angewendet.
Die Zweiteilung der Gesamtstichprobe in „klinisch unauffällige" und „klinisch auffällige" Personen dient als externes Teilungskriterium.
Die Aufteilung in Hoch- und Niedrigscorer beim internen Teilungskriterium „Testleistung" erfolgt durch den Mediansplit der Gesamtstichprobe. Hierbei werden Personen mit einem Rohscorewert von ≥ 13 den „Hochscorern" und ≤ 12 den „Niedrigscorern" zugeordnet.

5.2.1.1 Überprüfung des Raschmodells mit graphischen Modelltestungen

Die Itemparameter σ_i der jeweiligen Teilstichproben des Bentontests (*Tabelle A-34* im Anhang) werden für die graphischen Modellkontrollen verwendet. Gleiche Schwierigkeitsunterschiede bedeuten in der logarithmierten, summennormierten Form gleich große Abstände. Die Parameterschätzungen aus der Gesamtstichprobe „Benton Gesamt" sind in den Darstellungen auf der Abszisse, die Schätzungen aus den Teilstichproben auf der Ordinate aufgetragen.
Für das externe Kriterium „diagnostische Zugehörigkeit" wird die graphische Modellkontrolle in *Abbildung 5-5* gezeigt. *Abbildung 5-6* zeigt den graphischen Modelltest für das interne Kriterium „Testleistung".
Wenn sich sämtliche Punkte exakt auf der durchgezogenen Geraden y=x befinden, dann bedeutet dies in beiden *Abbildungen* den graphischen „Beweis" optimaler Modellpassung.
Modellverletzungen sind vor allem in den Extrembereichen, insbesondere beim Teilungskriterium „Testleistung" erkennbar.
Die zusätzlich berechneten Korrelationen der Itemparameter der Teilstichproben sind unter den *Abbildungen* aufgeführt. Sie bieten eine Vergleichsmöglichkeit des externen und internen Teilungskriteriums. Dabei ergeben sich kleinere Produkt-Moment-Korrelationen des internen Teilungskriteriums „Testleistung".

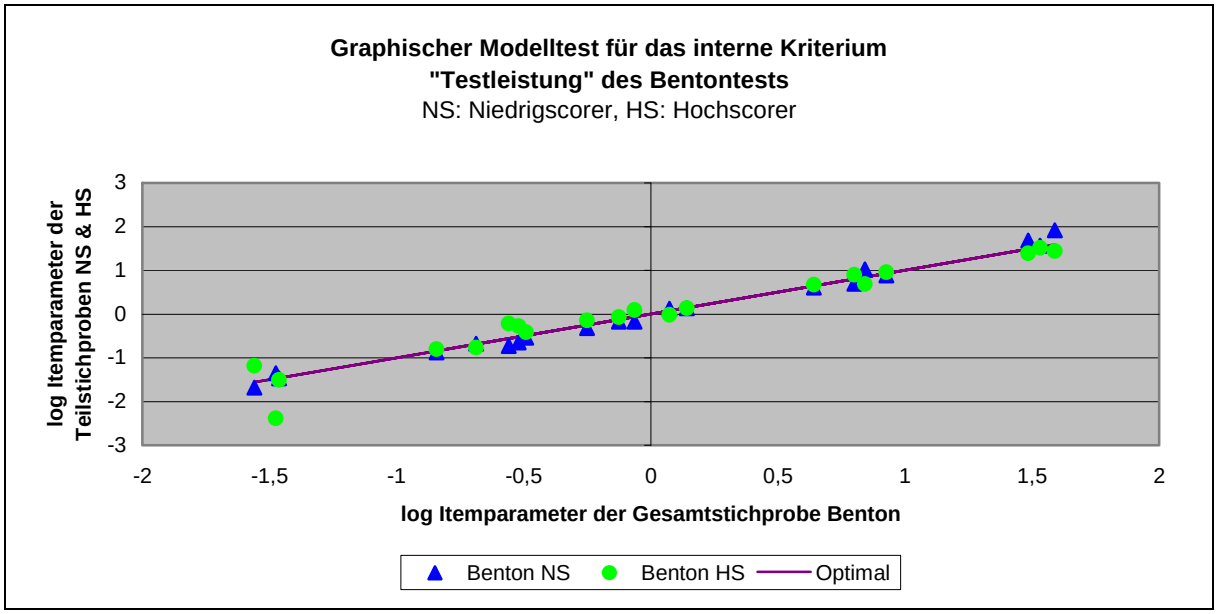
Es ist zu vermuten, dass dieses Resultat, analog zum Raventest, auf größere Unterschiede der Itemschwierigkeiten in den Hoch- und Niedrigscorer-Klassen hinweist.

Abbildung 5-5 Graphischer Modelltest für das externe Kriterium beim Bentontest



Die Produkt-Moment-Korrelation zwischen Benton klU und klA beträgt: 0,987

Abbildung 5-6 Graphischer Modelltest für das interne Kriterium beim Bentontest



Die Produkt-Moment-Korrelation zwischen Benton NS und HS beträgt: 0,943

5.2.1.2 Überprüfung des Raschmodells mit der „Andersen-Testmethode“

Die statistische Signifikanzprüfung des Raschmodells beim Bentontest mittels „Andersen-Methode“ wird analog zum graphischen Modelltest, mit einem externen und internen Kriterium durchgeführt.

5.2.1.2.1 Modelltestung anhand des externen Teilungskriteriums „diagnostische Zugehörigkeit“

Tabelle 5-15 fasst die marginalen Likelihoods beim Bentontest und die daraus resultierende statistische Prüfgröße für das externe Teilungskriterium „diagnostische Zugehörigkeit“ zusammen.

Tabelle 5-15 „Andersen-Methode“ (Teilungskriterium „diagnostische Zugehörigkeit“) beim Bentontest

Stichprobe	N	ln(mL)	n_p
Benton klU	327	-3634,21	21
Benton klA	273	-3034,81	21
Benton Gesamt	610	-6801,24	21
-2ln(LR)=264,44		FG=21	

Die ermittelte Prüfgröße beträgt 264,44. Sie liegt damit oberhalb der kritischen Grenze von 32,7 der χ^2-Verteilung (α=5%) bei 21 Freiheitsgraden.

Der Bentontest verhält sich für das externe Teilungskriterium „diagnostische Zugehörigkeit“ nicht modellkonform. Die Null-Hypothese „Raschhomogenität“ muss verworfen werden.

5.2.1.2.2 Modelltestung anhand des internen Teilungskriteriums „Testleistung“

Tabelle 5-16 fasst die marginalen Likelihoods beim Bentontest und die daraus resultierende statistische Prüfgröße für das interne Teilungskriterium „Testleistung“ zusammen.

Tabelle 5-16 „Andersen-Methode“ (Teilungskriterium „Testleistung“) beim Bentontest

Stichprobe	N	ln(mL)	n_p
BentonNS	326	-3748,87	21
BentonHS	284	-2692,76	21
BentonGesamt	610	-6801,24	21
-2ln(LR)= 719,22		FG=21	

Die ermittelte Prüfgröße beträgt 719,22. Sie liegt damit oberhalb der kritischen Grenze von 32,7 der χ^2-Verteilung (α =5%) bei 21 Freiheitsgraden.

Der Bentontest verhält sich für das interne Teilungskriterium „Testleistung“ nicht modellkonform. Die Null-Hypothese „Raschhomogenität“ muss verworfen werden.

5.2.1.3 Geltungsprüfung unterschiedlicher probabilistischer Modelle mittels informationstheoretischer Maße und Likelihoodquotiententests

Die berechneten Kennwerte der informationstheoretischen Maße sind für die Gesamtstichprobe in *Tabelle 5-17* zusammengefasst.

Tabelle 5-17 Informationstheoretische Maße beim Bentontest

N=610	loglikeli	AIC	BIC	n_p
BENTON RM	-6801,24	13644,47	13737,16	21
BENTON MRM2Kl	-6772,92	13631,84	13821,62	43
BENTON MRM3Kl	-6740,86	13611,75	13898,62	65
BENTON MRM4Kl	-6724,77	13623,55	14007,52	87
BENTON LCA1Kl	-7543,06	15126,12	15214,39	20
BENTON LCA2Kl	-6883,45	13848,90	14029,85	41
BENTON LCA3Kl	-6774,33	13672,67	13946,30	62
BENTON LCA4Kl	-6744,26	13654,51	14020,83	83
BENTON LCA5Kl	-6719,87	13647,74	14106,74	104
BENTON LCA6Kl	-6696,02	13642,04	14193,72	125
BENTON LCA7Kl	-6684,22	13660,44	14304,80	146
saturiertes Modell	-3877,89	2104905,78	6732748,51	1048575
Startwert 3690 abgesichert mit 4321 und 3321				

Die informationstheoretischen Maße AIC (13611,75 < 13642,04) und BIC (13737,16 < 13946,30) geben in der *Tabelle 5-17* jeweils Rasch- und Mixed-Rasch-Modellen den Vorzug. Beim Bentontest wird vom AIC-Wert (13611,75) die 3-Klassen-Lösung des Mixed-Rasch-Modells bevorzugt. Die Einklassen-Lösung, also das Rasch-Modell, wird vom BIC-Wert (13737,16) favorisiert.

Den weiteren Betrachtungen wird die 3-Klassen-Lösung zugrunde gelegt, weil die Einklassen-Lösung beim exakteren Andersen-Test in *Abschnitt 5.2.1.2* verworfen wurde.

Die 3-Klassen-Lösung wird inferenzstatistisch mittels Likelihoodquotiententests (vgl. *Tabelle 5-18*) abgesichert. Bei diesen Prüfungen ist zu beachten, dass der Likelihoodquotientenvergleich zwischen Klassen-Modell und saturiertem Modell angesichts der großen Parameteranzahl des saturierten Modells (n_p=1048575) und der vergleichsweise geringen Stichprobengröße (N=610) nur begrenzt aussagekräftig ist.

Tabelle 5-18 Likelihoodquotiententests Benton

LR	-2ln(LR)	FG	Prob> χ^2
MRM3Kl/sat.Modell	5725,94	1048510	1
RM/MRM3Kl	120,76	44	<,0001***

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die Ergebnisse der beiden gebildeten Likelihoodquotienten bedeuten, dass die vom AIC indizierte 3-Klassen-Lösung des Mixed-Rasch-Modells mit den vorliegenden Daten vereinbar ist. Diese passt besser als das Raschmodell und nicht schlechter als das saturierte Modell. Die 3-Klassen-Lösung ist für die Daten somit ein geeignetes Modell und die Grundlage der weiteren Berechnungen.

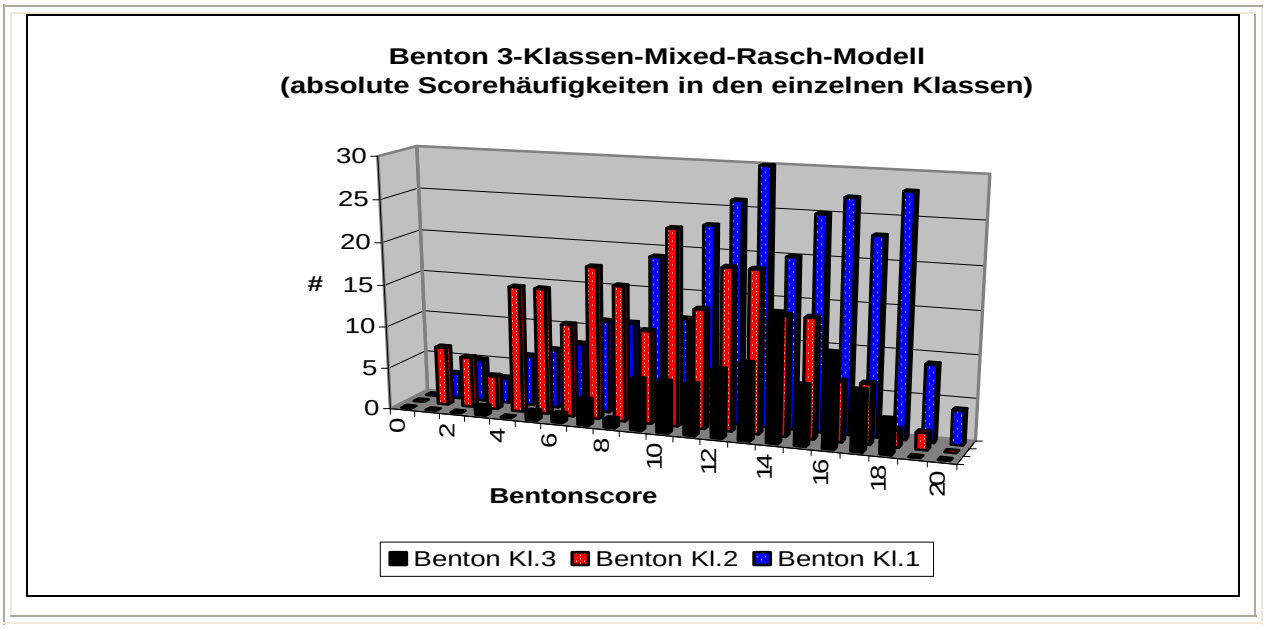
5.2.2 Interpretation latenter Klassen

Die vom AIC indizierte 3-Klassen-Lösung des Mixed-Rasch-Modells wird in diesem *Abschnitt* differenzierter betrachtet. Die einzelnen Klassen werden aufgrund der erhobenen Merkmale der zugeordneten Probanden voneinander abgegrenzt.

5.2.2.1 Summenscoreverteilung der 3-Klassen-Lösung des Mixed-Rasch-Modells

Die *Abbildung 5-7* zeigt klassenspezifische Zusammenhänge mit der Bentontestleistung. Es gibt zwei Klassen mit höheren und eine Klasse mit eher niedrigen Leistungen.

Abbildung 5-7 Bentonscores in den Klassen



5.2.2.2 Charakterisierung der 3-Klassen-Lösung des Mixed-Rasch-Modells

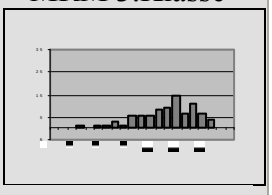
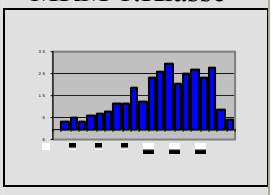
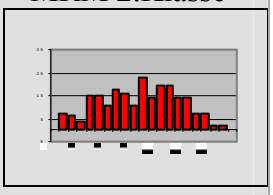
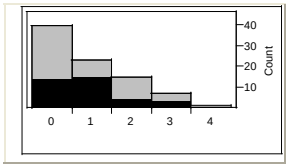
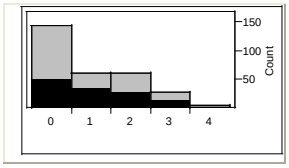
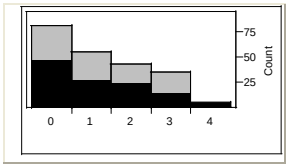
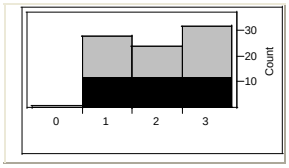
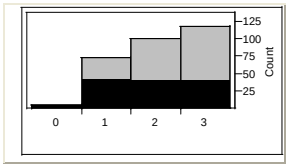
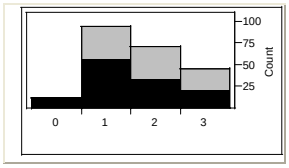
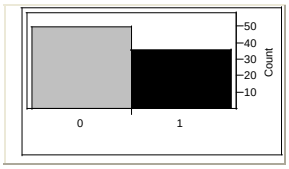
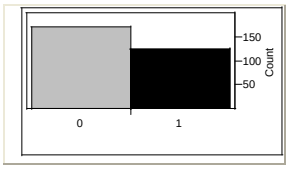
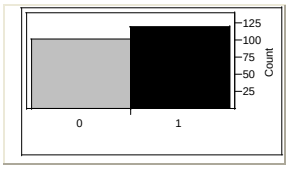
Die drei Klassen werden bezüglich der mittleren Benton-Scoreleistung geordnet und mit den Bezeichnungen „Könner A“ für Klasse 3, „Könner B“ für Klasse 1 und „Nicht-Könner“ für Klasse 2 versehen.

Die einzelnen Klassen werden in *Tabelle 5-19* charakterisiert.

Eine Zusammenfassung der Klassenkontraste gibt *Tabelle 5-20*.

Die ausführlichen Statistiken sind in der *Tabelle A-35* des Anhangs angegeben.

Tabelle 5-19 Die 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest

3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest			
	hoch ← Quantitative Testleistung → niedrig		
Bentontest (N=610) Scoreverteilung Maßstab x=Score: 0-20 y=N: 0-35	MRM 3.Klasse  **„Könner A“**	MRM 1.Klasse  **„Könner B“**	MRM 2.Klasse  **„Nicht-Könner“**
N [%-Anteil gesamt]	86 (14%)	300 (49%)	224 (37%)
mittlere klassenspezifische Zuordnungssicherheit	0,810	0,796	0,824
Cronbach-α #	0,69	0,83	0,83
Reliabilität ##	0,67	0,79	0,82
Bentonscore [Ø]	13,0	12,57	9,62
Bentonscore [σ]	3,22	4,39	4,32
Bentonfehler (BTF) [Ø]	9,40	10,72	16,37
Bentonfehler (BTF) [σ]	5,36	8,35	9,53
Alter [Ø]	40,65	40,30	45,20
Alter [σ]	12,94	15,22	14,30
Altersverteilung Hell: klinisch unauffällig Dunkel: klinisch auffällig 0=15-39LJ, 1=39-49LJ, 2=49-59LJ, 3=59-70LJ, 4= >70LJ			
Bildungsverteilung Hell: klinisch unauffällig Dunkel: klinisch auffällig 0=ohne Schulabschluss, 1=Hauptschule, 2=Realschule, 3=Abitur			
Klinische Auffälligkeit Hell=0: klinisch unauffällig Dunkel=1: klinisch auffällig			

#=Klassische Testtheorie; ## =Probabilistische Testtheorie (Reliabilität *vgl. Abschnitt 4.7.1*)

Aus den *Tabellen 5-19* und *5-20* geht hervor:

- Die Klasse 3 „Könner A“ und die Klasse 1 „Könner B“ sind im Hinblick auf die untersuchten Variablen nahezu identisch. Es sind keine statistisch signifikanten Klassenunterschiede für Alter, Bentonscore, Bentonfehler, Bildungsgrad und den Anteil klinisch auffälliger Personen nachweisbar.
  Das Mixed-Rasch-Modell ermöglicht dennoch die klassenspezifische Unterscheidung der Itemparameter, weil testleistungsunabhängige Zusatzinformation in den Daten vorhanden ist.

- Die Klasse 2 „Nicht-Könner“ unterscheidet sich mit signifikant schlechterer Testleistung, mehr Fehlern, geringerer Bildung und höherem Alter von den beiden Könner-Klassen.
  Im Vergleich mit der größten Klasse 1 „Könner B“ wird der Anteil klinisch auffälliger Personen signifikant.

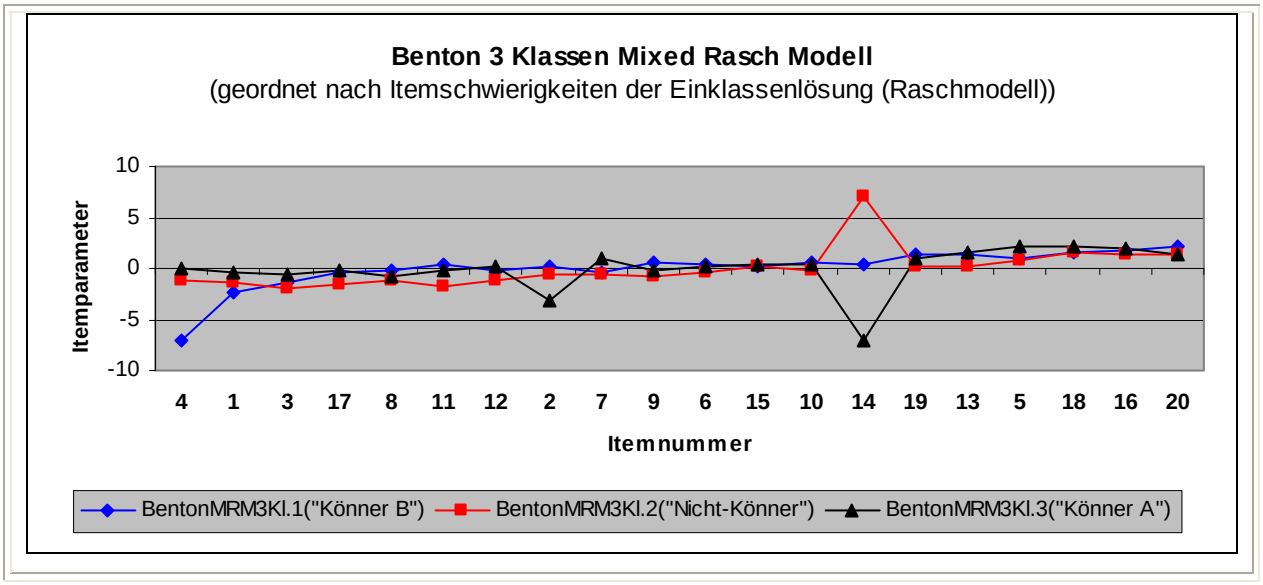
Tabelle 5-20 Zusammenfassung der Kontraste der 3-Kl.-Lösung des Mixed-Rasch-Modells beim Bentontest

Kontraste der Scores beim Bentontest	KönnerA	KönnerB	Nichtkönner
KönnerA		ns	***
KönnerB	ns		***
Nichtkönner	***	***	
Kontraste der Fehler (BTF) beim Bentontest	KönnerA	KönnerB	Nichtkönner
KönnerA		ns	***
KönnerB	ns		***
Nichtkönner	***	***	
Kontraste des Lebensalters beim Bentontest	KönnerA	KönnerB	Nichtkönner
KönnerA		ns	*
KönnerB	ns		***
Nichtkönner	*	***	
Kontraste der Altersgruppen beim Bentontest	KönnerA	KönnerB	Nichtkönner
KönnerA		ns	ns
KönnerB	ns		*
Nichtkönner	ns	*	
Kontraste der Bildungsgrade beim Bentontest	KönnerA	KönnerB	Nichtkönner
KönnerA		ns	*
KönnerB	ns		***
Nichtkönner	*	***	
Kontraste der Kontingenztafeln (klin. Auff. ja/nein) beim Bentontest	KönnerA	KönnerB	Nichtkönner
KönnerA		ns	ns
KönnerB	ns		*
Nichtkönner	ns	*	
*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur) entnommen aus *Tab.A-35*			

5.2.2.3 Die Profile der Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells

In der *Abbildung 5-8* sind die Itemparameter der 3-Klassen-Lösung des Bentontests nach den Itemschwierigkeiten des einfachen Raschmodells[51] geordnet.
Die zugehörigen Itemparameter sind in der *Tabelle A-36* des Anhangs angegeben.

Abbildung 5-8 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest



Bewertung der *Abbildung 5-8* und Fazit:

- Obwohl bei der Datenerhebung die Reihenfolge 1-20 der Itemschwierigkeiten nach den Ergebnissen von KINDERMANN (2001, S.72) neu angeordnet worden war, lässt sie sich in der vorliegenden Untersuchung nicht replizieren (vgl. Itemnummern der Abszisse).
- Insbesondere die Items 4 und 14 des Bentontests beeinflussen wegen erkennbarer Inhomogenitäten stark die Klassenbildungen.
  Nur die Klasse 1 „Könner B“ zeichnet sich durch ihren Lösungserfolg bei Item 4 aus.
  Das Item 14 ist für die Klasse 3 „Könner A“ am leichtesten und für die Klasse 2 „Nicht-Könner“ am schwersten.
- Die signifikanten und die positiven Rangkorrelationen der klassenspezifischen Itemparameter (vgl. im Anhang *Tabelle A-37*) zeigen an, dass weitgehend monotone Zusammenhänge existieren.

[51] von links nach rechts in aufsteigender Schwierigkeit

Die metrischen Informationen in den Itemparametern ergeben mit Produkt-Moment-Korrelationen keine signifikant positiven, vielmehr in einem Fall signifikant negative Zusammenhänge (vgl. *Tabelle A-37*).

Die differenziertere Analyse letzterer Besonderheit zeigt, dass sich diese negative Produkt-Moment-Korrelation der Klassen „Könner A“ und „Nicht-Könner“ auf das Item 14 mit extrem unterschiedlichen klassenspezifischen Lösungswahrscheinlichkeiten zurückführen lässt (vgl. auch *Abbildung 5-8)*.

Wie zu erwarten wird unter Ausschluss dieses Ausreißer-Items auch diese Produkt-Moment-Korrelation mit 0,70 positiv. Dann ergibt sich ein hochsignifikanter Zusammenhang, der die große Ähnlichkeit der restlichen Itemparameter der leistungsschwächsten Klasse „Nicht-Könner“ und der leistungsstärksten Klasse „Könner A“ bestätigt (ebenso *Tabelle A-37*).

Fazit: Hinweise auf klassenspezifische Testbearbeitungsstrategien, wie sie KINDERMANN (2001) mit anderem Datenpool aufzeigen konnte, können der *Abbildung 5-8* nicht entnommen werden (vgl. *Abbildung 1-1* zum direkten Vergleich).

5.2.3 Diagnostische Relevanz der latenten Klassen

Der klinische Nutzen des Bentontests soll genau analysiert werden. Hierzu werden die Scores innerhalb der einzelnen Klassen verglichen. Unter *Abschnitt 5.2.3.1* von klinisch auffälligen und unauffälligen Personen, zusätzlich unter *Abschnitt 5.2.3.2* von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung.

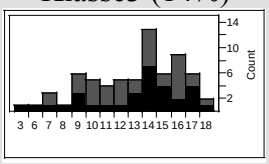
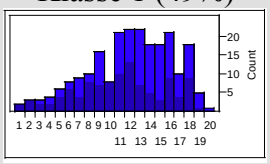
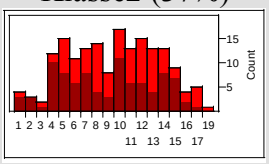
Für die Beurteilung der Vorhersagevalidität klinischer Kriterien mit klassischen und probabilistischen Prädiktoren wird ab *Abschnitt 5.2.3.3* die logistische Regression verwendet.

5.2.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen

Zur Beurteilung der Scoreunterschiede klinisch auffälliger und unauffälliger Personen in den einzelnen Klassen wird die parallelisierte Stichprobe (N=464) verwendet (vgl. *Tabelle 5-21*).

Innerhalb der Klassen ist der Scorevergleich der klinisch auffälligen und unauffälligen Gruppe möglich, weil keine signifikanten Unterschiede der Alters- und Bildungsgrade nachweisbar sind.

Tabelle 5-21 Differenzierte Betrachtung der einzelnen Klassen beim Bentontest (klinische Auffälligkeit)

Differenzierte Betrachtung der klinisch auffälligen und unauffälligen Stichprobe in den einzelnen Klassen des Bentontests						
Parallelisierte Stichprobe	hoch ← Quantitative Testleistung → niedrig					
Bentontest (N=464) Scoreverteilung [ungleicher Maßstab] Hell: klinisch unauffällig (klU) Dunkel: klinisch auffällig (klA)	Klasse3 (14%)  „Könner A“		Klasse 1 (49%)  „Könner B“		Klasse2 (37%)  „Nicht-Könner“	
	klU	klA	klU	klA	klU	klA
N (%-Anteil in Klasse)	36 (54%)	31 (46%)	118 (52%)	107 (48%)	81 (47%)	91 (53%)
t-Tests der Scores innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
Bentonscore [Ø]	12,94	12,87	13,12	11,01	10,10	8,89
Bentonscore [σ]	2,96	3,66	3,71	4,50	3,79	4,23
t-Test	0,091		3,852		1,964	
FG	65		223		170	
Prob > |t|	0,9278		0,0002***		0,0511	
t-Tests des Lebensalters innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
Alter [Ø]	42,47	42,42	41,90	44,64	47,59	48,40
Alter [σ]	13,90	9,54	16,06	12,82	13,50	12,11
t-Test	0,018		-1,404		-0,411	
FG	65		223		170	
Prob > |t|	0,9858		0,1618		0,6814	
Wilcoxon-Tests der Altersgruppen innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
χ^2 approx.	0,4333		0,1693		0,4032	
FG	1		1		1	
Prob> χ^2	0,5104		0,6808		0,5254	
Wilcoxon-Tests der Bildungsgrade innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)						
χ^2 approx.	0,9819		0,0000		0,0028	
FG	1		1		1	
Prob> χ^2	0,3217		0,9991		0,9575	

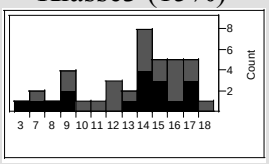
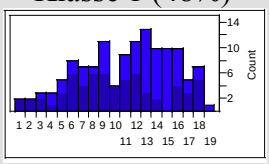
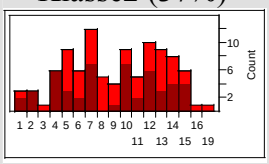
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die Bentonscores innerhalb der einzelnen Klassen unterscheiden sich für klinisch auffällige und unauffällige Personen hochsignifikant nur in der größten Klasse 1 „Könner B“.
Diese umfasst 49 Prozent aller Personen der parallelisierten Stichprobe.

5.2.3.2 Scorevergleiche von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung

Zur Beurteilung der Bentonscoreunterschiede von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung in den einzelnen Klassen wird die Stichprobe `parallel-cerebral` (N=265) verwendet (vgl. *Tabelle 5-22*).
Innerhalb der Klassen ist der Vergleich der Bentonscores dieser Personengruppen möglich, weil keine signifikanten Unterschiede der Alters- und Bildungsgrade feststellbar sind.
Es ergibt sich dasselbe Resultat wie bei klinisch auffälligen und unauffälligen Personen. Die Scores von Patienten mit cerebraler Beteiligung sind beim Bentontest nur in der größten Klasse 1 „Könner B“ signifikant kleiner als die von klinisch unauffälligen Probanden.
Diese Klasse 1 umfasst 48 Prozent aller Personen der Stichprobe `parallel-cerebral`.

Tabelle 5-22 Differenzierte Betrachtung der einzelnen Klassen beim Bentontest (cerebrale Beteiligung)

Differenzierte Betrachtung der Patienten mit cerebraler Beteiligung und klinisch unauffälliger Personen in den einzelnen Klassen des Bentontests						
Stichprobe `parallel-cerebral`	hoch ← Quantitative Testleistung → niedrig					
Bentontest (N=265) Scoreverteilung [ungleicher Maßstab] Hell: klinisch unauffällig (klU) Dunkel: klinisch auffällig mit cerebraler Beteiligung (mcB)	Klasse3 (15%)  „Könner A“		Klasse 1 (48%)  „Könner B“		Klasse2 (37%)  „Nicht-Könner“	
	klU	mcB	klU	mcB	klU	mcB
N (%-Anteil in Klasse)	22 (56%)	17 (44%)	64 (50%)	64 (50%)	48 (49%)	50 (51%)
t-Tests der Scores innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
Bentonscore [Ø]	13,5	12,76	12,31	9,94	9,58	8,72
Bentonscore [σ]	2,96	4,07	3,80	4,64	3,88	4,23
t Test	0,654		3,164		1,052	
FG	37		126		96	
Prob > |t|	0,5172		0,0019**		0,2955	
t-Tests des Lebensalters innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
Alter [Ø]	48,77	44,94	50,59	49,55	50,83	52,94
Alter [σ]	10,89	7,09	12,24	12,06	11,73	12,08
t Test	1,257		0,487		-0,876	
FG	37		126		96	
Prob > |t|	0,2166		0,6269		0,3834	
Wilcoxon-Tests der Altersgruppen innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
χ^2 approx.	1,0617		0,1044		0,7967	
FG	1		1		1	
Prob> χ^2	0,3028		0,7466		0,3721	
Wilcoxon-Tests der Bildungsgrade innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)						
χ^2 approx.	0,6439		0,3763		0,1218	
FG	1		1		1	
Prob> χ^2	0,4223		0,5396		0,7271	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

5.2.3.3 Vorhersagevalidität

Die Vorhersagevalidität der klinischen Kriterien „klinische Auffälligkeit“ und „cerebrale Beteiligung“ wird mit Kennwerten aus nominal logistischen Regressionen beurteilt. Hierzu werden die einzelnen und die in Kombination verwendeten klassischen und probabilistischen Prädiktoren des Bentontests verwendet.
Der Prädiktor Bentonscore wird in den *Abschnitten 5.2.3.3.1* und *5.2.3.3.2* verwendet.
Die Prädiktionen die mit dem Fehlermaß BTF berechnet werden, sind in den *Abschnitten 5.2.3.3.3* und *5.2.3.3.4* aufgeführt.

5.2.3.3.1 Prädiktion „klinische Auffälligkeit“

Der Bentontest wird hinsichtlich der Prädiktion „klinische Auffälligkeit“ untersucht.
Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die Statistiken sind zusammen mit den zugehörigen Entropiereduktionen der *Tabelle 5-23* zu entnehmen.
Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests, sind in *Tabelle 5-24* wiedergegeben. Diese Kombination wird hier als „Benton kombiniert“ bezeichnet.
In *Tabelle 5-25* sind die ROC-Kurven der Prädiktion „klinische Auffälligkeit“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-23* bis *5-25*:
Mit dem klassischen Prädiktor Bentonscore verbessert sich die Prädiktion von „klinische Auffälligkeit“ im Vergleich zur Zufallsentscheidung. Durch die Testung ergibt sich zwar hochsignifikanter, jedoch wenig relevanter Informationsgewinn.
Die klassisch-probabilistische Prädiktorenkombination trägt, verglichen mit dem separat verwendeten klassischen Prädiktor, kaum zu brauchbareren Prädiktionen bei.

Tabelle 5-23 Entropiereduktionen mit dem Bentontest (Kriterium: klinische Auffälligkeit)

Kriterium: Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test (N=464)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
Benton Score	1	17,232	<,0001***	2,68	(2,37)
probabilistischer Prädiktor					
Benton Kl.zugehörigkeit (MRM 3Kl)	2	1,416	0,493	0,22	(0,00)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Benton kombiniert	5	19,657	0,001**	3,06	(1,50)

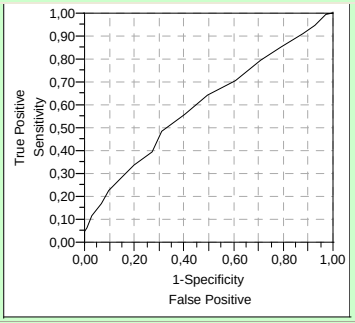
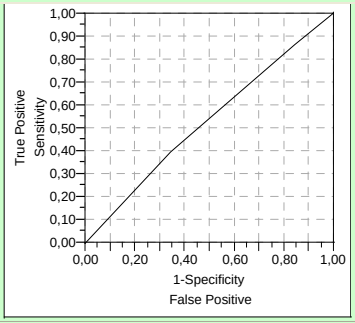
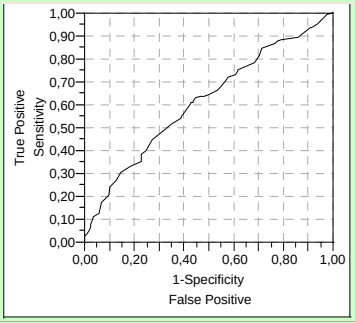
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-24 Wald Effekt Tests beim Bentontest (Kriterium: klinische Auffälligkeit)

Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Benton kombiniert	Bentonscore	1	5,117	0,0237*
	BentonMRM3	2	0,228	0,8922
	Interaktion	2	2,384	0,3037

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-25 ROC-Kurven beim Bentontest (Kriterium: klinische Auffälligkeit)

ROC Receiver Operating Curves (klinische Auffälligkeit = positiv)		
Bentonscore	Benton-Klassenzugehörigkeit	Benton kombiniert
		
Fläche unter der Kurve: 0,603	Fläche unter der Kurve: 0,528	Fläche unter der Kurve: 0,615

5.2.3.3.2 Prädiktion „cerebrale Beteiligung“

Ergänzend zur Prädiktion „klinische Auffälligkeit“ wird der Bentontest in diesem *Abschnitt* hinsichtlich der Prädiktion des Kriteriums „cerebrale Beteiligung“ untersucht.
Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die Statistiken sind zusammen mit den zugehörigen Entropiereduktionen der *Tabelle 5-26* zu entnehmen.
Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests sind in *Tabelle 5-27* wiedergegeben.
In *Tabelle 5-28* sind die ROC-Kurven der Prädiktion „cerebrale Beteiligung“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-26* bis *5-28*:
Das Ergebnis entspricht grundsätzlich den Resultaten bei der Prädiktion „klinische Auffälligkeit“ unter *Abschnitt 5.2.3.3.1*.
Der klassische Prädiktor Bentonscore ist der Zufallsentscheidung bei der Prädiktion von „cerebrale Beteiligung“ überlegen. Durch die Testung ergibt sich zwar sehr signifikanter, jedoch wenig relevanter Informationsgewinn.
Die klassisch-probabilistische Prädiktorenkombination bringt kaum zusätzliche Informationen.

Tabelle 5-26 Entropiereduktionen mit dem Bentontest (Kriterium: cerebrale Beteiligung)

Kriterium: cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test(N=265)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
Benton Score	1	10,149	0,0014**	2,76	(2,22)
probabilistischer Prädiktor					
Benton Kl.zugehörigkeit (MRM 3Kl)	2	0,650	0,7227	0,18	(0,00)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Benton kombiniert	5	11,911	0,0360*	3,24	(0,52)

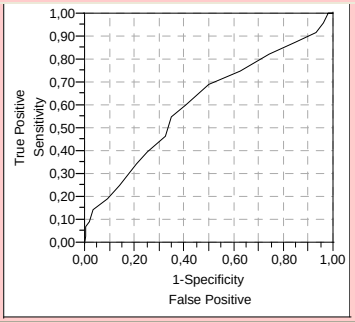
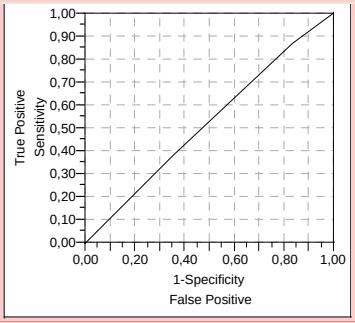
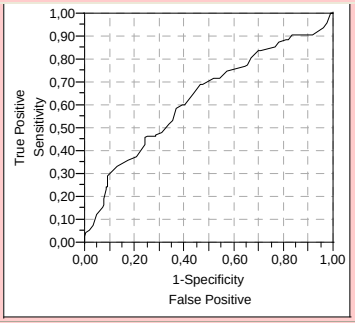
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-27 Wald Effekt Tests beim Bentontest (Kriterium: cerebrale Beteiligung)

Cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Benton kombiniert	Bentonscore	1	4,524	0,0334*
	BentonMRM3	2	0,211	0,8999
	Interaktion	2	1,473	0,4788

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-28 ROC-Kurven beim Bentontest (Kriterium: cerebrale Beteiligung)

ROC Receiver Operating Curves (cerebrale Beteiligung = positiv)		
Bentonscore	Benton-Klassenzugehörigkeit	Benton kombiniert
		
Fläche unter der Kurve: 0,609	Fläche unter der Kurve: 0,521	Fläche unter der Kurve: 0,630

5.2.3.3.3 Prädiktion „klinische Auffälligkeit“ und Fehlermaß BTF

Das Fehlermaß BTF wurde in *Abschnitt 4.4* schon kurz erläutert. Es gibt die Gesamtanzahl der Fehler in sämtlichen Items an. Im Vergleich zum Score des Bentontests, der nur die Anzahl richtiger Itemreproduktionen zählt, gilt es als reliablerer Auswertungsmodus (STECK et al., 1990). Die Vorhersagevalidität wird in den folgenden *Abschnitten* auch mit dem Fehlermaß BTF überprüft.

Das Fehlermaß BTF wird zunächst hinsichtlich der Prädiktion „klinische Auffälligkeit“ untersucht.

Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die Statistiken sind zusammen mit den zugehörigen Entropiereduktionen in *Tabelle 5-29* aufgeführt.

Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests sind in *Tabelle 5-30* wiedergegeben.

In *Tabelle 5-31* sind die ROC-Kurven der Prädiktion „klinische Auffälligkeit“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den Tabellen *5-29* bis *5-31*:

Mit dem klassischen Prädiktor BTF verbessert sich die Prädiktion von „klinische Auffälligkeit“ im Vergleich zur Zufallsentscheidung.

Die Kombination der Prädiktoren bringt keine zusätzlichen Informationen im Vergleich zum separat verwendeten klassischen Prädiktor.

Die Werte der Entropiereduktionen zeigen an, dass das Fehlermaß BTF (3,31% und adjustiert 3,00%) sicherere Prädiktionen von „klinische Auffälligkeit“ ermöglicht, als der Bentonscore (2,68% und adjustiert 2,37%) (vgl. *Abschnitt 5.2.3.3.1*).

Tabelle 5-29 Entropiereduktionen mit dem Bentontest (BTF) (Kriterium: klinische Auffälligkeit)

Kriterium: Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
BTF	1	21,284	<,0001***	3,31	(3,00)
probabilistischer Prädiktor					
Benton Kl.zugehörigkeit (MRM 3Kl)	2	1,416	0,493	0,22	(0,00)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Benton kombiniert (BTF)	5	22,878	0,0004***	3,56	(2,00)

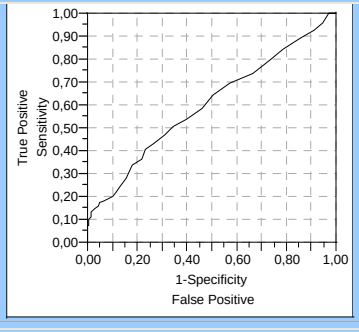
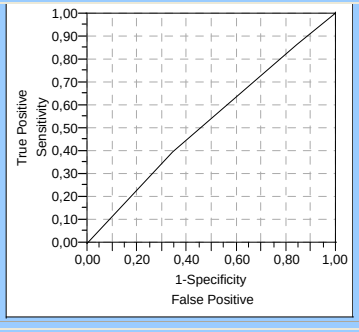
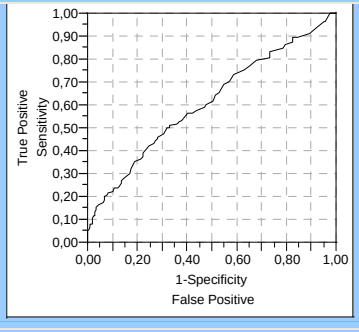
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-30 Wald Effekt Tests beim Bentontest (BTF) (Kriterium: klinische Auffälligkeit)

Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Benton kombiniert (BTF)	BTF	1	5,254	0,0219*
	BentonMRM3	2	0,215	0,8981
	Interaktion	2	1,471	0,4794

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-31 ROC-Kurven beim Bentontest (BTF) (Kriterium: klinische Auffälligkeit)

ROC Receiver Operating Curves (klinische Auffälligkeit = positiv)		
BTF	Benton-Klassenzugehörigkeit	Benton kombiniert (BTF)
		
Fläche unter der Kurve: 0, 600	Fläche unter der Kurve: 0, 528	Fläche unter der Kurve: 0, 609

5.2.3.3.4 Prädiktion „cerebrale Beteiligung“ und Fehlermaß BTF

Ergänzend zur Prädiktion „klinische Auffälligkeit“ wird das Fehlermaß BTF in diesem *Abschnitt* hinsichtlich der Prädiktion des Kriteriums „cerebrale Beteiligung“ untersucht.
Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Zusammen mit den zugehörigen Entropiereduktionen sind die Statistiken der *Tabelle 5-32* zu entnehmen.
Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests, sind in *Tabelle 5-33* wiedergegeben.
In *Tabelle 5-34* sind die ROC-Kurven der Prädiktion „cerebrale Beteiligung“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-32* bis *5-34*:
Die Ergebnisse entsprechen grundsätzlich den Resultaten bei der Prädiktion „klinische Auffälligkeit“ (vgl. *Abschnitt 5.2.3.3.3*).
Mit dem klassischen Prädiktor BTF ergibt sich eine hochsignifikante jedoch wenig relevante Verbesserung der Prädiktion von „cerebrale Beteiligung“ im Vergleich zur Zufallsentscheidung.
Die Kombination der Prädiktoren bringt im Vergleich zum separat verwendeten klassischen Prädiktor keine zusätzlichen Informationen.
Unbedeutende Flächenunterschiede der ROC-Kurven ergeben sich im direkten Vergleich mit den berechneten Werten beim Bentonscore (vgl. *Abschnitt 5.2.3.3.2*).
Das Fehlermaß BTF erweist sich in den berechneten Entropiereduktionswerten 3,08% und adjustiert 2,54% zur Vorhersage „cerebrale Beteiligung“ vorteilhafter als der Bentonscore mit 2,76% und adjustiert 2,22%.

Tabelle 5-32 Entropiereduktionen mit dem Bentontest (BTF) (Kriterium: cerebrale Beteiligung)

Kriterium: cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell					
Prädiktorvariable	**Whole Model Test**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
BTF	1	11,317	0,0008***	3,08	(2,54)
probabilistischer Prädiktor					
Benton Kl.zugehörigkeit (MRM 3Kl)	2	0,650	0,7227	0,18	(0,00)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Benton kombiniert (BTF)	5	12,453	0,0291*	3,39	(0,67)

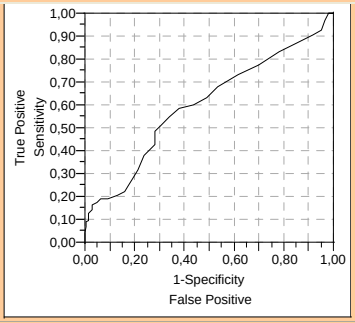
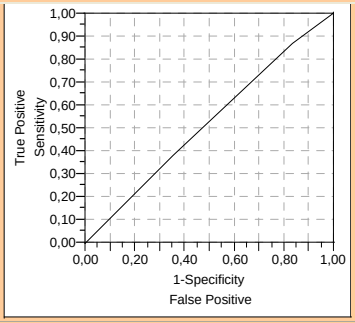
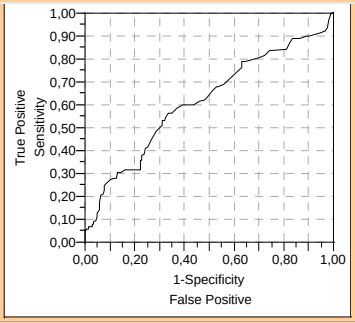
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-33 Wald Effekt Tests beim Bentontest (BTF) (Kriterium: cerebrale Beteiligung)

cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Benton kombiniert (BTF)	BTF	1	4,747	0,0293*
	BentonMRM3	2	0,198	0,9053
	Interaktion	2	0,860	0,6505

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-34 ROC-Kurven beim Bentontest (BTF) (Kriterium: cerebrale Beteiligung)

ROC Receiver Operating Curves (cerebrale Beteiligung = positiv)		
BTF	Benton-Klassenzugehörigkeit	Benton kombiniert (BTF)
		
Fläche unter der Kurve: 0, 600	Fläche unter der Kurve: 0, 521	Fläche unter der Kurve: 0, 615

5.3 Cattelltest

5.3.1 Identifikation latenter Klassen

Der als Cattelltest bezeichnete Speed-Power-Test unterscheidet sich von den übrigen Verfahren[52] durch die zusätzliche Speedkomponente. Die Anwendbarkeit des Raschmodells zur Prüfung der Personenhomogenität ist a priori nicht gegeben, weil beim Cattelltest durch die Art der Testvorgabe eine Konfundierung der Leistungsfähigkeit mit der Leistungsgeschwindigkeit vorliegt[53]. Das Modell der Latenten-Klassen-Analyse (LCA) wird deshalb bei diesem Test als Alternative zum Raschmodell verwendet. Es ermöglicht die Identifizierung qualitativ unterschiedlicher latenter Lösungsklassen, ohne die notwendige Voraussetzung des Raschmodells, dass alle Versuchspersonen sämtliche Items bearbeitet haben.

Beim Cattelltest ist zu beachten, dass die mittels LCA aus den Daten identifizierte latente Klassenzugehörigkeit einer Versuchsperson in engerer Verbindung zu ihrem ermittelten Summenscore steht, als es bei den Klassen der Mixed-Rasch-Modelle bei Raven- und Bentontest der Fall ist. Die stärkere Trennung nach Scores folgt bei der LCA aus der diagnostischen Relevanz der Summenscores.

5.3.1.1 Geltungsprüfung unterschiedlicher probabilistischer Modelle mittels informationstheoretischer Maße und Likelihoodquotiententests

Die berechneten Kennwerte der informationstheoretischen Maße sind für die Gesamtstichprobe in *Tabelle 5-35* zusammengefasst.

[52] Raven- und Bentontest

[53] möglichst schnelle und richtige Bearbeitung in drei Minuten

Tabelle 5-35 Informationstheoretische Maße beim Cattelltest

N=607	loglikeli	AIC	BIC	n_p
CATTELL LCA1Kl	-3248,89	6523,77	6581,08	13
CATTELL LCA2Kl	-3095,32	6244,65	6363,68	27
CATTELL LCA3Kl	-3073,91	6229,83	6410,58	41
CATTELL LCA4Kl	-3052,48	6214,96	6457,43	55
CATTELL LCA5Kl	-3040,23	6218,47	6522,66	69
CATTELL LCA6Kl	-3031,63	6229,26	6595,17	83
saturiertes Modell	-2791,83	21965,65	58075,91	8191
Startwert 4321 abgesichert mit 3321 und 2321				

Aus *Tabelle 5-35* geht hervor, dass vom BIC-Wert (6363,68) eine 2-Klassen-Lösung, vom AIC-Wert (6214,96) eine 4-Klassen-Lösung bevorzugt wird. Mit inferenzstatistischem Likelihoodquotiententest wird die 4-Klassen-Lösung abgesichert (vgl. *Tabelle 5-36*).

Die große Parameteranzahl des saturierten Modells (n_p=8191) und die vergleichsweise geringe Stichprobengröße (N=607) bedingen, dass die Likelihoodquotientenvergleiche nur begrenzt aussagekräftig sind.

Tabelle 5-36 Likelihoodquotiententests Cattell

LR	-2ln(LR)	FG	Prob> χ^2
LCA2Kl/sat.Modell	606,98	8164	1
LCA4Kl/sat.Modell	521,30	8136	1
LCA1Kl/LCA4Kl	392,82	42	<,0001***

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die Ergebnisse der gebildeten Likelihoodquotienten sagen aus, dass die vom AIC indizierte 4-Klassen-Lösung mit den Daten vereinbar ist. Diese passt besser als die Einklassen-Lösung und nicht schlechter als das saturierte Modell. Demnach ist die 4-Klassen-Lösung für die erhobenen Daten ein geeignetes Modell, obwohl auch die vom BIC-Wert angezeigte 2-Klassen-Lösung die Daten nicht schlechter beschreibt als das saturierte Modell.

Die 4-Klassen-Lösung ist die Grundlage der weiteren Berechnungen, da mit einem Modell größerer Klassenanzahl differenziertere diagnostische Aussagen möglich werden.

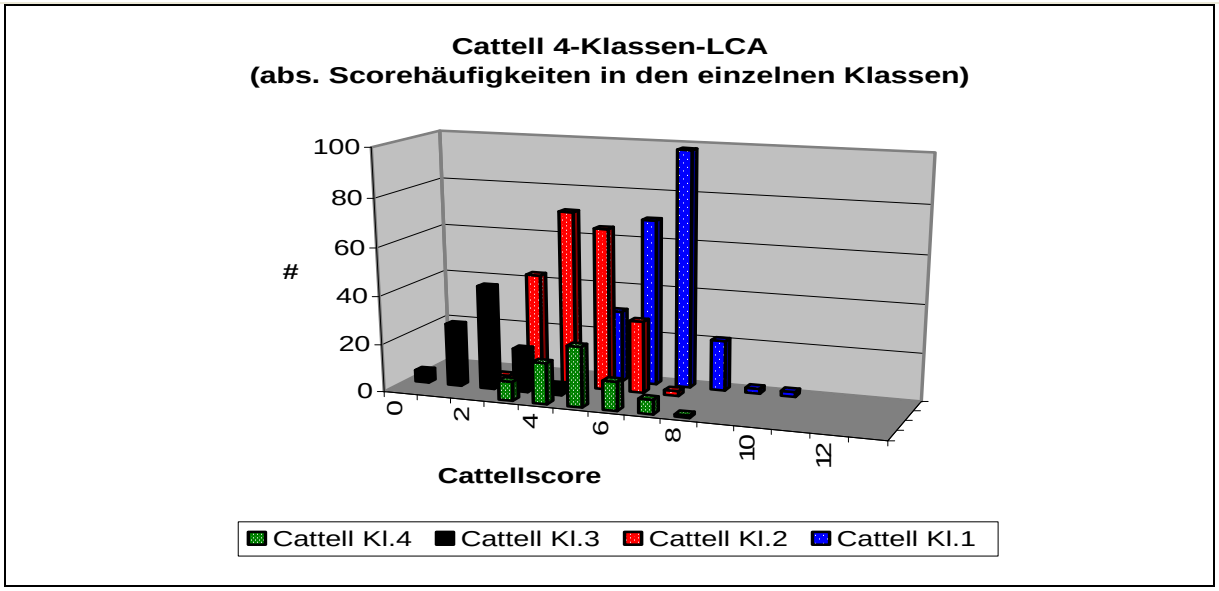
5.3.2 Interpretation latenter Klassen

Die vom AIC indizierte 4-Klassen-Lösung der Latenten-Klassen-Analyse wird genauer untersucht. Die einzelnen Klassen werden aufgrund sämtlicher erhobener Merkmale der zugeordneten Probanden voneinander abgegrenzt.

5.3.2.1 Summenscoreverteilung der 4-Klassen-Lösung der Latenten-Klassen-Analyse

Die unterschiedlichen Testscoreleistungen der vier indizierten Klassen beim Cattelltest stellt *Abbildung 5-9* dar.

Abbildung 5-9 Cattellscores in den Klassen

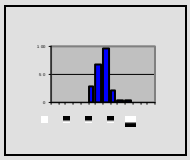
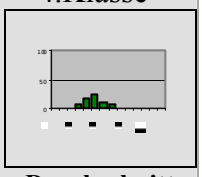
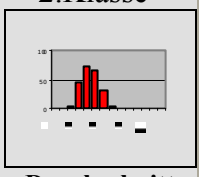
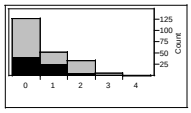
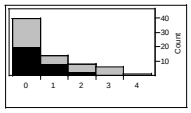
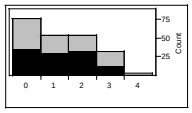
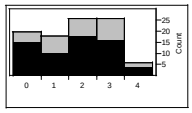
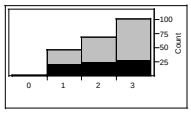
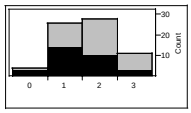
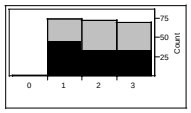
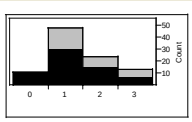
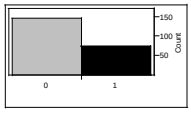
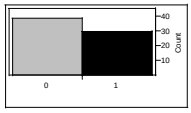
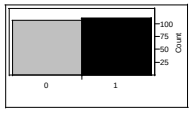
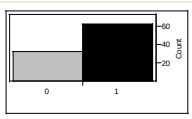


5.3.2.2 Charakterisierung der 4-Klassen-Lösung der Latenten-Klassen-Analyse

Die vier Klassen des Cattelltests werden nach dem Score-Mittelwert geordnet und mit den Bezeichnungen „Könner “ für Klasse 1, „Durchschnitt A“ für Klasse 4, „Durchschnitt B“ für Klasse 2 und „Nicht-Könner“ für Klasse 3 versehen.

Diese Klassen werden in *Tabelle 5-37* charakterisiert und durch die zugehörigen Kontraste in *Tabelle 5-38* ergänzt.

Tabelle 5-37 Die 4-Klassen-Lösung der LCA beim Cattelltest

4-Klassen-Lösung der Latenten-Klassen-Analyse beim Cattelltest				
	hoch ← Quantitative Testleistung → niedrig			
Cattelltest (N=607) Scoreverteilung Maßstab x=Score: 0-13 y=N: 0-100	LCA4, 1.Klasse  **„Könner“**	LCA4, 4.Klasse  **„Durchschnitt A“**	LCA4, 2.Klasse  **„Durchschnitt B“**	LCA4, 3.Klasse **„Nicht-Könner“**
N (%-Anteil gesamt)	222 (37%)	69 (11%)	220 (36%)	96 (16%)
mittlere klassenspezifische Zuordnungssicherheit	0,813	0,912	0,773	0,859
Cattellscore [Ø]	6,56	4,91	4,38	1,90
Cattellscore [σ]	0,93	1,17	1,02	0,91
bearbeitete Items [Ø]	9,17	10,17	7,54	7,93
bearbeitete Items [σ]	1,85	1,85	2,06	2,76
Alter [Ø]	36,42	38,22	44,74	51,60
Alter [σ]	12,65	13,67	14,34	13,88
Altersverteilung Hell: klinisch unauffällig Dunkel: klinisch auffällig 0=15-39LJ, 1=39-49LJ, 2=49-59LJ, 3=59-70LJ, 4= >70LJ				
Bildungsverteilung Hell: klinisch unauffällig Dunkel: klinisch auffällig 0=ohne Schulabschluss, 1=Hauptschule, 2=Realschule, 3=Abitur				
Klinische Auffälligkeit Hell=0: klinisch unauffällig Dunkel=1: klinisch auffällig				

Aus den *Tabellen 5-37* und *5-38* geht hervor:

- Sämtliche Scoreunterschiede der vier Klassen sind mindestens sehr signifikant.
- Die ähnlichen[54] Scoremittelwerte, insbesondere der beiden identifizierten latenten Klassen „Durchschnitt A“ mit 4,91 und „Durchschnitt B“ mit 4,38 geben Hinweise darauf, dass Klassenzuteilungen nicht nur aus quantitativen sondern auch aus qualitativen Informationen resultieren.

[54] dennoch sehr signifikant unterschiedlichen

Qualitative Bearbeitungsunterschiede ergeben sich aus der separaten Betrachtung der *Tabelle 5-37* zu entnehmenden Anzahl bearbeiteter Cattell-Items.

In der vorgegebenen Zeitspanne von drei Minuten bearbeiten die Personen der Klasse „Durchschnitt A“ mit 10,17 Items hochsignifikant mehr Items als „Durchschnitt B“ mit 7,54 Items. Gleichzeitig ist die „leistungsschwächere“ Klasse „Durchschnitt B“ mit 0,58 richtigen Lösungen pro bearbeitetem Item, der „leistungsstärkeren“ Klasse „Durchschnitt A“ mit 0,48 richtigen Lösungen pro bearbeitetem Item überlegen.

Die mittlere Anzahl bearbeiteter Items erweist sich nur in den beiden Klassen „Durchschnitt B“ mit 7,54 Items und „Nichtkönner“ mit 7,93 Items als nicht signifikant verschieden. Das tendenziell schnellere Arbeitstempo der „Nichtkönner“ kann mit Rateversuchen bei der Itemlösung gedeutet werden.

- In den zugehörigen Grafiken lässt sich eine nicht durchgehend signifikante Alters- und Bildungsabhängigkeit der Testleistungen erkennen. In den Klassen mit niedrigerer Testleistung zeigt sich außerdem eine Zunahme klinisch auffälliger Personen. Die Einzelvergleiche der Klassen mit hoher und niedriger Testleistung ergeben signifikante Unterschiede der Klassenzusammensetzungen aus klinisch auffälligen und unauffälligen Personen.

Sämtliche Kontraste in *Tabelle 5-38* sind ausführlicher im Anhang in *Tabelle A-38* dargestellt.

Tabelle 5-38 Zusammenfassung der Kontraste der 4-Kl.-Lösung der LCA beim Cattelltest

Zusammenfassung der Kontraste beim Cattelltest				
Kontraste der Scores beim Cattelltest	Könner	DurchschnittA	DurchschnittB	Nichtkönner
Könner		***	***	***
DurchschnittA	***		**	***
DurchschnittB	***	**		***
Nichtkönner	***	***	***	
Kontraste bearbeiteter Items beim Cattelltest	Könner	DurchschnittA	DurchschnittB	Nichtkönner
Könner		***	***	***
DurchschnittA	***		***	***
DurchschnittB	***	***		ns
Nichtkönner	***	***	ns	
Kontraste des Lebensalters beim Cattelltest	Könner	DurchschnittA	DurchschnittB	Nichtkönner
Könner		ns	***	***
DurchschnittA	ns		*	***
DurchschnittB	***	*		***
Nichtkönner	***	***	***	
Kontraste der Altersgruppen beim Cattelltest	Könner	DurchschnittA	DurchschnittB	Nichtkönner
Könner		ns	***	***
DurchschnittA	ns		**	***
DurchschnittB	***	**		**
Nichtkönner	***	***	**	
Kontraste der Bildungsgrade beim Cattelltest	Könner	DurchschnittA	DurchschnittB	Nichtkönner
Könner		***	**	***
DurchschnittA	***		ns	ns
DurchschnittB	**	ns		***
Nichtkönner	***	ns	***	
Kontraste der Kontingenztafeln (klin. Auff. ja/nein) beim Cattelltest	Könner	DurchschnittA	DurchschnittB	Nichtkönner
Könner		ns	**	***
DurchschnittA	ns		ns	*
DurchschnittB	**	ns		ns
Nichtkönner	***	*	ns	
*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur)				
entnommen aus *Tab.A-38*				

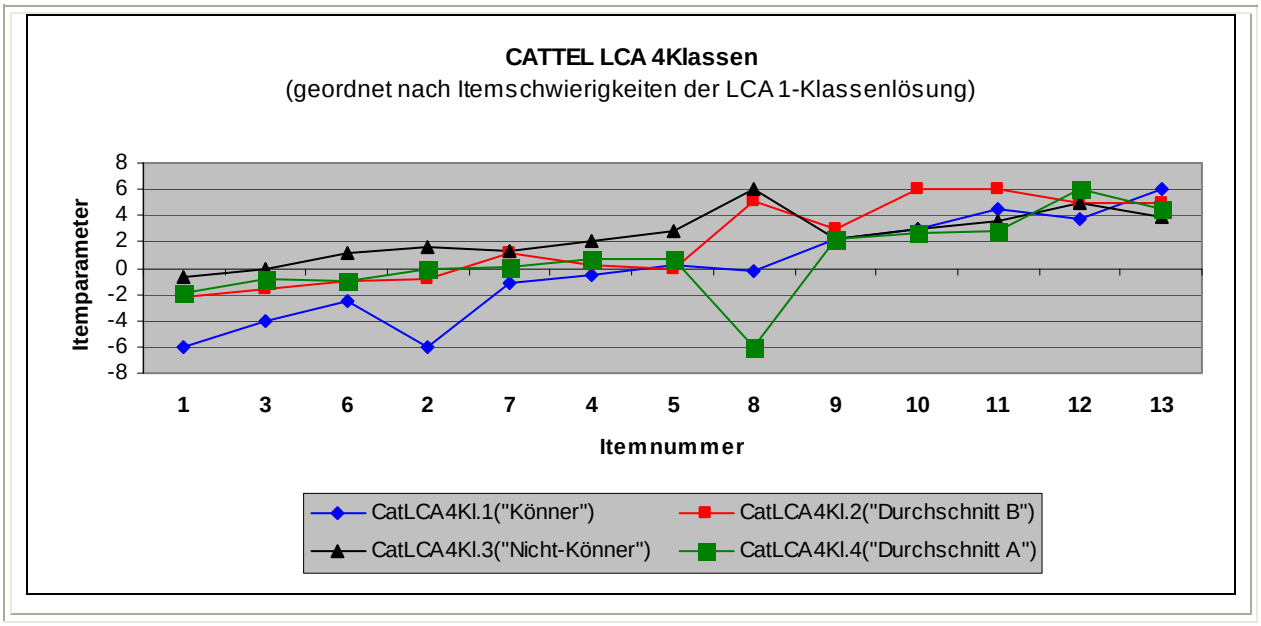
5.3.2.3 Die Profile der Itemparameter der 4-Klassen-Lösung der Latenten-Klassen-Analyse

Die Itemparameter der 4-Klassen-Lösung des Cattelltests sind in *Abbildung 5-10* dargestellt. Die zugehörigen Werte sind *Tabelle A-39* im Anhang zu entnehmen.

Die Anordnung der Itemparameter entlang der Abszisse resultiert aus der Reihenfolge der Schwierigkeiten[55] wie sie die Einklassen-Lösung der Latenten-Klassen-Analyse vorgibt.

In der *Abbildung* erweist sich die Abfolge der Items mit den Nummern 8, 9, 10, 11, 12 und 13 im rechten Teil der Abszisse als Auswirkung der Speedkomponente des Cattelltests, weil Testpersonen spätere Items aus Zeitmangel zunehmend seltener bearbeiten.

Abbildung 5-10 Itemparameter der 4-Klassen-Lösung der LCA beim Cattelltest



Bewertung der *Abbildung 5-10* und Fazit:

- Beim Cattelltest sind Inhomogenitäten außer bei Item 2, noch offensichtlicher bei Item 8 erkennbar.

  Dieses Item 8 bereitet der Klasse 4 „Durchschnitt A“ mit dem schnellsten Arbeitstempo keinerlei Lösungsschwierigkeiten. Es ist jedoch für die beiden schwächeren Klassen 2 „Durchschnitt B“ und 3 „Nicht-Könner“ nahezu unlösbar. Zeitmangel wäre eine mögliche Erklärung dafür, dass dieses Item von den Personen der beiden letzteren Klassen nur in seltenen Fällen bearbeitet wird.

  Auch die leistungsstärkste Klasse 1 „Könner“ hat bei der Lösung von Item 8 größere Schwierigkeiten als die leistungsschwächere, jedoch vom Arbeitstempo schnellere

[55] von links nach rechts in ansteigender Schwierigkeit

Klasse 4 „Durchschnitt A“. Eine Erklärungsmöglichkeit dafür wäre, dass „Könner“ im Vergleich zu den Personen von „Durchschnitt A“ gegen Ende des Tests unter Zeitdruck kommen und aus Zeitmangel bei diesem „späten“ Item zu Fehlern neigen.

Fazit: Der *Abbildung 5-10* können keine Hinweise auf klassenspezifische Bearbeitungsstrategien entnommen werden, die den gesamten Testverlauf beeinflussen.

Die Auswirkung der Speedkomponente kommt erst im Testverlauf zu ihrer vollen Entfaltung. So ergibt sich beim Cattelltest erst nach dem vierten vorgelegten Item eine deutlicher werdende Überlagerung der Itemparameterprofile, die jedoch bis auf die Items 2 und 8 in allen Klassen grundsätzlich ähnliche Verläufe haben (vgl. *Abbildung 5-10*). Diesen Schluss belegen die positiven, weitgehend monotonen Rangkorrelationen der Itemparameter der einzelnen Klassen. Die Zusammenhänge werden bis auf eine Ausnahme (Rangkorrelation „Durchschnitt A“ - „Nichtkönner“: 0,54) mindestens signifikant (vgl. im Anhang *Tabelle A-40*).

Die metrischen Informationen in den Itemparametern ergeben in den Produkt-Moment-Korrelationen bis auf eine Ausnahme (Produkt-Moment-Korrelation „Könner“ - „Durchschnitt B“: 0,87) deutlich kleinere Korrelationen mit geringeren Signifikanzen. Die kleineren Werte lassen sich auf die klassenspezifischen Unterschiede der Itemparameter bei Item 8 zurückführen (vgl. *Abbildung 5-10*). Werden die Produkt-Moment-Korrelationen ohne dieses problematische Item berechnet, ergeben sich für sämtliche Klassenvergleiche mit Korrelationswerten >0,83 hohe positive und hochsignifikante Zusammenhänge der Itemparameter (ebenso *Tabelle A-40*).

5.3.2.4 Homogenität der Items

i) Graphisches Vorgehen

Entsprechend dem in *Abschnitt 3.2 b)* erläuterten Vorgehen soll anhand der Itemprofile der LCA die Homogenität der ersten vier vorgelegten Cattelltest-Items geprüft werden. Hierzu müssen **zwei Monotoniebedingungen** erfüllt sein:

1) Geordnete Profile der latenten Klassen
2) Identische Rangordnung der Antwortwahrscheinlichkeiten

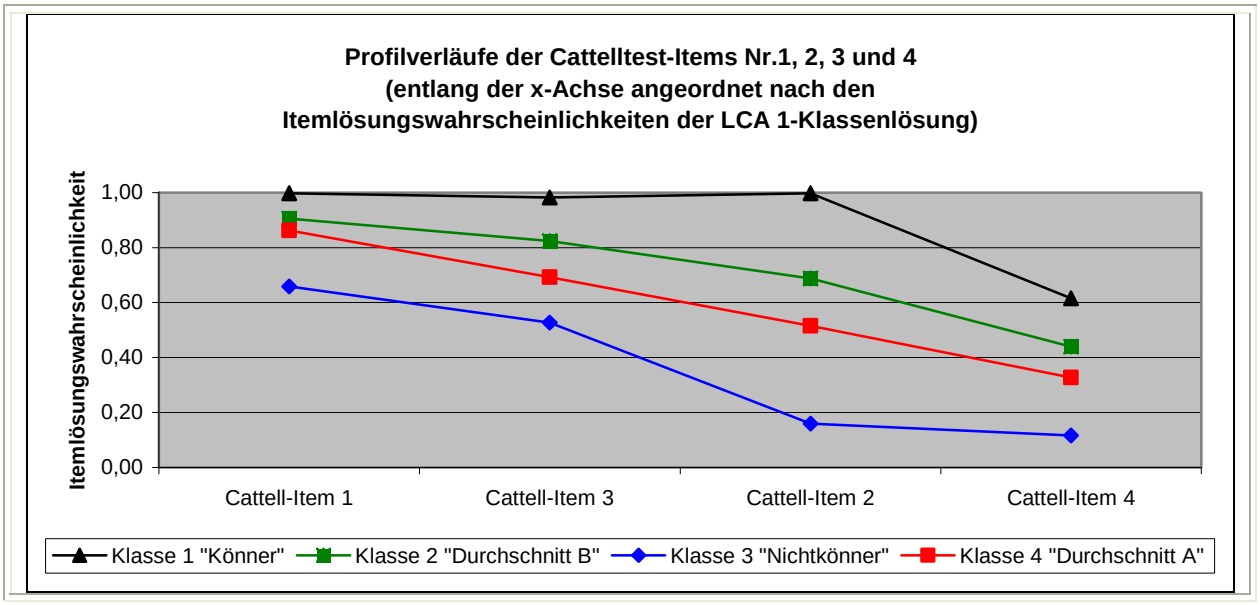
Die klassenspezifischen Lösungswahrscheinlichkeiten der Cattelltest-Items Nr. 1, 2, 3 und 4 sind in *Abbildung 5-11* dargestellt. Dieser *Abbildung* sind geordnete überschneidungsfreie Profile der vier latenten Klassen zu entnehmen. Das Niveau des Lösungserfolgs wird

demnach direkt durch die Klassenzugehörigkeit bestimmt. Die „Nichtkönner“ haben bei der Lösung der vier betrachteten Items durchgängig die größten und die „Könner“ die geringsten Schwierigkeiten. Somit ist die **1. Monotoniebedingung** erfüllt.
Identische Rangordnung der Antwortwahrscheinlichkeiten, also die **2. Monotoniebedingung**, wird wegen Item 2 nicht erfüllt. Dieses Item ist nur für die „Könner“ leichter zu lösen als das Item 3.

ii) Likelihoodquotiententest
Im Likelihoodquotiententest des einfachen Raschmodells ($\ln(mL_{RM})$=-1251,29) mit der 2-Klassen-Lösung des Mixed-Rasch-Modells ($\ln(mL_{2Kl.MRM})$=-1241,26) beträgt die ermittelte Prüfgröße 20,06. Sie liegt damit oberhalb der kritischen Grenze von 12,59 der χ^2- Verteilung (α=5%) bei sechs Freiheitsgraden.
Ebenso wie beim graphischen Vergleich bestätigt sich auch hier, dass die Annahme der Raschhomogenität der ersten vier Cattelltest-Items verworfen werden muss.

Somit bleibt die gültige Anwendung eines quantitativen Testmodells (Raschmodell) für diese vier Items nicht abgesichert, weil heterogene Personenpopulationen nachgewiesen werden.
Abbildung 5-11 ist zu entnehmen, dass mit dem Likelihoodquotiententest aufgedeckte Inhomogenitäten weniger aus der, wegen Item 2, nicht erfüllten **2. Monotoniebedingung**, sondern vielmehr aus der im Klassenvergleich zu hohen Lösungswahrscheinlichkeit bei Item 3 in der Klasse der „Nichtkönner“ resultieren.

Abbildung 5-11 Profilverläufe der ersten vier Cattelltest-Items



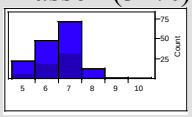
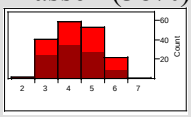
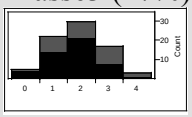
5.3.3 Diagnostische Relevanz der latenten Klassen

Der klinische Nutzen des Cattelltests soll genau analysiert werden. Hierzu werden die Scores innerhalb der einzelnen Klassen verglichen. Unter *Abschnitt 5.3.3.1* von klinisch auffälligen und unauffälligen Personen, zusätzlich unter *Abschnitt 5.3.3.2* von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung.
Für die Beurteilung der Vorhersagevalidität klinischer Kriterien mit klassischen und probabilistischen Prädiktoren wird ab *Abschnitt 5.3.3.3* die logistische Regression verwendet.

5.3.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen

Zur Beurteilung der Scoreunterschiede klinisch auffälliger und unauffälliger Personen in den einzelnen Klassen des Cattelltests wird die parallelisierte Stichprobe (N=463) verwendet.
Bei vergleichbaren Zusammensetzungen der Alters- und Bildungsgrade der klinisch auffälligen und unauffälligen Personengruppe innerhalb der einzelnen Klassen des Cattelltests, ergeben sich in *Tabelle 5-39* keine statistisch signifikanten Scoreunterschiede.

Tabelle 5-39 Differenzierte Betrachtung der einzelnen Klassen beim Cattelltest (klinische Auffälligkeit)

Differenzierte Betrachtung der klinisch auffälligen und unauffälligen Stichprobe in den einzelnen Klassen des Cattelltests								
Parallelisierte Stichprobe	hoch ← Quantitative Testleistung → niedrig							
Cattelltest (N=463) Scoreverteilung [ungleicher Maßstab] Hell: klinisch unauffällig (klU) Dunkel: klinisch auffällig (klA)	Klasse1 (34%)  „Könner“		Klasse4 (11%) „Durchschnitt A“		Klasse2 (38%)  „DurchschnittB“		Klasse3 (17%)  „Nicht-Könner“	
	klU	klA	klU	klA	klU	klA	klU	klA
N (%-Anteil in Klasse)	95 (61%)	62 (39%)	30 (59%)	21 (41%)	81 (46%)	97 (54%)	29 (38%)	48 (62%)
t-Tests der Scores innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)								
Cattellscore [Ø]	6,53	6,53	4,93	4,81	4,43	4,21	2,10	1,75
Cattellscore [σ]	0,97	0,80	1,26	0,87	1,07	0,96	1,01	0,91
t-Test	-0,040		0,390		1,485		1,582	
FG	155		49		176		75	
Prob > |t|	0,9680		0,6985		0,1393		0,1179	
t-Tests des Lebensalters innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)								
Alter [Ø]	39,79	38,95	41,67	39,24	46,17	47,06	53,72	54,58
Alter [σ]	13,27	8,99	17,34	8,33	15,44	11,21	11,73	12,47
t-Test	0,436		0,594		-0,444		-0,299	
FG	155		49		176		75	
Prob > |t|	0,6635		0,5550		0,6575		0,7654	
Wilcoxon-Tests der Altersgruppen innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)								
χ^2 approx.	5,5016		2,8581		0,0019		0,1025	
FG	1		1		1		1	
Prob> χ^2	0,0190*		0,0909		0,9650		0,7489	
Wilcoxon-Tests der Bildungsgrade innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig)								
χ^2 approx.	0,1651		0,0041		0,6863		0,0099	
FG	1		1		1		1	
Prob> χ^2	0,6845		0,9487		0,4074		0,9206	

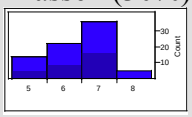
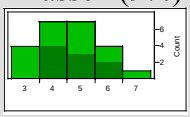
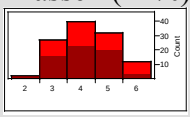
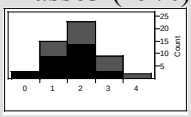
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

5.3.3.2 Scorevergleiche von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung

Zur Beurteilung der Scoreunterschiede von klinisch unauffälligen Personen und Patienten mit cerebraler Beteiligung in den einzelnen Klassen des Cattelltests wird die Stichprobe `parallel-cerebral` (N=265) verwendet.

Bei vergleichbarer Zusammensetzung der Alters- und Bildungsgrade sind in *Tabelle 5-40* nur bei den Patienten mit cerebraler Beteiligung in der leistungsschwächsten Klasse 3 „Nicht-Könner“ signifikant niedrigere Scores nachweisbar.

Tabelle 5-40 Differenzierte Betrachtung der einzelnen Klassen beim Cattelltest (cerebrale Beteiligung)

Differenzierte Betrachtung der Patienten mit cerebraler Beteiligung und klinisch unauffälliger Personen in den einzelnen Klassen des Cattelltests								
Stichprobe `parallel-cerebral`	hoch ← Quantitative Testleistung → niedrig							
Cattelltest (N=265) Scoreverteilung [ungleicher Maßstab] Hell: klinisch unauffällig (klU) Dunkel: klinisch auffällig mit cerebraler Beteiligung (mcB)	Klasse1 (30%)  „Könner“		Klasse4 (9%)  „Durchschnitt A“		Klasse2 (42%)  „DurchschnittB“		Klasse3 (19%)  „Nicht-Könner“	
	klU	mcB	klU	mcB	klU	mcB	klU	mcB
N (%-Anteil in Klasse)	47 (61%)	30 (39%)	14 (61%)	9 (39%)	50 (44%)	63 (56%)	23 (44%)	29 (56%)
t-Tests der Scores innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)								
Cattellscore [Ø]	6,45	6,37	4,50	4,77	4,34	4,13	2,17	1,59
Cattellscore [σ]	0,93	0,77	1,29	0,83	1,08	0,90	0,94	0,82
t-Test	0,395		-0,573		1,139		2,404	
FG	75		21		111		50	
Prob > |t|	0,6941		0,5729		0,2571		0,0200*	
t-Tests des Lebensalters innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)								
Alter [Ø]	45,57	42,13	51,71	41,33	52,26	49,71	55,30	60,79
Alter [σ]	8,82	7,82	13,05	9,96	12,71	10,24	11,49	9,61
t-Test	1,742		2,030		1,179		-1,875	
FG	75		21		111		50	
Prob > |t|	0,0857		0,0552		0,2408		0,0666	
Wilcoxon-Tests der Altersgruppen innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)								
χ^2 approx.	3,541		3,9179		0,5651		2,7906	
FG	1		1		1		1	
Prob> χ^2	0,0599		0,0478*		0,4522		0,0948	
Wilcoxon-Tests der Bildungsgrade innerhalb der einzelnen Klassen (Klinisch unauffällig - Klinisch auffällig mit cerebraler Beteiligung)								
χ^2 approx.	0,0075		0,6844		0,2451		0,4313	
FG	1		1		1		1	
Prob> χ^2	0,9311		0,4081		0,6206		0,5114	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

5.3.3.3 Vorhersagevalidität

Die Vorhersagevalidität der klinischen Kriterien „klinische Auffälligkeit“ und „cerebrale Beteiligung“ wird mit Kennwerten aus nominal logistischen Regressionen beurteilt.
Hierzu werden die einzelnen und die in Kombination verwendeten klassischen und probabilistischen Prädiktoren des Cattelltests verwendet.

5.3.3.3.1 Prädiktion „klinische Auffälligkeit“

Der Cattelltest wird hinsichtlich der Prädiktion des Kriteriums „klinische Auffälligkeit“ untersucht.
Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die zugehörigen Statistiken und die Entropiereduktionen sind *Tabelle 5-41* zu entnehmen.
Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests sind in *Tabelle 5-42* wiedergegeben. Diese Kombination wird hier als „Cattell kombiniert“ bezeichnet.
In *Tabelle 5-43* sind die ROC-Kurven der Prädiktion „klinische Auffälligkeit“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-41* bis *5-43*:
Beim Cattelltest werden sämtliche Modelltests mindestens sehr signifikant.
Die Sicherheit der Prädiktion „klinische Auffälligkeit“ ist für alle Prädiktoren vergleichbar.
Durch die Kombination des klassischen und probabilistischen Prädiktors ergeben sich keine relevanten Verbesserungen im Vergleich zu den separat verwendeten Prädiktoren.
Die ähnlichen Prädiktionskennwerte weisen darauf hin, dass der nur vier Klassen umfassende probabilistische Prädiktor im Vergleich zum klassischen Prädiktor kaum Information einbüßt, obwohl dieser dreizehn mögliche Scores beinhaltet. Dieses Ergebnis unterstreicht die Qualität und Relevanz der Klassenbildung mittels LCA beim Cattelltest.

Tabelle 5-41 Entropiereduktionen mit dem Cattelltest (Kriterium: klinische Auffälligkeit)

Kriterium: Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test (N=463)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
Cattell Score	1	15,560	<,0001***	2,42	(2,11)
probabilistischer Prädiktor					
Cattell Kl.zugehörigkeit (LCA 4Kl)	3	14,656	0,002**	2,28	(1,35)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Cattell kombiniert	7	19,545	0,007**	3,05	(0,86)

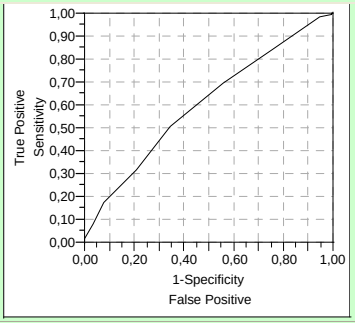
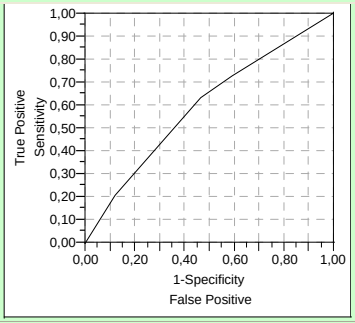
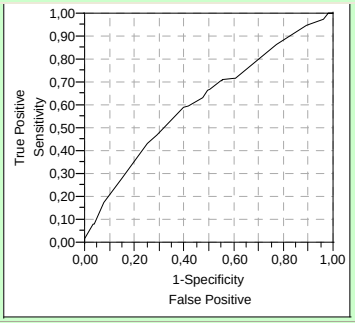
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-42 Wald Effekt Tests beim Cattelltest (Kriterium: klinische Auffälligkeit)

Klinische Auffälligkeit parallelisierte Stichprobe - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Cattell kombiniert	Cattellscore	1	2,720	0,0991
	Cattell LCA4	3	3,397	0,3344
	Interaktion	3	1,937	0,5855

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-43 ROC-Kurven beim Cattelltest (Kriterium: klinische Auffälligkeit)

ROC Receiver Operating Curves (klinische Auffälligkeit = positiv)		
Cattellscore	Cattell-Klassenzugehörigkeit	Cattell kombiniert
		
Fläche unter der Kurve: 0,601	Fläche unter der Kurve: 0,596	Fläche unter der Kurve: 0,615

5.3.3.3.2 Prädiktion „cerebrale Beteiligung“

Ergänzend zur Prädiktion „klinische Auffälligkeit“ wird der Cattelltest in diesem *Abschnitt* hinsichtlich der Prädiktion des Kriteriums „cerebrale Beteiligung“ untersucht.
Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die zugehörigen Statistiken und die Entropiereduktionen sind *Tabelle 5-44* zu entnehmen.
Signifikanzprüfungen einzelner Beiträge aus der Kombination klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests sind in *Tabelle 5-45* wiedergegeben.
In *Tabelle 5-46* sind die ROC-Kurven der Prädiktion „cerebrale Beteiligung“ für die Prädiktoren abgebildet und die Größen der Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-44* bis *5-46*:
Beim Cattelltest haben die Modelltests zur Prädiktion „cerebrale Beteiligung“ unterschiedliche Signifikanzniveaus. Diese sind sehr signifikant beim klassischen Prädiktor, signifikant bei der Kombination der Prädiktoren und nicht signifikant beim probabilistischen Prädiktor.
Die größte Entropiereduktion, also die sicherste Prädiktion von „cerebrale Beteiligung“ ergibt sich mit der klassisch-probabilistischen Prädiktorenkombination.

Tabelle 5-44 Entropiereduktionen mit dem Cattelltest (Kriterium: cerebrale Beteiligung)

Kriterium: cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell					
Prädiktorvariable	**Modell-Test (N=265)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassischer Prädiktor					
Cattell Score	1	7,192	0,0073**	1,96	(1,41)
probabilistischer Prädiktor					
Cattell Kl.zugehörigkeit (LCA 4Kl)	3	7,039	0,0707	1,92	(0,28)
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)					
Cattell kombiniert	7	14,523	0,0426*	3,95	(0,14)

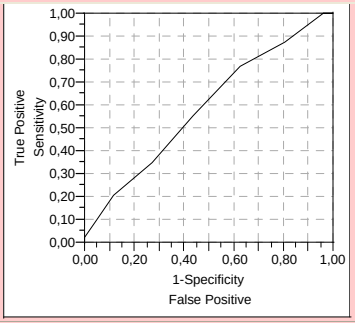
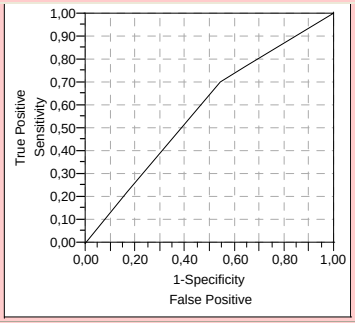
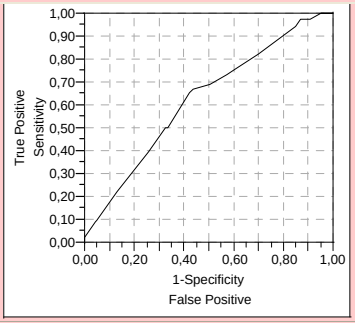
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-45 Wald Effekt Tests beim Cattelltest (Kriterium: cerebrale Beteiligung)

Cerebrale Beteiligung Stichprobe `parallel-cerebral` - nominal logistisches Modell				
Wald Effekt Tests				
	Variable	FG	Wald- χ^2	Prob> χ^2
klassischer und probabilistischer Prädiktor (einschließlich Interaktion)				
Cattell kombiniert	Cattellscore	1	1,983	0,1591
	Cattell LCA4	3	5,841	0,1196
	Interaktion	3	4,011	0,2602

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-46 ROC-Kurven beim Cattelltest (Kriterium: cerebrale Beteiligung)

ROC Receiver Operating Curves (cerebrale Beteiligung = positiv)		
Cattellscore	Cattell-Klassenzugehörigkeit	Cattell kombiniert
		
Fläche unter der Kurve: 0,588	Fläche unter der Kurve: 0,579	Fläche unter der Kurve: 0,623

5.4 Raven-, Benton- und Cattelltest gemeinsam

Nachdem bislang jeder Test separat betrachtet wurde, wird nun die Vorhersagevalidität einer aus Raven-, Benton- und Cattelltest zusammengesetzten Testbatterie untersucht.
Zur Bestimmung der Prädiktion „klinische Auffälligkeit“ und „cerebrale Beteiligung“ werden alle klassischen, probabilistischen und die aus klassischen und probabilistischen kombinierten Prädiktoren in nominal logistische Modelle einbezogen.

5.4.1 Prädiktion „klinische Auffälligkeit“

Die Testbatterie wird hinsichtlich der Prädiktion des Kriteriums „klinische Auffälligkeit“ untersucht.
Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die zugehörigen Statistiken und die Entropiereduktionen sind *Tabelle 5-47* zu entnehmen.
In *Tabelle 5-48* sind Signifikanzprüfungen einzelner Beiträge aus den Kombinationen klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests wiedergegeben.
In *Tabelle 5-49* sind die ROC-Kurven für die Prädiktoren abgebildet und die Größen der zugehörigen Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-47* bis *5-49*:
Die Werte der Entropiereduktionen und die Größen der Flächen unter den ROC-Kurven zeigen eine verbesserte Prädiktion von „klinische Auffälligkeit“ durch die Kombination klassischer und probabilistischer Prädiktoren an[56]. Die Werte der Entropiereduktionen steigen von 3,41% klassisch und 2,29% probabilistisch, auf 5,58% bei kombinierten Prädiktoren.
Die Wald-Effekt-Tests der klassischen Prädiktorenkombination belegen dabei mit sehr signifikanten bzw. signifikanten Beiträgen die Relevanz des Benton- und des Cattellscores.
In der probabilistischen Kombination erweist sich nur die sehr signifikante 4-Klassen-Lösung der LCA des Cattelltests als nützlich.
Wald-Effekt-Tests der klassisch-probabilistischen Prädiktorenkombination zeigen keine signifikanten Ergebnisse. Zur Prädiktion von „klinische Auffälligkeit“ trägt hier nur der Bentonscore bei, allerdings mit marginalen Signifikanzen ($p \leq 0{,}1$).

[56] jedoch muss von der großen Anzahl hierzu nötiger Schätzparameter abgesehen werden!

Tabelle 5-47 Entropiereduktionen bei drei verwendeten Tests (Kriterium: klinische Auffälligkeit)

Kriterium: Klinische Auffälligkeit (parallelisierte Stichprobe - nominal logistisches Modell)					
Prädiktorvariablen-kombination	**Modell-Test (N=455)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassische Prädiktoren					
Score R,C&B	3	21,418	<,0001***	3,41	(2,46)
probabilistische Prädiktoren					
Klassenzug. R,C&B	7	14,413	0,044*	2,29	(0,06)
klassische und probabilistische Prädiktoren (einschließlich Interaktionen)					
Kombiniert	17	35,174	0,006**	5,58	(0,19)

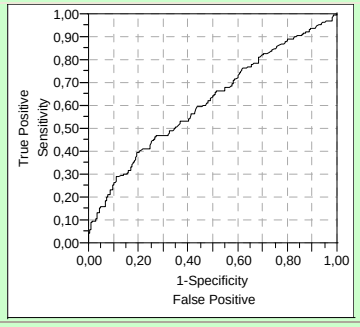
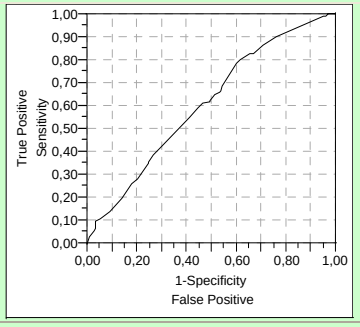
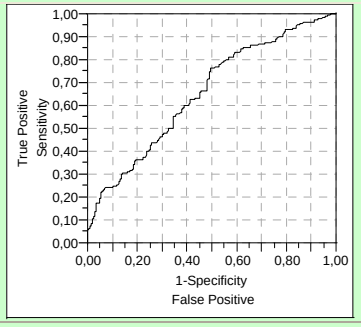
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-48 Wald Effekt Tests bei drei verwendeten Tests (Kriterium: klinische Auffälligkeit)

Klinische Auffälligkeit (parallelisierte Stichprobe - nominal logistisches Modell)				
Prädiktorvariablen-kombination	**Wald Effekt Tests**			
	Variable	FG	Wald- χ^2	Prob> χ^2
klassische Prädiktoren				
Score R,C&B	Ravenscore	1	3,500	0,0614
	Cattellscore	1	6,489	0,0109*
	Bentonscore	1	7,959	0,0048**
probabilistische Prädiktoren				
Klassenzug. R,C&B	RavenMRM3	2	2,032	0,3620
	Cattell LCA4	3	11,912	0,0077**
	BentonMRM3	2	0,642	0,7253
klassische und probabilistische Prädiktoren (einschließlich Interaktionen)				
Kombiniert	Ravenscore	1	0,851	0,3562
	RavenMRM3	2	1,705	0,4263
	Interaktion Ravenscore & MRM3	2	4,311	0,1158
	Cattellscore	1	2,488	0,1147
	Cattell LCA4	3	2,914	0,4050
	Interaktion Cattellscore & LCA4	3	3,013	0,3896
	Bentonscore	1	3,198	0,0737
	BentonMRM3	2	0,116	0,9438
	Interaktion Bentonscore & MRM3	2	2,235	0,3270

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-49 ROC-Kurven bei drei verwendeten Tests (Kriterium: klinische Auffälligkeit)

ROC Receiver Operating Curves (klinische Auffälligkeit = positiv)		
R,C & B Scores	R,C & B Klassenzugehörigkeiten	Kombiniert
		
Fläche unter der Kurve: 0,616	Fläche unter der Kurve: 0,605	Fläche unter der Kurve: 0,654

5.4.2 Prädiktion „cerebrale Beteiligung“

Ergänzend wird die Testbatterie hinsichtlich der Prädiktion des Kriteriums „cerebrale Beteiligung“ untersucht.

Die Modellgültigkeit der logistischen Regressionen wird mit Likelihood-Ratio-Tests geprüft. Die zugehörigen Statistiken und die Entropiereduktionen sind *Tabelle 5-50* zu entnehmen.

In *Tabelle 5-51* sind die Signifikanzprüfungen einzelner Beiträge aus den Kombinationen klassischer und probabilistischer Prädiktoren auf der Grundlage der Wald Effekt Tests wiedergegeben.

In *Tabelle 5-52* sind die ROC-Kurven für die Prädiktoren abgebildet und die zugehörigen Flächen unter den Kurven angegeben.

Erläuterung der Ergebnisse in den *Tabellen 5-50* bis *5-52*:

Die Prädiktion „cerebrale Beteiligung“ wird durch die Kombination klassischer und probabilistischer Prädiktoren deutlich verbessert[57]. Die Werte der Entropiereduktionen steigen von 3,71% klassisch und 2,49% probabilistisch, auf beachtliche 8,13% bei kombinierten Prädiktoren.

Signifikante Wald-Effekt-Tests bleiben für das Kriterium „cerebrale Beteiligung“ bei dieser klassisch-probabilistischen Prädiktorenkombination und auch bei der separat durchgeführten probabilistischen Prädiktorenkombination aus.

Dagegen veranschaulichen die Wald-Effekt-Tests der klassischen Prädiktorenkombination

[57] ohne Berücksichtigung der großen Anzahl hierzu nötiger Schätzparameter!

mit sehr signifikanten Beiträgen die Relevanz des Bentonscores und - immerhin signifikant - des Ravenscores.

Dies ist ein bemerkenswerter Nebenbefund, da der Ravenscore bei sonstigen Prädiktionen in dieser Arbeit nirgends eine statistisch signifikante Rolle spielt. Zur Prädiktion des Kriteriums „klinische Auffälligkeit“ hatten sich in *Abschnitt 5.4.1* stattdessen Bentonscore und Cattellscore als wichtig erwiesen.

Die Analysen der einzelnen Koeffizienten in der Regressionsgleichung zeigen, dass der Ravenscore zur Vorhersage des dichotomen Kriteriums „cerebrale Beteiligung“ für konstante Benton- und Cattellscores in die entgegengesetzte Richtung differenziert.

Für die verwendeten Prädiktoren Ravenscore, Bentonscore und Cattellscore lautet die zugehörige Regressionsgleichung des nominal logistischen Modells:

Prädiktion „cerebrale Beteiligung“ für das Modell Raven-, Cattell- und Bentonscore

logit (p(cerebr. Bet.=ja))=0,712 + 0,094 Ravenscore - 0,158 Cattellscore - 0,108 Bentonscore

Aus den Vorzeichen der Koeffizienten folgt, dass die Wahrscheinlichkeit eine Auffälligkeit mit cerebraler Beteiligung zu diagnostizieren mit größer werdenden Ravenscores steigt, sofern die Scores in Cattell- und Bentontest konstant bleiben.

Tabelle 5-50 Entropiereduktionen bei drei verwendeten Tests (Kriterium: cerebrale Beteiligung)

Kriterium: cerebrale Beteiligung (Stichprobe `parallel-cerebral` - nominal logistisches Modell)					
Prädiktorvariablen-Kombination	**Modell-Test (N=260)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
klassische Prädiktoren					
Score R,C&B	3	13,366	0,0039**	3,71	(2,04)
probabilistische Prädiktoren					
Klassenzug. R,C&B	7	8,982	0,2540	2,49	(0,00)
klassische und probabilistische Prädiktoren (einschließlich Interaktionen)					
Kombiniert	17	29,296	0,0319*	8,13	(0,00)

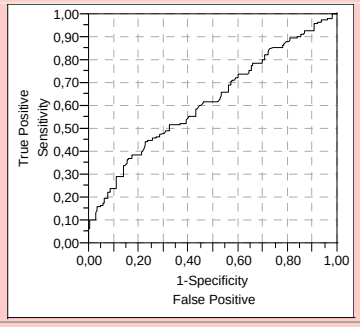
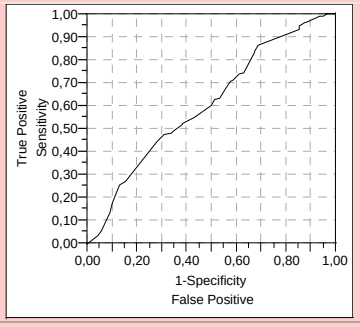
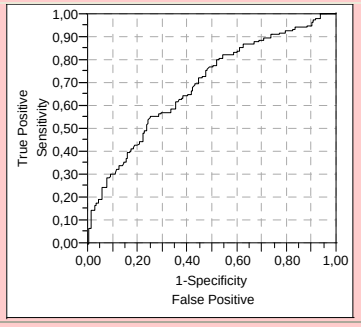
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-51 Wald Effekt Tests bei drei verwendeten Tests (Kriterium: cerebrale Beteiligung)

cerebrale Beteiligung (Stichprobe `parallel-cerebral` - nominal logistisches Modell)				
Prädiktorvariablen-kombination	**Wald Effekt Tests**			
	Variable	FG	Wald- χ^2	Prob> χ^2
klassische Prädiktoren				
Score R,C&B	Ravenscore	1	4,295	0,0382*
	Cattellscore	1	3,122	0,0773
	Bentonscore	1	7,119	0,0076**
probabilistische Prädiktoren				
Klassenzug. R,C&B	RavenMRM3	2	2,651	0,2657
	Cattell LCA4	3	6,179	0,1032
	BentonMRM3	2	0,603	0,7398
klassische und probabilistische Prädiktoren (einschließlich Interaktionen)				
Kombiniert	Ravenscore	1	1,475	0,2246
	RavenMRM3	2	0,827	0,6612
	Interaktion Ravenscore & MRM3	2	3,707	0,1567
	Cattellscore	1	2,341	0,1260
	Cattell LCA4	3	5,201	0,1576
	Interaktion Cattellscore & LCA4	3	4,633	0,2007
	Bentonscore	1	2,583	0,1080
	BentonMRM3	2	0,699	0,7049
	Interaktion Bentonscore & MRM3	2	2,427	0,2972

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-52 ROC-Kurven bei drei verwendeten Tests (Kriterium: cerebrale Beteiligung)

ROC Receiver Operating Curves (cerebrale Beteiligung = positiv)		
R,C & B Scores	R,C & B Klassenzugehörigkeiten	Kombiniert
		
Fläche unter der Kurve: 0,618	Fläche unter der Kurve: 0,604	Fläche unter der Kurve: 0,689

5.5 Zusammenfassung der Kennwerte zur Prädiktion „klinische Auffälligkeit“ und „cerebrale Beteiligung“

Die bislang durchgeführten Modelltests der logistischen Regressionen sowie die auf Grundlage dieser Tests berechneten Prädiktionskennwerte Entropiereduktion und adjustierte Entropiereduktion sind in *Tabelle 5-53* zusammengefasst[58].

Bei den signifikanten Modelltests liefern die vielen Testscoreabstufungen (klassische Prädiktoren) validere Prädiktionen als die wenigen Klassen (probabilistische Prädiktoren). Dieser Befund ist bei anähernd eindimensionalen, psychometrischen Tests zu erwarten.
Die Verwendung der klassisch-probabilistischen Prädiktorenkombination führt jeweils zur sichersten Prädiktion der Kriteriumsvariablen „klinische Auffälligkeit“ oder „cerebrale Beteiligung“. Im Hinblick auf die Ergebnisse von KINDERMANN (2001), die schon in *Abschnitt 1-10* zusammengefasst worden sind, bleiben die Vorteile der Parameterkombination zur Verbesserung der Vorhersagevalidität in der vorliegenden Kreuzvalidierung beim Bentontest und bei den beiden zusätzlich analysierten Tests Raven- und Cattelltest hinter den Erwartungen zurück.

Der Vergleich der Kennwerte weist darauf hin, dass sich die Vorhersagevalidität klinischer Kriterien durch kombinierte Anwendung der drei Tests als Testbatterie im Vergleich zu einzelnen Testdurchführungen kaum verbessern lässt.

[58] In *Tabelle 5-53* ist zu beachten, dass zwei unterschiedlich große Stichproben als Datengrundlage gedient haben, die einen direkten Vergleich der Werte von „klinische Auffälligkeit“ und „cerebraler Beteiligung“ nicht zulassen.

Tabelle 5-53 Zusammenfassung sämtlicher Entropiereduktionswerte

		Prädiktion „klinische Auffälligkeit“ in der „parallelisierten Stichprobe“ (N=468)		**Prädiktion „cerebrale Beteiligung“ in der „Stichprobe `parallel-cerebral`“ (N=269)**	
Raventest					
Prädiktor(en)	Variable	Modelltest Signifikanz	Entropie-reduktion % (adj.%)	Modelltest Signifikanz	Entropie-reduktion % (adj.%)
klassisch	Score	n.s.	0,57 (0,26)	n.s.	0,22 (0,00)
probabilistisch (MRM)	3 Klassen	n.s.	0,29 (0,00)	n.s.	0,62 (0,00)
klass. & prob. & Interaktion	Kombiniert	n.s.	1,51 (0,00)	n.s.	2,19 (0,00)
Bentontest					
Prädiktor(en)	Variable	Modelltest Signifikanz	Entropie-reduktion % (adj.%)	Modelltest Signifikanz	Entropie-reduktion % (adj.%)
klassisch	Score (BTR)	***	2,68 (2,37)	**	2,76 (2,22)
probabilistisch (MRM)	3 Klassen	n.s.	0,22 (0,00)	n.s.	0,18 (0,00)
klass. & prob. & Interaktion	Kombiniert	**	3,06 (1,50)	*	3,24 (0,52)
Fehleranzahl (BTF) im Bentontest					
klassisch	Fehlermaß (BTF)	***	3,31 (3,00)	***	3,08 (2,54)
probabilistisch (MRM)	3 Klassen	n.s.	0,22 (0,00)	n.s.	0,18 (0,00)
klass. & prob. & Interaktion	Kombiniert	***	3,56 (2,00)	*	3,39 (0,67)
Cattelltest					
Prädiktor(en)	Variable	Modelltest Signifikanz	Entropie-reduktion % (adj.%)	Modelltest Signifikanz	Entropie-reduktion % (adj.%)
klassisch	Score	***	2,42 (2,11)	**	1,96 (1,41)
probabilistisch (LCA)	4 Klassen	**	2,28 (1,35)	n.s.	1,92 (0,28)
klass. & prob. & Interaktion	Kombiniert	**	3,05 (0,86)	*	3,95 (0,14)
Drei Tests gemeinsam					
Prädiktoren	Variable	Modelltest Signifikanz	Entropie-reduktion % (adj.%)	Modelltest Signifikanz	Entropie-reduktion % (adj.%)
klassisch	Scores	***	3,41(2,46)	**	3,71(2,04)
probabilistisch	Klassen	*	2,29 (0,06)	n.s.	2,49 (0,00)
klass. & prob. & Interaktion	Kombiniert	**	5,58 (0,19)	*	8,13 (0,00)

LCA= Latente-Klassen-Analyse; MRM= Mixed-Rasch-Modell; *p≤ 0,05;**p≤ 0,01;***p≤ 0,001

In Abhängigkeit von den ermittelten Entropiereduktionen mit Testscores und Klassenzugehörigkeiten ergeben sich für die Prädiktionen die in den *Tabellen 5-54* und *5-55* angegebenen Rangfolgen der verwendeten Tests:

Tabelle 5-54 Entropiereduktionen geordnet nach Tests (klinische Auffälligkeit)

„größte“ Entropiereduktion	„mittlere“ Entropiereduktion	„kleinste“ Entropiereduktion	**Prädiktion von „klinische Auffälligkeit“**	
Benton	Cattell	Raven	mit Score	Klassische Prädiktorvariablen
2,68%	2,42%	0,57%		
Cattell (LCA)	Raven (MRM)	Benton (MRM)	mit Klassen	Probabilistische Prädiktorvariablen
2,28%	0,29%	0,22%		
Benton	Cattell	Raven	mit Score und Klassen einschl. Interaktion	Prädiktorvariablen-kombination
3,06%	3,05%	1,51%		

Tabelle 5-55 Entropiereduktionen geordnet nach Tests (cerebrale Beteiligung)

„größte“ Entropiereduktion	„mittlere“ Entropiereduktion	„kleinste“ Entropiereduktion	**Prädiktion von „cerebrale Beteiligung“**	
Benton	Cattell	Raven	mit Score	Klassische Prädiktorvariablen
2,76%	1,96%	0,22%		
Cattell (LCA)	Raven (MRM)	Benton (MRM)	mit Klassen	Probabilistische Prädiktorvariablen
1,92%	0,62%	0,18%		
Cattell	Benton	Raven	mit Score und Klassen einschl. Interaktion	Prädiktorvariablen-kombination
3,95%	3,24%	2,19%		

LCA= Latente-Klassen-Analyse; MRM= Mixed-Rasch-Modell

Von den drei untersuchten klassischen Prädiktorvariablen erweist sich der Bentonscore als das sicherste Maß für die Prädiktion „klinische Auffälligkeit“. Für den Cattellscore ergeben sich hier geringfügig kleinere Prädiktionsgrößen.

Die Überlegenheit des Bentonscores über den Cattellscore wird insbesondere bei der Prädiktion „cerebrale Beteiligung“ deutlich. Der Bentontest hat hier die größere diagnostische Aussagekraft und stellt damit seine Qualitäten als „Hirnschadenstest“ unter Beweis.

Für den Ravenscore werden weder signifikante noch relevante Zusammenhänge mit klinischen Auffälligkeiten aufgezeigt. Er erfasst somit keine für pathologische Prozesse anfällige Fähigkeit, wenn er - wie in dieser Arbeit geschehen - ohne Zeitbegrenzung vorgelegt wird.

Die größte Entropiereduktion der probabilistischen Prädiktoren zeigt die mit Latenter-Klassen-Analyse berechnete Klassen-Lösung beim Cattelltest. Es gilt zu beachten, dass die mittels LCA identifizierten Klassen nicht direkt mit den Klassen der Mixed-Rasch-Modelle bei Raven- und Bentontest verglichen werden dürfen, weil der Zusammenhang der Klassen mit den Testscores anders ist. Im Unterschied zu den Mixed-Rasch-Modellen gilt, dass sich mit der Latenten-Klassen-Analyse dann eine stärkere Trennung der Klassen nach Scores ergibt, wenn mit den Testscores an sich schon diagnostisch relevante Aussagen möglich sind.
Die Prädiktion mit dem probabilistischen Prädiktor gelingt daher in ähnlicher Größenordnung wie mit dem Cattellscore. Schon im Theorieteil (*Abschnitt 3.2)* wurde darauf hingewiesen, dass die ähnlichen Werte hier - beim Speed-Power-Test Cattell - zu erwarten sind.

Die Klassen der Mixed-Rasch-Modelle liefern zur Prädiktion „klinische Auffälligkeit“ oder „cerebrale Beteiligung“ keine relevanten Beiträge. Die vom Mixed-Rasch-Modell identifizierten zusätzlichen Informationen resultieren somit bei Raven- und Bentontest aus zahlreichen, aufgrund der vorliegenden Daten nicht näher identifizierbaren Einflüssen. Sollte es sich hierbei um reine Störinformationen handeln, ist die annähernde Gültigkeit des Summenscores der beiden Power-Tests als Indikator für testspezifische Fähigkeiten zusätzlich belegt. Solche Störinformationen sind in einer Testsituation nie ganz auszuschließen.

Mit dem zweiten Testmodell, der Latenten-Klassen-Analyse, lassen sich analog auch die deutlich höheren Entropiereduktionswerte beim Cattelltest, unter Verwendung der klassisch-probabilistischen Prädiktorenkombination erklären.
Ein gültiger Vergleich mit Raven- und Bentontest, die mit Mixed-Rasch-Modellen analysiert wurden, ist auch hier wegen der unterschiedlichen zugrunde liegenden Modellannahmen nicht möglich.

5.6 Latente-Klassen-Analyse der gesamten Testbatterie

In den vorausgegangenen *Abschnitten* wurde gezeigt, dass sich verbesserte Prädiktionen von „klinische Auffälligkeit“ und „cerebrale Beteiligung“ erzielen lassen, wenn die Ergebnisse von Raven-, Benton- und Cattelltest als Testbatterie genutzt werden. Die berechneten Verbesserungen der Vorhersage sind jedoch kaum von klinischer Relevanz, weil die Tests hochkorrelierte, also ähnliche Fähigkeiten quantifizieren.

Probabilistische Methoden bieten indes andere Ansatzpunkte, mit denen in den Bearbeitungen der Testbatterie nach diagnostisch verwertbarer Information gesucht werden kann.

In den folgenden *Abschnitten* wird mittels Latenter-Klassen-Analyse den Fragen nachgegangen,

1.) ob sich in den Bearbeitungen der Testbatterie typische Profilverläufe, also qualitativ unterschiedliche Lösungsmuster ergeben und wie diese interpretiert werden können?

2.) ob und in welchem Maße die als latente Klassen identifizierten Profilverläufe diagnostische Relevanz haben?

3.) ob sich die identifizierten Zusammenhänge beim „klassischen Auswertungsmodus“ von denen des „probabilistischen Auswertungsmodus“ unterscheiden?

Für die Latenten-Klassen-Analysen werden beim **„klassischen Auswertungsmodus“** in *Abschnitt 5.6.1* die herkömmlichen Summenscores der Testpersonen als Ausgangsdaten verwendet.

Beim **„probabilistischen Auswertungsmodus“** in *Abschnitt 5.6.2* werden zur Brechnung zusätzlich zu den klassenspezifischen Summenscores auch die Klassenzugehörigkeiten der Testpersonen genutzt.

Die Analyse dieser Auswertungsmodi wird mit der Teilstichprobe „Testbatterie“ (N=590) durchgeführt. Alle darin enthaltenen Personen haben die gesamte Testbatterie, bestehend aus Raven-, Cattell- und Bentontest, bearbeitet.

Diesem Vorgehen liegt die Annahme zugrunde, dass Testleistungen unabhängig von der vorgelegten Testreihenfolge konstant bleiben. Derartige Ermüdungs- und Motivationseffekte sind, wie in *Abschnitt 4.7.2* statistisch gezeigt, nicht nachweisbar.

5.6.1 Klassischer Auswertungsmodus der Testbatterie

5.6.1.1 Identifikation latenter Klassen

Der klassische Auswertungsmodus verwendet als Ausgangsdaten die Testleistungen in Form der Summenscores. Diese werden über eine Quartilszuteilung umgewandelt, weil für Latente-Klassen-Analysen kategoriale Daten erforderlich sind.

Im klassischen Auswertungsmodus ergeben sich für die Testscores die in *Tabelle 5-56* dargestellten Quartile:

Tabelle 5-56 Klassischer Auswertungsmodus: Testscorequartile

Klassischer Auswertungs-modus	1. Quartil	2. Quartil	3. Quartil	4. Quartil
Ravenscore	1-9	10-13	14-16	17-20
Cattellscore	0-4	5	6	7-10
Bentonscore	1-8	9-12	13-15	16-20

Die Berechnungen der Latenten-Klassen-Analysen des klassischen Auswertungsmodus erfolgen auf der Grundlage der in *Tabelle 5-57* dargestellten vierkategoriellen Variablen.

Tabelle 5-57 Klassischer Auswertungsmodus: Variablenkategorien

Klassischer Auswertungsmodus	1.Variable	2.Variable	3.Variable
	Ravenscore Quartil	**Cattellscore Quartil**	**Bentonscore Quartil**
Kategorie 1	1.	1.	1.
Kategorie 2	2.	2.	2.
Kategorie 3	3.	3.	3.
Kategorie 4	4.	4.	4.

Die Ergebnisse der durchgeführten Latenten-Klassen-Analysen des klassischen Auswertungsmodus stellen sich folgendermaßen dar (vgl. *Tabelle 5-58*):

Tabelle 5-58 Klassischer Auswertungsmodus: Informationstheoretische Maße

N=590	loglikeli	AIC	BIC	n_p
klassisch LCA1Kl	-2412,61	4843,21	4882,64	9
klassisch LCA2Kl	-2203,11	4444,22	4527,44	19
klassisch LCA3Kl	-2167,01	4392,02	4519,04	29
klassisch LCA4Kl	-2161,49	4400,98	4571,80	39
klassisch LCA5Kl	-2158,93	4415,85	4630,48	49
saturiertes Modell	-2150,96	4427,92	4703,87	63

Startwert 4321
abgesichert mit
5321 und 321

Der AIC (4392,02) und auch der BIC-Wert (4519,04) indizieren eine 3-Klassen-Struktur.

5.6.1.1.1 Inferenzstatistische Prüfung

Die 3-Klassen-Lösung wird durch die inferenzstatistische Prüfung mittels Likelihoodquotiententests (vgl. *Tabelle 5-59*) abgesichert.

Tabelle 5-59 Klassischer Auswertungsmodus: Likelihoodquotiententests

LR	-2ln(LR)	FG	Prob> χ^2
Klas.LCA3Kl/sat.Modell	32,1	34	0,561
Klas.LCA1Kl/Klas.LCA3Kl	491,2	20	<,001***

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die Ergebnisse der beiden gebildeten Likelihoodquotienten zeigen an, dass die vom AIC und BIC indizierte 3-Klassen-Lösung der Latenten-Klassen-Analyse mit den Daten vereinbar ist. Diese passt besser als das Einklassenmodell und nicht schlechter als das saturierte Modell. Die 3-Klassen-Lösung wird somit den weiteren Betrachtungen als geeignetes Modell zugrunde gelegt.

5.6.1.2 Interpretation der indizierten Klassen

In diesem *Abschnitt* werden die Testleistungen und die im Fragebogen erhobenen Merkmale der Testpersonen genutzt, um die indizierten drei latenten Klassen voneinander abzugrenzen.

5.6.1.2.1 Charakterisierung der indizierten Klassen

Die Klassen werden entsprechend ihrer Leistungen in der Testbatterie benannt und in der *Tabelle 5-60* unter Angabe von Klassengrößen, Zuordnungssicherheiten, Testleistungen und Altersdurchschnitten angeordnet.
Die zugefügten Balkendiagramme zeigen die klassenspezifische Zusammensetzung der Altersgruppen und Bildungsgrade.
Die Klassenzuordnung der Personen erfolgt anhand der vom Modell vorgegebenen größten Zuordnungswahrscheinlichkeit.
Die gemittelten klassenspezifischen Testscores werden in *Abbildung 5-12* dargestellt.

Sämtliche Klassenunterschiede der *Tabelle 5-60* sind statistisch hochsignifikant und bedeuten eine Bildungs- und Altersabhängigkeit der Leistungen in der Testbatterie. Die Zusammenfassung der Kontraste ist der *Tabelle 5-61* und ausführlich der *Tabelle A-41* im Anhang zu entnehmen.

Tabelle 5-60 Klassischer Auswertungsmodus: Die 3-Klassen-Lösung der LCA

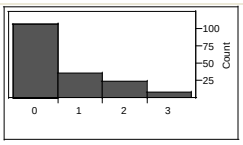
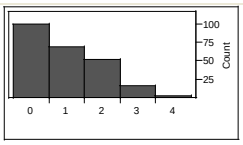
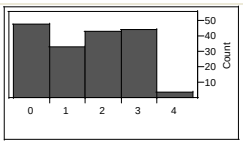
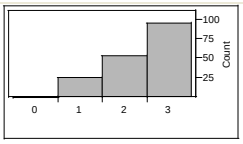
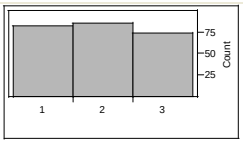
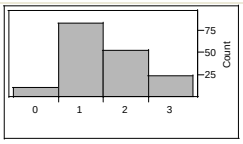
Teilstichprobe „Testbatterie“ (N=590)	hoch ← Testleistungsfähigkeit → niedrig		
Klassischer Auswertungsmodus	**Klasse 3** **„Hohes Leistungsniveau“**	**Klasse 1** **„Mittleres Leistungsniveau“**	**Klasse 2** **„Niedriges Leistungsniveau“**
Vgl. *Abschnitt*	*5.6.1.2.1.1 a)*	*5.6.1.2.1.1 b)*	*5.6.1.2.1.1 c)*
Klassengröße N (%Anteil gesamt)	175 (30%)	243 (41%)	172 (29%)
mittlere klassenspezifische Zuordnungssicherheit	0,873	0,835	0,934
Altersmittelwerte in den Klassen			
Alter [Ø]	35,55	41,67	49,17
Alter [σ]	13,37	13,43	14,05
Altersgruppen in den Klassen (ungleicher Maßstab)			
Altersgruppe 0=15-39LJ Altersgruppe 1=39-49LJ Altersgruppe 2=49-59LJ Altersgruppe 3=59-70LJ Altersgruppe 4= >70LJ			
Bildungsgrade in den Klassen (ungleicher Maßstab)			
Bildungsgrad 0=kein Abschluss Bildungsgrad 1=Hauptschule Bildungsgrad 2=Realschule Bildungsgrad 3=Abitur			
Klassenspezifische Leistungen im Raventest			
Ravenscore [Ø]	16,74	12,86	7,19
Ravenscore [σ]	1,58	1,91	2,64
Raven-Leistungsreihenfolge *(lt. Kontrasten Tab.5-61)*	***1.***	***2.***	***3.***
Klassenspezifische Leistungen im Cattelltest			
Cattellscore [Ø]	6,42	4,94	3,27
Cattellscore [σ]	1,06	1,35	1,74
Cattell-Leistungsreihenfolge *(lt. Kontrasten Tab.5-61)*	***1.***	***2.***	***3.***
bearbeitete Cattell-Items [Ø]	8,90	8,58	7,94
bearbeitete Cattell-Items [σ]	1,94	2,18	2,58
Klassenspezifische Leistungen im Bentontest			
Bentonscore [Ø]	15,62	12,02	7,29
Bentonscore [σ]	2,43	2,85	3,58
Benton-Leistungsreihenfolge *(lt. Kontrasten Tab.5-61)*	***1.***	***2.***	***3.***
Bentonfehler (BTF) [Ø]	***5,4***	***11,02***	***21,05***
Bentonfehler (BTF) [σ]	***3,40***	***5,10***	***8,99***

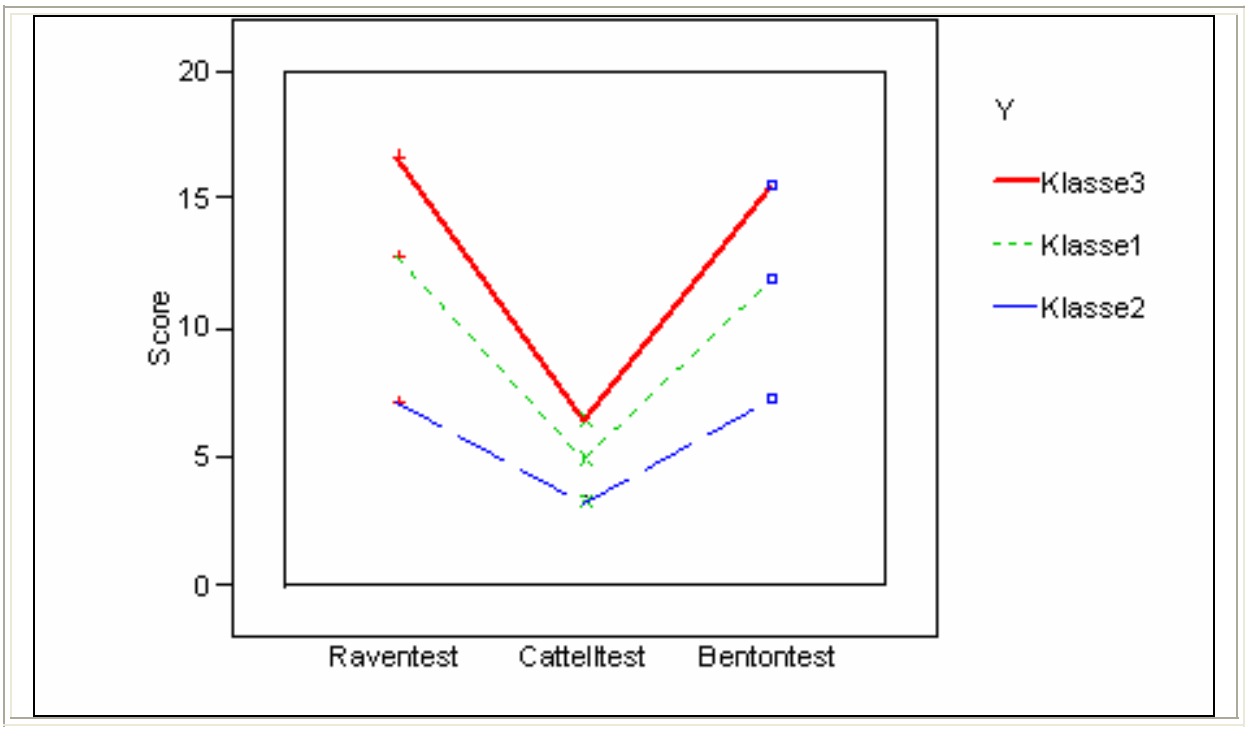
Tabelle 5-61 Klassischer Auswertungsmodus: Zusammenfassung der Kontraste

Kontraste des Alters	Hohes Leistungs-niveau	Mittleres Leistungs-niveau	Niedriges Leistungs-niveau
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	
Kontraste der Altersgruppen	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	
Kontraste der Bildungsgrade	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	
Kontraste der Ravenscores	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	
Kontraste der Cattellscores	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	
Kontraste der bearbeiteten Cattell-Items	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		ns	***
Mittl. Leistungsniveau	ns		*
Niedr. Leistungsniveau	***	*	
Kontraste der Bentonscores	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	
Kontraste der Bentonfehler (BTF)	**Hohes Leistungs-niveau**	**Mittleres Leistungs-niveau**	**Niedriges Leistungs-niveau**
Hohes Leistungsniveau		***	***
Mittl. Leistungsniveau	***		***
Niedr. Leistungsniveau	***	***	

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)

entnommen aus *Tab.A-41*

Abbildung 5-12 Klassischer Auswertungsmodus: klassenspezifische Durchschnittsscores



Die Darstellung der Scoremittelwerte aus *Tabelle 5-60* in *Abbildung 5-12* zeigt, dass sich klar geordnete - also überschneidungsfreie - Klassen ergeben. Im klassischen Auswertungsmodus kann das Lösungsverhalten von Versuchspersonen demnach als Ausprägung einer einzigen latenten Variable entlang des eindimensionalen Konstruktes „Testleistungsfähigkeit" beschrieben werden.

Es ist davon auszugehen, dass dieses Resultat aus den hochkorrelierten Fähigkeiten der drei verwendeten Tests folgt (vgl. auch *Tabelle 4-17*).

5.6.1.2.1.1 Merkmalsprofile der indizierten Klassen

In diesem *Abschnitt* werden die Merkmalsprofile der drei latenten Klassen des klassischen Auswertungsmodus graphisch dargestellt und erläutert.

Die Variablenkategorien in diesen Darstellungen werden von der Einteilung in *Tabelle 5-57* übernommen. Diese sind in *Tabelle 5-62* nochmals zusammengefasst.

Die Anteile klinisch unauffälliger Personen, sowie von Patienten **mit** und **ohne** cerebraler Beteiligung in den klassenspezifischen Variablenkategorien geben die *Abbildungen 5-13* bis *5-15* als Balkendiagramme wieder.

Die klassenspezifischen Verteilungen der Altersgruppen und Bildungsgrade sind in den *Tabellen 5-63* bis *5-65* aufgeführt.

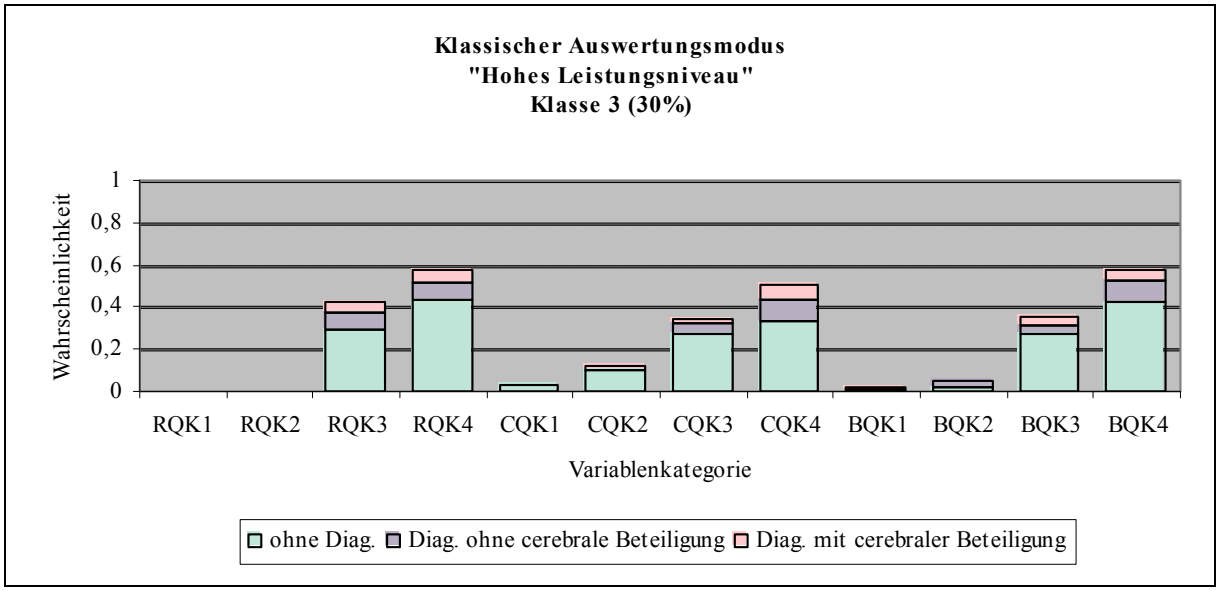
Tabelle 5-62 Klassischer Auswertungsmodus: Abkürzungen der Variablenkategorien

Raventest	
1. Testscorequartil im Raventest	RQK1
2. Testscorequartil im Raventest	RQK2
3. Testscorequartil im Raventest	RQK3
4. Testscorequartil im Raventest	RQK4
Cattelltest	
1. Testscorequartil im Cattelltest	CQK1
2. Testscorequartil im Cattelltest	CQK2
3. Testscorequartil im Cattelltest	CQK3
4. Testscorequartil im Cattelltest	CQK4
Bentontest	
1. Testscorequartil im Bentontest	BQK1
2. Testscorequartil im Bentontest	BQK2
3. Testscorequartil im Bentontest	BQK3
4. Testscorequartil im Bentontest	BQK4

Beispiel: RQK1 = **R**avenscore **Q**uartil **K**ategorie **1**

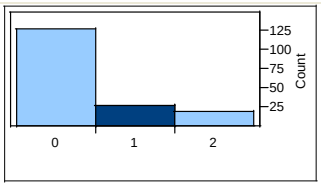
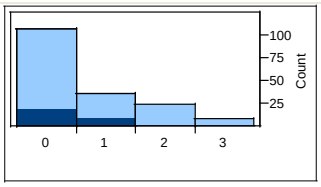
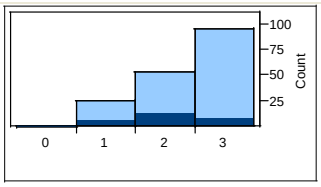
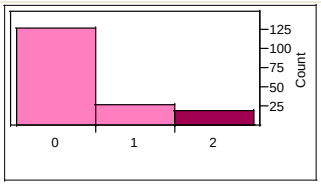
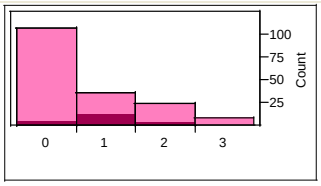
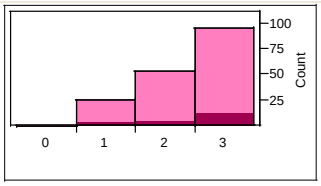
a) Die Klasse 3 „Hohes Leistungsniveau“

Abbildung 5-13 Merkmalsprofil: Klasse 3 „Hohes Leistungsniveau“



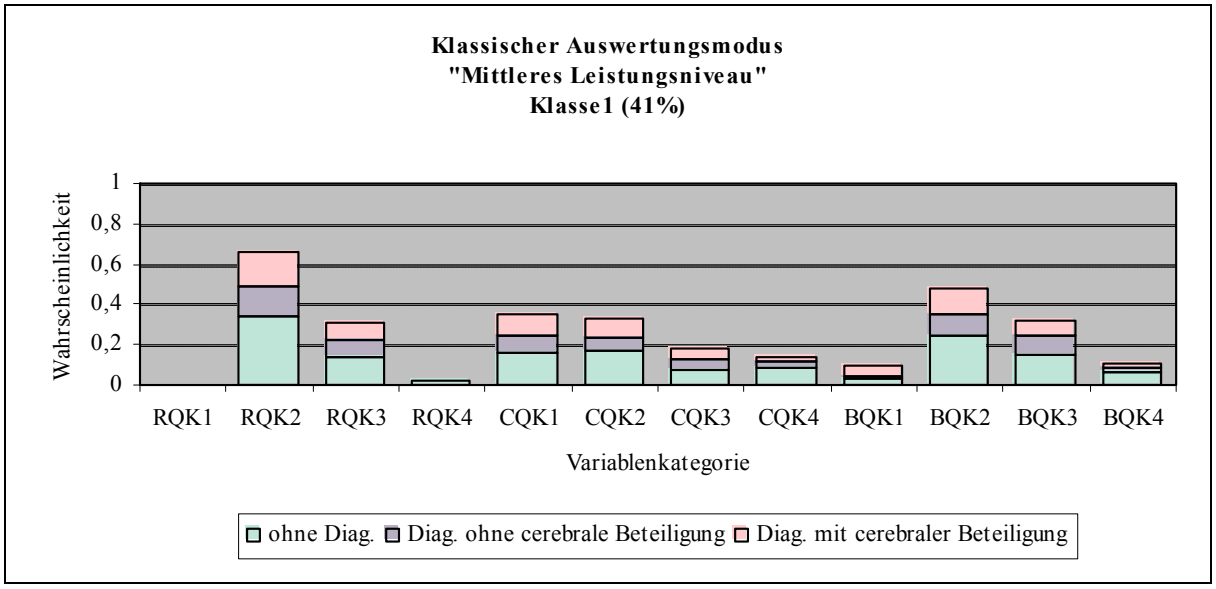
Die zweitgrößte Klasse 3 beinhaltet die jüngsten Personen und höchsten Bildungsgrade. In den drei Tests werden hohe Scores (vor allem 3. und 4. Quartil) erreicht.

Tabelle 5-63 Alters- und Bildungsverteilung in Klasse 3 „Hohes Leistungsniveau“

Klasse 3	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70LJ	0=kein Abschluss 1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

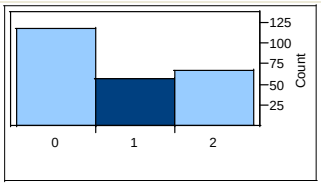
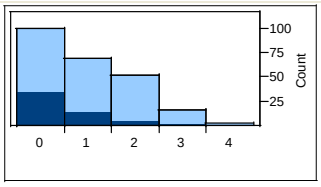
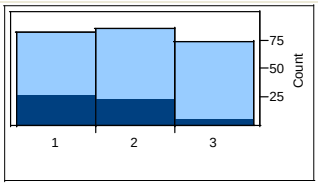
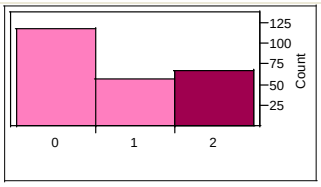
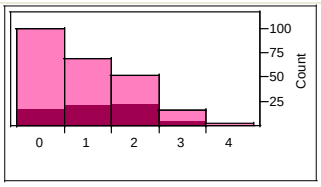
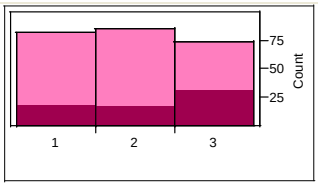
b) Die Klasse 1 „Mittleres Leistungsniveau“

Abbildung 5-14 Merkmalsprofil: Klasse 1 „Mittleres Leistungsniveau“



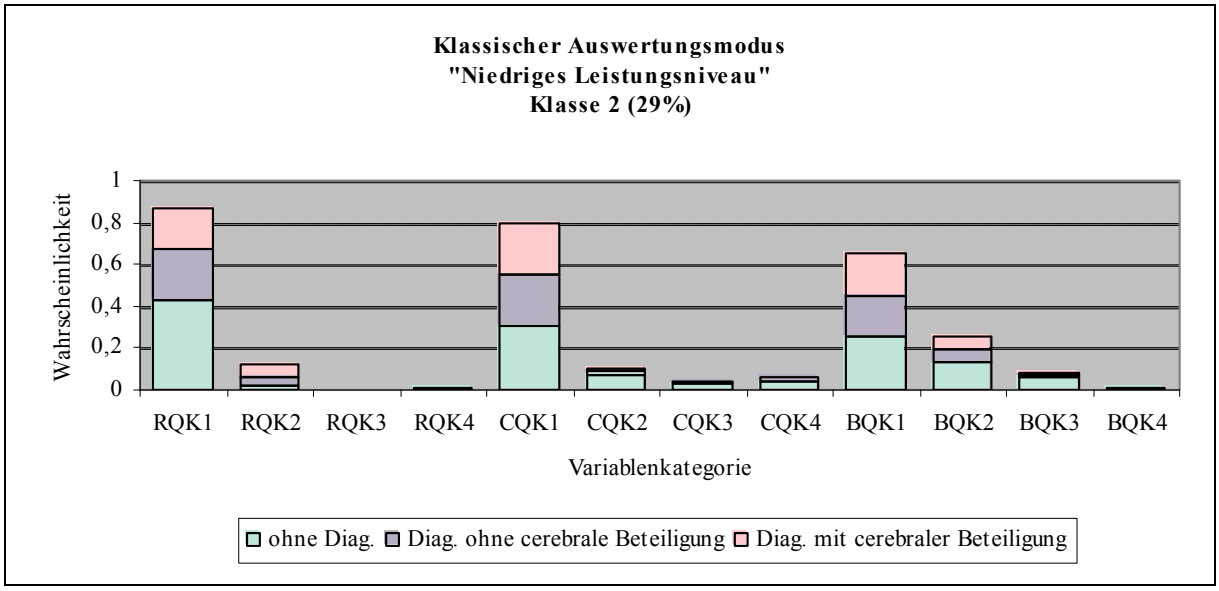
In der größten Klasse 1 „Mittleres Leistungsniveau“ befinden sich überwiegend Personen mittleren Alters und mittleren Bildungsgrades. Die Klasse erreicht Scores, die um das 2. Quartil streuen.

Tabelle 5-64 Alters- und Bildungsverteilung in Klasse 1 „Mittleres Leistungsniveau“

Klasse 1	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70LJ 4=>70LJ	1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

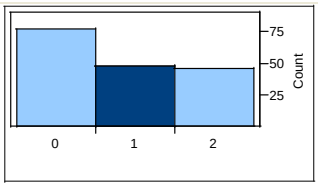
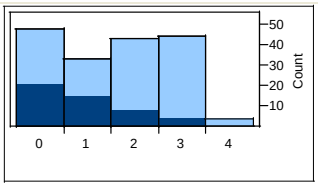
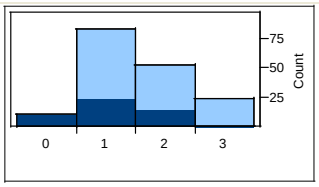
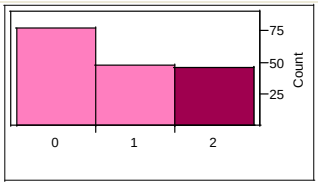
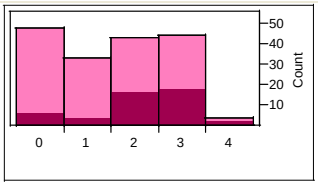
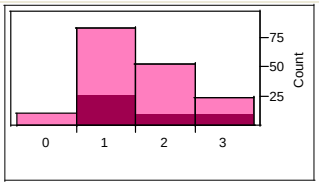
c) Die Klasse 2 „Niedriges Leistungsniveau“

Abbildung 5-15 Merkmalsprofil: Klasse 2 „Niedriges Leistungsniveau“



In der kleinsten Klasse 2 „Niedriges Leistungsniveau“ sind bei schwachen Testleistungen (häufig 1. Quartil) überwiegend ältere, sowie Personen mit niedrigerer Bildung vertreten. Auf diagnostische Implikationen dieser Klasse wird in *Abschnitt 5.6.1.3.1* eingegangen.

Tabelle 5-65 Alters- und Bildungsverteilung in Klasse 2 „Niedriges Leistungsniveau“

Klasse 2	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70LJ 4=>70LJ	0=kein Abschluss 1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

5.6.1.3 Diagnostische Relevanz der latenten Klassen

Die drei latenten Klassen der Testbatterie werden in die beinhalteten Diagnosegruppen „ohne Diagnose“, „Diagnose **ohne** cerebrale Beteiligung“ und „Diagnose **mit** cerebraler Beteiligung“ unterteilt. Um zusätzliche diagnostisch verwertbare Information zu erfassen, wird jede latente Klasse separat analysiert. Die Scoreunterschiede der Diagnosegruppen in den Klassen werden in *Abschnitt 5.6.1.3.1* betrachtet. Die Klassenzusammensetzungen werden in *Abschnitt 5.6.1.3.2* aufgezeigt und die Validität der 3-Klassen-Lösung des klassischen Auswertungsmodus wird in *Abschnitt 5.6.1.3.3* untersucht.

5.6.1.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen

Die Scoreunterschiede der drei Diagnosegruppen[59] werden klassenspezifisch mit Varianzanalysen untersucht. Signifikante Ergebnisse werden mit Kontrasten weitergehend ausgewertet. Die resultierenden Klassenzusammensetzungen finden sich in *Tabelle 5-66*, die Varianzanalysen der Scores in den *Tabellen 5-67* und *5-68*. Die Einzelvergleiche wurden in *Tabelle 5-69* zusammengefasst, die ausführlichen Statistiken sind *Tabelle A-42* im Anhang zu entnehmen.

Tabelle 5-66 Klassischer Auswertungsmodus: Klassenzusammensetzung

Klassenspezifische Betrachtung der diagnostischen Gruppen									
	hoch ← Testleistungsfähigkeit → niedrig								
Teilstichprobe „Testbatterie“ (N=590)	**Klasse 3 „Hohes Leistungsniveau“**			**Klasse 1 „Mittleres Leistungsniveau“**			**Klasse 2 „Niedriges Leistungsniveau“**		
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
N	128 (73%)	28 (16%)	19 (11%)	119 (49%)	57 (23%)	67 (28%)	78 (45%)	48 (28%)	46 (27%)
Alter [Ø]	34,68	34,00	43,68	41,20	37,11	46,39	49,12	42,56	56,17
Alter [σ_n]	1,16	2,48	3,01	1,20	1,73	1,60	1,49	1,90	1,94

σ_n = Standardfehler

[59] klU= ohne Diagnose, Diag.o.cB.= Diagnose **ohne** cerebrale Beteiligung & Diag.m.cB.= Diagnose **mit** cerebraler Beteiligung

Tabelle 5-67 Klassischer Auswertungsmodus: Varianzanalysen der Raven- und Cattell-Scores

Klassenspezifische Betrachtung der diagnostischen Gruppen									
	hoch ← Testleistungsfähigkeit → niedrig								
Teilstichprobe „Testbatterie“ (N=590)	**Klasse 3 „Hohes Leistungsniveau“**			**Klasse 1 „Mittleres Leistungsniveau“**			**Klasse 2 „Niedriges Leistungsniveau“**		
Varianzanalysen der Raven-Scores innerhalb der einzelnen Klassen									
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
Ravenscore [Ø]	16,80	16,71	16,42	12,76	12,93	12,97	7,19	6,69	7,70
Ravenscore [σ_n]	0,14	0,30	0,36	0,18	0,25	0,23	0,30	0,38	0,39
	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total
FG	2	172	174	2	240	242	2	169	171
QS	2,363	431,06	433,43	2,37	879,59	881,96	23,88	1168,17	1192,05
$\hat{\sigma}^2$	1,18	2,51		1,18	3,66		11,94	6,91	
F	0,47			0,32			1,73		
Prob > F	0,625			0,724			0,1809		
Varianzanalysen der Cattell-Scores innerhalb der einzelnen Klassen									
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
Cattellscore [Ø]	6,36	6,75	6,37	5,08	4,96	4,67	3,73	3,06	2,70
Cattellscore [σ_n]	0,09	0,20	0,24	0,12	0,18	0,16	0,19	0,24	0,25
	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total
FG	2	172	174	2	240	242	2	169	171
QS	3,57	193,14	196,71	7,33	435,87	443,19	33,80	481,90	515,70
$\hat{\sigma}^2$	1,78	1,12		3,66	1,82		16,90	2,85	
F	1,59			2,02			5,93		
Prob > F	0,2071			0,1352			0,0033 **		

klU = ohne Diagnose, Diag.o.cB. = Diagnose **ohne** cerebrale Beteiligung
und Diag.m.cB. = Diagnose **mit** cerebraler Beteiligung;
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001
σ_n = Standardfehler

Tabelle 5-68 Klassischer Auswertungsmodus: Varianzanalysen der Benton-Scores

Klassenspezifische Betrachtung der diagnostischen Gruppen									
	hoch ← Testleistungsfähigkeit → niedrig								
Teilstichprobe „Testbatterie“ (N=590)	**Klasse 3 „Hohes Leistungsniveau“**			**Klasse 1 „Mittleres Leistungsniveau“**			**Klasse 2 „Niedriges Leistungsniveau“**		
Varianzanalysen der Benton-Scores innerhalb der einzelnen Klassen									
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
Bentonscore [Ø]	15,74	15,39	15,10	12,26	12,23	11,31	8,49	6,35	6,24
Bentonscore [σ_n]	0,21	0,46	0,56	0,26	0,37	0,35	0,39	0,49	0,51
	Gru-ppe	Fehler	Total	Gru-ppe	Fehler	Total	Gru-ppe	Fehler	Total
FG	2	172	174	2	240	242	2	169	171
QS	8,39	1014,96	1023,35	45,93	1918,00	1963,93	204,63	1984,84	2189,47
$\hat{\sigma}^2$	4,19	5,90		22,96	7,99		102,32	11,75	
F	0,71			2,87			8,71		
Prob > F	0,4927			0,0585			0,0003 ***		

klU = ohne Diagnose, Diag.o.cB. = Diagnose **ohne** cerebrale Beteiligung
und Diag.m.cB. = Diagnose **mit** cerebraler Beteiligung;
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001
σ_n = Standardfehler

Tabelle 5-69 Klassischer Auswertungsmodus: Zusammenfassung der Kontraste in Klasse 2 „Niedriges Leistungsniveau“

Kontraste der Cattellscores	klU	Diag.o.cB	Diag.m.cB
klU		ns	**
Diag.o.cB	ns		ns
Diag.m.cB	**	ns	
Kontraste der Bentonscores	**klU**	**Diag.o.cB**	**Diag.m.cB**
klU		**	**
Diag.o.cB	**		ns
Diag.m.cB	**	ns	
Kontraste des Lebensalters	**klU**	**Diag.o.cB**	**Diag.m.cB**
klU		*	*
Diag.o.cB	*		***
Diag.m.cB	*	***	
Kontraste der Bildungsgrade	**klU**	**Diag.o.cB**	**Diag.m.cB**
klU		*	ns
Diag.o.cB	*		ns
Diag.m.cB	ns	ns	

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)
klU = ohne Diagnose, Diag.o.cB. = Diagnose ohne cerebrale Beteiligung
und Diag.m.cB. = Diagnose mit cerebraler Beteiligung
entnommen aus *Tab.A-42*

Erläuterung der *Tabellen 5-66* bis *5-69*:

In den Varianzanalysen ist die **Klasse 2 „Niedriges Leistungsniveau“** die einzige Klasse mit signifikanten Gruppenunterschieden der Testleistungen. Unterschiede ergeben sich bei den Leistungsmaßen Cattellscore (*Tabelle 5-67)* und Bentonscore (*Tabelle 5-68)*, nicht jedoch beim Ravenscore (ebenfalls *Tabelle 5-67)*.

Den Einzelvergleichen in *Tabelle 5-69* ist zu entnehmen, dass der **Cattellscore** erfolgreich zwischen klinischer Unauffälligkeit und den Diagnosen **mit** cerebraler Beteiligung differenziert und der **Bentonscore** eine Unterscheidung beider klinischer Gruppen von klinischer Unauffälligkeit ermöglicht.

Die Aussagekraft dieser Unterschiede bleibt jedoch fraglich, weil gleichzeitig signifikante Alters- und Bildungsunterschiede der Diagnosegruppen aufgezeigt werden (vgl. *Tabelle 5-69)*.

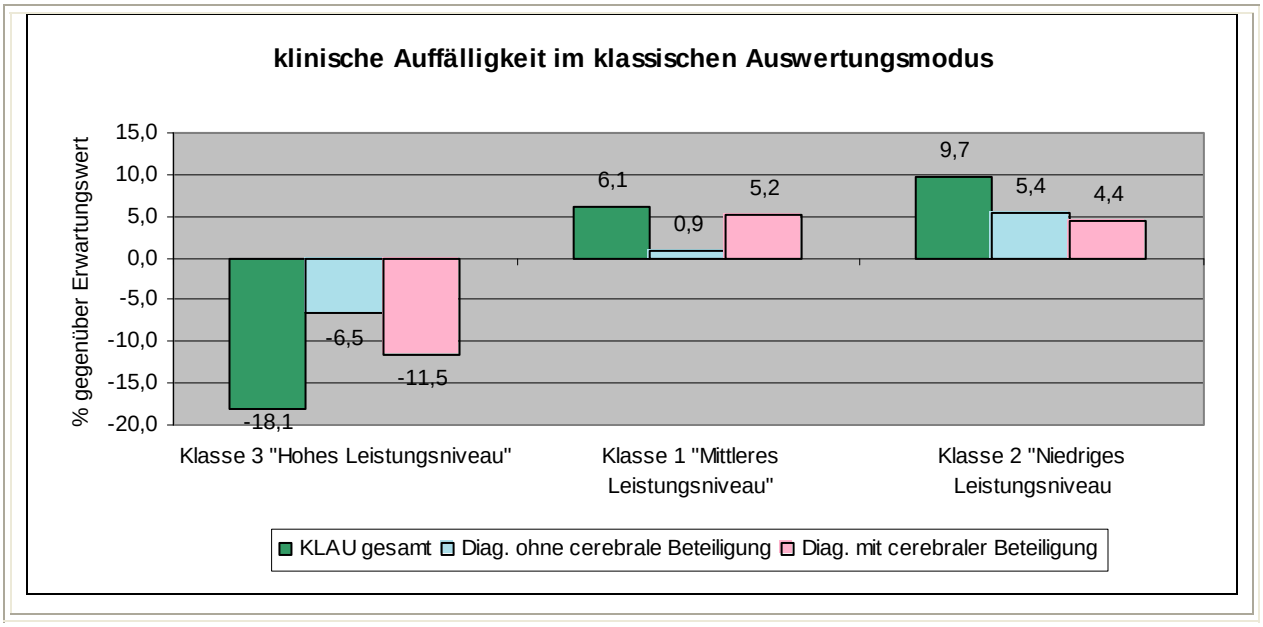
5.6.1.3.2 Vergleich der Klassenzusammensetzungen

Die Verteilungen der Diagnosegruppen und der Krankheitsbilder in den latenten Klassen werden in den beiden folgenden *Abschnitten* dargestellt.

5.6.1.3.2.1 Diagnostische Gruppen

Die prozentualen Abweichungen vom klassenspezifischen Erwartungswert nach Diagnosegruppen[60] sind in *Abbildung 5-16* dargestellt. Die zugehörigen Zellbesetzungen sind der *Tabelle 5-70* zu entnehmen.

Abbildung 5-16 Klassischer Auswertungsmodus: Diagnosegruppen innerhalb der Klassen



KLAU gesamt = Diagnose **ohne** cerebrale Beteiligung + Diagnose **mit** cerebraler Beteiligung

Erläuterung der *Abbildung 5-16*:

Der Anteil klinisch auffälliger Personen nimmt proportional zu abnehmender Testleistung zu. Der Klasse 3 „hohes Leistungsniveau“ ist eine geringere Anzahl klinisch auffälliger Personen zugeordnet als vom Erwartungswert prognostiziert.

Klinisch auffällige Personen finden sich leicht gehäuft in den leistungsschwächeren Klassen 1 „mittleres Leistungsniveau“ und 2 „niedriges Leistungsniveau“.

[60] klinisch auffällig allgemein (KLAU), Diagnose **mit** und **ohne** cerebrale Beteiligung

Tabelle 5-70 Klassischer Auswertungsmodus: Zellbesetzungen der Diagnosegruppen

Teilstichprobe „Testbatterie“ (N=590)	**Klasse 3 „Hohes Leistungsniveau“**		**Klasse 1 „Mittleres Leistungsniveau“**		**Klasse 2 „Niedriges Leistungsniveau“**		(Klassen geordnet nach Testleistungs fähigkeit)	
Diagnose-Gruppe	N	+/-%	N	+/-%	N	+/-%	**Σ**	Erw.%
Diagnose ohne cerebrale Beteiligung	28	-6,5	57	+0,9	48	+5,4	**133**	22,5
Diagnose mit cerebraler Beteiligung	19	-11,5	67	+5,2	46	+4,4	**132**	22,4
klinisch auffällige Personen insgesamt (KLAU) **Σ**	47	-18,1	124	+6,1	94	+9,7	265	44,9
klinisch unauffällige Personen	128	+18,1	119	-6,1	78	-9,7	325	55,1
Σ	**175**		**243**		**172**		**590**	100

5.6.1.3.2.2 Krankheitsbilder

Wird die klassenspezifische Verteilung der Krankheitsbilder aus *Tabelle 5-71* betrachtet, so ergibt sich, analog zu *Abbildung 5-16,* die *Abbildung 5-17*. Diesem Schaubild können keine zusätzlichen differentialdiagnostischen Informationen entnommen werden.

Die Beträge auf der Ordinate entsprechen den Abweichungen vom klassenspezifischen Erwartungswert der Häufigkeit jedes einzelnen Krankheitsbildes. Einzelne Zellen sind jedoch gering besetzt, weshalb die *Abbildung* nur unter Vorbehalt interpretiert werden kann.

Abbildung 5-17 Klassischer Auswertungsmodus: Krankheitsbilder innerhalb der Klassen

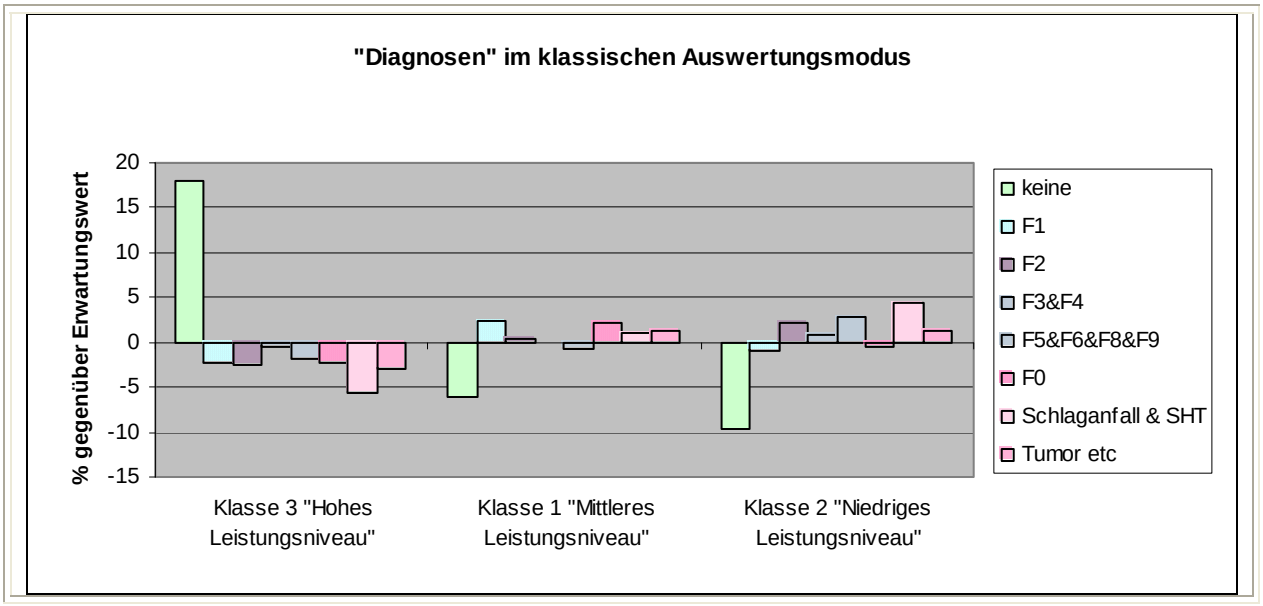


Tabelle 5-71 Klassischer Auswertungsmodus: Zellbesetzungen der Krankheitsbilder

	Klasse 3 „Hohes Leistungsniveau“		**Klasse 1 „Mittleres Leistungsniveau“**		**Klasse 2 „Niedriges Leistungsniveau“**		**Teilstichprobe „Testbatterie“ (N=590)**	
Krankheitsbild	N	+/-%	N	+/-%	N	+/-%	**Σ**	Erw.%
ohne	128	+18,1	119	-6,1	78	-9,7	**325**	55,1
F1	15	-2,3	32	+2,3	17	-1,0	**64**	10,8
F2	3	-2,5	11	+0,3	11	+2,2	**25**	4,2
F3&F4	13	-0,5	19	-0,1	15	+0,8	**47**	8,0
F5&F6&F8&F9	1	-1,8	4	-0,7	9	+2,9	**14**	2,4
F0	5	-2,4	18	+2,2	8	-0,6	**31**	5,3
Schlaganfall & SHT	8	-5,6	27	+0,9	25	+4,4	**60**	10,2
Tumor etc	2	-2,9	13	+1,3	9	+1,2	**24**	4,1
Σ	**175**		**243**		**172**		**590**	100

5.6.1.3.3 Validierung der 3-Klassen-Lösung

Der Einfluss der mehrkategoriellen Prädiktorvariablen Altersgruppe, Bildungsgrad, Diagnosegruppe einschließlich Interaktionen auf die Klassenbildung wird mit nominal logistischen Regressionen untersucht. Dabei wird folgendermaßen vorgegangen:
Die wenig besetzte Altersgruppe 4 (>70.LJ) wird der Altersgruppe 3 (60.-70.LJ) zugeordnet. Die kleine Personengruppe mit Bildungsgrad 0 (ohne Schulabschluss) wird dem Bildungsgrad 1 (Hauptschule) zugerechnet. Dadurch wird die Anzahl der Freiheitsgrade reduziert.
Die Berechnungen werden mit der Prädiktorvariablenkombination ohne Berücksichtigung der dreifachen Interaktion aus Altersgruppe, Bildungsgrad und Diagnosegruppe durchgeführt, um eine Explosion der Parameteranzahl zu vermeiden.
Die Prädiktorvariable „Diagnosegruppe“ ist dreifach unterteilt in keine Diagnose, Diagnose **mit** cerebraler Beteiligung und Diagnose **ohne** cerebrale Beteiligung.

Die Prüfungen der Modellgültigkeit der logistischen Regressionen mit Likelihood-Ratio-Tests für separate und kombinierte Prädiktoren samt Entropiereduktionen fasst *Tabelle 5-72* zusammen. Die zugehörigen Wald-Effekt-Tests der Prädiktorvariablenkombination sind *Tabelle 5-73* zu entnehmen.

Tabelle 5-72 Klassischer Auswertungsmodus: Entropiereduktion (Kriterium: 3-Klassen-Lösung)

Kriterium: 3-Klassen-Lösung des klassischen Auswertungsmodus Stichprobe Testbatterie - nominal logistisches Modell					
Prädiktor	**Modell-Test (N=590)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
Altersgruppe	6	73,01	<,0001***	5,70	(4,76)
Bildungsgrad	4	89,84	<,0001***	7,02	(6,39)
Diagnosegruppe	4	37,02	<,0001***	2,89	(2,26)
Prädiktorvariablen-kombination Altersgruppe, Bildungsgrad, Diagnosegruppe und Interaktionen	46	238,55	<,0001***	18,63	(11,44)

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-73 Klassischer Auswertungsmodus: Wald-Effekt-Tests (Kriterium: 3-Klassen-Lösung)

Kriterium: 3-Klassen-Lösung des klassischen Auswertungsmodus Stichprobe Testbatterie - nominal logistisches Modell			
Wald Effekt Tests der Prädiktorvariablenkombination			
Variable	FG	Wald- χ^2	Prob> χ^2
Altersgruppe	6	8,965	0,1755
Bildungsgrad	4	27,461	0,0000***
Diagnosegruppe	4	4,942	0,2933
Altersgruppe × Bildungsgrad	12	14,250	0,2851
Altersgruppe × Diagnosegruppe	12	18,575	0,0993
Bildungsgrad × Diagnosegruppe	8	4,252	0,8337

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Aus den Wald Effekt Tests der Prädiktorvariablenkombination folgt unter hochsignifikanter Geltung des logistischen Modells:

Nur der Bildungsgrad der Versuchspersonen beeinflusst hochsignifikant die Prädiktion der drei indizierten Klassen des klassischen Auswertungsmodus. Weder die Altersgruppe, die Diagnosegruppe noch die Interaktionen sind von statistischer Signifikanz.

Dieses Resultat bedeutet, dass die Position einer Testperson entlang des von den Klassen beschriebenen eindimensionalen Konstruktes „Testleistungsfähigkeit“ am sichersten durch ihren Bildungsgrad vorhergesagt wird (vgl. *Abbildung 5-12*).

5.6.2 Probabilistischer Auswertungsmodus der Testbatterie

5.6.2.1 Identifikation latenter Klassen

Im komplexeren probabilistischen Auswertungsmodus können - im Unterschied zum klassischen - jeweils klassenspezifische Testscorequartile aus den Summenscores berechnet werden, weil Klassenzugehörigkeitsparameter für jede Testperson bestimmt worden sind.
Die Verwendung des Testscores ist innerhalb jeder einzelnen Klasse der Mixed-Rasch-Modelle zulässiger als bei der Einklassenlösung Raschmodell. Dies wurde für die erhobenen Daten beim Raven- und Bentontest in den *Abschnitten 5.1.1.3* und *5.2.1.3* statistisch gezeigt.
Das zweite und dritte Quartil wird beim Cattelltest zusammengefasst, weil die Varianzen der Scores in den Klassen sehr klein sind.
Die klassenspezifischen Quartile sind in *Tabelle 5-74* dargestellt.

Tabelle 5-74 Probabilistischer Auswertungsmodus: Testscorequartile

Probabilistischer Auswertungsmodus		1.Quartil	2.Quartil	3.Quartil	4.Quartil
Ravenscore					
MRM 1.Klasse „Könner“		6-13	14-15	16-17	18-20
MRM 2.Klasse „Durchschnittliche“		4-9	10-11	12-13	14-19
MRM 3.Klasse „Nicht-Könner“		1-4	5-6	7-8	9-11
Cattellscore					
LCA 1.Klasse „Könner“		5-6	7		8-10
LCA 2.Klasse „Durchschnitt B“		2-4	5		6-7
LCA 3.Klasse „Nicht-Könner“		0-1	2		3-4
LCA 4.Klasse „Durchschnitt A“		3-4	5-6		7-8
Bentonscore					
MRM 1.Klasse „Könner B“		1-10	11-13	14-16	17-20
MRM 2.Klasse „Nicht-Könner“		1-6	7-10	11-13	14-19
MRM3.Klasse„Könner A“		3-11	12-14	15-16	17-18

Die Berechnungen der Latenten-Klassen-Analysen erfolgen auf der Grundlage der Variablen der *Tabelle 5-75*, also den indizierten Klassen und klassenspezifischen Testscorequartilen der Versuchspersonen.

Die Klassen der Testmodelle[61] wurden für jeden Test nach Testleistung geordnet. Dabei steht die Kategorie 1 für die jeweils beste Leistungsklasse[62].

Den Berechnungen wird die Teilstichprobe „Testbatterie“ (N=590) zugrunde gelegt.

Tabelle 5-75 Probabilistischer Auswertungsmodus: Variablenkategorien

Probab. Aus-wertungs-modus	1.Variable	2.Variable	3.Variable	4.Variable	5.Variable	6.Variable
	Raven MRM3	**Raven Score Quartil***	**Cattell LCA4**	**Cattell Score „Quartil“***	**Benton MRM3**	**Benton Score Quartil***
Kate-gorie1	1.Klasse „Könner“	1.	1.Klasse „Könner“	1.	3.Klasse „Könner A“	1.
Kate-gorie2	2.Klasse „Durch-schnittlich“	2.	4.Klasse „Durch-schnitt A“	2. und 3.	1.Klasse „Könner B“	2.
Kate-gorie3	3.Klasse „Nicht-Könner“	3.	2.Klasse „Durch-schnitt B“	4.	2.Klasse „Nicht-Könner“	3.
Kate-gorie4		4.	3.Klasse „Nicht-Könner“			4.

*klassenspezifische Quartile

Die Ergebnisse der durchgeführten Latenten-Klassen-Analysen nach probabilistischem Auswertungsmodus stellen sich folgendermaßen dar (*Tabelle 5-76*):

[61] aus den *Abschnitten 5.1.2.2, 5.2.2.2* und *5.3.2.2* - dargestellt in der 1., 3. und 5. Variablen -

[62] („Könner“ bei Raven- und Cattell-, „Könner A“ beim Bentontest)

Tabelle 5-76 Probabilistischer Auswertungsmodus: Informationstheoretische Maße

N=590	loglikeli	AIC	BIC	n_p
probabilistisch LCA1Kl	-4106,11	8248,22	8327,06	15
probabilistisch LCA2Kl	-3943,58	7961,16	8123,22	31
probabilistisch LCA3Kl	-3893,07	7898,14	8143,43	47
probabilistischLCA4Kl	-3870,56	7891,13	8219,64	63
probabilistisch LCA5Kl	-3834,11	7856,23	8267,96	79
probabilistisch LCA6Kl	-3814,39	7854,78	8349,73	95
probabilistisch LCA7Kl	-3802,48	7868,96	8447,13	111
saturiertes Modell	-3340,30	10134,60	17699,07	1727

Startwert 321 abgesichert mit 4321 und 5321

Der AIC (7854,78) bevorzugt eine 6-Klassen-, der BIC-Wert (8123,22) eine 2-Klassen-Struktur.

5.6.2.1.1 Inferenzstatistische Prüfung

Die Ergebnisse werden in *Tabelle 5-77* mittels Likelihoodquotiententests abgesichert.

Tabelle 5-77 Probabilistischer Auswertungsmodus: Likelihoodquotiententests

LR	-2ln(LR)	FG	Prob> χ^2
Probabilistisch LCA2Kl/sat.Modell	1206,56	1696	1
Probabilistisch LCA6Kl/sat.Modell	948,18	1632	1
Probabilistisch LCA1Kl/Prob.LCA6Kl	583,44	80	<,001***

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die Ergebnisse der gebildeten Likelihoodquotienten in *Tabelle 5-77* zeigen an, dass die vom AIC indizierte 6-Klassen-Lösung mit den Daten vereinbar ist. Diese passt besser als das Einklassenmodell und nicht schlechter als das saturierte Modell.

Obwohl auch die vom BIC-Wert indizierte 2-Klassen-Lösung die Daten nicht signifikant schlechter beschreibt als das saturierte Modell, wird die vom AIC angezeigte größere 6-Klassen-Lösung wegen differenzierterer diagnostischer Möglichkeiten als Grundlage der nachfolgenden Betrachtungen ausgewählt.

Die große Parameteranzahl des saturierten Modells (n_p=1727) und die im Vergleich dazu geringe Stichprobengröße (N=590) bedingen, dass die Likelihoodquotientenvergleiche

zwischen den Klassen-Modellen und dem saturierten Modell nur begrenzt aussagekräftig sind.

5.6.2.2 Interpretation der indizierten Klassen

In diesem *Abschnitt* werden die Testleistungen und die im Fragebogen erhobenen Merkmale der zugehörigen Testpersonen genutzt, um die indizierten sechs latenten Klassen voneinander abzugrenzen.

5.6.2.2.1 Charakterisierung der indizierten Klassen

Die Klassen werden entsprechend ihrer Leistungen in der Testbatterie benannt und in den *Tabellen 5-78* und *5-79* unter Angabe von Klassengrößen, mittleren klassenspezifischen Zuordnungssicherheiten, Testleistungen und Altersdurchschnitten angeordnet.
Die zugefügten Balkendiagramme verdeutlichen die Zusammensetzung der Altersgruppen und der Bildungsgrade über alle Klassen.
Die Klassenzuordnung der Personen erfolgt anhand der vom Modell vorgegebenen größten Zuordnungswahrscheinlichkeit.

Die Zusammenfassung der Kontraste von Testleistungen, Bildungsgraden, Alter und Altersgruppen zeigt *Tabelle 5-80*. Ausführlich sind die zugehörigen Kontraste der *Tabelle A-43* des Anhangs zu entnehmen.
Die gemittelten klassenspezifischen Testscores sind in *Abbildung 5-18* dargestellt.

Tabelle 5-78 Probabilistischer Auswertungsmodus: Die 6-Klassen-Lösung der LCA (Teil 1)

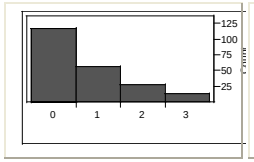
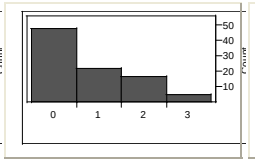
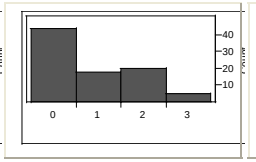
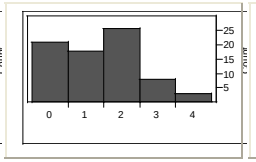
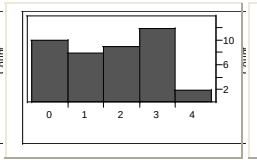
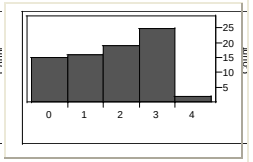
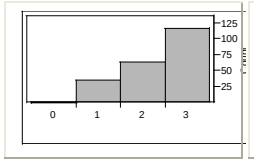
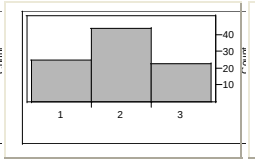
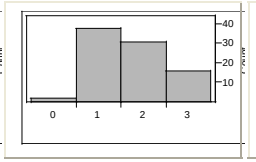
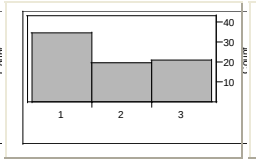
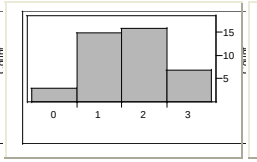
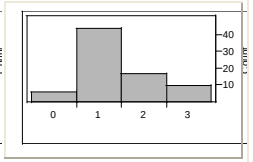
Teilstichprobe „Testbatterie“ (N=590)	hoch ← Testleistungsfähigkeit → niedrig					
Probabilistischer Auswertungsmodus	**Klasse 1** **„jüngere Personen mit hohem Leistungsniveau“**	**Klasse 4** **„jüngere Personen mit leicht reduziertem Leistungsniveau“**	**Klasse 3** **„jüngere Personen mit reduziertem Leistungsniveau“**	**Klasse 2** **„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“**	**Klasse 6** **„ältere Personen mit stark reduziertem Leistungsniveau & rel. Leistungseinbruch im Cattelltest“**	**Klasse 5** **„ältere Personen mit Leistungseinbrüchen in allen drei Tests“**
Vgl. *Abschnitt*	*5.6.2.2.1.1 a)*	*5.6.2.2.1.1 b)*	*5.6.2.2.1.1 c)*	*5.6.2.2.1.1 d)*	*5.6.2.2.1.1 e)*	*5.6.2.2.1.1 f)*
Klassengröße N (%-Anteil gesamt)	217 (37%)	92 (16%)	87 (15%)	76 (12%)	41 (7%)	77 (13%)
mittlere klassenspezifische Zuordnungssicherheit	0,902	0,866	0,738	0,900	0,939	0,924
Altersmittelwerte in den Klassen						
Alter [Ø]	37,12	39,38	40,56	46,86	51,54	50,96
Alter [σ]	13,58	13,41	12,64	14,04	13,85	13,84
Altersgruppen in den Klassen (ungleicher Maßstab)						
Altersgruppe 0=15-39LJ Altersgruppe 1=39-49LJ Altersgruppe 2=49-59LJ Altersgruppe 3=59-70LJ Altersgruppe 4= >70LJ						
Bildungsgrade in den Klassen (ungleicher Maßstab)						
Bildungsgrad 0=kein Abschluss Bildungsgrad 1=Hauptschule Bildungsgrad 2=Realschule Bildungsgrad 3=Abitur						

Tabelle 5-79 Probabilistischer Auswertungsmodus: Die 6-Klassen-Lösung der LCA (Teil 2)

Teilstichprobe „Testbatterie“(N=590)	hoch ← Testleistungsfähigkeit → niedrig					
Probabilistischer Auswertungsmodus	**Klasse 1** **„jüngere Personen mit hohem Leistungsniveau“**	**Klasse 4** **„jüngere Personen mit leicht reduziertem Leistungsniveau“**	**Klasse 3** **„jüngere Personen mit reduziertem Leistungsniveau“**	**Klasse 2** **„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“**	**Klasse 6** **„ältere Personen mit stark reduziertem Leistungsniveau & rel. Leistungseinbruch im Cattelltest“**	**Klasse 5** **„ältere Personen mit Leistungseinbrüchen in allen drei Tests“**
Klassenspezifische Leistungen im Raventest						
Ravenscore [Ø]	16,13	12,28	11,14	11,39	9,59	5,60
Ravenscore [σ]	1,90	2,47	2,70	3,12	2,96	2,17
Raven-Leistungsreihenfolge (lt. Kontrasten Tab.5-80)	***1.***	***(2.)***	***(2.)***	***(2.)***	***3.***	***4.***
Klassenspezifische Leistungen im Cattelltest						
Cattellscore [Ø]	6,12	5,14	4,68	4,95	2,34	2,70
Cattellscore [σ]	1,24	1,39	1,28	1,26	1,26	1,69
Cattell-Leistungsreihenfolge (lt. Kontrasten Tab.5-80)	***1.***	***2.***	***2.***	***2.***	***3.***	***3.***
bearbeitete Cattell-Items [Ø]	8,72	8,79	8,69	8,13	7,61	8,09
bearbeitete Cattell-Items [σ]	2,05	2,20	2,19	2,32	2,77	2,53
Klassenspezifische Leistungen im Bentontest						
Bentonscore [Ø]	15,27	12,71	11,18	8,58	8,43	5,90
Bentonscore [σ]	2,39	2,81	2,81	2,87	3,99	3,02
Benton-Leistungsreihenfolge (lt. Kontrasten Tab.5-80)	***1.***	***2.***	***3.***	***4.***	***4.***	***5***
Bentonfehler (BTF) [Ø]	5,88	9,89	12,90	17,20	18,71	24,18
Bentonfehler (BTF) [σ]	3,34	4,93	5,29	6,76	9,66	8,84

Tabelle 5-80 Probabilistischer Auswertungsmodus: Zusammenfassung der Kontraste

Kontraste des Alters	„jüngere Personen mit hohem Leistungs-niveau“ (Kl. 1)	„jüngere Personen mit leicht reduziertem Leistungs-niveau“ (Kl. 4)	„jüngere Personen mit reduziertem Leistungs-niveau“ (Kl. 3)	„ältere Personen mit reduziertem Leistungs-niveau & relativem Leistungs-einbruch im Bentontest“ (Kl. 2)	„ältere Personen mit stark reduziertem Leistungsni veau & rel. Leistungs-einbruch im Cattelltest“(Kl. 6)	„ältere Personen mit Leistungs-einbrüchen in allen drei Tests“ (Kl. 5)
Kl. 1 „jPmhL“		ns	ns	***	***	***
Kl. 4 „jPmlrL“	ns		ns	**	***	***
Kl. 3 „jPmrL“	ns	ns		*	***	***
Kl. 2 „äPmrL&rLiB“	***	**	*		ns	ns
Kl. 6 „äPmsrL&rLiC“	***	***	***	ns		ns
Kl. 5 „äPmLiadT“	***	***	***	ns	ns	
Kontraste der Altersgruppen	„jPmhL“	jPmlrL“	„jPmrL“	„äPmrL&LiB“	„äPmsrL&LiC“	„äPmLiadT“
Kl. 1 „jPmhL“		ns	ns	***	***	***
Kl. 4 „jPmlrL“	ns		ns	**	***	***
Kl. 3 „jPmrL“	ns	ns		**	**	***
Kl. 2 „äPmrL&rLiB“	***	**	**		ns	ns
Kl. 6 „äPmsrL&rLiC“	***	***	**	ns		ns
Kl. 5 „äPmLiadT“	***	***	***	ns	ns	
Kontraste der Bildungsgrade	„jPmhL“	jPmlrL“	„jPmrL“	„äPmrL&LiB“	„äPmsrL&LiC“	„äPmLiadT“
Kl. 1 „jPmhL“		***	***	***	***	***
Kl. 4 „jPmlrL“	***		ns	ns	ns	***
Kl. 3 „jPmrL“	***	ns		ns	ns	ns
Kl. 2 „äPmrL&rLiB“	***	ns	ns		ns	ns
Kl. 6 „äPmsrL&rLiC“	***	ns	ns	ns		ns
Kl. 5 „äPmLiadT“	***	***	ns	ns	ns	

(*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur))

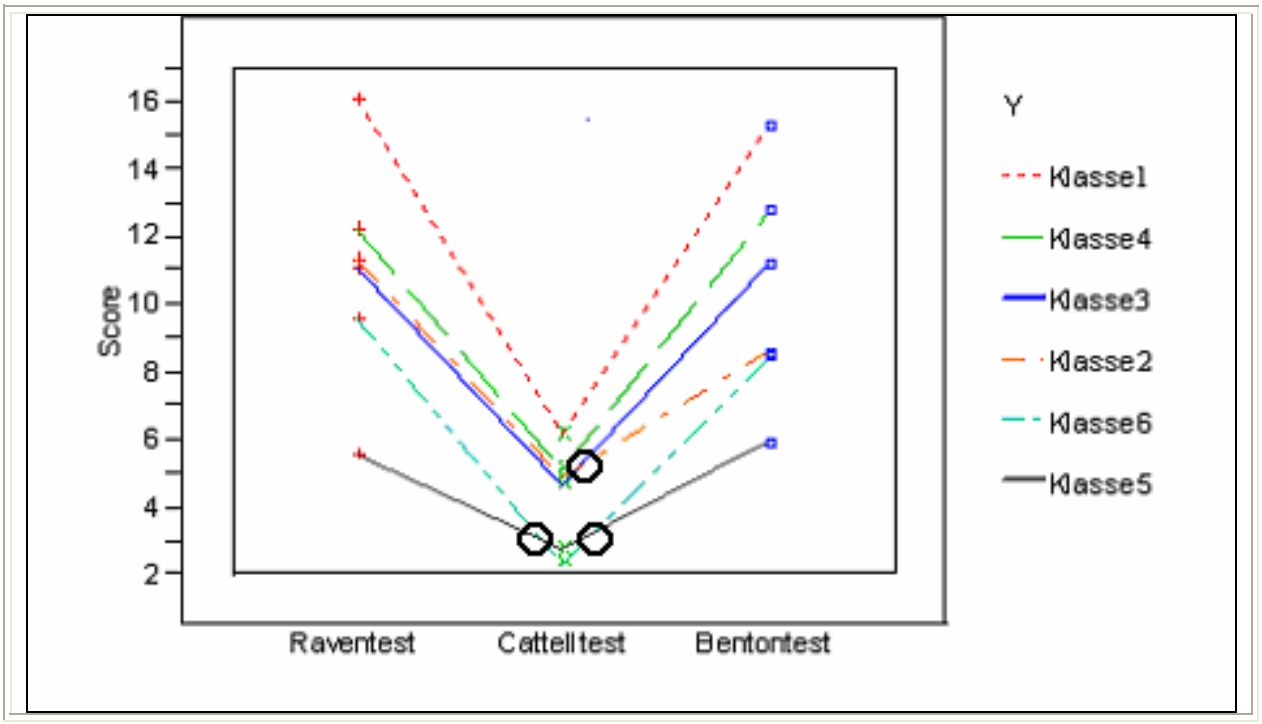
entnommen aus *Tab.A-43*

Fortsetzung der Tabelle auf der nächsten Seite

Fortsetzung Tabelle 5-80 Probabilistischer Auswertungsmodus: Zusammenfassung der Kontraste

Kontraste der Ravenscores	„jüngere Personen mit hohem Leistung-niveau“ (Kl. 1)	„jüngere Personen mit leicht reduziertem Leistungs-niveau“ (Kl. 4)	„jüngere Personen mit reduziertem Leistungs-niveau“ (Kl. 3)	„ältere Personen mit reduziertem Leistungs-niveau & relativem Leistungs-einbruch im Bentontest“ (Kl. 2)	„ältere Personen mit stark reduziertem Leistungsniveau & rel. Leistungs-einbruch im Cattelltest“(Kl. 6)	„ältere Personen mit Leistungs-einbrüchen in allen drei Tests“ (Kl. 5)
Kl. 1 „jPmhL“		***	***	***	***	***
Kl. 4 „jPmlrL“	***		*	ns	***	***
Kl. 3 „jPmrL“	***	*		ns	*	***
Kl. 2 „äPmrL&rLiB“	***	ns	ns		*	***
Kl. 6 „äPmsrL&rLiC“	***	***	*	*		***
Kl. 5 „äPmLiadT“	***	***	***	***	***	
Kontraste der Cattellscores	„jPmhL“	jPmlrL“	„jPmrL“	„äPmrL&LiB“	„äPmsrL&LiC“	„äPmLiadT“
Kl. 1 „jPmhL“		***	***	***	***	***
Kl. 4 „jPmlrL“	***		ns	ns	***	***
Kl. 3 „jPmrL“	***	ns		ns	***	***
Kl. 2 „äPmrL&rLiB“	***	ns	ns		***	***
Kl. 6 „äPmsrL&rLiC“	***	***	***	***		ns
Kl. 5 „äPmLiadT“	***	***	***	***	ns	
Kontraste der bearbeiteten Cattell-Items	signifikant (*) nur bei Kl. 1„jüngere Personen mit hohem Leistungsniveau“ & Kl. 6„ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“					
Kontraste der Bentonscores	„jPmhL“	jPmlrL“	„jPmrL“	„äPmrL&LiB“	„äPmsrL&LiC“	„äPmLiadT“
Kl. 1 „jPmhL“		***	***	***	***	***
Kl. 4 „jPmlrL“	***		**	**	***	***
Kl. 3 „jPmrL“	***	**		***	***	***
Kl. 2 „äPmrL&rLiB“	***	**	***		ns	***
Kl. 6 „äPmsrL&rLiC“	***	***	***	ns		**
Kl. 5 „äPmLiadT“	***	***	***	***	**	
Kontraste der Bentonfehler (BTF)	bis auf Kl. 2 „äPmrL&rLiB“& Kl. 6 „äPmsrL&rLiC“ sind sämtliche Unterschiede mindestens signifikant					
(*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur)) entnommen aus *Tab.A-43*						

Abbildung 5-18 Probabilistischer Auswertungsmodus: klassenspezifische Durchschnittsscores



Auf der Grundlage der *Tabellen 5-78* bis *5-80* lassen sich die folgenden Feststellungen machen:

1) <u>Klassen:</u> Die Latente-Klassen-Analyse des probabilistischen Auswertungsmodus indiziert drei Klassen mit eher jüngeren (Ø ca. 40. Lebensjahr) und drei Klassen mit eher älteren Personen (Ø ca. 50. Lebensjahr).
2) <u>Sicherheit der Zuordnung:</u> Personen werden der leistungsschwächsten der jüngeren Klassen „jüngere Personen mit reduziertem Leistungsniveau“ mit deutlich geringerer Sicherheit (p=0,738) zugeordnet, als den restlichen fünf Klassen ($p \geq 0{,}866$).
3) <u>Bildungsgrade:</u> Die leistungsstärkste Klasse „jüngere Personen mit hohem Leistungsniveau“ unterscheidet sich in ihren Bildungsgraden hochsignifikant von sämtlichen anderen Klassen.

   Eine Bildungsabhängigkeit der Testleistungen ist in den zugehörigen Abbildungen in *Tabelle 5-78* deutlicher bei den drei jüngeren Klassen zu erkennen.
4) <u>Dimensionalität:</u> Den in *Abbildung 5-18* dargestellten klassenspezifischen Durchschnittscores aus *Tabelle 5-79* sind nahezu geordnete Klassen zu entnehmen.

   Ebenso wie im klassischen Auswertungsmodus wird eine Struktur der latenten Variable als eindimensionales Konstrukt „Testleistungsfähigkeit“ nahe gelegt.

5) Diagnostische Implikationen: Diagnostisch aufschlussreich sind insbesondere jene Klassen die mit Profillinienschnittpunkten Hinweise auf qualitative Bearbeitungsunterschiede geben. Diese sind in *Abbildung 5-18* durch Kreise markiert.

Die Profile überschneiden sich bei den Klassen 2 und 3, sowie den Klassen 6 und 5:

a) In der Klasse 2 „ältere Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“ macht sich im Vergleich zu den jüngeren Klassen ein altersbedingter Leistungsabfall beim Bentontest bemerkbar (vgl. *Abschnitt 5.6.2.2.1.1 d))*.

b) In der Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“ weisen die kleinsten Cattellscores auf die Relevanz der Speedkomponente zur testdiagnostischen Erkennung bestimmter Hirnverletzungen hin (vgl. *Abschnitt 5.6.2.3.2.2*).

5.6.2.2.1.1 Merkmalsprofile der indizierten Klassen

Für die Merkmalsprofile der sechs Klassen des probabilistischen Auswertungsmodus, welche detailliert in den *Abbildungen 5-19* bis *5-24* dargestellt und erläutert werden, wurden die Variablenkategorien der in *Tabelle 5-75* gemachten Einteilung übernommen. Die verwendeten Abkürzungen werden in *Tabelle 5-81* zusammengefasst.

Die Anteile klinisch unauffälliger Personen, sowie von Patienten **mit** und **ohne** cerebrale Beteiligung in sämtlichen Variablenkategorien sind in den *Abbildungen* als Balkendiagramme dargestellt.

Die klassenspezifischen Verteilungen der Alters- und Bildungsgrade sind in den *Tabellen 5-82* bis *5-87* aufgeführt.

Tabelle 5-81 Probabilistischer Auswertungsmodus: Abkürzungen der Variablenkategorien

Raventest	
Klasse der „Könner“ beim Raventest	RKlK1
Klasse der „Durchschnittlichen“ beim Raventest	RKlK2
Klasse der „Nicht-Könner“ beim Raventest	RKlK3
1. Testscorequartil im Raventest	RQK1
2. Testscorequartil im Raventest	RQK2
3. Testscorequartil im Raventest	RQK3
4. Testscorequartil im Raventest	RQK4
Cattelltest	
Klasse der „Könner“ beim Cattelltest	CKlK1
Klasse „Durchschnitt A“ beim Cattelltest	CKlK2
Klasse „Durchschnitt B“ beim Cattelltest	CKlK3
Klasse der „Nicht-Könner“ beim Cattelltest	CKlK4
1. Testscorequartil im Cattelltest	CQK1
2. und 3. Testscorequartil im Cattelltest	CQK2
4. Testscorequartil im Cattelltest	CQK3
Bentontest	
Klasse der „Könner A“ beim Bentontest	BKlK1
Klasse der „Könner B“ beim Bentontest	BKlK2
Klasse der „Nicht-Könner“ beim Bentontest	BKlK3
1. Testscorequartil im Bentontest	BQK1
2. Testscorequartil im Bentontest	BQK2
3. Testscorequartil im Bentontest	BQK3
4. Testscorequartil im Bentontest	BQK4

Beispiele: RQK1 = **R**aventestscore-**Q**uartil **K**ategorie **1;** RKlK1 = **R**aven **Kl**assenbezeichnung **K**ategorie **1**

a) Die Klasse 1 der „jüngeren Personen mit hohem Leistungsniveau“

Die Klasse 1 im probabilistischen Auswertungsmodus

Abbildung 5-19 Merkmalsprofil: Klasse 1

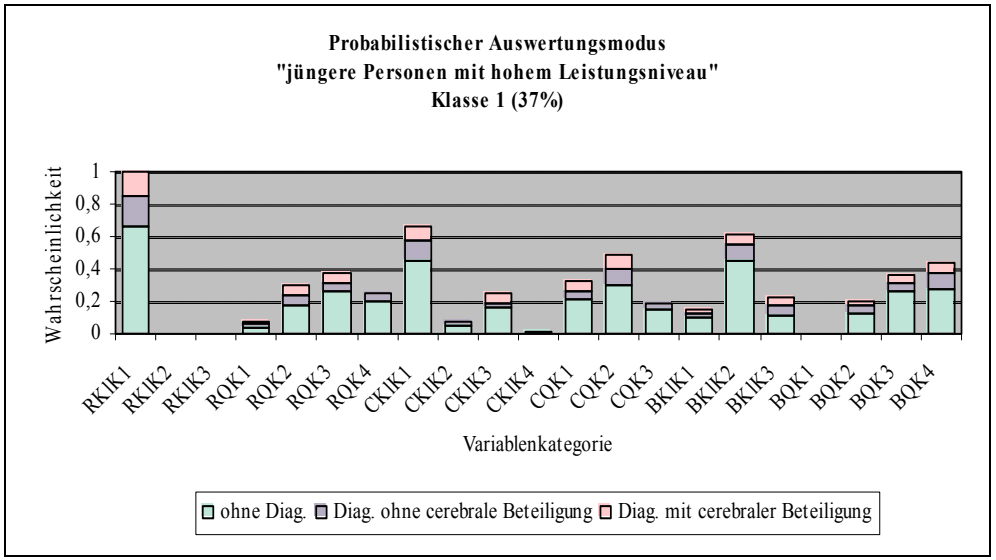
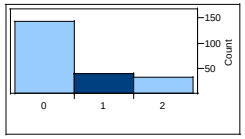
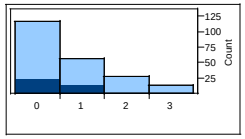
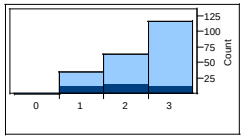
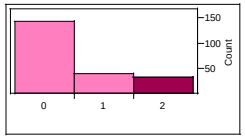
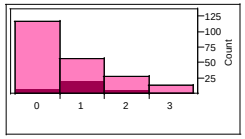
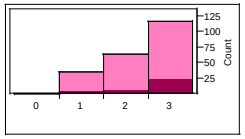


Tabelle 5-82 Alters- und Bildungsverteilung in Klasse 1

Klasse 1	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70LJ	1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

Die größte Klasse 1 „jüngere Personen mit hohem Leistungsniveau“ erzielt in allen drei Tests die besten Leistungen. Alle darin enthaltenen Personen kommen beim Raventest aus der Klasse der „Könner“ RKlK1. Beim Cattelltest ist die Klasse der „Nicht-Könner“ CKlK4 unbesetzt. Zellbesetzungen der untersten klassenspezifischen Testscorequartile fehlen beim Bentontest völlig (BQK1) und sind auch beim Raventest selten (RQK1).

b) Die Klasse 4 der „jüngeren Personen mit leicht reduziertem Leistungsniveau“

Die Klasse 4 im probabilistischen Auswertungsmodus

Abbildung 5-20 Merkmalsprofil: Klasse 4

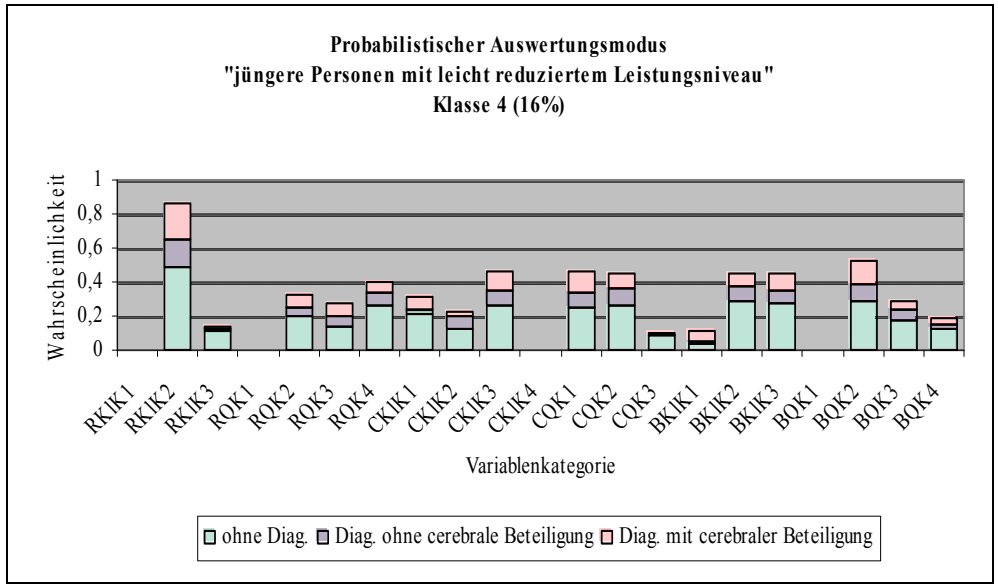
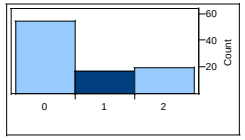
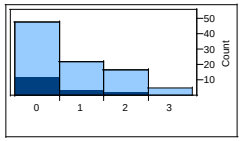
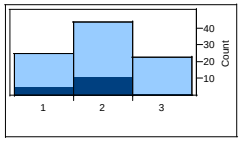
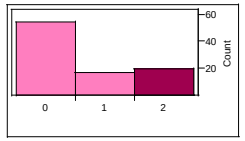
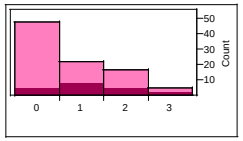
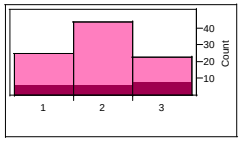


Tabelle 5-83 Alters- und Bildungsverteilung in Klasse 4

Klasse 4	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70LJ	1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

Die mittelgroße Klasse 4 „jüngere Personen mit leicht reduziertem Leistungsniveau“ erzielt die zweitbesten Leistungen in allen Tests. Alle darin enthaltenen Personen kommen beim Raventest nicht aus der Klasse der „Könner“ RKlK1, dagegen stammen die Personen zu über 80% aus den „Durchschnittlichen“ RKlK2.

Die Klasse 4 erreicht die höheren klassenspezifischen Testscorequartile beim Raventest (RQK2 bis RQK4) und auch beim Bentontest (BQK2 bis BQK4). Beim Cattelltest wird keine Person den „Nicht-Könnern“ CKlK4 zugeordnet.

c) Die Klasse 3 der „jüngeren Personen mit reduziertem Leistungsniveau“

Die Klasse 3 im probabilistischen Auswertungsmodus

Abbildung 5-21 Merkmalsprofil: Klasse 3

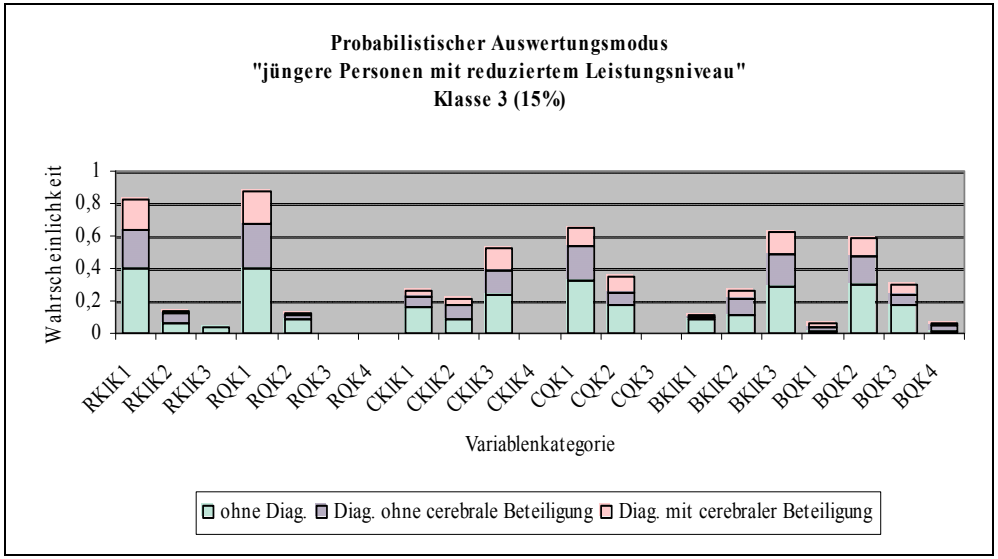
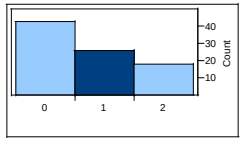
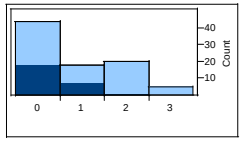
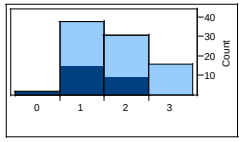
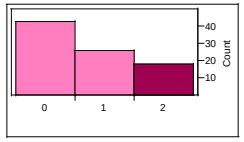
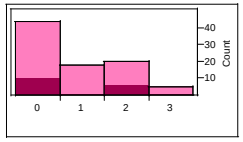
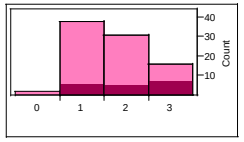


Tabelle 5-84 Alters- und Bildungsverteilung in Klasse 3

Klasse 3	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70LJ	0=kein Abschluss 1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

Die mittelgroße, ebenso mit jungen Personen besetzte Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“ hat im Vergleich zur leistungsstärkeren Klasse zuvor, signifikante Leistungseinbrüche bei Raven- und Bentontest. Ein großer Personenanteil kommt beim Raventest aus den untersten Quartilen RQK1 der Klasse der „Könner“ RKlK1, beim Bentontest aus dem 2. und 3. Quartil (BQK2 und BQK3) der Klasse der „Nicht-Könner“ BKlK3. Beim Cattelltest sind in der Klasse der „Nicht-Könner“ CKlK4 und in den obersten klassenspezifischen Testscorequartilen CQK3 keine Personen vertreten.

Auf diagnostische Implikationen dieser Klasse wird in *Abschnitt 5.6.2.3.2.2* näher eingegangen.

d) Die Klasse 2 der „älteren Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“

Die Klasse 2 im probabilistischen Auswertungsmodus

Abbildung 5-22 Merkmalsprofil: Klasse 2

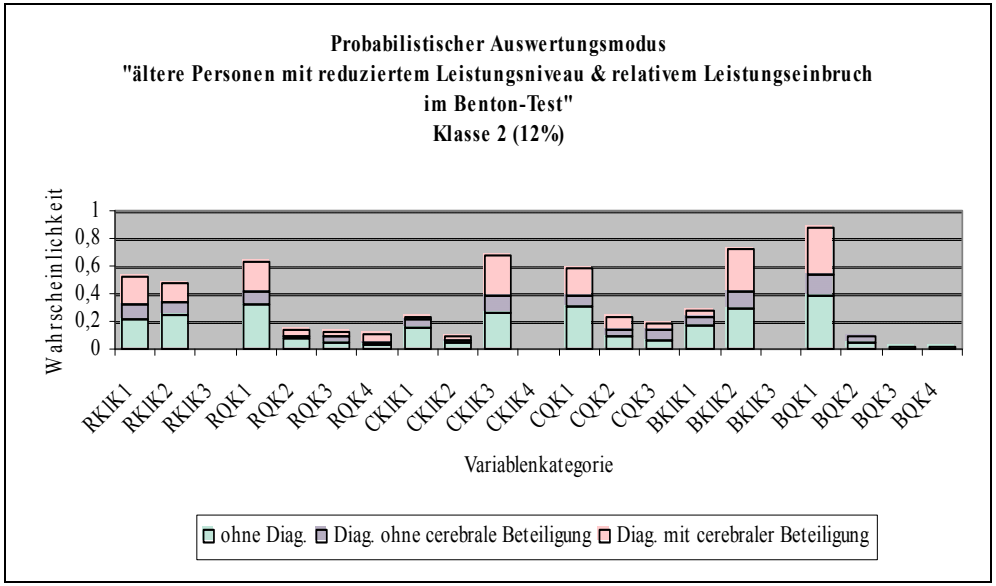
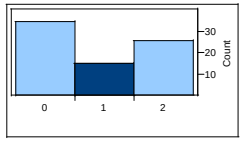
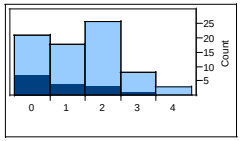
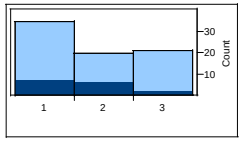
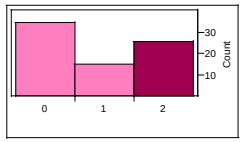
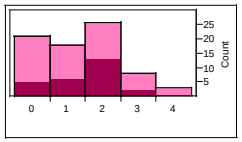
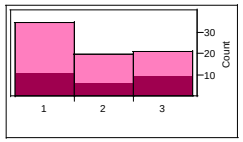


Tabelle 5-85 Alters- und Bildungsverteilung in Klasse 2

	Diagnosegruppe	Verteilung der Altersgruppen	Verteilung der Bildungsgrade
Klasse 2	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70L 4=>70LJ	1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

Klasse 2 „ältere Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“ ist ebenfalls mittelgroß und die leistungsstärkste der drei mit eher älteren Personen besetzten Klassen. Diese hat verglichen mit der jüngeren und leistungsstärkeren Klasse zuvor, signifikante Leistungseinbrüche nur beim Bentontest. Der Befund bestätigt den aus früheren Untersuchungen bekannten altersabhängigen Leistungsknick bei diesem Test. Keine Person dieser Klasse wird den „Nicht-Könner“-Klassen der drei Tests zugeordnet. Im Bentontest wird zudem von über 80% der Personen nur das unterste klassenspezifische Testscorequartil BQK1 erreicht. Auf diagnostische Implikationen dieser Klasse wird in *Abschnitt 5.6.2.3.1* näher eingegangen.

e) Die Klasse 6 der „älteren Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“

Die Klasse 6 im probabilistischen Auswertungsmodus

Abbildung 5-23 Merkmalsprofil: Klasse 6

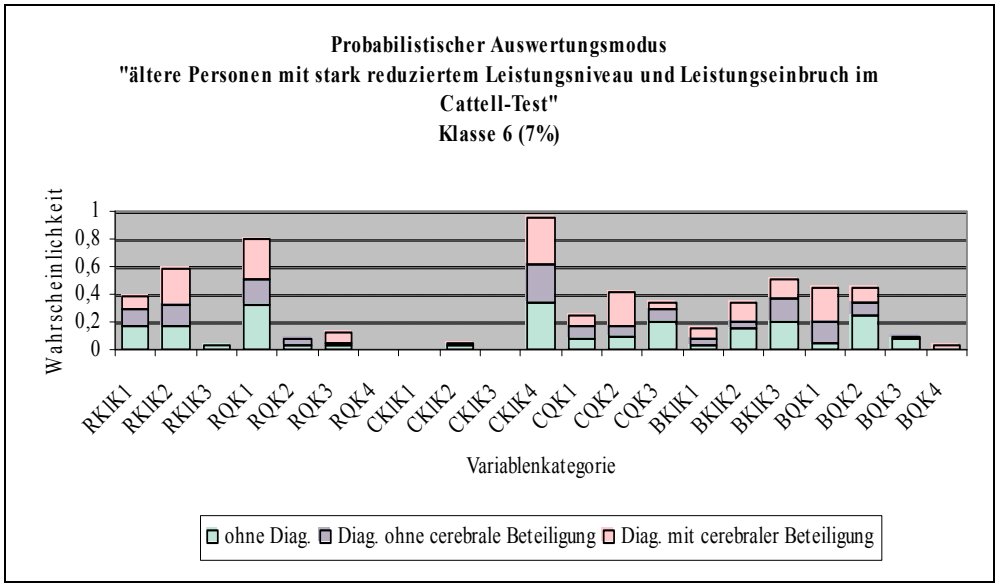
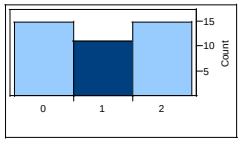
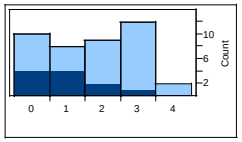
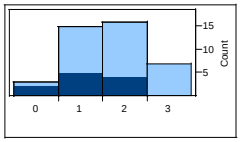
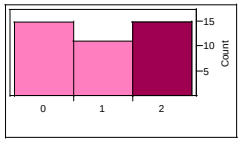
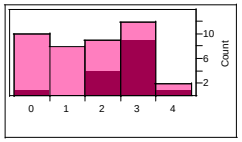
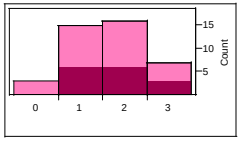


Tabelle 5-86 Alters- und Bildungsverteilung in Klasse 6

Klasse 6	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70L 4=>70LJ	0=kein Abschluss 1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

Die kleinste Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“ verzeichnet im Vergleich zur leistungsstärkeren Klasse zuvor, signifikante Leistungseinbrüche bei Raven- und Cattelltest.

80 Prozent der Personen erreichen im Raventest nur das unterste klassenspezifische Testscorequartil RQK1.

Im Bentontest werden nur äußerst selten die höheren klassenspezifische Testscorequartile BQK3 und BQK4 erreicht.

Fast alle Personen der Klasse 6 kommen beim Cattelltest aus der „Nicht-Könner“-Klasse CKlK4 mit den tendenziell schlechtesten Leistungen in diesem Test.

Auf diagnostische Implikationen dieser Klasse wird in *Abschnitt 5.6.2.3.2.2* näher eingegangen.

f) Die Klasse 5 der „älteren Personen mit Leistungseinbrüchen in allen drei Tests“

Die Klasse 5 im probabilistischen Auswertungsmodus

Abbildung 5-24 Merkmalsprofil: Klasse 5

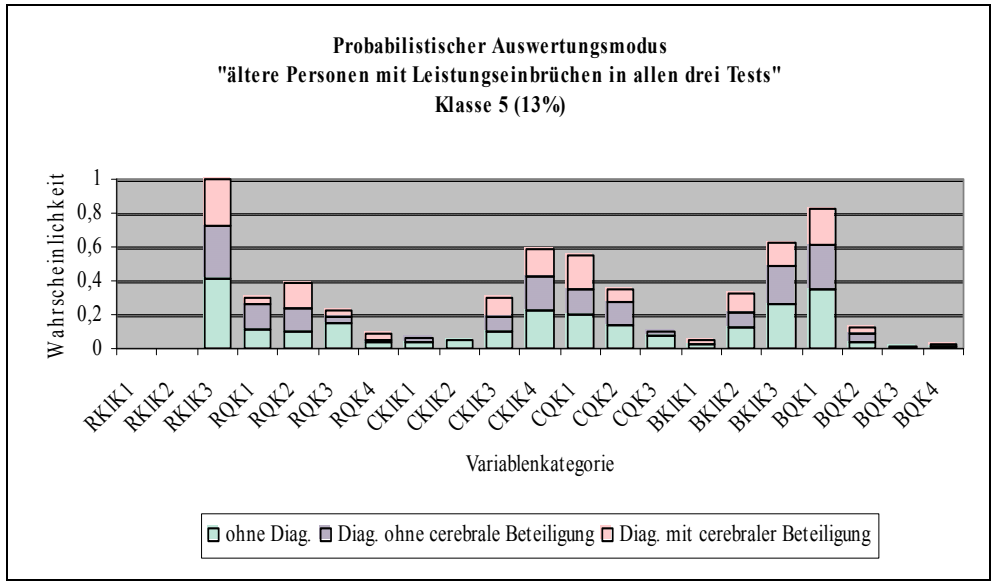
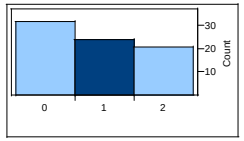
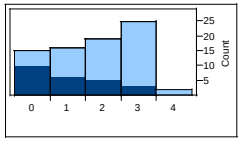
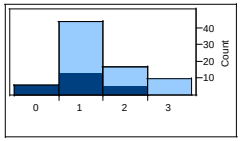
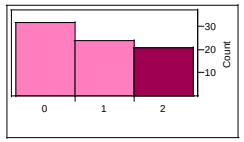
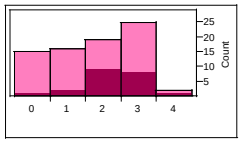
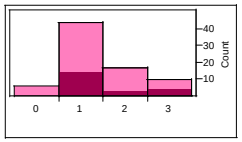


Tabelle 5-87 Alters- und Bildungsverteilung in Klasse 5

Klasse 5	**Diagnosegruppe**	**Verteilung der Altersgruppen**	**Verteilung der Bildungsgrade**
	0=ohne Diagnose 1=Diagnose ohne cerebrale Beteiligung 2=Diagnose mit cerebraler Beteiligung	0=15-39LJ 1=39-49LJ 2=49-59LJ 3=59-70L 4=>70LJ	0=kein Abschluss 1=Hauptschule 2=Realschule 3=Abitur
Dunkel= Diagnose ohne cerebrale Beteiligung Hell= Rest			
Dunkel= Diagnose mit cerebraler Beteiligung Hell= Rest			

Die mittelgroße Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“ verzeichnet im Vergleich zur leistungsstärkeren Klasse zuvor eine nochmals signifikante Leistungsreduktion bei Raven- und Bentontest. Fast alle Personen kommen aus den Klassen der „Nicht-Könner“.

Insbesondere beim Cattell- und Bentontest werden nur die untersten klassenspezifischen Testscorequartile CQK1 und BQK1 erreicht.

Auf diagnostische Implikationen dieser Klasse wird in *Abschnitt 5.6.2.3.2* näher eingegangen.

5.6.2.3 Diagnostische Relevanz der latenten Klassen

In diesem *Abschnitt* werden die sechs latenten Klassen in Diagnosegruppen aufgeteilt, um diagnostisch verwertbare Information aufzuzeigen.

5.6.2.3.1 Scorevergleiche klinisch auffälliger und unauffälliger Personen

Die Scoreunterschiede der diagnostischen Gruppen werden klassenspezifisch mit Varianzanalysen geprüft und mit Kontrasten weitergehend ausgewertet. In den *Tabellen* sind die Klassenzusammensetzung (*5-88)*, die Varianzanalysen der Scores (*5-89* bis *5-91)* und die Einzelvergleiche (5-92) aufgeführt.

Tabelle 5-88 Probabilistischer Auswertungsmodus: Klassenzusammensetzung

Klassenspezifische Betrachtung der diagnostischen Gruppen																		
hoch ← Testleistungsfähigkeit → niedrig																		
Teil-stichprobe „Test-batterie“ (N=590)	**Klasse 1 „jüngere Personen mit hohem Leistungsniveau“**			**Klasse 4 „jüngere Personen mit leicht reduziertem Leistungsniveau“**			**Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“**			**Klasse 2 „ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“**			**Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“**			**Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“**		
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
N [% Kl-anteil]	145 (67%)	40 (18%)	32 (15%)	55 (60%)	17 (18%)	20 (22%)	43 (49%)	26 (30%)	18 (21%)	35 (46%)	15 (20%)	26 (34%)	15 (37%)	11 (27%)	15 (37%)	32 (42%)	24 (31%)	21 (27%)
Alter [Ø]	35,99	35,28	44,56	37,71	36,94	46,05	41,37	36,42	44,61	48,34	40,73	48,38	46,40	44,40	44,73	53,56	42,54	56,62
Alter [σ_n]	1,10	2,10	2,35	1,76	3,17	2,93	1,89	2,44	2,93	2,35	3,59	2,72	3,03	3,53	3,03	2,25	2,59	2,77

klU = ohne Diagnose, Diag.o.cB. = Diagnose **ohne** cerebrale Beteiligung und Diag.m.cB. = Diagnose **mit** cerebraler Beteiligung, σ_n = Standardfehler

Tabelle 5-89 Probabilistischer Auswertungsmodus: Varianzanalysen der Raven-Scores

Klassenspezifische Betrachtung der diagnostischen Gruppen																		
hoch ← Testleistungsfähigkeit → niedrig																		
Teil-stichprobe „Test-batterie“ (N=590)	**Klasse 1 „jüngere Personen mit hohem Leistungsniveau“**			**Klasse 4 „jüngere Personen mit leicht reduziertem Leistungsniveau“**			**Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“**			**Klasse 2 „ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“**			**Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“**			**Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“**		
Varianzanalysen der Raven-Scores innerhalb der einzelnen Klassen																		
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
Ravenscore [Ø]	16,37	15,73	15,56	12,18	12,41	12,45	11,05	10,92	11,67	10,66	12,53	11,73	9,67	9,45	9,60	6,00	4,58	6,14
Ravenscore [σ_n]	0,16	0,30	0,33	0,34	0,61	0,56	0,41	0,53	0,64	0,52	0,79	0,60	0,78	0,91	0,78	0,37	0,42	0,45
	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total
FG	2	214	216	2	89	91	2	84	86	2	73	75	2	38	40	2	74	76
QS	25,38	757,74	783,12	1,40	555,25	556,65	6,59	619,75	626,34	41,42	690,73	732,16	0,29	349,66	349,95	36,11	320,40	356,52
$\hat{\sigma}^2$	12,69	3,54		0,70	6,24		3,30	7,38		20,71	9,46		0,15	9,20		18,06	4,33	
F	3,58			0,11			0,45			2,19			0,02			4,17		
Prob > F	0,0294*			0,8938			0,6412			0,1193			0,9843			0,0192*		

klU = ohne Diagnose, Diag.o.cB. = Diagnose **ohne** cerebrale Beteiligung und Diag.m.cB. = Diagnose **mit** cerebraler Beteiligung;
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001; $\boldsymbol{\sigma_n}$ = Standardfehler

Tabelle 5-90 Probabilistischer Auswertungsmodus: Varianzanalysen der Cattell-Scores

Klassenspezifische Betrachtung der diagnostischen Gruppen																		
hoch ← Testleistungsfähigkeit → niedrig																		
Teil-stichprobe „Test-batterie“ (N=590)	**Klasse 1 „jüngere Personen mit hohem Leistungsniveau“**			**Klasse 4 „jüngere Personen mit leicht reduziertem Leistungsniveau“**			**Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“**			**Klasse 2 „ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“**			**Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“**			**Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“**		
Varianzanalysen der Cattell-Scores innerhalb der einzelnen Klassen																		
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
Cattellscore [Ø]	6,17	6,33	5,63	5,32	4,71	5,00	4,72	4,65	4,61	5,09	5,47	4,46	2,73	2,09	2,13	3,16	2,50	2,24
Cattellscore [σ_n]	0,10	0,19	0,22	0,19	0,33	0,31	0,20	0,25	0,31	0,21	0,32	0,24	0,32	0,38	0,32	0,29	0,34	0,36
	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total
FG	2	214	216	2	89	91	2	84	86	2	73	75	2	38	40	2	74	76
QS	9,82	322,30	332,12	5,52	169,64	175,16	0,1750	140,81	140,99	10,852	108,94	119,79	3,64	59,58	63,22	12,10	206,03	218,13
$\hat{\sigma}^2$	4,91	1,51		2,76	1,91		0,09	1,68		5,43	1,49		1,82	1,57		6,05	2,78	
F	3,26			1,45			0,05			3,64			1,16			2,17		
Prob > F	0,0403*			0,24			0,9492			0,0312*			0,3237			0,1210		

klU = ohne Diagnose, Diag.o.cB. = Diagnose **ohne** cerebrale Beteiligung und Diag.m.cB. = Diagnose **mit** cerebraler Beteiligung;
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001; $\boldsymbol{\sigma_n}$ = Standardfehler

Tabelle 5-91 Probabilistischer Auswertungsmodus: Varianzanalysen der Benton-Scores

Klassenspezifische Betrachtung der diagnostischen Gruppen																		
hoch ← Testleistungsfähigkeit → niedrig																		
Teil-stichprobe „Test-batterie“ (N=590)	**Klasse 1 „jüngere Personen mit hohem Leistungsniveau“**			**Klasse 4 „jüngere Personen mit leicht reduziertem Leistungsniveau“**			**Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“**			**Klasse 2 „ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“**			**Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“**			**Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“**		
Varianzanalysen der Benton-Scores innerhalb der einzelnen Klassen																		
	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB	klU	Diag. o.cB	Diag. m.cB
Bentonscore [Ø]	15,48	14,95	14,69	12,80	12,58	12,55	11,58	11,04	10,44	9,37	9,73	6,84	10,20	7,45	7,40	6,50	4,67	6,38
Bentonscore [σ_n]	0,20	0,38	0,42	0,38	0,69	0,64	0,43	0,55	0,66	0,44	0,67	0,51	1,00	1,16	1,00	0,52	0,60	0,64
	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total	Gruppe	Fehler	Total
FG	2	214	216	2	89	91	2	84	86	2	73	75	2	38	40	2	74	76
QS	21,52	1217,0	1238,5	1,21	717,87	719,08	17,19	661,87	679,06	120,04	496,49	616,53	73,37	564,73	638,10	52,88	638,29	691,17
$\hat{\sigma}^2$	10,76	5,69		0,60	8,07		8,59	7,87		60,02	6,80		36,69	14,86		26,44	8,63	
F	1,89			0,07			1,09			8,82			2,47			3,07		
Prob > F	0,1533			0,9279			0,3407			0,0004***			0,0982			0,0526		

klU = ohne Diagnose, Diag.o.cB. = Diagnose **ohne** cerebrale Beteiligung und Diag.m.cB. = Diagnose **mit** cerebraler Beteiligung;
*p≤ 0,05;**p≤ 0,01;***p≤ 0,001; $\boldsymbol{\sigma_n}$ = Standardfehler

Tabelle 5-92 Probabilistischer Auswertungsmodus:Zusammenfassung der Kontraste in den Klassen 1, 2 und 5

Klasse 1 „jüngere Personen mit hohem Leistungsniveau“			
Kontraste der Ravenscores	klU	Diag.o.cB	Diag.m.cB
klU		ns	ns
Diag.o.cB	ns		ns
Diag.m.cB	ns	ns	
Kontraste der Cattellscores	klU	Diag.o.cB	Diag.m.cB
klU		ns	ns
Diag.o.cB	ns		ns
Diag.m.cB	ns	ns	
Klasse 2 „ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“			
Kontraste der Cattellscores	klU	Diag.o.cB	Diag.m.cB
klU		ns	ns
Diag.o.cB	ns		*
Diag.m.cB	ns	*	
Kontraste der Bentonscores	klU	Diag.o.cB	Diag.m.cB
klU		ns	***
Diag.o.cB	ns		**
Diag.m.cB	***	**	
Kontraste des Lebensalters	klU	Diag.o.cB	Diag.m.cB
klU		ns	ns
Diag.o.cB	ns		ns
Diag.m.cB	ns	ns	
Kontraste der Bildungsgrade	klU	Diag.o.cB	Diag.m.cB
klU		ns	ns
Diag.o.cB	ns		ns
Diag.m.cB	ns	ns	
Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“			
Kontraste der Ravenscores	klU	Diag.o.cB	Diag.m.cB
klU		ns	ns
Diag.o.cB	ns		ns
Diag.m.cB	ns	ns	

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)
klU = ohne Diagnose, Diag.o.cB. = Diagnose ohne cerebrale Beteiligung
und Diag.m.cB. = Diagnose mit cerebraler Beteiligung

entnommen aus *Tab.A-44*

Erläuterung der *Tabellen 5-88* bis *5-92*:

In den Varianzanalysen der Testleistungsmaße ergeben sich nur für die **Klasse 2 „ältere Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“** signifikante Gruppenunterschiede beim Cattellscore (*Tabelle 5-90*) und beim Bentonscore (*Tabelle 5-91*), nicht jedoch beim Ravenscore (*Tabelle 5-89*).

Die Klasse 2 ist aufgrund dieser festgestellten Scoreunterschiede von besonderem klinischem Interesse.

Der **Cattellscore** kann in den Einzelvergleichen dieser Klasse[63] erfolgreich zwischen den beiden Patientengruppen differenzieren.

Mit dem **Bentonscore** gelingt die Unterscheidung der leistungsschwächeren Patienten **mit** cerebraler Beteiligung von den beiden leistungsstärkeren Gruppen Patienten **ohne** cerebrale Beteiligung und Personen ohne klinische Auffälligkeiten. Das Fehlen signifikanter Alters- und Bildungsunterschiede validiert dieses Resultat, dass sich beim Bentontest im Einklang mit der landläufigen Meinung befindet (z.B. BERG, FRANZEN & WEDDING, 1987). Es muss aber als schwacher Effekt mit wenig diagnostischer Relevanz bezeichnet werden.

5.6.2.3.2 Vergleich der Klassenzusammensetzungen

Die Zusammensetzungen der latenten Klassen des probabilistischen Auswertungsmodus werden in Balkendiagrammen veranschaulicht. Die enthaltenen Diagnosegruppen[64] sind in *Abbildung 5-25* dargestellt. Die Unterteilung nach Krankheitsbildern ergibt *Abbildung 5-26*.

5.6.2.3.2.1 Diagnostische Gruppen

Die prozentualen Abweichungen vom klassenspezifischen Erwartungswert nach Diagnosegruppen sind in *Abbildung 5-25* wiedergegeben. Die Zellbesetzung der Diagnosegruppen ist in *Tabelle 5-93* aufgeführt.

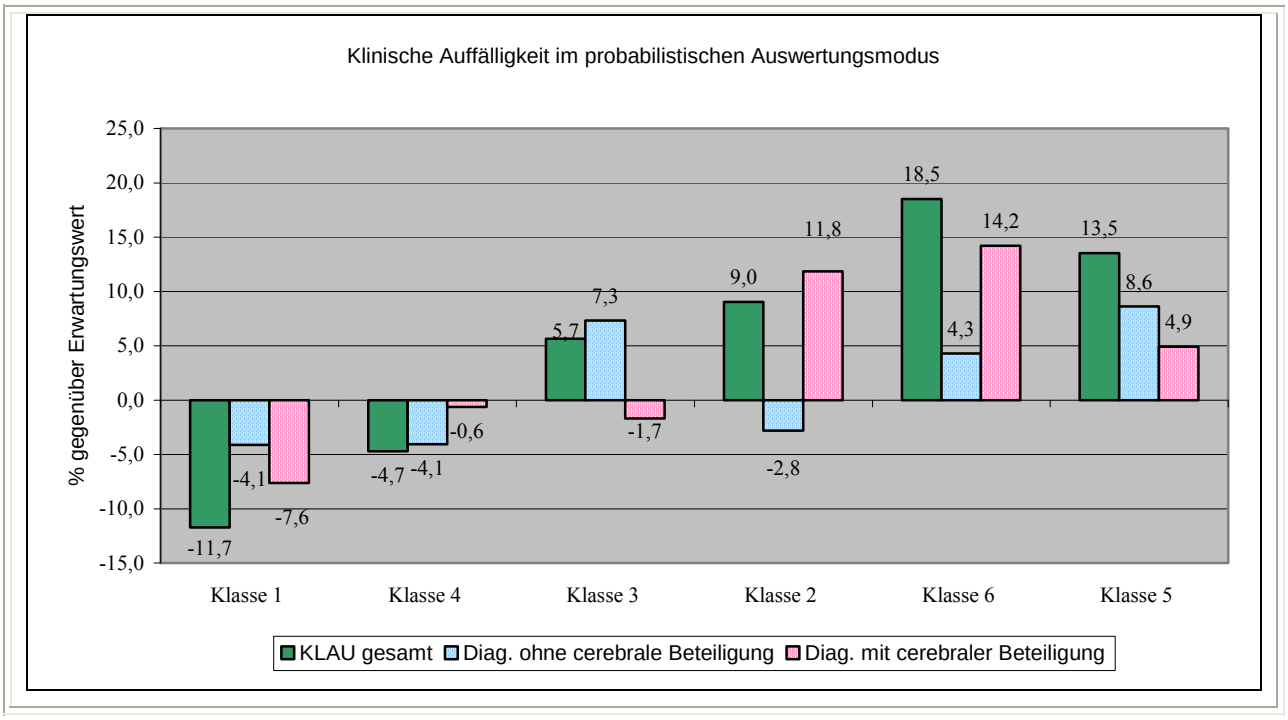
Die zugehörigen Klassenbezeichnungen des probabilistischen Auswertungsmodus sind der *Tabelle 5-94* zu entnehmen.

---

[63] *Tabelle 5-92,* ausführlich im Anhang in *Tabelle A-44*

[64] klinische Auffälligkeit allgemein (KlAU), Diagnose **ohne** und **mit** cerebraler Beteiligung

Abbildung 5-25 Probabilistischer Auswertungsmodus: Diagnosegruppen innerhalb der Klassen



KLAU gesamt = Diagnose **ohne** cerebrale Beteiligung + Diagnose **mit** cerebraler Beteiligung

Tabelle 5-93 Probabilistischer Auswertungsmodus: Zellbesetzungen der Diagnosegruppen

Teilstichprobe „Testbatterie“ (N=590)	**Klasse 1**		**Klasse 4**		**Klasse 3**		**Klasse 2**		**Klasse 6**		**Klasse 5**		(Klassen geordnet nach Testleistungs fähigkeit)	
Diagnose-gruppe	N	+/- %	N	+/- %	N	+/- %	N	+/- %	N	+/- %	N	+/- %	**Σ**	**%**Erw
Diagnose ohne cerebrale Beteiligung	40	-4,1	17	-4,1	26	+7,3	15	-2,8	11	+4,3	24	+8,6	**133**	22,5
Diagnose mit cerebraler Beteiligung	32	-7,6	20	-0,6	18	-1,7	26	+ 11,8	15	+ 14,2	21	+4,9	**132**	22,4
klinisch auffällige Personen (KLAU) insgesamt **Σ**	72	-11,7	37	-4,7	44	+5,6	41	+9,0	26	+ 18,5	45	+ 13,5	**268**	44,9
klinisch unauffällige Personen	145	+11,7	55	+4,7	43	-5,6	35	-9,0	15	- 18,5	32	- 13,5	**325**	55,1
Σ	**217**		**92**		**87**		**76**		**41**		**77**		**590**	100%

Tabelle 5-94 Probabilistischer Auswertungsmodus: Klassenbezeichnungen

Klasse 1	„jüngere Personen mit hohem Leistungsniveau“
Klasse 4	„jüngere Personen mit leicht reduziertem Leistungsniveau
Klasse 3	„jüngere Personen mit reduziertem Leistungsniveau
Klasse 2	„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“
Klasse 6	„ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“
Klasse 5	„ältere Personen mit Leistungseinbrüchen in allen drei Tests“

Erläuterung der *Abbildung 5-25*:

Der Anteil klinisch auffälliger Personen nimmt mit Ausnahme der leistungsschwächsten Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“ proportional zu abnehmender Testleistung zu.

In der leistungsstärksten Klasse 1 „jüngere Personen mit hohem Leistungsniveau“ finden sich deutlich weniger klinisch auffällige Personen als vom Erwartungswert prognostiziert.

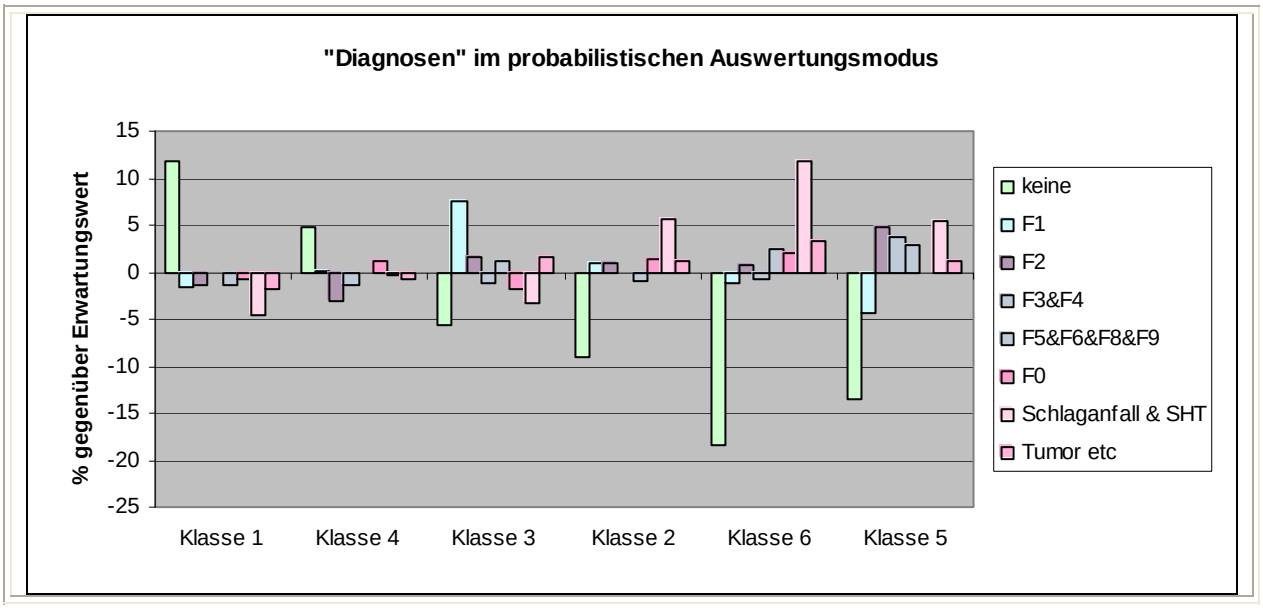
In der Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“ befinden sich klinisch auffällige Personen prozentual am häufigsten.

5.6.2.3.2.2 Krankheitsbilder

Die *Abbildung 5-26* zeigt die Krankheitsbilder[65] in den Klassen des probabilistischen Auswertungsmodus. Die Beträge der Ordinate entsprechen den Abweichungen vom klassenspezifischen Erwartungswert der Häufigkeit jedes einzelnen Krankheitsbildes. Die Zellbesetzung der Krankheitsbilder ist der *Tabelle 5-95* zu entnehmen.

[65] einzelne Zellen sind nur gering besetzt, deshalb ist die *Abbildung 5-26* nur unter Vorbehalten interpretierbar

Abbildung 5-26 Probabilistischer Auswertungsmodus: Krankheitsbilder innerhalb der Klassen



Erläuterung der *Abbildung 5-26*:

In Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“ ist ein hoher Patientenanteil mit Störungen durch psychotrope Substanzen (F1 v.a. Alkohol)[66] enthalten.

Der Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“ sind überdurchschnittlich viele Schlaganfall- und Schädelhirntraumapatienten[67] zugeordnet.

Obwohl die Klasse 5 die leistungsschwächste Klasse ist, finden sich weniger Patienten **mit** cerebralen Beeinträchtigungen als in Klasse 6, welche die mit Abstand besseren Leistungen in Raven- und Bentontest erzielt[68]. Tendenziell unterscheiden sich beide Klassen in den Maßen Score und Anzahl bearbeiteter Items beim Speed-Power-Test Cattell (vgl. auch *Abbildung 5-18*). Diese Maße sind konträr zu den Leistungen in Raven- und Bentontest (vgl. *Tabelle 5-79*). Die Scoreunterschiede der Klassen 5 und 6 bestätigen den hinlänglich bekannten Befund, dass Personen nach Hirnverletzungen deutliche Leistungseinbußen, insbesondere bei Tests mit ausgeprägter Speedkomponente z.B. dem Cattelltest aufweisen, weil aufgrund der Störungen eine allgemeine Verlangsamung der psychischen Abläufe resultiert (vgl. SCHÖTTKE, 1988).

Tabelle 5-95 gibt die Zellbesetzungen der Krankheitsbilder in den einzelnen Klassen wieder.

---

66 explizite Diagnosen in *Tabelle A-45* im Anhang

67 explizite Diagnosen in *Tabelle A-46* im Anhang

68 Im direkten Vergleich ist Klasse 6 im Ravenscore hochsignifikant besser, und erreicht ebenso beim Bentontest sehr signifikant höhere Scores als Klasse 5. Damit zusammenhängend - wie zu erwarten – ist auch die signifikant geringere Benton-Fehleranzahl (BTF) (vgl. *Tabelle 5-79*)

Tabelle 5-95 Probabilistischer Auswertungsmodus: Zellbesetzungen der Krankheitsbilder

Teilstichprobe „Testbatterie“ (N=590)	**Klasse 1**		**Klasse 4**		**Klasse 3**		**Klasse 2**		**Klasse 6**		**Klasse 5**			
Krankheitsbild	N	+/- %	N	+/- %	N	+/- %	N	+/- %	N	+/- %	N	+/- %	**Σ**	**%**Erw
ohne	145	+ 11,7	55	+4,7	43	-5,7	35	-9,0	15	- 18,5	32	- 13,5	**325**	55,1
F1	20	-1,6	10	0,0	16	+7,5	9	+1,0	4	-1,1	5	-4,4	**64**	10,8
F2	6	-1,5	1	-3,2	5	+1,5	4	+1,0	2	+0,6	7	+4,9	**25**	4,2
F3&F4	17	-0,1	6	-1,4	6	-1,1	6	-0,1	3	-0,6	9	+3,7	**47**	8,0
F5&F6&F8&F9	2	-1,5	2	-0,2	3	+1,1	1	-1,1	2	+2,5	4	+2,8	**14**	2,4
F0	10	-0,6	6	+1,3	3	-1,8	5	+1,3	3	+2,1	4	-0,1	**31**	5,3
Schlaganfall & SHT	12	-4,6	9	-0,4	6	-3,3	12	+5,6	9	+ 11,8	12	+5,4	**60**	10,2
Tumor etc	5	-1,8	3	-0,8	5	+1,7	4	+1,2	3	+3,2	4	+1,1	**24**	4,1
Σ	**217**		**92**		**87**		**76**		**41**		**77**		**590**	100

5.6.2.3.3 Validierung der 6-Klassen-Lösung

Der Einfluss der Prädiktorvariablen Altersgruppe, Bildungsgrad, Diagnosegruppe einschließlich Interaktionen auf die Klassenbildung wird mit nominal logistischen Regressionen untersucht. Das Vorgehen ist analog zum klassischen Auswertungsmodus.

Die Zuordnung der wenig besetzten Altersgruppe 4 (>70.LJ) zur Altersgruppe 3 (60.-70.LJ) und der wenigen Personen aus Bildungsgrad 0 (ohne Schulabschluss) zum Bildungsgrad 1 (Hauptschule) reduziert die Anzahl der Freiheitsgrade.

Die Nichtberücksichtigung der dreifachen Interaktion aus Altersgruppe, Bildungsgrad und Diagnosegruppe verhindert eine Explosion der Parameteranzahl bei der Berechnung mit der Prädiktorvariablenkombination.

Die Prädiktorvariable „Diagnosegruppe“ beinhaltet die Dreiteilung der Diagnosen in keine Diagnose, Diagnose **mit** cerebraler Beteiligung und Diagnose **ohne** cerebrale Beteiligung.

Tabelle 5-96 fasst die Ergebnisse der logistischen Regressionen mit Likelihood-Ratio-Tests samt Entropiereduktionen zusammen. Die Wald-Effekt-Tests der Prädiktorvariablen-kombination sind der *Tabelle 5-97* zu entnehmen.

Tabelle 5-96 Probabilistischer Auswertungsmodus: Entropiereduktion (Kriterium: 6-Klassen-Lösung)

Kriterium: 6-Klassen-Lösung des probabilistischen Auswertungsmodus Stichprobe Testbatterie - nominal logistisches Modell					
Prädiktor	**Modell-Test (N=590)**			**Entropiereduktion**	
	FG	χ^2	Prob> χ^2	%	(adj.%)
Altersgruppe	15	93,27	<,0001***	4,78	(3,24)
Bildungsgrad	10	110,06	<,0001***	5,64	(4,61)
Diagnosegruppe	10	34,55	0,0001***	1,77	(0,00)
Prädiktorvariablen-kombination Altersgruppe, Bildungsgrad, Diagnosegruppe und Interaktionen	115	323,04	<,0001***	16,54	(4,76)

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle 5-97 Probabilistischer Auswertungsmodus: Wald-Effekt-Tests (Kriterium: 6-Klassen-Lösung)

Kriterium: 6-Klassen-Lösung des probabilistischen Auswertungsmodus Stichprobe Testbatterie - nominal logistisches Modell			
Wald Effekt Tests der Prädiktorvariablenkombination			
Variable	FG	Wald- χ^2	Prob> χ^2
Altersgruppe	15	22,138	0,1042
Bildungsgrad	10	23,229	0,0099**
Diagnosegruppe	10	14,228	0,1629
Altersgruppe × Bildungsgrad	30	21,737	0,8634
Altersgruppe × Diagnosegruppe	30	28,325	0,5532
Bildungsgrad × Diagnosegruppe	20	9,235	0,9800

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

In den Wald Effekt Tests der Prädiktorvariablenkombination zeigt sich unter hochsignifikanter Modellgeltung separater und kombinierter Prädiktoren ein Analogon zum klassischen Auswertungsmodus:

Wie dort wird auch im probabilistischen Auswertungsmodus nur die Prädiktorvariable „Bildungsgrad“ der Versuchspersonen sehr signifikant. Zur Prädiktion der sechs indizierten Klassen liefern in der Kombination der Prädiktorvariablen weder die Altersgruppe, die Diagnosegruppe noch die Interaktionen relevante Beiträge. Auch im probabilistischen Auswertungsmodus kann infolgedessen die Position einer Testperson entlang des durch die

Klassen gebildeten nahezu eindimensionalen Konstruktes „Testleistungsfähigkeit“ am sichersten durch ihren Bildungsgrad vorhergesagt werden (vgl. *Abbildung 5-18*).
Im Vergleich zum klassischen Auswertungsmodus ist die Anzahl der Freiheitsgrade beim probabilistischen Auswertungsmodus mehr als doppelt so groß, was zu ungenaueren Schätzungen der Modellparameter und der Entropiereduktionen führt. Dieser Effekt zeigt sich in den vergleichsweise kleinen Werten der adjustierten Entropiereduktionen.

Im nächsten *Abschnitt 5.6.*3 folgt eine zusammenfassende Gegenüberstellung der Ergebnisse beider Auswertungsmodi in *Tabelle 5-98*.

5.6.3 Zusammenfassung der Prädiktionskennwerte der Klassen der Testbatterie beider Auswertungsmodi

Die *Tabelle 5-98* fasst die Kennwerte zur Prädiktion der Klassen des klassischen und des probabilistischen Auswertungsmodus zusammen (entnommen aus *Tabelle 5-72* und *5-73* sowie *Tabelle 5-96* und *5-97*).

Tabelle 5-98 Zusammenfassung der Entropiereduktionen und Wald-Effekt-Tests beider Auswertungsmodi

Kriterium: 3-Klassen-Lösung des klassischen Auswertungsmodus Stichprobe Testbatterie - nominal logistisches Modell			
Prädiktor	Signifikanz des nominal logistischen Modells	Entropiereduktion % (adj.%)	Signifikante Wald-Effekt-Tests
Altersgruppe	***	5,70 (4,76)	-
Bildungsgrad	***	7,02 (6,39)	-
Diagnosegruppe	***	2,89 (2,26)	-
Kombination einschließlich Interaktionen	***	18,63 (11,44)	Bildungsgrad***
Kriterium: 6-Klassen-Lösung des probabilistischen Auswertungsmodus Stichprobe Testbatterie - nominal logistisches Modell			
Prädiktor	Signifikanz des nominal logistischen Modells	Entropiereduktion % (adj.%)	Signifikante Wald-Effekt-Tests
Altersgruppe	***	4,78 (3,24)	-
Bildungsgrad	***	5,64 (4,61)	-
Diagnosegruppe	***	1,77 (0,00)	-
Kombination einschließlich Interaktionen	***	16,54 (4,76)	Bildungsgrad**

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Die latenten Klassen der Testbatterie in beiden Auswertungsmodi bilden ein nahezu eindimensionales Konstrukt (vgl. *Abschnitt 5.6.1* und *Abschnitt 5.6.2*). Dieses wird bei beiden Auswertungsmodi als „Testleistungsfähigkeit“ bezeichnet. Einschränkungen der Linearität des Konstruktes ergeben sich lediglich beim probabilistischen Auswertungsmodus.

Die Prädiktion der Position einer Testperson auf dem Kontinuum „Testleistungsfähigkeit“ gelingt laut Entropiereduktionen und Wald-Effekt-Tests der Prädiktorkombination am sichersten mit dem personenspezifisch erreichten Bildungsgrad.

Weder die Altersgruppe, die Diagnosegruppe noch die Interaktionen spielen in der Prädiktorvariablenkombination eine signifikante Rolle.

Die im Vergleich mit den unkorrigierten Entropiereduktionen kleineren Werte der adjustierten Entropiereduktionen resultieren aus der großen Anzahl nötiger Schätzparameter. Insbesondere beim probabilistischen Auswertungsmodus werden deshalb die Werte der Entropiereduktionen immens nach unten korrigiert.

6 DISKUSSION

Die klassischen psychometrischen Verfahren Raven-, Benton- und Cattelltest wurden separat und als Testbatterie eingehenden testtheoretischen Prüfungen unterzogen.
Die zentrale Frage war, ob eine von Mischverteilungsmodellen identifizierte Art der Testbearbeitung klinisch-prognostisch verwertbaren Informationsgewinn bietet.
Die analysierten Tests sind historisch unter klassischer Perspektive konstruiert worden und messen eine latente Fähigkeit per Summenscore. Im Mittelpunkt des Interesses steht die Frage, ob aufwendige probabilistische Auswertungsmethodik der erhobenen Testdaten eine verbesserte Prädiktion klinischer Kriterien ermöglicht.
Die Fragestellung der Arbeit ergab sich unter dem Aspekt der Kreuzvalidierung testdiagnostisch „sensationeller“ Befunde von KINDERMANN (2001). Dort wurde mit dem Mixed-Rasch-Modell bei einer Analyse einer neuen Form des Bentontests eine 2-Klassen-Lösung indiziert, wobei in der leistungsschwächeren der beiden Klassen 93 Prozent sämtlicher Patienten enthalten waren. Die seinerzeit verwendete N=1124 Versuchspersonen umfassende Gesamtstichprobe wurde im Verlauf der Neuentwicklung des Bentontests nach und nach seit 1985 erhoben.
Für die vorliegende Arbeit wurde wegen unklarer Datenlage der damaligen Befunde eine völlig neue, N=615 klinisch auffällige und unauffällige Versuchspersonen umfassende Gesamtstichprobe herangezogen. Siebenundneunzig Prozent der Versuchspersonen hatten alle drei genannten Verfahren bearbeitet. Dies ermöglichte zusätzlich eine Testprofilanalyse der gesamten Testbatterie mittels Latenter-Klassen-Analyse im klassischen und probabilistischen Auswertungsmodus.

Im Folgenden werden die erhaltenen Resultate der Untersuchung im Hinblick auf die einzelnen Fragestellungen überprüft. Abschließend folgt eine allgemeine Diskussion.

6.1 Fragestellung 1: Werden von stochastischen Testmodellen bei den einzelnen Testverfahren unterschiedliche Arten der Testbearbeitung (latente Klassen) identifiziert?

Die beiden Power-Tests Raven und Benton erweisen sich beim „Andersen-Test“, mit den manifesten Teilungskriterien „Testleistung“ und „diagnostische Zugehörigkeit“ als nicht raschhomogen. Die Testbearbeitungen lassen sich in mehrere latente raschhomogene Subpopulationen unterteilen, weil innerhalb der Daten testleistungsunabhängige qualitative Zusatzinformation nachgewiesen werden kann.

Die beiden informationstheoretischen Maße AIC und BIC zeigen beim **Raventest** eine 3-Klassen-Lösung des Mixed-Rasch-Modells an.

Beim **Bentontest** indiziert der AIC eine 3-Klassen-, der BIC die Einklassen-Lösung Raschmodell.

Die 3-Klassen-Lösungen werden bei beiden Verfahren mit Likelihoodquotiententests abgesichert und den weiteren Analysen als mögliche Modelle zugrunde gelegt.

Der **Cattelltest** ist im Gegensatz zu Raven- und Bentontest als Speed-Power-Test für Raschanalysen ungeeignet. Hier indiziert das informationstheoretische Maß AIC eine 4-Klassen-, der BIC eine 2-Klassen-Lösung der Latenten-Klassen-Analyse. Weil der Fokus dieser Arbeit auf die möglichst differenzierte Diagnostik von Testpersonen gerichtet ist, wurde für die vom AIC angezeigte Klassenanzahl entschieden.

6.2 Fragestellung 2: Wie können die von stochastischen Testmodellen identifizierten Arten der Testbearbeitung bei den drei Tests interpretiert werden?

Die vom jeweiligen Modell identifizierten latenten Klassen wurden bei allen verwendeten Tests den durchschnittlich erreichten Testscores nach angeordnet und benannt.

Eine Alters- und Bildungsabhängigkeit dieser klassenspezifischen Testleistungen wird aufgezeigt und lässt sich bis auf wenige Ausnahmen auch statistisch belegen.

Unterschiede der Klassenzusammensetzungen aus klinisch auffälligen und unauffälligen Personen werden jedoch nicht deutlich, was insbesondere die in Frage stehenden Ergebnisse beim Bentontest von KINDERMANN (2001) nicht kreuzvalidiert. Die berechneten klassenspezifischen Itemparameter ergeben bei keinem der Tests Hinweise auf unterschiedliche Bearbeitungsstrategien. Es zeigen sich lediglich einzelne auffällige

Testitems, deren Lösung oder Nichtlösung für die latente Klassenzuteilung von Versuchspersonen in besonderem Maße wichtig ist.

Im Einzelnen:

Die drei latenten Klassen des **Raventests** werden entsprechend ihrer Testleistung als „Könner", „Durchschnittliche" und „Nicht-Könner" bezeichnet, wobei die „Könner" die leistungsstärkste und die „Nicht-Könner" die leistungsschwächste Klasse sind. Diese Klassen unterscheiden sich durch signifikante Score-, Alters- und Bildungsunterschiede.

Raven-Item 7 zeigt klassenspezifische, graphisch auffällige Itemparameter und ist für die kleinste Klasse der „Nicht-Könner" das mit Abstand schwerste Item.

Beim **Bentontest** sind zwei der drei vom Mixed-Rasch-Modell identifizierten Klassen nahezu identisch. Unterschiede zwischen den Klassen der „Könner A" und „Könner B" beruhen lediglich auf unterschiedlichen Itemparametern und können nicht auf statistisch signifikante Score-, Bildungs-, Alters-, oder Klassenzusammensetzungsdifferenzen zurückgeführt werden.

Die Personen der „Könner B" lösen verglichen mit den „Könnern A" besonders häufig Benton-Item 4.

Die Klasse der „Nicht-Könner" weist im Vergleich zu den beiden „Könner-Klassen" eine signifikant schlechtere Bentontestleistung, ein niedrigeres Bildungsniveau und ein höheres Alter auf. In dieser leistungsschwächsten Bentonklasse können Unterschiede der Klassenzusammensetzung auch statistisch nachgewiesen werden. Der Klasse der „Nicht-Könner" sind im Vergleich mit den „Könnern B" prozentual mehr klinisch auffällige Personen zugeordnet.

Für die leistungsschwächste Klasse der „Nicht-Könner" ist Benton-Item 14 am schwierigsten. Für die leistungsstärkste Klasse der „Könner A" ist dasselbe Item am leichtesten.

In der vorliegenden Arbeit geben die Ergebnisse der Kreuzvalidierung im Gegensatz zu KINDERMANN (2001) keine entscheidenden Hinweise, die sich mit unterschiedlichen Benton-Bearbeitungsstrategien in Zusammenhang bringen lassen. Erklärungen für dieses negative Resultat werden in *Abschnitt 6.7* der allgemeinen Diskussion behandelt.

Beim **Cattelltest** werden die vier identifizierten Klassen entsprechend der Testleistung geordnet und als „Könner", „Durchschnitt A", „Durchschnitt B" und „Nicht-Könner" bezeichnet. Die Scoreunterschiede aller Klassen und die Unterschiede der Bildungs- und Alterszusammensetzungen können bis auf wenige Ausnahmen auch statistisch belegt werden.

Auch beim Cattelltest ergeben sich keine Hinweise auf unterschiedliche Bearbeitungsstrategien. Nur Cattell-Item 8 zeichnet sich durch besonders unterschiedliche klassenspezifische Lösungswahrscheinlichkeiten aus.

Die ersten vier Items des Cattelltests zeigen geordnete Lösungsprofile der Latenten Klassen. Die Auswirkung der Speedkomponente des Cattelltests auf die Klassenbildungen wird erst im späteren Testverlauf deutlich. Die gültige Anwendung eines quantitativen Testmodells mit Summenscore kann jedoch auch für diese vier Items nicht belegt werden, weil die Itemschwierigkeiten in den einzelnen Klassen zu stark variieren.

6.3 Fragestellung 3: Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen der einzelnen Tests?

Innerhalb der Klassen wurde geprüft, ob sich die Testscores der Patientengruppen „klinische Auffälligkeit“ und „cerebrale Beteiligung“ von klinisch unauffälligen Personen unterscheiden. Bei vergleichbarer Alters- und Bildungsgradezusammensetzung zeigt sich Folgendes:

Beim **Raventest** erreichen klinisch unauffällige Personen nur in der größten und leistungsstärksten Klasse der „Könner“ mindestens signifikant höhere Testscores als die beiden Patientengruppen.

Beim **Bentontest** unterscheiden sich die klinisch unauffälligen Personen von den zuvor erwähnten Patientengruppen mindestens sehr signifikant mit deutlich besseren Leistungen ebenfalls in der größten, jedoch zweitbesten Klasse „Könner B“.

Die Scores im **Cattelltest** sind nur in der leistungsschwächsten Klasse „Nichtkönner“ bei den zugeordneten Patienten mit „cerebraler Beteiligung“ und <u>nicht</u> beim Merkmal „klinische Auffälligkeit“ signifikant kleiner.

6.4 Fragestellung 4: Können die identifizierten Arten der Testbearbeitung die Vorhersagevalidität der drei Tests verbessern?

Mit logistischen Regressionen wurden Kennwerte berechnet, um Verbesserungen der Vorhersagevalidität aufzuzeigen, wenn klassische und probabilistische Prädiktoren nicht nur separat, sondern auch kombiniert betrachtet werden.

Die Prädiktion der klinischen Kriterien „klinische Auffälligkeit“ und „cerebrale Beteiligung“ gelingt nach den Testdurchführungen nur unwesentlich besser als mit Zufallsentscheidungen.

Der Einfluss der klassischen und probabilistischen Prädiktoren auf den Prädiktionserfolg erweist sich in den vorliegenden Fällen als voneinander unabhängig, d.h. Interaktionen aus

Summenscore und Klassenzugehörigkeit spielen bei keinem der analysierten Tests eine signifikante Rolle.

Beim **Raventest** ergibt sich weder unter Einbezug des klassischen Prädiktors Testscore, des probabilistischen Prädiktors Klassenzugehörigkeit, noch kombinierter Prädiktoren, eine signifikant verbesserte Prädiktion „klinische Auffälligkeit“ oder „cerebrale Beteiligung“.

Der Raventest erfasst demnach eine für pathologische Prozesse wenig anfällige, äußerst resistente Fähigkeit, wenn er ohne Zeitbegrenzung bearbeitet wird.

Benton- und Cattelltest zeigen im Gegensatz dazu zumindest mit klassischen und kombinierten Prädiktoren signifikante, aber letztlich wenig relevante Verbesserungen der Prädiktion „klinische Auffälligkeit“ und „cerebrale Beteiligung“.

Bei den klassischen Prädiktoren bescheinigen die Ergebnisse der logistischen Regressionen dem Bentonscore für den klinischen Gebrauch durchgehend solide Vorteile im direkten Vergleich mit dem Cattellscore. Die Kennwertunterschiede verdeutlichen sich, wenn es um die Aufdeckung allgemeiner hirnorganischer Störungen geht, was dem Kriterium „cerebrale Beteiligung“ entspricht.

Werden klassische und probabilistische Prädiktoren im logistischen Modell kombiniert, ergeben sich gegenüber den separat betrachteten Testscores unwesentliche Zuwächse der Vorhersagevalidität.

Für die als Indikator der Vorhersagevalidität verwendete prozentuale Entropiereduktion lassen sich zwei Resultate bei der Berechnung mit kombinierten Prädiktoren unterscheiden:

Resultat 1: Die Vorhersage des Kriteriums „klinische Auffälligkeit“ gelingt bei Benton- und Cattelltest in vergleichbarer Größenordnung.

Resultat 2: Die Vorhersage des Kriteriums „cerebrale Beteiligung“ gelingt mit dem Cattelltest besser als mit dem Bentontest.

Dieser vermeintliche Vorteil des Cattelltests zur Prädiktion von Hirnorganikern bleibt jedoch aus zwei Gründen fraglich:

(1) Die Klassenparameter wurden aus der Gesamtstichprobe (N=615) bestimmt. Die zwecks Parallelisierung geschichtete Teilstichprobe `parallel-cerebral` (N=269) ist möglicherweise zu klein, um eindeutige Effekte der Klassenparameter bei der separaten Betrachtung der Patienten mit cerebraler Beteiligung zu erhalten.

(2) Ein gültiger direkter Vergleich der Entropiewerte ist nicht gestattet, weil die Klassenzugehörigkeiten von unterschiedlichen Modellen bestimmt worden sind[69].

Letztendlich tragen die insbesondere von den Mixed-Rasch-Modellen identifizierten Klassen der Tests weder zu einer signifikanten noch zu einer relevanten Verbesserung der Vorhersagevalidität bei.

6.5 Fragestellung 5: Ergibt sich durch eine Kombination der Einzeltests zu einer Testbatterie eine Verbesserung der Vorhersagevalidität?

Die Vorhersagevalidität verbessert sich durch den Einsatz der drei Einzeltests als Testbatterie im Vergleich mit den separat betrachteten Testergebnissen. Diese Vorteile, die sich aus der Zusammenlegung sämtlicher klassischer, probabilistischer oder kombinierter Prädiktoren als Testbatterie ergeben, sind letztlich nur marginal und werden durch die größere Anzahl nötiger Schätzparameter relativiert.

Die auffallend größten Entropiewerte resultieren bei der Prädiktion „klinischer Auffälligkeit“ mit 5,58 Prozent und „cerebraler Beteiligung“ mit 8,13 Prozent, wenn die Kombinationen aus klassischen Parametern, probabilistischen Parametern und deren Interaktionen verwendet werden.

6.6 Fragestellung 6: Können Lösungsmuster oder Typen der Testbatteriebearbeitung identifiziert und differentialdiagnostisch genutzt werden?

Bearbeitungstypen der gesamten Testbatterie wurden unter Anwendung eines klassischen[70] und probabilistischen Auswertungsmodus[71] mit Latenten Klassen Analysen identifiziert. Bei beiden Auswertungsmodi zeigen sich ähnliche Ergebnisse.

Im probabilistischen Auswertungsmodus werden auch wegen der größeren vom informationstheoretischen Maß AIC angezeigten Klassenanzahl zusätzlich zu quantitativen auch qualitative Bearbeitungsaspekte der Testbatterie aufgezeigt, die in *Abschnitt 6.6.2* behandelt werden.

---

[69] beim Power-Test Benton Mixed-Rasch, beim Speed-Power-Test Cattell Latente-Klassen-Analyse

[70] LCA der Testscorequartile

[71] LCA der Klassen und der klassenspezifischen Testscorequartile

6.6.1 Fragestellung 6.1: Werden unterschiedliche Bearbeitungstypen (latente Klassen) der gesamten Testbatterie identifiziert?

Bei der Testbatterie wurde mit Latenter-Klassen-Analyse geprüft, ob unterschiedliche Bearbeitungstypen aufgezeigt werden.

Für beide Auswertungsmodi wurde die gleiche Stichprobe „Testbatterie“ (N=590) verwendet.

Die Latente-Klassen-Analyse indiziert beim **klassischen Auswertungsmodus** mit beiden informationstheoretischen Maßen AIC und BIC eine 3-Klassen-Lösung.

Beim **probabilistischen Auswertungsmodus** zeigt der AIC eine 6-Klassen-, der BIC eine 2-Klassen-Lösung an.

Die vom AIC indizierte Typenanzahl wurde den weiteren Berechnungen für die differenziertere Betrachtung der Testpersonen zugrunde gelegt.

Die klassische 3-Klassen- und die probabilistische 6-Klassen-Lösung wurde als mögliche Beschreibung der Daten mit Likelihoodquotiententests abgesichert.

6.6.2 Fragestellung 6.2: Wie können die von der LCA identifizierten latenten Klassen interpretiert werden?

Im **klassischen Auswertungsmodus** ergeben sich signifikante und geordnete Testscoreunterschiede der drei indizierten Klassen. Diese wurden nach aufsteigender Testleistung folgendermaßen benannt:

„hohes Leistungsniveau“,
„mittleres Leistungsniveau“ und
„niedriges Leistungsniveau“

Die Klassen befinden sich entlang des eindimensionalen Konstruktes „Testleistungsfähigkeit“. Aufgrund der hohen Interkorrelationen der drei verwendeten Tests ist dieser Befund zu erwarten.

Die Testleistung steigt, je jünger das Lebensalter und je höher das Bildungsniveau der Testpersonen ist. In beiden Fällen sind signifikante klassenspezifische Unterschiede nachweisbar.

Die differentialdiagnostischen Möglichkeiten die sich aus den drei Klassen des klassischen Auswertungsmodus ergeben, werden in *Abschnitt 6.6.4* behandelt.

Auch mit dem **probabilistischen Auswertungsmodus** können weitgehend geordnete Klassen identifiziert werden. Wie im klassischen Auswertungsmodus befinden sich die sechs Klassen

entlang des nahezu eindimensionalen Konstruktes „Testleistungsfähigkeit“.

In Form von zwei Leistungsprofilüberlagerungen werden außer quantitativen, im Gegensatz zum klassischen Auswertungsmodus, jedoch auch zwei qualitative Bearbeitungsunterschiede deutlich, die sich folgendermaßen interpretieren lassen:

(1) Beim **Bentontest** zeigt sich ein signifikanter Leistungsknick bei den älteren, ansonsten testleistungsstarken Personen.

(2) Beim **Cattelltest** werden die schlechtesten Leistungen für die Klasse angezeigt, die in Raven- und Bentontest deutlich besser abschneidet als die leistungsschwächste der sechs Klassen.

Nach aufsteigender Testleistung angeordnet ergeben sich drei Klassen statistisch gleichen, eher jüngeren Alters (Altersdurchschnitt ca. 40. Lebensjahr) und drei ebenfalls statistisch altersgleiche Klassen mit eher älteren Personen (Altersdurchschnitt ca. 50. Lebensjahr).

Diese sechs Klassen wurden folgendermaßen bezeichnet:

„jüngere Personen mit hohem Leistungsniveau“,

„jüngere Personen mit leicht reduziertem Leistungsniveau“,

„jüngere Personen mit reduziertem Leistungsniveau“,

„ältere Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“,

„ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“ und

„ältere Personen mit Leistungseinbrüchen in allen drei Tests“

Die differentialdiagnostischen Möglichkeiten die sich aus den sechs Klassen des probabilistischen Auswertungsmodus ergeben, werden in *Abschnitt 6.6.4* behandelt.

6.6.3 Fragestellung 6.3: Wie unterscheiden sich klinisch auffällige von unauffälligen Probanden innerhalb der latenten Klassen in den einzelnen Testergebnissen?

Im **klassischen Auswertungsmodus** ergeben sich signifikante Scoreunterschiede nur in der leistungsschwächsten Klasse „niedriges Leistungsniveau“, die sich folgendermaßen darstellen:

Der Score des **Raventests** ermöglicht keine Unterscheidung klinisch auffälliger oder unauffälliger Personen.

Der Score des **Cattelltests** differenziert zwischen klinisch unauffälligen Personen und

Patienten **mit** cerebraler Beteiligung. Die Cattellscores der Patienten **mit** cerebraler Beteiligung sind sehr signifikant kleiner.
Der Score des **Bentontests** unterscheidet beide Patientengruppen von Personen ohne klinische Auffälligkeiten. Die Bentonscores der Patienten sind sehr signifikant kleiner.
Sämtliche Scoreunterschiede müssen vorsichtig interpretiert werden, weil unterschiedliche Zusammensetzungen der Bildungsgrade und Altersverteilungen in den Gruppen der untersuchten Klasse statistisch nachweisbar sind.

Im **probabilistischen Auswertungsmodus** ergeben sich signifikante Scoreunterschiede analog zum klassischen Auswertungsmodus ebenfalls nur in einer Klasse der „älteren Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest“. Die Differenzierung klinischer Gruppen ist in dieser Klasse mit dem Score des **Raventests** nicht möglich.
Der Score des **Cattelltests** ermöglicht signifikante Unterscheidungen klinisch auffälliger Personen **mit** cerebraler Beteiligung von den leistungsstärkeren klinisch auffälligen Personen **ohne** cerebrale Beteiligung.
Mit dem Score des **Bentontests** gelingt in dieser Klasse die hochsignifikante Differenzierung klinisch auffälliger Personen **mit** cerebraler Beteiligung von den beiden anderen Gruppen. Diese Resultate werden durch die statistisch vergleichbaren Alters- und Bildungsgrade in den Gruppen der untersuchten Klasse abgesichert.

6.6.4 Fragestellung 6.4: Bieten die identifizierten latenten Klassen der Testbatterie diagnostisch verwertbaren Informationsgewinn?

Im **klassischen Auswertungsmodus** zeigt sich in den Klassen eine relative Zunahme klinisch auffälliger Personen proportional zu abnehmender Testleistung. Klinisch auffällige Personen befinden sich deutlich seltener als erwartet in der Klasse mit dem höchsten, ein erhöhter Anteil ist in der Klasse mit dem niedrigsten Leistungsniveau zu verzeichnen.
Differentialdiagnostischer Nutzen ergibt sich aus der identifizierten 3-Klassenzuteilung der LCA nicht, da die Abweichungen vom Erwartungswert für sämtliche Diagnosen ähnlich und zudem kaum mehr als 5 Prozent übersteigen (vgl. *Abbildung 5-17* und *Tabelle 5-71*).
Der Einfluss auf die Klassenbildung der Testbatterie ist nur für die Prädiktorvariable „Bildungsgrad“ hochsignifikant, nicht jedoch für die Variablen „Altersgruppe“,

„Diagnosegruppe“[72] oder deren Interaktionen nachweisbar.
Diese Resultate weisen darauf hin, dass die Position einer Testperson entlang des eindimensionalen Konstruktes „Testleistungsfähigkeit“ am sichersten durch den Bildungsgrad bestimmt wird.

Im **probabilistischen Auswertungsmodus** ergibt sich wegen der größeren Klassenanzahl ein differenzierterer Befund.
Bis auf eine Ausnahme sind auch hier in den Klassen zunehmend mehr klinisch auffällige Personen proportional zu abnehmender Testleistung enthalten. Dieser Sonderfall betrifft den Testleistungsvergleich der leistungsschwächsten Klasse „ältere Personen mit Leistungseinbrüchen in allen drei Tests“ mit der leistungsstärkeren Klasse „ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“.
Die leistungsschwächste Klasse erzielt mindestens sehr signifikant niedrigere Scores in Benton- und Raventest. Entgegen der Erwartung sind klinisch auffällige Personen in dieser Klasse mit den schlechtesten Testleistungen seltener vertreten. Stattdessen befinden sich viele Patienten in der leistungsstärkeren Klasse „ältere Personen mit stark reduziertem Leistungsniveau und relativem Leistungseinbruch im Cattelltest“, insbesondere nach Schlaganfall- oder Schädelhirntrauma. Dies verdeutlicht die hinlänglich bekannte Tatsache, dass drastische Leistungseinbußen bei Tests mit deutlicher Geschwindigkeitskomponente z.B. Cattelltest validere Indikatoren bestimmter Hirnverletzungen darstellen als Normabweichungen bei reinen Powertests.

Differentialdiagnostisches Interesse verdient der verhältnismäßig höhere Patientenanteil mit Störungen durch psychotrope Substanzen F1 nach ICD-10 in der leistungsschwächsten der drei jüngeren Klassen „jüngere Personen mit reduziertem Leistungsniveau“. Im Vergleich mit der benachbarten zweitbesten Klasse „jüngere Personen mit leicht reduziertem Leistungsniveau“ wird der Bentontest mit signifikant kleineren Scores zum Test mit der größten diagnostischen Aussagekraft.

Auch im probabilistischen Auswertungsmodus ist ein sehr signifikanter Zusammenhang der identifizierten Klassen nur für die Prädiktorvariable „Bildungsgrad“, nicht jedoch für die Variablen „Altersgruppe“, „Diagnosegruppe“[73] oder den daraus gebildeten Interaktionen

[72] (drei Gruppen: klinisch unauffällig, klinisch auffällig **mit** und klinisch auffällig **ohne** cerebrale Beteiligung)

[73] (drei Gruppen: klinisch unauffällig, klinisch auffällig **mit** und klinisch auffällig **ohne** cerebrale Beteiligung)

nachweisbar. Dieses Resultat bedeutet, dass sich auch bei den sechs Klassen des probabilistischen Auswertungsmodus die Position einer Testperson entlang des annähernd eindimensionalen Konstruktes „Testleistungsfähigkeit“ am sichersten durch den Bildungsgrad bestimmt.

6.7 Allgemeine Diskussion

Die wissenschaftliche Psychologie hat sich gegen Ende des 19. Jahrhunderts aus dem Verbund mit der Philosophie gelöst und ist seither insbesondere methodologisch den eigenen, jedoch an exakten Wissenschaften angelehnten Wegen gefolgt. Diese bergen die Möglichkeit physikalisch-technischer Genauigkeit auf der Grundlage quantitativer Methodik in sich, als Konsequenz aber auch das Risiko eines Menschenbildes, dem der Mensch als Mensch weder entsprechen kann noch entsprechen darf.

Diesbezüglich machte der frühere Psychologe und spätere Philosoph Karl Jaspers Anfang des 20. Jahrhunderts darauf aufmerksam, dass zur Beschreibung von Seelischem niemals Kausalgleichungen wie in Physik und Chemie aufgestellt werden können. „Dies würde eine völlige Quantifizierung der untersuchten Vorgänge voraussetzen, die beim Seelischen, das seinem Wesen nach immer qualitativ bleibt, prinzipiell nie möglich ist, ohne den eigentlichen, nämlich den seelischen Untersuchungsgegenstand zu verlieren“ (JASPERS, 1953, S.251).

Die Entwicklung der Psychologie als Wissenschaft führte dessen ungeachtet zur unkritischen Übernahme klassischer, aus der Physik übernommener Messvorstellungen, deren Anwendung in der Psychologie nur dem Anschein nach problemlos ist. Bei genauer Betrachtung haben sich die folgenden anfangs unerkannten oder vernachlässigten Problemfelder ergeben:

Das klassische Messprinzip muss hinterfragt werden, weil es physikalisch-technische Vorstellungen ohne falsifizierbare Theoriegrundlage auf den Menschen überträgt. Eine solche Messmethodik verhindert möglichen testdiagnostischen Fortschritt und somit auch psychologische Erkenntnisse. Die Mängel des klassischen Ansatzes wurden von LUMSDEN (1976) in einer „Folterkammeranalogie“ am deutlichsten aufgezeigt.

Die Alternative von FISCHER (1974) war ein naturwissenschaftlicher Psychologieansatz. Dieser zielte darauf ab, dass psychologische Charakteristika des einzelnen Menschen in einer Testsituation unter einem theoretisch gültigen probabilistischen Ansatz mit dem Raschmodell empirisch abgesichert vermessen werden. Sämtliche Testverfahren, die in den *Tabellen A-47* und *A-48* aufgeführt sind, haben dem Modell von Rasch theoretisch entsprochen. Dabei darf nicht übersehen werden, dass die Absicherung der klassischen Strategie mittels Raschmodell auch bei scheinbar homogenen Probandenstichproben in der praktischen Anwendung unbefriedigend ist. Selbst falls nach Prüfung des Raschmodells in einer Personenstichprobe von einer gültigen Messung per Summenscore ausgegangen werden kann, sind die

Kreuzvalidierungen solcher Ergebnisse regelmäßig erfolglos geblieben. Die Interpretation von Testscores bleibt deshalb fraglich.

Mit POPPERS (2004) dreistufigem Schema aktiver wissenschaftlicher Methode, welches allgemein in Problem [1], Lösungsversuche [2] und Problemelimination [3] untergliedert wird, kann festgehalten werden:

Mit dem Lösungsversuch Summenscore der klassischen Theorie [2a] ist keine stimmige Beschreibung des einzelnen Menschen möglich. Mangels Falsifikationsmöglichkeit ist dieses Problem [1] nicht zu eliminieren. Die Ausprägungen latenter psychologischer Fähigkeiten und die Anzahl der beantworteten Testitems sind a priori nicht dasselbe.

Erst mit den Lösungsversuchen probabilistischer Theorie - insbesondere Raschmodell [2b] - wurde dieses störanfällige und unpräzise Zusammenspiel mathematisch benennbar und veränderbar. Dennoch ist es nicht gelungen die messtheoretischen Unzulänglichkeiten der klassischen Testtheorie zu überwinden[74]. Eine Gleichsetzung von Fähigkeit und gelöster Itemzahl wird in der Praxis immer problematisch bleiben.

Punktum, das Testleistungsmaß Summenscore [2b] ist für die anwendungsorientierte wissenschaftliche Psychologie auch künftig ein diagnostisch zweifelhafter, der Wirklichkeit nicht gerecht werdender Problemlösungsversuch.

Wenn die Aufstellung von befriedigenden Theorien als Zielsetzung von Wissenschaft gilt und das Falsifizierbarkeitskriterium als entscheidendes Charakteristikum wissenschaftlicher Methode (POPPER, 2004) in der Psychologie bedacht wird, dann gibt es für studierte und denkende Psychologen bislang weder einen Grund die Suche nach neuen, alternativen und dem Mensch angemesseneren Problemlösungsversuchen [2] einzustellen, noch einen *Grund* im philosophischen Sinne Heideggers[75].

Es besteht weiterhin Grund zu der Annahme, dass auch die „probabilistischen Testmodelle den gesellschaftlich hervorgetriebenen Messansprüchen nicht genügen. Eher schon den Bedürfnissen nach Legitimation (....)“ (GRUBITZSCH, 1991, S. 49).

Moderne Testtheorie hat diese Funktion für die klassische Testperspektive in zweifacher Hinsicht[76] übernommen:

---

[74] (also noch keine Problemelimination [3], d.V.)

[75] Der Satz vom Grund ist *der* Grundsatz aller Grundsätze. Dieser lautet „Nichts ist ohne Grund“. In dem von Heidegger erläuterten ursprünglichen Sinn dieses Satzes (HEIDEGGER, 1997) gehört es zum Bedenklichsten, wie studierte Psychologen mit jeder quantitativ bestimmten latenten Merkmalsausprägung oder jeder anderen bloß technisch konstruierten Vorstellung vom Menschen, am Wesen des Mensch-*Seins* und somit am Grundsatz des *Grundes* vorbeidenken.

[76] wenn klassischen Testanwendungen mehr als nur Unterhaltungswert zugeschrieben werden soll

i) Sie ist überlebenswichtiger „Airbag“ (MOLENAAR, 1997, S. 49) für die Testperson bei klassischen Testanwendungen.

ii) Sie ist außerdem der vermeintliche Rettungsring der testhistorisch gewachsenen, ins Vorbild der Physik verrannten Vermessung so genannter „wahrer“ Fähigkeitsausprägungen.

Zwei Überlegungen sollen die Konsequenzen dieses Schattendaseins moderner Testtheorie für die wissenschaftliche Psychologie verdeutlichen:

1.) Falls die probabilistische Testtheorie bislang nur zur Legitimation von Messungen nach physikalisch-technischem Vorbild benötigt wird, dann wird ihre Relevanz für die psychologische Diagnostik a priori falsch eingeschätzt.

2.) Wenn sich physikalisch-technische Messvorstellungen für die Untersuchung psychologischer Fragestellungen als unangemessen erweisen, dann gründet die wissenschaftliche Form der Psychologie bei der Vermessung latenter Konstrukte mit Testitems bislang auf theoretischen Annahmen, die angewendet auf den Menschen in ihrer bisherigen Interpretation falsch sind und korrigiert werden müssen.

WALTER (1991, S. 123) zieht bezüglich dieser Problematik folgende Schlussfolgerung:
„Eine psychologische Innovation, die diesen Namen verdiente, wäre erst erreicht, wenn die Erarbeitung und Anwendung probabilistischer Testmodelle mit einer expliziten (nicht bloß kosmetischen) Revision psychologischer Theorie einherginge“.

Die Richtung der nötigen Revision psychologischer Theorie soll im Folgenden umrissen werden. Sie entspricht der von Friedensreich Hundertwasser im Jahre 1958 als „Verschimmelungsmanifest“ verfassten Kritik an der herkömmlichen Architektur (HUNDERTWASSER, 2004). Dieselben Argumente werden im Rahmen der vorliegenden Arbeit auf herkömmliche Testkonstruktion und Testanwendung als unstimmigen Ausgangspunkt naturwissenschaftlicher Psychologie vom Menschen übertragen.

Das klassische Richtmaß Summenscore, mit dem psychologische Erkenntnisse über den Menschen festgestellt werden, erhält eine völlig andere Bedeutung, wenn es von einer Meta-Ebene betrachtet wird. Hinter diesem technischen Maß versteckt sich in Worten die folgende Denkrichtung:

„Bis wohin fährt der Test mit der Testperson?“[77]

Angewendet auf den Menschen beinhaltet dieser Ansatz einen folgenschweren Denkfehler. Paradoxerweise muss jede mit Testitems quantitativ bestimmte Fähigkeit einer Testperson aus der übergeordneten Perspektive anders aufgefasst werden, obwohl die „Testleistungsfähigkeit“ naturwissenschaftlich augenscheinlich als solide Größe angesehen wird und in der psychodiagnostischen Anwendung unheimlicherweise als brauchbares Konstrukt zu funktionieren scheint. Jede quantitativ bestimmte Ausprägung einer „Testleistungsfähigkeit“ lässt es zu, in einem positiven Rahmen als „Sklavenniveau“ (vgl. Fußnote Nr.77) und in einem negativen Rahmen als „bequemheitslüsterne Gehirnlosigkeit“ interpretiert zu werden. Folgerichtig wird dieses Maß zwar der „bequemheitslüsternen, gehirnlosen Massenameise“ gerecht, es bleibt aber jedem Menschen wesensfremd (vgl. HUNDERTWASSER, 2004).

Es ist das Konzept der modernen, probabilistischen Testtheorie, das mit Mischverteilungsmodellen und dem Prinzip der spezifischen Objektivität zur Klassenunterscheidung, die methodischen Voraussetzungen für eine Abkehr aus dieser messtechnischen Fata Morgana einer Wissenschaft vom Menschen bietet.

Der Leitgedanke eines neuen, gewandelten Test- und auch Menschenverständnisses muss umgekehrt zum klassischen Ansatz, folgendermaßen richtig gestellt werden:

„~~Bis~~ Wohin fährt die Testperson mit dem Test?“.

Weil die strenge Form der spezifischen Objektivität, die vom einfachen Raschmodell gefordert ist, nur im technischen Sachbereich erfüllt werden kann, aber im geistig-menschlich-psychologischen Kontext nicht erreicht werden darf[78], ist die Berücksichtigung leistungsunabhängiger Aspekte einer Testbearbeitung hier abhängig von der diagnostischen

[77] Ein bedeutsamer Druckfehler in RAUCHFLEISCH (1989, S. 59) charakterisiert diese testdiagnostische Situation der Testperson folgendermaßen: „Die Art der Normen wird sich jeweils nach der Art des zu erfassenden Testverhaltens und dem Grad seiner Quantifizierbarkeit richten müssen, wodurch das *Sklavenniveau* determiniert wird“ (Hervorhebung durch den Verfasser).

[78] d.h. die Übereinstimmung der Aussagen „Bis wohin fährt die Testperson mit dem Test“ und „Bis wohin fährt der Test mit der Testperson“.

Fragestellung ausdrücklich gefordert. Der Ausgangspunkt für die Entwicklung neuartiger diagnostischer Instrumente ist der im Test fest verankerte aktive Gestaltungsspielraum. Der zugestandene Messfehler spielt bei der Suche nach diagnostischer Information die wesentliche Rolle. Als elementarer Bestandteil des Testverfahrens erzeugt dieser das unabdingbare Reservoir für qualitative Zusatzinformationen. Den daraus entnommenen entscheidungsrelevanten Informationsanteil gibt die Testperson selbst vor. Dieser hängt vom Testverhalten ab und kann aktiv in unterschiedliche, diagnostisch aufschlussreiche Richtungen, latente Klassen, gelenkt werden. Ihm gilt, aufgezeigt mit Mischverteilungsmodellen, das eigentliche testdiagnostische Interesse. Nur wenn es aus wissenschaftlicher Sicht für bestimmte diagnostische Zwecke erforderlich bleibt, dann kann an zweiter Stelle zusätzlich für die annähernd sichere Bestimmung von statisch-sterilen quantitativen Leistungsgesichtspunkten - im Sinne moderner Schädelmessung (GOULD, 1999) - gesorgt werden.

Die vorliegende Arbeit gibt einen Beitrag für die von Wissenschaftlern unter Umständen angestrebte, aber bislang kaum erkennbare Zukunft von Testdiagnostik und daraus abgeleiteter psychologischer Erkenntnisse. Diese wissenschaftliche Erkundung einer Alternative zum klassischen Ansatz, die auch eine Abkehr davon bedeutet, war ein Abtasten eines neuen und in der Literatur bislang kaum durchdachten Testverständnisses, das sich aus zweifelhaftem Befund und Ansporn (KINDERMANN, 2001) ergeben hatte. Für die erfolgreiche Anwendung der eingesetzten Methode zum vorgestellten qualitativ orientierten Testformat waren die hierzu verwendeten Patienten- und Personenstichproben zu heterogen[79] und die hochkorrelierten Prototypen ursprünglichster klassischer Testtradition mit den großen Namen Raven-, Benton- und Cattell-Test keine günstige Ausgangsbasis.
Die nachweisbaren Effekte bei der Prädiktion klinischer Auffälligkeiten mit den verwendeten Testverfahren sind in dieser Studie sehr gering ausgefallen[80]. Dies könnte aus einer zu leistungsstarken oder testdiagnostisch geübten Patientenpopulation resultieren, stellt aber auch den entscheidungstheoretischen Ansatz von CRONBACH & GLESER (1965) unter Beweis. Diese Autoren halten es für erforderlich, den klassischen Ansatz entscheidend zu erweitern. CRONBACH & GLESER leugnen nicht die Nützlichkeit genauer Fähigkeitsschätzungen vor dem Hintergrund klassisch entwickelter Tests mit möglichst hoher Reliabilität und Validität. Der eigentliche Zweck aller Diagnostik ist es für die Autoren jedoch qualifizierte

[79] auch in Relation zum Umfang der in dieser Arbeit verwendeten Stichproben

[80] abzulesen an den kleinen Werten der Entropiereduktionen (PRU) und den Flächen unter den ROC-Kurven

Entscheidungen zu treffen. „Eine angemessene Testtheorie muss daher von einer systematischen Analyse der Entscheidungssituationen ausgehen“ (MICHEL & MAI, 1968, S.100).

Die vorliegenden Ergebnisse bestätigen, dass Mischverteilungsmodelle Personengruppen aus der Gesamtstichprobe identifizieren können, in denen unterschiedliche statistische Modelle Geltung beanspruchen. Die identifizierten latenten Klassen beruhen bei allen drei Tests auf der testtheoretisch nicht trivialen Tatsache, dass zusätzliche Information in den Daten vorliegt. Diese wäre im herkömmlichen Testleistungsmaß Summenscore nicht beachtet worden. Im ungünstigsten Fall, kann dies zu groben Verzerrungen der Testleistungsinterpretation wegen heterogener Personenpopulationen führen.

Bei den analysierten Testverfahren hat sich dieses klassisch kaum kalkulierbare Risiko nicht deutlich gezeigt. So kann zwar keinem der drei analysierten Tests die Raschhomogenität bescheinigt werden, die für die Personenmehrheit zur reliablen Fähigkeitsbestimmung erforderliche Eindimensionalität des Merkmalsraums und der manifesten Variablen (Items) ist jedoch mit Ausnahme einzelner Items gegeben. Das Maß Summenscore erweist sich für die Mehrzahl der Testpersonen als gute Annäherung an die Ausprägung einer jeweils vom Test definierten und der Testperson diktierten latenten Fähigkeitsdimension.

Die bislang theoretische Vermutung, dass probabilistische Methoden latente Klassen identifizieren können, um in unterschiedlichsten Anwendungsbereichen wichtige Informationen für diagnostische Entscheidungen bereitzustellen, ist im vorliegenden Fall bei ursprünglich klassisch konstruierten, also nahezu eindimensionalen Tests misslungen. Der Fokus war in dieser Arbeit auf die Vorhersage klinischer Kriterien gerichtet. Bei den drei analysierten Tests konnte keine valide Prädiktion von „klinische Auffälligkeit“ und „cerebrale Beteiligung“ unter Berücksichtigung der Klassenzugehörigkeiten belegt werden, weil in den Daten zu wenig verwertbare zusätzliche systematische und spezifische Varianzanteile verfügbar waren. Als Indikatoren pathologischer Prozesse erweisen sich die beim Raven-, Benton- und Catteltest mit Mischverteilungsmodellen identifizierten latenten Klassen den Summenscores unterlegen. Dieses Ergebnis ist bei annähernd eindimensionalen Tests zu erwarten.

Vorausgesetzt, dass die etwa dreiprozentige Verbesserung der Prädiktion für wünschenswert gehalten wird[81,82], bewähren sich auch die Summenscores in dieser Arbeit kaum, obwohl sie

---

[81] Cohen (1988) ist das PRU-Maß noch unbekannt. Für η^2 gibt er folgende Daumenregeln an: $\geq 1\%$ kleine, $\geq 6\%$ mittlere, $\geq 14\%$ große Effektgrößen.

[82] vgl. auch *Abschnitt 1.11* Kritische Anmerkungen zu Untersuchungen der Vorhersagevalidität von

sich speziell beim Bentontest in der Mehrzahl der Studien zur Vorhersage allgemeiner kognitiver und hirnorganischer Beeinträchtigungen als hilfreich erwiesen haben (KASCHEL, 1994).

Eine „allgemeine Störungssensibilität“ ergibt sich in der vorliegenden Arbeit analog zu den Ergebnissen von STECK et al. (1990, S.48) beim Bentontest, aber ebenso beim Cattelltest. Beide Tests erfüllen die unerlässliche Validitätsmindestanforderung für „klinische Tests“ nur wegen statistisch signifikanter Absicherung. Demzufolge sollte die Testanwendung eine bessere Voraussage eines Kriteriums ermöglichen als ihre Unterlassung (LIENERT & RAATZ, 1994).

Zwei Komponenten von Benton- und Cattelltest stellen sich zur Abklärung von klinischen Auffälligkeiten als besonders wichtig heraus:

i) Die **Zeitkomponente**[83].

ii) Der Faktor „**Ausmaß der Beteiligung des Arbeitsgedächtnisses**“[84].

Der **Cattelltest** beinhaltet als Speed-Power-Test die ausgeprägteste Geschwindigkeitskomponente der analysierten Testverfahren. Eine verlangsamte Bearbeitungsgeschwindigkeit erweist sich auch in dieser Arbeit als wichtiger Indikator bestimmter cerebraler Störungen. Einige Ergebnisse der vorliegenden Untersuchung lassen vermuten, dass mit dem Cattelltest Beeinträchtigungen nach Schädel-Hirn-Trauma oder Schlaganfall deutlicher aufgewiesen werden als mit dem Power-Test Benton.

Der **Bentontest** erfordert bei der Bearbeitung die deutlichste Beteiligung des Arbeitsgedächtnisses[85]. Im direkten Vergleich mit dem Cattelltest sind solide diagnostische Vorteile bei jüngeren Personen mit cerebral bedingtem Leistungsabbau im gesamten Fähigkeitsbereich, speziell nach Substanzmissbrauch erkennbar[86]. In Übereinstimmung mit

---

Testverfahren

[83] Bentontest: Power-Test mit 10 Sekunden Einprägezeit pro Item; Cattelltest: Speed-Power-Test mit dreiminütiger Bearbeitungszeit

[84] Beim Benton-, jedoch auch beim Cattelltest sind auf dieser Fähigkeitskomponente in der in *Abschnitt 4.7.6* dargestellten Hauptkomponentenanalyse hohe Faktorladungen zu verzeichnen.

[85] vgl. die Ergebnisse der Hauptkomponentenanalyse in *Abschnitt 4.7.6*

[86] Es sei darauf hingewiesen, dass dieser Vorteil auch aus der Itemanzahl (höhere Reliabilität) oder der Testdauer (größere Ermüdungserscheinungen der klinisch auffälligen Personen vom Einprägen der Benton-Items) resultieren kann: Der Bentontest beinhaltet 20 Items und dauert circa 15 Minuten, der Cattelltest beinhaltet nur 13 Items und dauert nur 3 Minuten.

anderen klinischen Studien (SCHEURER et al., 1994; KESSLER & KASCHEL, 1997; STECK, 2002) wird die in der Literatur erwähnte spezielle Eignung des Bentontests zur Unterscheidung hirnorganischer von psychiatrischen Erkrankungen deutlich überschätzt. Eine solche Differenzierungsfähigkeit des Bentontests zeigt sich in den vorliegenden Resultaten aber doch als schwacher Effekt mit wenig diagnostischer Relevanz in der latenten Klasse „ältere Personen mit reduziertem Leistungsniveau und relativem Leistungseinbruch im Bentontest" der Testbatterie nach probabilistischem Auswertungsmodus.

Für die klinische Anwendung von Cattel- und Bentontest bleibt klärungsbedürftig, ob sich relevante Leistungseinbußen nach krankheitsbedingten Störungen auch mit Längsschnittstudien belegen lassen.

Der **Raventest** hat sich zur testdiagnostischen Abklärung klinischer Auffälligkeiten als nicht hilfreich erwiesen. Der negative Befund resultiert vermutlich daraus, dass dieser in der vorliegenden Arbeit ohne Zeitdruck bearbeitet wurde. Wird der Raventest auf diese Weise gelöst, eignet er sich somit im klinischen Bereich für rechtspsychologische Einschätzungen eines prämorbiden Leistungsniveaus intellektueller Fähigkeiten von Opfern oder Straftätern. Seinem ursprünglichen Verwendungszweck im Bereich der Intelligenzmessung kommt diese Aufgabe sehr nahe, zumal die vorliegenden Ergebnisse bestätigen, dass mit dem Raventest ein überdauerndes, weil krankheitsresistentes Konstrukt gemessen wird.

Auch dieses Ergebnis bedarf der Überprüfung mit Längsschnittstudien.

Die identifizierten latenten Klassen der aus Raven-, Benton- und Cattelltest zusammengesetzten **Testbatterie** und die daraus resultierende Position der klinisch auf- und unauffälligen Testpersonen entlang des im klassischen Auswertungsmodus eindimensionalen und im probabilistischen Auswertungsmodus nahezu eindimensionalen Konstruktes „Testleistungsfähigkeit" werden deutlicher durch den Bildungsgrad als durch Alterseffekte oder klinische Handicaps bestimmt. Kognitive Operationen die Testpersonen eher ***er***lernen als ***ver***lernen, sind demnach entscheidend für die erfolgreiche Lösungskonstruktion der Testmaterialien.

Bei allen drei Tests zeigen sich mit den eingesetzten probabilistischen Methoden nur vage Hinweise auf interindividuelle Verarbeitungsmechanismen der Testitems, im Sinne unterschiedlicher Bearbeitungsstrategien, die einer gültigen Scoreinterpretation im Wege stehen (vgl. PUTZ-OSTERLOH, 1981).

Die zwei diagnostisch relevanten Strategien beim Bentontest, die der Ausgangspunkt dieser Arbeit gewesen sind (KINDERMANN, 2001), haben sich in der Kreuzvalidierung nicht bestätigt. Die aufgezeigten zwei „Bentonteststrategien" könnten sich dort deshalb ergeben

haben, weil die Itemreihenfolge im Laufe der jahrelangen Konstruktion der neuen 20-Item-Form des Bentontests geändert worden war. Anzunehmen ist, dass sämtliche Patienten schon früh in der Testentstehung mit einer anderen, eventuell auch von 20 auf 30 Items verlängerten Abfolge getestet wurden (STECK et al., 1990). Hinweise darauf gibt auch *Abbildung 1-1*. Diese *Abbildung* zeigt, dass der Klasse mit der „gezackteren" Kurve fast ausnahmslos Patienten zugeordnet wurden. Vermutlich kann dies als Ausdruck der neu angeordneten Itemabfolge interpretiert werden. Sie sind nicht in der eher stetig ansteigenden Itemschwierigkeitskurve beinhaltet, mit anscheinend regulärer Itemabfolge.
Das bemerkenswerte Ergebnis von KINDERMANN (2001) darf somit als Stichprobenartefakt bezeichnet werden. Beim Bentontest bietet sich folgender Erklärungsversuch an:
Die Schwierigkeit der Benton-Items wird für Testpersonen aus dem Konglomerat von der Itemschwierigkeit im eigentlichen Sinne und der Itemposition bestimmt. Es ist zu vermuten, dass bei späteren Itempositionen die Lösungswahrscheinlichkeiten auch von Aufmerksamkeitsveränderungen[87] beeinträchtigt werden. Diese Ansicht würde auch die neu berechnete Reihenfolge der Itemschwierigkeiten erklären, obwohl diese den früheren Resultaten von KINDERMANN (2001) entsprechend angeordnet worden war.

Im Rahmen dieser Arbeit soll noch einmal betont werden:
Ein Informationsverlust bei der Beschreibung von Testleistungen in einer Personenstichprobe ist testtheoretisch nur dann ausgeschlossen, wenn die Gültigkeit des einfachen Raschmodells nachgewiesen werden kann. Die damit verknüpfte Forderung der strengen „spezifischen Objektivität" ist für psychologische Daten aus störanfälligen Testbearbeitungen äußerst streng. Das Raschmodell deckt jede Form testleistungsunabhängiger Information auf und zwar mit äußerster Präzision. Geeignete Mischverteilungsmodelle bieten Erklärungsansätze für diese neben der Testleistung stetig vorhandenen aber bislang selten interpretierbaren Zusatzinformationen.
Im vorliegenden Fall erweist sich die Klassenzuteilung für diagnostische Entscheidungen als nicht valide. Die vom Mixed-Rasch-Modell erkannten Informationen können aufgrund der verfügbaren Daten nur als „Rauschen" interpretiert werden. Ungerichtete Information im Sinne klassischer Störinformation wird in einer Testsituation nie ganz auszuschalten sein.
Nur systematische Forschung wird zeigen, ob Alternativen zum klassischen Testverständnis mit Summenscore zu finden sind. Für die hier aufgezeigte Testidee müssen Konzepte gefunden werden, wie sinnvolle Zusatzinformationen für unterschiedlichste, spezifische

[87] z.B. zunehmende Ermüdung im zeitlichen Verlauf

diagnostische Fragestellungen in ausreichender Menge durch Testitems ausgedrückt werden können.

Der erste Versuch die Profiltypen einer Testbatterie bei klinischen Fragestellungen zu nutzen, wurde von KUBINGER (1983a) unternommen. Bei der Zuordnung von knapp 100 Kindern *mit* und *ohne* Cerebralschädigung zu den mittels Latenter-Klassen-Analyse ermittelten vier Profiltypen des HAWIK wurden keine signifikanten Häufungen festgestellt. KUBINGER (1983b, S. 99) resümiert „auf keinen Fall realisieren cerebralgeschädigte Kinder denjenigen Typ häufiger als nicht cerebral geschädigte, welcher dem von Wechsler vermeinten Profil ´organisch Gehirnerkrankter´ entspricht“ und geht davon aus, dass dieses negative Resultat keine Absage an die grundsätzliche diagnostische Brauchbarkeit von Typisierungen bedeutet. So ist bei Cerebralschädigungen mit einer großen Symptomvielfalt zu rechnen, die bei KUBINGERS Untersuchung mit dem HAWIK keine spezifischen Testleistungsprofile ergeben haben.

Ebenso erfolglos waren GIEGLER & ROST (1993). Diese Autoren kommen bei der Latenten-Klassen-Analyse des Persönlichkeitsinventars „Gießentest“ zu dem Schluss, dass es mit großen Schwierigkeiten verbunden ist, „stabile Typen zu identifizieren, die mehr sind als nur graduelle Abstufungen psychosozialer Beeinträchtigungen (S. 151)“.

Das mit der Testbatterie in der vorliegenden Arbeit erfasste latente Konstrukt „Testleistungsfähigkeit“ kann auch durch die drei hochkorrelierten Testfähigkeiten erklärt werden[88]. Es bestätigt andererseits jedoch die konstante Beobachtung der empirischen Psychologie, dass eine Person die in einem Test viele Items löst, auch in einem anderen Test eher gute Ergebnisse erzielen wird.

Der zukunftsweisende, gemeinsame Aspekt der zuvor genannten drei Arbeiten ist folgender:

Bei der Suche nach neuartigen Konzeptionen darf **nicht** von Tests und Testskalen ausgegangen werden, die auf der Grundlage einer klassisch-eindimensionalen, also technischen Ansprüchen gerecht werdenden, exakten Konstruktionsperspektive[89] entstanden sind. Stattdessen muss von Anfang an in der Testkonstruktion auf vielfältige, qualitativ verschiedene und auch DIF-belastete Items abgezielt werden[90]. Der Weg zu neuen Ansätzen führt deshalb in jedem Fall zu einer **Abkehr vom klassisch unverzichtbaren Itemhomogenitätsideal** quantitativer Testkonstruktionen[91], weil der individuellen

[88] vgl. die *Tabellen 4-17, 4-18* sowie *Abbildung 4-7*

[89] diese war: Bis wohin fährt der Test mit der Testperson?

[90] DIF=differential item functioning, vgl. *Abschnitt 1.8.3*

[91] Stichwort „Verschimmelung“ als psychologische und testdiagnostische Herausforderung der Zukunft

Testbearbeitung durch jede einzelne Testperson von Anfang an Rechnung getragen werden muss. Ein denkbarer Testentwurf mit dem Fokus auf qualitativen Bearbeitungsunterschieden bei probabilistisch analysierbarer Itemanzahl ist die Zusammenfügung von Mini-Subskalen aus jeweils höchstens zwei bis drei homogenen Items zu einem Gesamttest.

Testdiagnostik ist in der herkömmlichen quantitativen Form als Kreuzchenauszählung[92] mit dem Raschmodell entweder am äußersten Limit angekommen oder steht an einer überwindbaren Schallmauer ihrer Möglichkeiten. Der qualitative Ansatz[93] mittels moderner Testtheorie besitzt die Voraussetzungen für die Entwicklung sinnvoller neuartiger diagnostischer Instrumente, damit auch künftig nicht alles beim Alten bleibt:

Im Bereich der kognitiven Entwicklung (MAY, 1979; KOLB, 1993; BALOGH, 1994; SPIEL et al., 1997 & 2004) und im Fragebogeneinsatz (ROST, 2002) ist es mit Mischverteilungsmodellen gelungen, unterschiedlichste psychodiagnostisch aufschlussreiche Aspekte qualitativer Gruppenunterschiede und heterogener Antwortstile zu beleuchten. Während Erstere die Theorien Piaget`s und entsprechender Erweiterungen zur Grundlage hatten, waren Ärgererleben (EID et al., 1996), Depression (KELLER & KEMPF, 1997; FRICK et al., 1997; MÜLLER & WETZEL, 1998), Sonnenschutzverhalten (EID, 1997), Self-Monitoring (DAVIER & ROST, 1997), Physikinteresse bei Schülern (ROST et al., 1999) und die Identifizierung unskalierbarer Beantworter (WAGNER-MENGHIN, 2002b) im Fokus von Fragebogenstudien. ARENDASY et al. (2004) nutzen in ihrer Studie zum Einfluss von Bearbeitungsstrategien von Textaufgaben bei Zweitklässlern auch Mischverteilungsmodelle. Sie resümieren in Bezug auf die Fragestellung, dass interindividuellen, differenziellen Unterschieden bei der Bewertung von Leistungen stärker als bisher Beachtung geschenkt werden müsse.

Wenn probabilistische Ansätze zur Klärung struktureller Leistungszusammenhänge künftig auch in anderen Testbereichen gelingen[94] und derartige Erkenntnisse bei der Testentwicklung in Konstruktion inklusive Auswertung einfließen, dann wäre es nach den Überlegungen in dieser Arbeit als Quantensprung klinisch-psychologischer Diagnostik und als längst fällige, auffällige Integration moderner Testtheorie in das anwendungsorientierte Fach, aber besonders in das Bewusstsein seiner Anwender zu bezeichnen.

---

[92] Anzahl der richtigen Kreuzchen bei Fähigkeitstests, Anzahl der Kreuzchen bei Persönlichkeitstests

[93] „Welche Items wurden gelöst?“

[94] im klinischen Einsatz z.B. zur Aufdeckung von Demenzen im Frühstadium, zur Abklärung depressiver Verarbeitungsmuster oder zur Differentialdiagnostik neuropsychologischer Störungen

Die Feststellung von HILKE (1984), dass sich Methoden nicht am Menschen orientieren, sondern der Mensch an den Methoden orientiert wurde, ist unter der eindimensionalen, klassischen, ergebnisorientierten Testperspektive und normorientierten Leistungsauslegung richtig. Moderne Testtheorie hält jedoch mit Modellweiterentwicklungen auch vollkommen neue testdiagnostische Perspektiven bereit, die das was Menschen bei Tests tun[95] in den Mittelpunkt rücken und so dem methodentheoretischen Hoffnungsgedanken von SPRUNG & SPRUNG (2000, S. 47) Ausdruck verleihen, „dass das Subjekt-Methode-Objekt-Verhältnis - das jeder empirischen humanwissenschaftlichen Untersuchung zu Grunde liegt - in ein Subjekt-Methode-Subjekt-Verhältnis verwandelt wird".

Die wissenschaftliche Form der Psychologie besitzt mit stochastischen Testmodellen schon jetzt das theoretische Rüstzeug, um jenen gewaltigen Schritt zu tun, der die Medizin weg vom Aderlass, hin zu aufwendig entwickelten, spezifisch wirkenden Arzneimitteln gebracht hat. Bislang nutzt die Psychodiagnostik ihre Chance nicht, um in eine neue Ära zu gelangen. Diese befindet sich jenseits der üblichen leistungsbetonenden oder defizitorientierten Auswertemethoden und hilflos festgefahrenen, rigide verallgemeinernden Denkmustern entlang von ***einer*** Zahlengeraden, die ***keinem*** Mensch gerecht werden[96].

Im Fall der Testtheorie führte der rationale Weg dahin, das wissenschaftlich und psychodiagnostisch Sagbare über Menschen zu hinterfragen und in jeglicher Hinsicht nach außen mit einem Ungewissheits- oder Zufallsaspekt zu begrenzen, um es - und auch *Es* selbst - dadurch in aller Vorsicht zu erweitern.

[95] und nicht das, was Testpersonen als Objekt einer von Psychologen konstruierten, an technischen Vorstellungen orientierten künstlichen Realität, als Ergebnis in Form des Summenscores erreichen.

[96] es ist davon auszugehen, dass diese Denkmuster eine ganze Gesellschaft prägen.

6.8 Verschimmelungsmanifest für das Fach Psychologie als Wissenschaft vom Menschen[97]

...

Der Mensch muss seine kritisch-schöpferische Funktion wieder einnehmen,
die er verloren hat und ohne die er aufhört, als Mensch zu existieren!
Der Urwald der geraden Linien, der uns immer mehr wie
Gefangene in einem Gefängnis umstrickt, muss gerodet werden.

Der Mensch hat bisher immer noch die Dschungel gerodet, in denen er sich befand, und sich befreit. Allerdings muss er sich erst dessen bewusst werden, dass er sich in einem Dschungel befindet, denn dieser Dschungel hat sich schleicherisch gebildet, ohne dass die Bevölkerung davon etwas weiß. Und diesmal ist es ein Dschungel der geraden Linien.

FUNDAMENTALE MISSION:

Verankerung eines moralischen Verantwortungsgefühls gegenüber
der schöpferischen Verschimmelung und der kritischen Verwitterung!
Nur die Techniker und Wissenschaftler, die imstande sind, im Schimmel zu leben
und Schimmel schöpferisch zu erzeugen, werden die Herren von morgen sein.
Jeder lebendige Mensch hat nichts zu verlieren als seine Ketten.
Die wissenschaftliche Form der Psychologie muss ganz von vorne anfangen,
um zum freien, bewussten, intelligenten, vernünftigen und gesunden Wesen
des einzelnen Menschen zurückzufinden:
Zwischen Kontrolle und lebendigem Chaos!

...

[97] Stände nur Hundertwasser (HUNDERTWASSER, 2004) noch hier, um dies mit eigenen Augen zu sehen!

Der Verfasser.

6.9 Epilog

Schon JASPERS (1990, S. 168) schrieb mit Bezug auf eine im Jahre 1906 gemachte Studie von KRUEGER & SPEARMAN folgenden Wortlaut:

„Jemand könnte gegen diese Arbeit einwenden, es seien bei all der Mühe doch kaum Resultate gewonnen, die nicht längst bekannt seien (...). Dieser Einwand würde sehr ungerecht sein. Der Wert vorliegender Arbeit liegt in der neuen Benutzung einer Methode, die exakte, nicht mehr abzustreitende Feststellungen erlaubt. Treffen diese Feststellungen etwas der klinischen Erfahrung schon Bekanntes, so muss man bedenken, dass derselben klinischen Erfahrung vieles bekannt ist, was falsch oder unbeweisbar ist, und dass eine solche Bestätigung immer ein Resultat bedeutet“.

Ähnliches lässt sich wohl auch von der vorliegenden Arbeit sagen.

7 A) ZUSAMMENFASSUNG

Moderne stochastische Testmodelle ergänzen klassische testtheoretische Ansätze um das fehlende prüfbare, d.h. im Sinne Popper`s falsifizierbare Theoriegerüst. Damit eröffnen sie auch völlig neuartige Möglichkeiten der Testkonstruktion.

In der vorliegenden Arbeit wurden die drei psychometrischen Testverfahren Benton-, Raven- und Cattelltest[98] - getrennt und als Testbatterie - eingehenden klassischen und probabilistischen testtheoretischen Prüfungen unterzogen. Der Vergleich dieser beiden Methoden bezüglich der Vorhersagevalidität klinischer Kriterien war zur Kreuzvalidierung der testdiagnostisch bemerkenswerten Ergebnisse von KINDERMANN (2001) nötig geworden.

In dieser vorausgegangenen Arbeit sind bei einer neuen Form des Bentontests mit dem Mixed-Rasch-Modell zwei latente Klassen aufgezeigt worden. Auffallend war dabei, dass sich 93 Prozent aller getesteten Patienten aus der gemischten Personenstichprobe in nur einer der beiden identifizierten Klassen befanden. Dieses Ergebnis läßt sich als zwei qualitativ unterschiedliche und diagnostisch relevante Lösungsstrategien beim Bentontest interpretieren, die nur mit stochastischen Testmodellen moderner Theorie identifizierbar sind.

Für die vorliegende Arbeit wurde ein völlig neuer Datensatz erhoben. Die Gesamtstichprobe (N=615) setzt sich wie bei KINDERMANN (2001) aus klinisch unauffälligen und auffälligen Personen mit psychiatrischen und neurologischen Diagnosen zusammen.

Unter Anwendung von Mischverteilungsmodellen kann mit probabilistischen Methoden auch in der Kreuzvalidierung testleistungsunabhängige qualitative Information innerhalb der Daten jedes einzelnen Tests identifiziert werden.

In der klassischen, quantitativen Auswertung mittels Summenscore bleibt diese Information unberücksichtigt.

Die entscheidende Frage nach praktischen Konsequenzen aus diesen identifizierten Zusatzinformationen kann im vorliegenden Fall nicht positiv beantwortet werden. Durch die Kombination von klassischen und probabilistischen Prädiktoren, also den Summenscores und den Klassenzugehörigkeitsparametern, konnte eine relevante Verbesserung der Prädiktion

[98] Es handelt sich beim Bentontest um die von STECK (2005) verlängerte 20-Item-Form, beim Raventest um eine aus 20-Items der Advanced Progressive Matrices extra für diese Untersuchung zusammengestellte Form, beim Cattelltest um die 13 regulären Items des Untertests „Matrizen“ im CFT-3, 2. Teil .

klinischer Kriterien bei keinem der Tests nachgewiesen werden. Es zeigt sich zu der ohnehin schon geringen Prädiktionsverbesserung mittels klassischer Testleistungsmaße ein weiterer, jedoch unwesentlicher Prädiktionszuwachs durch die zusätzliche Berücksichtigung der Klassenparameter. Die identifizierten Klassen geben keine relevanten Hinweise über das Vorliegen klinischer Auffälligkeiten bei Testpersonen.

Latente-Klassen-Analysen (LCA) der aus Raven-, Benton- und Cattelltest zusammengesetzten Testbatterie nach klassischem (LCA der Summenscorequartile) und probabilistischem Auswertungsmodus (LCA der Klassenzugehörigkeiten samt klassenspezifischen Summenscorequartilen) weisen darauf hin, dass die Testbatterie im ersteren Fall ein eindimensionales und im letzteren Fall ein nahezu eindimensionales Konstrukt erfasst, das als „Testleistungsfähigkeit“ bezeichnet wurde. Die Variable Bildungsgrad bestimmt die Position einer Testperson auf dieser Ordinalskala „Testleistungsfähigkeit“ bei beiden Auswertungsmodi am deutlichsten.

Im Hinblick auf die vom Praktiker bislang kaum bemerkte oder nicht vorhandene Relevanz moderner Testtheorie im testdiagnostischen Alltag ist aus den Überlegungen in dieser Arbeit folgendes Fazit zu ziehen:

Der klassische Ansatz zur Messung psychologischer Fähigkeiten erweist sich in der derzeit gängigen Interpretation als nicht stimmig, auch wenn dieser testtheoretisch abgesichert erscheint. Quantitative Testdiagnostik misst vom Test definierte, in technischer Hinsicht per Raschmodell nahezu objektiv vermessbare Fähigkeiten. Völlig unberücksichtigt bleibt dabei jedoch der einzelne Mensch in seiner Subjektrolle bei Tests, die sich durch gedankliche Freiheit, Lebendigkeit und Einmaligkeit auszeichnet.

Wenn Psychologie darauf abzielt, die Wissenschaft vom bewusst handelnden Individuum Mensch zu werden, dann ist die Abkehr vom klassischen und somit quantitativen Ansatz entscheidend. Für die Realisierung diagnostisch hilfreicher Testverfahren, die dem menschlichen Wesen gerecht werden, müssen allerdings völlig neue und bislang wenig einladende Wege, unter radikaler Abkehr von gewohnten Denkweisen und testhistorisch lieb gewonnenen Testkonstruktionsprinzipien eingeschlagen werden.

Vorschläge und Anknüpfungspunkte werden in der Diskussion der vorliegenden Arbeit aufgezeigt.

7 B) ABSTRACT

Modern stochastic test models supplement the classic test-theoretical approach with the missing proof, i.e. in the sense of Popper falsifiable theory stand. Thus, they also open completely new possibilities for test construction.

In the following thesis, the three psychometric tests Benton, Raven and Cattell were subjected separately and as a test battery to detailed classical and probabilistic test-theoretical examinations. The comparison of these two methods regarding the forecast validity of clinical criteria had become necessary for the cross validation of test-diagnostically remarkable results from Kindermann (2001).

In this previous work, two latent classes were identified using a new form of the Benton test with the mixed Rasch-model. The surprising result was that 93% of tested patients in a mixed random sample fell into only one of the two identified classes. This result may suggest that two qualitatively different and diagnostically relevant solution strategies within the Benton test are identifiable only with stochastic test models of modern theory.

For the current thesis, a completely new set of data was collected. The total sample (N=615) was comprised, as in Kindermann (2001), of clinically symptomatic and asymptomatic persons with psychiatric and neurological diagnoses.

Utilizing mixed distribution models in the cross validation, it was possible to identify qualitative information within the data of each individual test, independent of the test performance. This information is not accounted for when using the classical, quantitative evaluation by means of test score.

The decisive question which arises in this case from the practical consequences of the additional identified data, cannot be answered positively. Using the combination of classical and probabilistic predictors, i.e. the testscores and the class parameters, does not result in any relevant improvement in the prediction of clinical criteria in any of the tests. The additional consideration of the class parameters pointed to a potential not significant improvement in prediction over classical testscores. The identified classes do not deliver any relevant data about the presence of existing clinical symptoms in the test persons.

Latent class analyses (LCA) of the test battery comprising of the Raven-, Benton- and Cattell-test, yielded for both, a classical (LCA of the testscore quartiles) and a probabilistic analysis mode (LCA of the class association including class-specific testscore quartiles), an almost linear construct called “test efficiency“. In both analysis modes, the variable level of

education most clearly determines the position of a test person on this ordinal scale for „test efficiency".

With regard to the previously almost unnoticeable or nearly non-existent relevance of modern test theory in practical test-diagnostic applications, the following conclusions can be made from the considerations in this thesis:

The classical approach for measuring psychological abilities in the present interpretation proves to be inconsistent, despite the fact that this theory appears to be test-theoretically correct.

Although quantitative test diagnostics, according to the theory of "Rasch homogeneity" technically measure theoretical given abilities in a person, they do not account for the subject role of each individual within the test, who is characterised by his or her freedom of thought, mental aliveness and uniqueness.

If psychology strives to become the science of consciously acting individual human beings, then the renunciation of the classical and thus quantitative approach is crucial. For realization of diagnostically helpful test procedures, which can do justice to human nature, it will be necessary to go down previously unexplored but also uninviting avenues and to turn our backs on the old "much loved" theories.

Suggestions and related factors will be discussed in the following thesis.

8 LITERATURVERZEICHNIS

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory*, S. 267-281. Budapest Academiai Kiado.

Amthauer, R. (1973). *Intelligenz-Struktur-Test, IST-70*. Göttingen: Hogrefe.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrica, 38*, 1, 123-140.

Anastasi, A. (1976). *Psychological Testing* (4th ed.). New York: Macmillan

Anastasi, A. (1985). Some Emerging Trends in Psychological Measurement: A Fifty-Year Perspective. *Applied Psychological Measurement, 9,* 2, 121-138.

Ansley, R.A., & Forsyth, T.N. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Arendasy, M., Sommer, M. & Glück, J. (2004). Dimensionalität und differenzielle Validität von Textaufgaben – Zum Einfluss von Bearbeitungsstrategien. *Zeitschrift für Pädagogische Psychologie, 18* (3/4), 231-243.

Artelt, C., & Baumert, J. (2004). Zur Vergleichbarkeit von Schülerleistungen bei Leseaufgaben unterschiedlichen sprachlichen Ursprungs. *Zeitschrift für Pädagogische Psychologie, 18* (3/4), 171-185.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). *Multivariate Analysemethoden-Eine anwendungsorientierte Einführung* (10. Auflage). Berlin: Springer.

Balogh, A. (1994). *Ansätze zur Verbesserung der Diagnostik bei Problemlöseprozessen am Beispiel von Balkenwaagenaufgaben unter Verwendung der Latent-Class-Analyse.* Psychologische Diplomarbeit. Universität Konstanz.

Bauer, H., Guttmann, G., Leodolter, M., & Leodolter, U. (2004). *Mentale Rotation (MR) Test zur Erfassung des Raumvorstellungsvermögens* (Computerprogramm). Mödling: Schuhfried.

Bechger, T.M., Maris, G., Verstralen, H.H.F.M., & Béguin, A.A. (2003). Using Classical Test Theory in Combination With Item Response Theory. *Applied Psychological Measurement, 27* (5), 319-334.

Becker, P. (2003). *TIPI. Trierer Persönlichkeitsinventar*. Göttingen: Hogrefe.

Beckmann, J.F., Guthke, J., & Vahle, H. (1997). Analysen zum Zeitverhalten bei computergestützten adaptiven Intelligenz-Lerntests. *Diagnostica, 43* (1), 40-62.

Beckmann, J.F., & Herzberg, P.Y. (Hrsg.) (2001). *Dynamik im Testen. Perspektiven auf die Zone der nächsten Entwicklung in der Psychodiagnostik. Festschrift zum 60. Geburtstag von Jürgen Guthke. (Psychologie, Bd. 40).* Landau: Verlag Empirische Pädagogik.

Bem, D.J., & Allen, A. (1974). On Predicting Some Of The People Some Of The Time: The Search For Cross-Situatuional Consistencies In Behavior. *Psychological Review, 81* (6), 506-520.

Benton, A.L. (1996). Vorwort. In Benton-Sivan, A., & Spreen, O., *Der Benton-Test (Handbuch)* (S. 7). Bern-Göttingen-Toronto-Seattle: Verlag Hans Huber.

Benton-Sivan, A., & Spreen, O. (1996). *Der Benton-Test. Handbuch* (7., vollständig überarbeitete Auflage). Bern-Göttingen-Toronto-Seattle: Verlag Hans Huber.

Berg, R., Franzen, M., & Wedding, D. (1987). *Screening For Brain Impairment- A Manual for Mental Health Practice.* New York: Springer.

Berk, R. (1982). *Handbook of methods for detecting test bias.* Baltimore: John Hopkins University Press.

Biehl, B. (1996). *Tachistoskopischer Verbalauffassungstest Mannheim* (Version 1.00) (Computerprogramme mit Manualen). Mödling: Schuhfried.

Biehl, B. (2000). *Alkoholiker Trenn Verfahren. ATV* (Version 22.00) (Computerprogramm). Mödling: Schuhfried.

Blough, P.M., & Slavin, L.K. (1987). Reaction time assessment of gender differences in visual-spatial performance. *Perception and Psychophysics, 41*, 276-281.

Booth, J.F. (1990). Computerdiagnostik: Errungenschaften und Möglichkeiten der Technologie. *Diagnostica, 36*, (2), 180-203.

Bortz, J. (1989). *Lehrbuch der Statistik. Für Sozialwissenschaftler* (3. Auflage). Berlin-Heidelberg-New York: Springer.

Bozdogan, H. & Ramirez, D.E. (1988). FACAIC: Model Selection Algorithm For The Orthogonal Factor Model Using AIC And CAIC. *Psychometrica, 53*, (3), 407-415.

Bruder, K.-J. (1982). Entwurf der Kritik bürgerlicher Psychologie. In Bruder, K.-J. (Hrsg.), *Kritik der bürgerlichen Psychologie* (S. 92-194). Frankfurt: Fischer.

Bukasa, B., Kisser, R., & Wenninger, U. (1990). Computergestützte Leistungsdiagnostik bei verkehrspsychologischen Eignungsuntersuchungen. *Diagnostica*, 36, 148-165.

Cattell, J.M. (1890). *Mental Tests and Measurements. Mind* (Vol. XV, S. 373-380). London: Williams and Norgate.

Carpenter, P.A., Just, M.A., & Shell, P. (1990). What One Intelligence Test Measures: A Theoretical Account of the Processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, (3), 404-431.

Cohen, J. (1988). Statistical Power Analysis fort he Behavioral Sciences (2nd Edition). Hillsdale, NJ: Erlbaum.

Conrad, W., Baumann, E., & Mohr, V. (1980). *Mannheimer Test zur Erfassung des physikalisch-technischen* Problemlösens *(MTP)*. Göttingen: Hogrefe.

Cochran, K.F., & Wheatley, G.H. (1989). Ability and sex-related differences in cognitive strategies on spatial tasks. *Journal of General Psychology, 116*, 43-55.

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

Cronbach, L.J., & Gleser, G.C. (1965). *Psychological Tests and Personnel Decisions* (2nd Edition). Urbana: University of Illinois Press.

Davier, M. v. (1996). *Methoden zur Prüfung probabilistischer Testmodelle.* Kiel: IPN

Davier, M. v. (2001). WINMIRA 2001 1.36. *Computerprogramm.* Kiel: IPN.

Davier, M. v., & Rost, J. (1997). Self Monitoring – A Class Variable ?. In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social siences* (S. 314-323). Münster: Waxmann.

Drasgow, F., Levine, M.V., & McLaughlin, M.E. (1987). Detecting Inappropriate Test Scores with Optimal and Practical Appropriateness Indices. *Applied Psychological Measurement, 11*, (1), 59-79.

Edjieh, D. (1980). *Transkultureller Vergleich des WMT von Formann und Rasch-Skalierung des CFT-3 und WMT für den Iran". Dissertation.* Wien: Universität.

Eid, M. (1997). Sonnenschutzverhalten: Ein typologischer Ansatz. *Zeitschrift für Gesundheitspsychologie, 5,* (2), 73-90.

Eid, M., Emmermann, A., & Schwenkmetzger, P. (1996). Sind Ärgererleben und Ärgerausdrucksstile universelle Konstrukte? Ergebnisse einer Analyse des State-Trait-Ärgerausdrucks-Inventars mit logistischen Mischverteilungsmodellen. *Diagnostica, 42*, (1), 3-26.

Etzel, S., & Küppers, A. (2000). *pro facts. professional assessment by computer for training and selection.* Aachen: RWTH, Institut für Psychologie.

Ferguson, A., Meyers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, N.R., Craik, K.J.W., Drever, J., Guild, J., Houstoun, R.A., Irwin, J.O., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby, J.H., Smith, T., Thouless, R.H., & Tucker, W.S. (1940). Quantitative estimates of sensory events. *The advancement of science. The report of the British Association for the Advancement of Science, 2*, 331-349.

Ferrando, P.J., & Chico, E. (2001). Detecting Dissimulation In Personality Test Scores: A Comparison Between Person-Fit Indices And Detection Scales. *Educational And Psychological Measurement, 61* (6), 997-1012.

Fischer, G.H. (1968). *Psychologische Testtheorie.* Bern: Huber.

Fischer, G.H. (1972). Conditional maximum-likelihood of item parameters for a linear logistic test model. *Research Bulletin, 9*. Wien: Psychologisches Institut der Universität Wien.

Fischer, G.H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 36*, 359-374.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

Fisher, R.A. (1921). On the mathematical foundations of theoretical statistics. Phil. Trans. Roy. Soc. London, (a) 222, 309-368. In: Fisher, R. A. (1950).*Contributions to mathematical statistics.* New York: Wiley.

Fiske, D.W., & Butler, J.M. (1963). The Experimental Conditions For Measuring Individual Differences. *Educational and Psychological Measurement, 23*, (2), 249-266.

Fisz, M. (1970). *Wahrscheinlichkeitsrechnung und mathematische Statistik* (5., erweiterte Auflage). Berlin: VEB Deutscher Verlag der Wissenschaften.

Flier, van der, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267-298.

Frary, R.B. (1982, March). *A comparison of person-fit measures.* Paper presented at the annual meeting of the American Educational Research Association, New York.

Frick, U., Rehm, J., & Thien, U. (1997). On the Latent Structure of the Beck Depression Inventory (BDI): Using the „Somatic" Subscale to Evaluate a Clinical Trial. In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social siences* (S. 305-313). Münster: Waxmann.

Frühauf, K. (1975). Zur psychometrischen Bestimmung von Verhaltensauffälligkeiten bei hirnorganischen Störungen. *Zeischrift für Psychologie, 183;* (3), 279-293.

Formann, A. K. (1984). *Die Latent-Class-Analyse*. Weinheim: Beltz.

Formann, A.K., & Piswanger, K. (1979). *Wiener Matrizen-Test (WMT)*. Weinheim: Beltz.

Gatternig, J., & Kubinger, K.D. (1994). *Erkennen von Meta-Regeln*. Test: Software und Manual. Frankfurt/M.: Swets.

Giegler, H., & Rost, J. (1993). Typenbildung und Responsesets beim Gießen-Test: Clusteranalyse versus Analyse latenter Klassen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14* (3), 137-152.

Gittler, G. (1990). *Dreidimensionaler Würfeltest (3DW)*. Weinheim: Beltz.

Gittler, G. (2000). *Adaptiver Dreidimensionaler Würfeltest. A3DW* (Version 22.00) (Computerprogramm). Mödling: Schuhfried.

Glück, J., Machat, R., Jirasco, M., & Rollett, B. (2002). Training-related changes in solution strategy in a spatial test: An application of item response models. *Learning and Individual Differences, 13*, 1-22.

Gould, S.J. (1999). *Der falsch vermessene Mensch* (3. Auflage). Frankfurt am Main: Suhrkamp.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Grössenbrunner, P. (1994). *Pilot's Spatial Test (PST)*. Software und Manual. Mödling: Dr.G.Schuhfried GmbH.

Grubitzsch, S. (1991). *Testtheorie- Testpraxis*. Reinbek bei Hamburg: Rowohlt Taschenbuch Verlag.

Grünwald, H. (1980). *Die sozialen Ursprünge psychologischer Diagnostik*. Darmstadt: Dr. Dietrich Steinkopff Verlag.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Gulliksen, H. (1961). Linear And Multidimensional Scaling. *Psychometrika, 26*, (1), 9-25.

Guthke, J., Böttcher, H.R., & Sprunge, **L.** (Hrsg.) (1990). *Psychodiagnostik. Ein Lehr- und Arbeitsbuch für Psychologen sowie empirisch arbeitende Humanwissenschaftler.* Berlin: Deutscher Verlag der Wissenschaften.

Guthke, J., & Wiedl, K.H. (1996). *Dynamisches Testen – Zur Psychodiagnostik der intraindividuellen Variabilität.* Göttingen: Hogrefe.

Guttman, L. (1950). The basis of scalogram analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, J.A. Clausen (eds.), *Studies in social psychology in world war II,* Vol. IV. Princeton/N.J.: Princeton Univ. Press.

Guttmann, L. (1953). A special review of Harold Gulliksen, theory of mental tests. *Psychometrika*, *18*, 123-130.

Harnisch, D.L. (1983). Item response patterns: Applications of educational practice. *Journal of Educational Measurement, 20*, 191-206.

Hartmann, D. (1993). *Naturwissenschaftliche Theorien – Wissenschaftstheoretische Grundlagen am Beispiel der Psychologie.* Mannheim: BI-Wissenschaftsverlag.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park: Sage.

Häcker, H.O., & Stapf, K.-H. (2004). *Dorsch, Psychologisches Wörterbuch* (14. Auflage). Bern: Verlag Hans Huber

Häusler, J. (2004). *Hyperkinetisches Syndrom Diagnostikum (HKSD)* (Computerprogramm). Mödling: Schuhfried.

Heaton, R.K., Chelune, G.J., Talley, J.L., Kay, G.G., & Curtis, G. (1993). *Card Sorting Test*. Odessa, FL: Psychological Assessment Ressources.

Heidegger, M. (1997). *Der Satz vom Grund* (8. Auflage). Stuttgart: Neske

Hergovich, A., & Hörndler, H. (1994). *Gestaltwahrnehmungstest*. Software und Manual. Frankfurt/M.: Swets.

Herle, M., Sommer, M., Wenzl, M., & Litzenberger, M. (2004). *Inventar verkehrsrelevanter Persönlichkeitseigenschaften (IVPE)* (Computerprogramm). Mödling: Schuhfried.

Heyde, G. (1995). *Inventar komplexer Aufmerksamkeit (INKA)*. Frankfurt/M.: Swets.

Hilke, R. (1980). *Grundlagen normorientierter und kriteriumorientierter Tests*. Bern: Huber.

Hilke, R. (1984).Handlungstheoretisch orientierte psychologische Diagnostik: Ausweg aus der Krise der psychologischen Diagnostik. In G. Jüttemann (Hrsg.), *Neue Aspekte klinisch-psychologischer Diagnostik* (S. 10-34). Göttingen: Hogrefe.

Holland, P.W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Holocher-Ertl, S., Kubinger, K.D., & Menghin, S. (2003). *Big Five Plus One Persönlichkeitsinventar* (Computerprogramm und Manual). Mödling: Schuhfried.

Hornke, L.F. (1997). Untersuchung von Itembearbeitungszeiten beim computergestützten adaptiven Testen. *Diagnostica, 43* (1), 27-39.

Hornke, L.F., & Etzel, S. (1999a). *VISGED. Visueller Gedächtnis Test. Ein computergestützter adaptiver Test zur Messung der visuellen Gedächtnisleistung* (Computerprogramm). Mödling: Schuhfried.

Hornke, L.F., & Etzel, S. (1999b). *VERGED. Verbaler Gedächtnis Test. Ein computergestützter adaptiver Test zur Messung der vebalen Gedächtnisleistung* (Computerprogramm). Mödling: Schuhfried.

Hornke, L.F., Etzel, S., Rettig, K. (2000). *Adaptiver Matrizentest. AMT (Version 21.00)* (Computerprogramm). Mödling: Schuhfried.

Hornke, L.F., Arendasy, M., Sommer, M., Häusler, J., Wagner-Menghin, M., Gittler, G., Bogner, B., & Wenzl, M. (2004). *Intelligenz Struktur Batterie. Eine Testbatterie zur Messung von Intelligenz* (Computerprogramm). Mödling: Schuhfried.

Hosenfeld, I.; Strauss, B., & Köller, **O.** (1997). Geschlechtsdifferenzen bei Raumvorstellungsaufgaben - Eine Frage der Strategie. *Zeitschrift für Pädagogische Psychologie*, *11* (2), 85-94.

Hosmer, D.W.Jr., & Lemeshow, S. (1989). *Applied Logistic Regression*. New York: Wiley.

Hundertwasser, F. (2004). Verschimmelungsmanifest gegen den Rationalismus in der Architektur. In W. Schurian (Hrsg.), *Hundertwasser - Schöne Wege, Gedanken über Kunst und Leben* (S. 216-222). München: Langen Müller Verlag.

Jäger, A.O. (1967). *Dimensionen der Intelligenz.* Göttingen: Hogrefe.

Jäger, R.S. (2003). *Praxis der Psychologischen Diagnostik.* Landau: Verlag Empirische Pädagogik.

Jäger, R.S., & Krieger, W. (1994). Zukunftsperspektiven der computerunterstützten Diagnostik, dargestellt am Beispiel der treatmentorientierten Diagnostik. *Diagnostica, 40* (3), 217-243.

Jaspers, K. (1953). Allgemeine Psychopathologie (6., unveränderte Aufl.). Berlin, Göttingen, Heidelberg: Springer.

Jaspers, K. (1990). Gesammelte Schriften zur Psychopathologie (Nachdruck der 1. Auflage von 1963). Berlin, Göttingen, Heidelberg: Springer.

Kahraman, N., & Kamata A. (2004). Increasing the Precision of Subscale Scores by Using Out-of-Scale Information. *Applied Psychological Measurement,28* (6), 407-426.

Kaschel, R. (1994). Der "Benton-Test" (Buch und Testbesprechungen). *Zeitschrift für Neuropsychologie, 5*, 187-199.

Keller, F., & Kempf, W. (1997). Some Latent Trait and Latent Class Analyses of the Beck-Depression-Inventory (BDI). In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social siences* (S. 314-323). Münster: Waxmann.

Kempf, W.F. (1972). Probabilistische Modelle experimentalpsychologischer Versuchssituationen. *Psychologische Beiträge*, 14, 16-37.

Kempf, W.F. (1974). Dynamische Modelle zur Messung sozialer Verhaltensdispositionen. In W. F. Kempf (Hrsg.), *Probabilistische Modelle in der Sozialpsychologie* (S. 13-55). Bern: Huber.

Kempf, W. (2003). *Psychologie- Zwischen naturwissenschaftlichem Experiment und sozialwissenschaftlicher Hermeneutik- Band 1: Theorie und Empirie.* Berlin: Regener.

Kempf, W.F. (2004). Dynamic Version 2.5. *Computerprogramm*

Kessler, J., & Kaschel, R. (1997). Testrezension zu Benton-Test. *Zeitschrift für Differentielle und Diagnostische Psychologie, 18*, 66-68.

Kindermann, F. (2001). *Rasch-Analyse einer neuen Form des Benton-Tests.* Psychologische Diplomarbeit. Universität Konstanz.

Klieme, E., & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. *European Journal of Psychology of Education, 16* (3), 385-402.

Kolb, J. (1993). *Rekonstruktion theoretischer Datenstrukturen mittels Latent-Class-Analyse – zur Eignung des Modells als theoriengenerierendes Verfahren.* Psychologische Diplomarbeit. Universität Konstanz.

Köller, O. (1994). Psychometrische und psychologische Betrachtungen des Rateverhaltens in Schulleistungstests. *Empirische Pädagogik, 8* (1), 59- 84.

Köller, O., Rost, J., & Köller, M. (1994). Individuelle Unterschiede beim Lösen von Raumvorstellungsaufgaben aus dem IST- bzw. IST-70-Untertest „Würfelaufgaben". *Zeitschrift für Psychologie, 202*, 65-85.

Kubinger, K.D. (1983a). Verfehlte Einwände gegen den HAWIK als Differential-Diagnostikum – Eine Antwort auf Wallasch & Dony. *Diagnostica,* 29, 118-135.

Kubinger, K.D. (1983b). *Der HAWIK-Möglichkeiten und Grenzen seiner Anwendung.* Weinheim & Basel: Beltz

Kubinger, K.D. (1986). Adaptive Intelligenzdiagnostik. *Diagnostica, 32* (4), 330-344.

Kubinger, K.D. (1988). *Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen.* Weinheim: Beltz.

Kubinger, K.D. (1996). *Einführung in die Psychologische Diagnostik* (2. Auflage). Weinheim: PVU.

Kubinger, K.D. (1998). Psychometric shortcomings of Wechsler´s intelligence scales - Results on the German WISC, conclusions for the WISC-R. *Review of Psychology*, 5, 3-12.

Kubinger, K.D., & Wurst, E. (1985; 1991). *Adaptives Intelligenz Diagnostikum (AID)*. Weinheim: Beltz.

Kubinger, K.D., Fischer, D., & Schuhfried, G. (1993). *Begriffs-Bildungs-Test (BBT)*. Software und Manual. Mödling: Dr.G.Schuhfried GmbH.

Kubinger, K.D., Formann, A.K., & Schubert, M. (1980). Typisierung von HAWIK-Profilen. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 1, 117-126.

Kühnlein, I., Hahlweg, K., Müller G., & Sevenitz B. (1983). Überprüfung der SPM von Raven für den psychiatrischen Bereich und Entwicklung von Kurzformen. *Diagnostica,* 29 (3), 236-246.

Langfeldt-Nagel, M. (1982). Untersuchungen zur Konstruktvalidität der Grundintelligenztests (CFT) von Cattell und Weiss. *Diagnostica, 28* (1), 65-79.

Lazarsfeld, P.F. (1950). Logical and mathematical foundations of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, J.A. Clausen (eds.), *Studies in social psychology in world war II,* Vol. IV. Princeton/N.J.: Princeton Univ. Press.

Lazarsfeld, P.F. (1959). Latent structure analysis. In S. Koch (Ed.): *Psychology: a study of a science*, Vol. 3. New York: McGraw-Hill.

Lazarsfeld, P.F., & Henry, N.W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin Co.

Levine, M.V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Li, M.F. & Olejnik, S. (1997). The Power of Rasch Person-Fit Statistics in Detecting Unusual Response Patterns. *Applied Psychological Measurement, 21* (3), 215-231

Liedl, K. (1999). *Mechanisch-technisches Auffassungsvermögen (MTA)*. Software und Manual. Mödling: Dr.G.Schuhfried GmbH.

Lienert, G.A., & Raatz, U. (1994). *Testaufbau und Testanalyse* (5., völlig neubearb. und erw. Aufl.). Weinheim: Beltz, Psychologie-Verlags-Union.

Linden, van der, W.J., & Hambleton R.K. (1997). *Handbook of modern item response theory.* New York : Springer.

Long, J.C. (1997). *Regression Models for Categorical and Limited Dependent Variables*. London: Sage.

Lord, F. M. (1977). Practical Applications of Item Characteristic Curve Theory. *Journal of Educational Measurement, 14* (2), 117-138.

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

Lumsden, J. (1976). Test theory. *Annual Review of Psychology, 27*, 251-280.

Madsen, H.S. (1987). *Utilizing Rasch analysis to detect cheating on language examinations*. (ERIC Document Reproduction Service No. ED 287 284).

May, R. (1979). *Wie entwickelt sich das Verständnis von Proportionalität? Strukturelle Modelle zum Problemlösen bei Balkenwaage-Aufgaben.* Psychologische Diplomarbeit. Universität Konstanz.

Mc Fadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarempka (ed.), *Frontiers in Econometrics* (S. 105-142). New York: Academic Press.

Mc Leod, L., Lewis, C., & Thissen, D. (2003). A Bayesian Method for the Detection of Item Preknowledge in Computerized Adaptive Testing. *Applied Psychological Measurement, 27* (2), 121-137.

Meijer, R.R., & Sijtsma, K. (2001). Methodology Review: Evaluating Person Fit. *Applied Psychological Measurement, 25* (2), 107-135.

Merten, T. (1999). Über den Sinn und Unsinn der Verwendung von Screening–Instrumenten in der neuropsychologischen Diagnostik. *Diagnostica, 45* (3), 154-162.

Messick, S. (1984). The psychology of educational Measurement. *Journal of Educational Measurement, 21*, 215-237.

Michel, L., & Mai, N. (1968). Entscheidungstheorie und Probleme der Diagnostik bei Cronbach & Gleser. *Diagnostica, 14*, 99-121.

Microsoft Corporation (2002). Excel.

Molenaar, I.W. (1997). Lenient or Strict Application of IRT with an Eye on Practical Consequences. In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social siences* (S. 38-49). Münster: Waxmann.

Müller, H. (1999). *Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen.* Bern: Huber.

Müller, M.J., & Wetzel, H. (1998). Dimensionality of depression in acute schizophrenia: a methodological study using the Bech-Rafaelsen Melancholia Scale (BRMES). *Journal of Psychiatric Research, 32*, 369-378.

Nagl, W. (2005). Einführung in die Statistik. *Vorlesungsscript-Download*: Universität Konstanz

Neyman, J., & Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica, 16*, 1-32.

Nietzsche, F. (1982). *Die fröhliche Wissenschaft*. Frankfurt am Main: Insel Verlag

Novick, M. (1966). The axioms and principal results of classical test theory. *Journal of mathematical Psychology, 3*; 1-18.

Ozer, D.J., & Reise, S.P. (1994). Personality Assessment. *Annual Review of Psychology, 45,* 357-388.

Pawlik, K. (1976). *Diagnose der Diagnostik.* Stuttgart: Klett

Pfanzagl, J. (1993). A case of asymptotic equivalence between conditional and marginal maximum likelihood estimators. *Journal of Statistical Planning and Inference. 35*, 301-307.

Ponocny, I., & Klauer, K.-C. (2002). Zur Identifizierung nichtskalierbarer Persönlichkeitsfragenbeantworter: Der Einsatz von Indizes exakter Personenpassung. *Psychologische Beiträge. 44* (1), 94-107.

Popper, K.R. (1982). *Logik der Forschung* (7.Aufl.). Tübingen: Mohr.

Popper, K.R. (2004). *Alles Leben ist Problemlösen-Über Erkenntnis, Geschichte und Politik* (JubiläumsEdition). München: Piper.

Putz-Osterloh, W. (1981). *Problemlöseprozesse und Intelligenztestleistung.* Bern: Huber.

Rammsayer, T., & Brandler, S. (2003). Zum Zeitverhalten beim computergestützten adaptiven Testen. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24* (1), 57-63.

Rasch, D. (1989). *Einführung in die mathematische Statistik – Wahrscheinlichkeitsrechnung und Grundlagen der mathematischen Statistik* (3., überarbeitete Auflage). Berlin: VEB Deutscher Verlag der Wissenschaften.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche (2nd Edition, Chicago, University of Chicago Press, 1980).

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology.* Berkeley: University of California Press.

Rasch, G. (1977). On specific objectivity. An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58-94.

Rauchfleisch, U. (1983). Itemanalyse zum Progressiven Matrizentest von Raven (SPM) an einer Stichprobe psychisch Kranker. *Diagnostica, 29* (4), 359-366.

Rauchfleisch, U. (1989). *Testpsychologie* (2. Auflage). Göttingen: Vandenhoeck & Ruprecht.

Raven, J.C. (1965). *Advanced Progressive Matrices Sets I and II.* London: Lewis and Co. Ltd.

Raven, J. (2000). The Raven`s Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology, 41*, 1-48.

Reckase, M.D., & Mc Kinley, R.L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

Reise, S.P., & Waller, N.G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology, 65*, 143-151.

Rexilius, G. (1978). Grenzen der Testerei. In S. Grubitzsch & G. Rexilius (Hrsg.), *Testtheorie-Testpraxis* (S. 112-167). Reinbeck bei Hamburg: Rowohlt.

Roskam, E.E. (1996). Beobachtung und Daten. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 3-22). Weinheim: Beltz.

Rost, J. (1989). Lacord 3.31. *Computerprogramm*

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 3, 271-282.

Rost, J. (1995). Die testdiagnostische Erfassung von Typen. In K. Pawlik (Hrsg.), *Bericht über den 39. Kongress der Deutschen Gesellschaft für Psychologie in Hamburg 1994* (S. 392-398). Göttingen: Hogrefe.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion*. Bern: Huber.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden?. *Psychologische Rundschau, 50* (3), 140-156.

Rost, J. (2002). When personality questionnaires fail to be unidimensional. *Psychologische Beiträge, 44*, 108-125.

Rost, J. (2004). *Lehrbuch Testtheorie, Testkonstruktion* (2. Auflage). Bern: Huber.

Rost, J., & Erdfelder, E. (1996). Mischverteilungsmodelle. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 333-348). Weinheim: Beltz.

Rost, J., Sievers, K., Häußler, P., Hoffmann, L., & Langeheine, R. (1999). Struktur und Veränderung des Interesses an Physik bei Schülern der 6. bis 10. Klassenstufe. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 31* (1), 18-31.

Sacks, P. (1999). *Standardized Minds – The high price of america`s testing culture and what we can do to change it.* Cambridge, Massachusetts: Perseus Publishing.

SAS Institute Inc. (2003). JMP, Release 5.0.1.2. *Computerprogramm*

Schallberger, U. (1992). Metrische Verfahren: Stand, Probleme, Perspektiven. In U. Imoberdorf, R. Käser & R. Zihlmann (Hrsg.), *Psychodiagnostik heute - Beiträge aus Theorie und Praxis* (86-95). Stuttgart: Hirzel.

Scheurer, H., Quast, A., Richter, P., Erbacher, H., & Kröber, H.-L. (1994). Testtheoretische Analyse des Benton-Tests aufgrund einer Stichprobe männlicher Gewalttäter. *Diagnostica, 40* (4), 363-374.

Schmitt, N., Cortina, J.M., & Whtney, D.J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17*, 143-150.

Schmidt, K.H., & Metzler, P. (1992). *Wortschatztest (WST)*. Weinheim: Beltz.

Schmotzer, C., Kubinger, K.D., & Maryschka, C. (1994). *Rechnen in Symbolen*. Test: Software und Manual. Frankfurt/M.:Swets.

Schöttke, H. (1988). *Das Schädel-Hirn-Trauma. Defizite der Informationsverarbeitung und ihre Diagnose.* Regensburg: Roderer Verlag.

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

Schuhfried, Fa.Dr.G. (1995). *Advanced Progressive Matrices. APM*. Mödling: Schuhfried.

Schulz, P.E.W. (1997). *Psychodiagnostik-fragwürdige Grundlagen-fragwürdige Praxis.* Berlin: Verlag Dr. Köster.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

Sixtl, F. (1993). *Der Mythos des Mittelwertes – Neue Methodenlehre der Statistik*. München: Oldenbourg

Sixtl, F. (1996). *Einführung in die Exakte Psychologie*. München: Oldenbourg

Sixtl, F. (1999). Der Abschied von Homme Moyen alias Average Person. In W. Hacker & Rinck, M. (Hrsg.), *Bericht über den 41. Kongress der Deutschen Gesellschaft für Psychologie 1998: Schwerpunktthema "Zukunft gestalten"* (S. 519-526). Lengerich: Pabst Science.

Spiel, C., Gittler, G., Sirsch, U., & Glück, J. (1997). Application of the Rasch Model for Testing Piaget`s Theory of Cognitive Development. In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social siences* (S. 314-323). Münster: Waxmann.

Spiel, C., Glück, J., & Gößler, H. (2001). Stability and change of unidimensionality: The sample case of deductive reasoning. *Journal of Adolescent Research, 16* (2), *150-168.*

Spiel, C., Glück, J., & Gößler, H. (2004). Schlussfolgerndes Denken – SDV. *Diagnostica, 50* (3), 145-15.

Spiel, C., Gößler, H., & Glück, J. (2005). Test zur Erfassung von Leistungsprofil und Leistungshöhe im schlussfolgernden Denken-Verbal (SDV). Mödling: Schuhfried.

Sprung, L., & Sprung, H. (2000). Methodenlehre der Psychologie: System und Geschichte. *Zeitschrift für Differentielle und Diagnostische Psychologie, 21* (1), 31-48.

Srp, G. (1994). *Syllogismen*. Test: Software und Manual. Frankfurt/M.: Swets.

Steck, P.H. (1981). Probleme der klinisch-psychologischen Diagnostik. In E. G. Wehner (Hrsg.), *Psychodiagnostik in Theorie und Praxis* (S. 55-84). Frankfurt: Verlag Peter D. Lang

Steck, P.H. (1991). Bemerkungen zu L. Tents Beitrag „Psychodiagnostische Verfahren und die minima scientifica". *Diagnostica, 37* (1), 89-92.

Steck, P.H. (1997). Aus der Arbeit des Testkuratoriums. Psychologische Testverfahren in der Praxis. Ergebnisse einer Umfrage unter Testanwendern. *Diagnostica, 43*, 267-284.

Steck, P.H. (2002). Eine modifizierte Anwendung des Visual Retention Test nach A. L. Benton. *Zeitschrift für Neuropsychologie, 13* (2), 127-138.

Steck, P.H. (2005). A revision of A. L. Benton`s Visual Retention Test (BVRT) in two parallel forms. *Archives of Clinical Neuropsychology, 20*, 409-416.

Steck, P.H., Rausche A., & Steege, F.W. (1975). Faktorenanalytischer Vergleich gebräuchlicher deutschsprachiger Intelligenztests. *Psychologie & Praxis, 19*, 145- 158.

Steck, P.H., Beer, U., Frey, A., Frühschütz, H.G., & Körner, A. (1990). Testkritische Überprüfung einer 30-Item-Version des Visual Retention Tests nach A. L. Benton. *Diagnostica, 36* (1), 38-49.

Stieglitz, R.-D., Smolka, M., Bech, P., & Helmchen, H. (1998). *Bech-Rafaelsen-Melancholie-Skala (BRMS)* (Handanweisung). Göttingen: Hogrefe.

Stoll, F. (1992). Bilanz zur Tagung. In U. Imoberdorf, R. Käser & R. Zihlmann (Hrsg.), *Psychodiagnostik heute - Beiträge aus Theorie und Praxis* (309-315). Stuttgart: Hirzel.

Stoll, F. (1998). Computergestützte Diagnostik: Eine Chance für neue Messmodelle. In U. Imoberdorf, R. Käser & R. Zihlmann (Hrsg.), *Psychodiagnostik von Individuen, Gruppen und Organisationen* (15-23). Stuttgart: Hirzel.

Stumpf , H. (1996). Klassische Testtheorie. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 411-430). Weinheim: Beltz.

Sturm, W., & Hartje, W. (1989). Aufgaben- und Untersuchungsverfahren der allgemeinen Psychodiagnostik bei Hirnschädigungen. In: K. Poeck (Hrsg.), *Klinische Neuropsychologie.* Stuttgart-New York: Thieme.

Styles, I., & Raven, M. (1998). *Standard Progressive Matrices. Plus Version*. Dumfries: J. C. Raven.

Tarnai, C., & Dotterweich, F. (1996). *Anwendung der Latent Class Analyse zur Identifikation typischer Interessenprofile*. Münster: Institut für sozialwissenschaftliche Forschung.

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*, 55-73.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond Group-Mean Differences: The Concept of Item Bias. *Psychological Bulletin, 99* (1), 118-128.

Thorndike, R.L. (1947). *Research Problems and Techniques*. Report No.3. AAF Aviation Psychology Program Research Reports. Washington: U.S. Government Printing Office.

Van Abswoude, A.A.H., Van der Ark, L.A., & Sijtsma, K. (2004). A Comparative Study of Test Data Dimensionality Assessment Procedures Under Nonparametric IRT Models. *Applied Psychological Measurement, 28* (1), 3-24.

Ven, van der, A.H.G.S., & Ellis, J.L. (2000). A Rasch analysis of Raven`s standard progressive matrices. *Personality and Individual Differences, 29,* 45-64.

Wagner-Menghin, M.M. (2002a). Der Lexikon-Wissens-Test (LEWITE) als Prototyp moderner Testkonzepte. *Psychologie in Österreich, 2-3*, 68-72.

Wagner-Menghin, M.M. (2002b). Towards the identification of non-scalable personality questionnaire respondents: Taking response time into account. *Psychologische Beiträge, 44*, 62-77.

Walcher, W. (1989). *Praktikum der Physik* (6. Auflage). Stuttgart: Teubner.

Waller, N.G., & Reise, S.P. (1989). Computerized Adaptive Personality Assessment: An Illustration with the Absorption Scale. *Journal of Pers. and Soc. Psychology,57,* 6, 1051-1058.

Walter, P. (1991). Die „Vermessung" des Menschen: Messtheoretische und methodologische Grundlagen psychologischen Testens. In S. Grubitzsch, *Testtheorie-Testpraxis* (S. 98-127). Reinbeck bei Hamburg: Rowohlt.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response models. *Psychometrica, 54,* 427-450.

Weise, G. (1975). *Psychologische Leistungstests-Band 1.* Göttingen: Hogrefe.

Weiß, R. (1971). Grundintelligenztest CFT-3, Skala 3. Braunschweig: Westermann

Westmeyer, H. (1972). *Logik der Diagnostik.* Stuttgart: Kohlhammer

Wittling, W. (1983). Neuropsychologische Diagnostik. In: K. J. Groffmann & L. Michel (Herausgeber), *Enzyklopädie der Psychologie*, Themenbereich B, Serie 2, Band 4: Verhaltensdiagnostik, 193-335. Göttingen: Hogrefe.

Wottawa, H., & Hossiep, R. (1997). *Anwendungsfelder psychologischer Diagnostik*. Göttingen: Hogrefe

Wytek, R., Opgenoorth, E., & Presslich, O. (1984). Development of a New Shortened Version of Raven`s Matrices Test for Application and Rough Assessment of Present Intellectual Capacity within Psychopathological Investigation. *Psychopathology, 17*, 49-58.

Zweschper, H.-D. (1992). *Mischverteilungs-Raschanalyse des Visual Retention-Test nach A. L. Benton.* Psychologische Diplomarbeit, Universität Konstanz.

9 ANHANG

9.1 Anhang A: Tabellenverzeichnis

Tabelle 1-1 mittlere klassenspezifische Itemlösungswahrscheinlichkeiten (KINDERMANN, 2001) *40*
Tabelle 4-1 Kodierung der Diagnosen in der Gesamtstichprobe..... *74*
Tabelle 4-2 Kodierung der Bildungsgrade in der Gesamtstichprobe *74*
Tabelle 4-3 Bildungsgrade der Diagnosegruppen in der Gesamtstichprobe..... *75*
Tabelle 4-4 Kodierung der Altersgruppen in der Gesamtstichprobe..... *75*
Tabelle 4-5 Altersverteilung der Diagnosegruppen in der Gesamtstichprobe..... *75*
Tabelle 4-6 Geschlechterverteilung in der Gesamtstichprobe..... *75*
Tabelle 4-7 Durchgeführte Tests in der Gesamtstichprobe *76*
Tabelle 4-8 klassische Testkennwerte *95*
Tabelle 4-9 Testreihenfolgen..... *96*
Tabelle 4-10 Varianzanalysen der Testscores bei unterschiedlicher Testreihenfolge *96*
Tabelle 4-11 Scoremittelwerte in den Diagnosegruppen der Altersgruppe 0 ..... *97*
Tabelle 4-12 Van der Waerden Tests der Scores sämtlicher und klinischer Diagnosegruppen *98*
Tabelle 4-13 Testscoremittelwerte in der Stichprobe „GPN parallel“..... *100*
Tabelle 4-14 Varianzanalysen der Testscores in der Stichprobe „GPN parallel“ *100*
Tabelle 4-15 Zusammenfassung der Kontraste in der Stichprobe „GPN parallel“..... *100*
Tabelle 4-16 nominal logistische Regressionen der Stichprobe „GPN parallel“..... *103*
Tabelle 4-17 Korrelationen unterschiedlicher Variablen in der Gesamtstichprobe..... *104*
Tabelle 4-18 Hauptkomponentenanalyse der Gesamtstichprobe..... *106*
Tabelle 5-1 „Andersen-Methode“ (Teilungskriterium „diagnostische Zugehörigkeit“) beim Raventest..... *111*
Tabelle 5-2 „Andersen-Methode“ (Teilungskriterium „Testleistung“) beim Raventest..... *112*
Tabelle 5-3 Informationstheoretische Maße beim Raventest *113*
Tabelle 5-4 Likelihoodquotiententests Raven..... *113*
Tabelle 5-5 Die 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest *115*
Tabelle 5-6 Zusammenfassung der Kontraste der 3-Kl.-Lösung des Mixed-Rasch-Modells beim Raventest *116*
Tabelle 5-7 Differenzierte Betrachtung der einzelnen Klassen beim Raventest (klinische Auffälligkeit) *119*
Tabelle 5-8 Differenzierte Betrachtung der einzelnen Klassen beim Raventest (cerebrale Beteiligung)..... *121*
Tabelle 5-9 Entropiereduktionen mit dem Raventest (Kriterium: klinische Auffälligkeit)..... *123*
Tabelle 5-10 Wald Effekt Tests beim Raventest (Kriterium: klinische Auffälligkeit)..... *123*
Tabelle 5-11 ROC-Kurven beim Raventest (Kriterium: klinische Auffälligkeit)..... *123*
Tabelle 5-12 Entropiereduktionen mit dem Raventest (Kriterium: cerebrale Beteiligung) *125*
Tabelle 5-13 Wald Effekt Tests beim Raventest (Kriterium: cerebrale Beteiligung) *125*

Tabelle 5-14 ROC-Kurven beim Raventest (Kriterium: cerebrale Beteiligung *125*
Tabelle 5-15 „Andersen-Methode“ (Teilungskriterium „diagnostische Zugehörigkeit“) beim Bentontest *129*
Tabelle 5-16 „Andersen-Methode“ (Teilungskriterium „Testleistung“) beim Bentontest *130*
Tabelle 5-17 Informationstheoretische Maße beim Bentontest *130*
Tabelle 5-18 Likelihoodquotiententests Benton *131*
Tabelle 5-19 Die 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest *133*
Tabelle 5-20 Zusammenfassung der Kontraste der 3-Kl.-Lösung des Mixed-Rasch-Modells beim Bentontest .. *135*
Tabelle 5-21 Differenzierte Betrachtung der einzelnen Klassen beim Bentontest (klinische Auffälligkeit) *138*
Tabelle 5-22 Differenzierte Betrachtung der einzelnen Klassen beim Bentontest (cerebrale Beteiligung) *140*
Tabelle 5-23 Entropiereduktionen mit dem Bentontest (Kriterium: klinische Auffälligkeit) *142*
Tabelle 5-24 Wald Effekt Tests beim Bentontest (Kriterium: klinische Auffälligkeit) *142*
Tabelle 5-25 ROC-Kurven beim Bentontest (Kriterium: klinische Auffälligkeit) *142*
Tabelle 5-26 Entropiereduktionen mit dem Bentontest (Kriterium: cerebrale Beteiligung) *144*
Tabelle 5-27 Wald Effekt Tests beim Bentontest (Kriterium: cerebrale Beteiligung) *144*
Tabelle 5-28 ROC-Kurven beim Bentontest (Kriterium: cerebrale Beteiligung) *144*
Tabelle 5-29 Entropiereduktionen mit dem Bentontest (BTF) (Kriterium: klinische Auffälligkeit) *146*
Tabelle 5-30 Wald Effekt Tests beim Bentontest (BTF) (Kriterium: klinische Auffälligkeit) *146*
Tabelle 5-31 ROC-Kurven beim Bentontest (BTF) (Kriterium: klinische Auffälligkeit) *146*
Tabelle 5-32 Entropiereduktionen mit dem Bentontest (BTF) (Kriterium: cerebrale Beteiligung) *148*
Tabelle 5-33 Wald Effekt Tests beim Bentontest (BTF) (Kriterium: cerebrale Beteiligung) *148*
Tabelle 5-34 ROC-Kurven beim Bentontest (BTF) (Kriterium: cerebrale Beteiligung) *148*
Tabelle 5-35 Informationstheoretische Maße beim Cattelltest *150*
Tabelle 5-36 Likelihoodquotiententests Cattell *150*
Tabelle 5-37 Die 4-Klassen-Lösung der LCA beim Cattelltest *152*
Tabelle 5-38 Zusammenfassung der Kontraste der 4-Kl.-Lösung der LCA beim Cattelltest *154*
Tabelle 5-39 Differenzierte Betrachtung der einzelnen Klassen beim Cattelltest (klinische Auffälligkeit) *159*
Tabelle 5-40 Differenzierte Betrachtung der einzelnen Klassen beim Cattelltest (cerebrale Beteiligung) *161*
Tabelle 5-41 Entropiereduktionen mit dem Cattelltest (Kriterium: klinische Auffälligkeit) *163*
Tabelle 5-42 Wald Effekt Tests beim Cattelltest (Kriterium: klinische Auffälligkeit) *163*
Tabelle 5-43 ROC-Kurven beim Cattelltest (Kriterium: klinische Auffälligkeit) *163*
Tabelle 5-44 Entropiereduktionen mit dem Cattelltest (Kriterium: cerebrale Beteiligung) *165*
Tabelle 5-45 Wald Effekt Tests beim Cattelltest (Kriterium: cerebrale Beteiligung) *165*
Tabelle 5-46 ROC-Kurven beim Cattelltest (Kriterium: cerebrale Beteiligung) *165*
Tabelle 5-47 Entropiereduktionen bei drei verwendeten Tests (Kriterium: klinische Auffälligkeit) *168*
Tabelle 5-48 Wald Effekt Tests bei drei verwendeten Tests (Kriterium: klinische Auffälligkeit) *168*
Tabelle 5-49 ROC-Kurven bei drei verwendeten Tests (Kriterium: klinische Auffälligkeit) *169*
Tabelle 5-50 Entropiereduktionen bei drei verwendeten Tests (Kriterium: cerebrale Beteiligung) *171*
Tabelle 5-51 Wald Effekt Tests bei drei verwendeten Tests (Kriterium: cerebrale Beteiligung) *171*
Tabelle 5-52 ROC-Kurven bei drei verwendeten Tests (Kriterium: cerebrale Beteiligung) *172*

Tabelle 5-53 Zusammenfassung sämtlicher Entropiereduktionswerte *174*
Tabelle 5-54 Entropiereduktionen geordnet nach Tests (klinische Auffälligkeit) *175*
Tabelle 5-55 Entropiereduktionen geordnet nach Tests (cerebrale Beteiligung) *175*
Tabelle 5-56 Klassischer Auswertungsmodus: Testscorequartile *178*
Tabelle 5-57 Klassischer Auswertungsmodus: Variablenkategorien *178*
Tabelle 5-58 Klassischer Auswertungsmodus: Informationstheoretische Maße *179*
Tabelle 5-59 Klassischer Auswertungsmodus: Likelihoodquotiententests *179*
Tabelle 5-60 Klassischer Auswertungsmodus: Die 3-Klassen-Lösung der LCA *181*
Tabelle 5-61 Klassischer Auswertungsmodus: Zusammenfassung der Kontraste *182*
Tabelle 5-62 Klassischer Auswertungsmodus: Abkürzungen der Variablenkategorien *184*
Tabelle 5-63 Alters- und Bildungsverteilung in Klasse 3 „Hohes Leistungsniveau“ *185*
Tabelle 5-64 Alters- und Bildungsverteilung in Klasse 1 „Mittleres Leistungsniveau“ *186*
Tabelle 5-65 Alters- und Bildungsverteilung in Klasse 2 „Niedriges Leistungsniveau“ *187*
Tabelle 5-66 Klassischer Auswertungsmodus: Klassenzusammensetzung *188*
Tabelle 5-67 Klassischer Auswertungsmodus: Varianzanalysen der Raven- und Cattell-Scores *189*
Tabelle 5-68 Klassischer Auswertungsmodus: Varianzanalysen der Benton-Scores *190*
Tabelle 5-69 Klassischer Auswertungsmodus:
Zusammenfassung der Kontraste in Klasse 2 „Niedriges Leistungsniveau“ *191*
Tabelle 5-70 Klassischer Auswertungsmodus: Zellbesetzungen der Diagnosegruppen *193*
Tabelle 5-71 Klassischer Auswertungsmodus: Zellbesetzungen der Krankheitsbilder *194*
Tabelle 5-72 Klassischer Auswertungsmodus: Entropiereduktion (Kriterium: 3-Klassen-Lösung) *195*
Tabelle 5-73 Klassischer Auswertungsmodus: Wald-Effekt-Tests (Kriterium: 3-Klassen-Lösung) *195*
Tabelle 5-74 Probabilistischer Auswertungsmodus: Testscorequartile *196*
Tabelle 5-75 Probabilistischer Auswertungsmodus: Variablenkategorien *197*
Tabelle 5-76 Probabilistischer Auswertungsmodus: Informationstheoretische Maße *198*
Tabelle 5-77 Probabilistischer Auswertungsmodus: Likelihoodquotiententests *198*
Tabelle 5-78 Probabilistischer Auswertungsmodus: Die 6-Klassen-Lösung der LCA (Teil 1) *200*
Tabelle 5-79 Probabilistischer Auswertungsmodus: Die 6-Klassen-Lösung der LCA (Teil 2) *201*
Tabelle 5-80 Probabilistischer Auswertungsmodus: Zusammenfassung der Kontraste *202*
Tabelle 5-81 Probabilistischer Auswertungsmodus: Abkürzungen der Variablenkategorien *206*
Tabelle 5-82 Alters- und Bildungsverteilung in Klasse 1 ..... *207*
Tabelle 5-83 Alters- und Bildungsverteilung in Klasse 4 ..... *208*
Tabelle 5-84 Alters- und Bildungsverteilung in Klasse 3 ..... *209*
Tabelle 5-85 Alters- und Bildungsverteilung in Klasse 2 ..... *210*
Tabelle 5-86 Alters- und Bildungsverteilung in Klasse 6 ..... *211*
Tabelle 5-87 Alters- und Bildungsverteilung in Klasse 5 ..... *212*
Tabelle 5-88 Probabilistischer Auswertungsmodus: Klassenzusammensetzung *213*
Tabelle 5-89 Probabilistischer Auswertungsmodus: Varianzanalysen der Raven-Scores *214*
Tabelle 5-90 Probabilistischer Auswertungsmodus: Varianzanalysen der Cattell-Scores *215*

Tabelle 5-91 Probabilistischer Auswertungsmodus: Varianzanalysen der Benton-Scores 216
Tabelle 5-92 Probabilistischer Auswertungsmodus:Zusammenfassung der Kontraste in den Klassen 1, 2 und 5 217
Tabelle 5-93 Probabilistischer Auswertungsmodus: Zellbesetzungen der Diagnosegruppen 219
Tabelle 5-94 Probabilistischer Auswertungsmodus: Klassenbezeichnungen 220
Tabelle 5-95 Probabilistischer Auswertungsmodus: Zellbesetzungen der Krankheitsbilder 222
Tabelle 5-96 Probabilistischer Auswertungsmodus: Entropiereduktion (Kriterium: 6-Klassen-Lösung) 223
Tabelle 5-97 Probabilistischer Auswertungsmodus: Wald-Effekt-Tests (Kriterium: 6-Klassen-Lösung) 223
Tabelle 5-98 Zusammenfassung der Entropiereduktionen und Wald-Effekt-Tests beider Auswertungsmodi 225
Tabelle A-1 Kodierung der Diagnosen in der Stichprobe „GPN parallel“ 287
Tabelle A-2 Kodierung der Bildungsgrade in der Stichprobe „GPN parallel“ 287
Tabelle A-3 Bildungsgrade der Diagnosegruppen in der Stichprobe „GPN parallel“ 287
Tabelle A-4 Kodierung der Altersgruppen in der Stichprobe „GPN parallel“ 288
Tabelle A-5 Altersverteilung der Diagnosegruppen in der Stichprobe „GPN parallel“ 288
Tabelle A-6 Geschlechterverteilung in der Stichprobe „GPN parallel“ 288
Tabelle A-7 Durchgeführte Tests in der Stichprobe „GPN parallel“ 288
Tabelle A-8 Kodierung der Diagnosen in der parallelisierten Stichprobe 289
Tabelle A-9 Kodierung der Bildungsgrade in der parallelisierten Stichprobe 289
Tabelle A-10 Bildungsgrade der Diagnosegruppen in der parallelisierten Stichprobe 289
Tabelle A-11 Kodierung der Altersgruppen in der parallelisierten Stichprobe 290
Tabelle A-12 Altersverteilung der Diagnosegruppen in der parallelisierten Stichprobe 290
Tabelle A-13 Geschlechterverteilung in der parallelisierten Stichprobe 290
Tabelle A-14 Durchgeführte Tests in der parallelisierten Stichprobe 290
Tabelle A-15 Kodierung der Diagnosen in der Stichprobe `parallel-cerebral` 291
Tabelle A-16 Kodierung der Bildungsgrade in der Stichprobe `parallel-cerebral` 291
Tabelle A-17 Bildungsgrade der Diagnosegruppen in der Stichprobe `parallel-cerebral` 291
Tabelle A-18 Kodierung der Altersgruppen in der Stichprobe `parallel-cerebral` 292
Tabelle A-19 Altersverteilung der Diagnosegruppen in der Stichprobe `parallel-cerebral` 292
Tabelle A-20 Geschlechterverteilung in der Stichprobe `parallel-cerebral` 292
Tabelle A-21 Durchgeführte Tests in der Stichprobe `parallel-cerebral` 292
Tabelle A-22 Kodierung der Diagnosen in der Teilstichprobe „Testbatterie“ 293
Tabelle A-23 Kodierung der Bildungsgrade in der Teilstichprobe „Testbatterie“ 293
Tabelle A-24 Bildungsgrade der Diagnosegruppen in der Teilstichprobe „Testbatterie“ 293
Tabelle A-25 Kodierung der Altersgruppen in der Teilstichprobe „Testbatterie“ 294
Tabelle A-26 Altersverteilung der Diagnosegruppen in der Teilstichprobe „Testbatterie“ 294
Tabelle A-27 Geschlechterverteilung in der Teilstichprobe „Testbatterie“ 294
Tabelle A-28 Durchgeführte Tests in der Teilstichprobe „Testbatterie“ 294
Tabelle A-29 Kontraste in der Stichprobe „GPN parallel“ 295
Tabelle A-30 logarithmierte und summennormierte Itemparameter des Raventests 296

Tabelle A-31 Die Kontraste der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest 297
Tabelle A-32 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest 298
Tabelle A-33 Korrelationen der Itemparameter der latenten Klassen beim Raventest 298
Tabelle A-34 logarithmierte und summennormierte Itemparameter des Bentontests 299
Tabelle A-35 Die Kontraste der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest 300
Tabelle A-36 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest 301
Tabelle A-37 Korrelationen der Itemparameter der latenten Klassen beim Bentontest 302
Tabelle A-38 Die Kontraste der 4-Klassen-Lösung der LCA beim Cattelltest 303
Tabelle A-39 Itemparameter der 4-Klassen-Lösung der LCA beim Cattelltests 304
Tabelle A-40 Korrelationen der Itemparameter der latenten Klassen des Cattelltests 305
Tabelle A-41 a-c) Klassischer Auswertungsmodus: Kontraste 306
Tabelle A-42 Klassischer Auswertungsmodus: Kontraste der Klasse 2 „Niedriges Leistungsniveau“ 307
Tabelle A-43 a-o) Probabilistischer Auswertungsmodus: Kontraste 308
Tabelle A-44 Probabilistischer Auswertungsmodus: Kontraste der Klassen 1, 2 und 5 312
Tabelle A-45 Probabilistischer Auswertungsmodus: F1-Diagnosen der Personen der Klasse 3 313
Tabelle A-46 Probabilistischer Auswertungsmodus: Diagnosen der Schlaganfall- und Schädelhirntraumapatienten der Klasse 6 313
Tabelle A-47 Raschhomogene Tests 314
Tabelle A-48 Adaptive Tests 315
Tabelle A-49 Berechnung der Zugehörigkeitswahrscheinlichkeiten zu latenten Klassen 315
Tabelle A-50 Reliabilitätsmaß der probabilistischen Testtheorie 315

9.2 Anhang B: Abbildungsverzeichnis

Abbildung 1-1 Itemschwierigkeitsprofile (KINDERMANN, 2001) 41
Abbildung 3-1 Wahrscheinlichkeit als odds-ratio 56
Abbildung 3-2 Logit der Lösungswahrscheinlichkeit 57
Abbildung 3-3 Logistische Funktion 58
Abbildung 3-4 Aufgabenkennlinien (ICC) zweier Testaufgaben nach dem Rasch-Modell 59
Abbildung 3-5 Kennlinie einer Aufgabe bei gleich großen Parametern und zugehörige Informationsfunktion der Testaufgabe 66
Abbildung 3-6 Prinzip einer adaptiven Testung 68
Abbildung 4-1 Darstellung der Items des Raventests 80
Abbildung 4-2 Darstellung der Items des Bentontests 82
Abbildung 4-3 Darstellung der Items des Cattelltests 84
Abbildung 4-4 ROC-Kurve 1: Test 1 92

Abbildung 4-5 ROC-Kurve 2:Test 2 ..... *92*
Abbildung 4-6 ROC-Kurve 3: Test 3 ..... *92*
Abbildung 4-7 Graphische Darstellung der Ladungsmatrix der Hauptkomponentenanalyse *106*
Abbildung 5-1 Graphischer Modelltest für das externe Kriterium beim Raventest *110*
Abbildung 5-2 Graphischer Modelltest für das interne Kriterium beim Raventest *111*
Abbildung 5-3 Ravenscores in den Klassen *114*
Abbildung 5-4 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest *117*
Abbildung 5-5 Graphischer Modelltest für das externe Kriterium beim Bentontest *128*
Abbildung 5-6 Graphischer Modelltest für das interne Kriterium beim Bentontest *128*
Abbildung 5-7 Bentonscores in den Klassen *132*
Abbildung 5-8 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest *136*
Abbildung 5-9 Cattellscores in den Klassen *151*
Abbildung 5-10 Itemparameter der 4-Klassen-Lösung der LCA beim Cattelltest *155*
Abbildung 5-11 Profilverläufe der ersten vier Cattelltest-Items *157*
Abbildung 5-12 Klassischer Auswertungsmodus: klassenspezifische Durchschnittsscores *183*
Abbildung 5-13 Merkmalsprofil: Klasse 3 „Hohes Leistungsniveau“ *185*
Abbildung 5-14 Merkmalsprofil: Klasse 1 „Mittleres Leistungsniveau“ *186*
Abbildung 5-15 Merkmalsprofil: Klasse 2 „Niedriges Leistungsniveau“ *187*
Abbildung 5-16 Klassischer Auswertungsmodus: Diagnosegruppen innerhalb der Klassen *192*
Abbildung 5-17 Klassischer Auswertungsmodus: Krankheitsbilder innerhalb der Klassen *193*
Abbildung 5-18 Probabilistischer Auswertungsmodus: klassenspezifische Durchschnittsscores *204*
Abbildung 5-19 Merkmalsprofil: Klasse 1 ..... *207*
Abbildung 5-20 Merkmalsprofil: Klasse 4 ..... *208*
Abbildung 5-21 Merkmalsprofil: Klasse 3 ..... *209*
Abbildung 5-22 Merkmalsprofil: Klasse 2 ..... *210*
Abbildung 5-23 Merkmalsprofil: Klasse 6 ..... *211*
Abbildung 5-24 Merkmalsprofil: Klasse 5 ..... *212*
Abbildung 5-25 Probabilistischer Auswertungsmodus: Diagnosegruppen innerhalb der Klassen *219*
Abbildung 5-26 Probabilistischer Auswertungsmodus: Krankheitsbilder innerhalb der Klassen *221*

9.3 Anhang C: Ergänzende Tabellen

Stichprobe „GPN parallel“ (N=210)

Tabelle A-1 Kodierung der Diagnosen in der Stichprobe „GPN parallel“

Kodierung der Diagnosen		#VP *(davon* **mit** *cerebraler Beteiligung)*
0	Klinisch unauffällige Personen	70 *(0)*
1	F1 (psychotrope Substanzen)	31*(3)*
2	F2 (Schizophrenie)	13 *(0)*
3	F3 & F4 (Affektive & Neurotische Störungen)	31*(8)*
4	F5 & F6 & F8 & F9 (Persönlichkeitsstörungen etc.)	5 *(0)*
5	F0 & neurologische Patienten (Schlaganfall, SHT, Tumor)	60(60)
		Σ 210 *(71)*

Tabelle A-2 Kodierung der Bildungsgrade in der Stichprobe „GPN parallel“

Kodierung der Bildungsgrade		#G	#P	#N	#VP
0	Kein Schulabschluss	-	-	-	-
1	Hauptschulabschluss	28	28	28	84
2	Mittlere Reife	26	25	27	78
3	Abitur	16	16	16	48
		Σ 70	**Σ 69**	**Σ 71**	**Σ 210**

Tabelle A-3 Bildungsgrade der Diagnosegruppen in der Stichprobe „GPN parallel“

Bildungsgrade der Diagnosegruppen					
Diagnose-gruppe	Bildungsgrad 0 # VP	Bildungsgrad 1 # VP	Bildungsgrad 2 # VP	Bildungsgrad 3 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	-	13	14	4	31*(3)*
2	-	5	2	6	13 *(0)*
3	-	12	10	9	31*(8)*
4	-	2	3	-	5 *(0)*
5	-	24	23	13	60 *(60)*
					Σ 140 *(71)*

Tabelle A-4 Kodierung der Altersgruppen in der Stichprobe „GPN parallel“

Kodierung der Altersgruppen		#G	#P	#N	#VP
0	≤ 39. Lebensjahr	28	28	28	84
1	40. – 49. Lebensjahr	26	26	26	78
2	50. – 59. Lebensjahr	12	11	13	36
3	60. – 70. Lebensjahr	4	4	4	12
4	> 70. Lebensjahr	-	-	-	-
		Σ 70	**Σ 69**	**Σ 71**	**Σ 210**

Tabelle A-5 Altersverteilung der Diagnosegruppen in der Stichprobe „GPN parallel“

Altersverteilung der Diagnosegruppen						
Diagnose-gruppe	Altersgruppe 0 # VP	Altersgruppe 1 # VP	Altersgruppe 2 # VP	Altersgruppe 3 # VP	Altersgruppe 4 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	9	13	9	-	-	31*(3)*
2	7	4	2	-	-	*13 (0)*
3	14	12	2	3	-	31*(8)*
4	4	-	-	1	-	5 *(0)*
5	22	23	11	4	-	*(60)*
						Σ 140 *(71)*

Tabelle A-6 Geschlechterverteilung in der Stichprobe „GPN parallel“

Geschlecht		#G	#P	#N	#VP
0	♀	21	20	28	69
1	♂	49	49	43	141
		Σ 70	**Σ 69**	**Σ 71**	**Σ 210**

Tabelle A-7 Durchgeführte Tests in der Stichprobe „GPN parallel“

Durchgeführte Tests	#G	#P	#N	#VP
Raven	70	67	71	208
Cattell	70	68	69	207
Benton	70	69	68	207
Alle drei	70	67	68	205

parallelisierte Stichprobe (N=468)

Tabelle A-8 Kodierung der Diagnosen in der parallelisierten Stichprobe

Kodierung der Diagnosen		#VP *(davon* **mit** *cerebraler Beteiligung)*
0	Klinisch unauffällige Personen	235 *(0)*
1	F1 (psychotrope Substanzen)	48*(3)*
2	F2 (Schizophrenie)	18 *(0)*
3	F3 & F4 (Affektive & Neurotische Störungen)	40*(13)*
4	F5 & F6 & F8 & F9 (Persönlichkeitsstörungen etc.)	8 *(0)*
5	F0 & neurologische Patienten (Schlaganfall, SHT, Tumor)	119*(119)*
		Σ 468 *(135)*

Tabelle A-9 Kodierung der Bildungsgrade in der parallelisierten Stichprobe

Kodierung der Bildungsgrade		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	Kein Schulabschluss	-	-	-
1	Hauptschulabschluss	85	83*(46)*	168
2	Mittlere Reife	86	86*(34)*	172
3	Abitur	64	64*(55)*	128
		Σ 235	**Σ 233 *(135)***	**Σ 468**

Tabelle A-10 Bildungsgrade der Diagnosegruppen in der parallelisierten Stichprobe

Bildungsgrade der Diagnosegruppen					
Diagnose-gruppe	Bildungsgrad 0 # VP	Bildungsgrad 1 # VP	Bildungsgrad 2 # VP	Bildungsgrad 3 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	-	15	31	1	47*(3)*
2	-	7	8	3	18 *(0)*
3	-	14	13	13	40*(13)*
4	-	4	4	0	8 *(0)*
5	-	42	30	47	119*(119)*
					Σ 232 *(135)*

Tabelle A-11 Kodierung der Altersgruppen in der parallelisierten Stichprobe

Kodierung der Altersgruppen		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	≤ 39. Lebensjahr	80	81*(29)*	161
1	40. – 49. Lebensjahr	64	64*(35)*	128
2	50. – 59. Lebensjahr	55	55*(41)*	110
3	60. – 70. Lebensjahr	31	29*(25)*	60
4	> 70. Lebensjahr	5	5*(5)*	10
		Σ 235	**Σ 234 *(135)***	**Σ 469**

Tabelle A-12 Altersverteilung der Diagnosegruppen in der parallelisierten Stichprobe

Altersverteilung der Diagnosegruppen						
Diagnose-gruppe	Alters-gruppe 0 # VP	Alters-gruppe 1 # VP	Alters-gruppe 2 # VP	Alters-gruppe 3 # VP	Alters-gruppe 4 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	19	18	11	0	0	81*(29)*
2	13	2	3	0	0	64*(35)*
3	18	13	6	3	0	55*(41)*
4	7	0	0	1	0	29*(25)*
5	23	31	35	25	5	5*(5)*
						Σ 233 *(135)*

Tabelle A-13 Geschlechterverteilung in der parallelisierten Stichprobe

Geschlecht		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	♀	94	81*(50)*	175
1	♂	141	153*(85)*	294
		Σ 235	**Σ 234 *(135)***	**Σ 469**

Tabelle A-14 Durchgeführte Tests in der parallelisierten Stichprobe

Durchgeführte Tests	#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
Raven	235	225*(131)*	460
Cattell	235	228*(131)*	463
Benton	235	229*(131)*	464
Alle drei	235	220*(126)*	455

Stichprobe `parallel-cerebral` (N=269):

Tabelle A-15 Kodierung der Diagnosen in der Stichprobe `parallel-cerebral`

Kodierung der Diagnosen		#VP *(davon* **mit** *cerebraler Beteiligung)*
0	Klinisch unauffällige Personen	134
1	F1 (psychotrope Substanzen)	*(3)*
2	F2 (Schizophrenie)	-
3	F3 & F4 (Affektive & Neurotische Störungen)	*(13)*
4	F5 & F6 & F8 & F9 (Persönlichkeitsstörungen etc.)	-
5	F0 & neurologische Patienten (Schlaganfall, SHT, Tumor)	*(119)*
		Σ 269 *(135)*

Tabelle A-16 Kodierung der Bildungsgrade in der Stichprobe `parallel-cerebral`

Kodierung der Bildungsgrade		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	Kein Schulabschluss	-	-	-
1	Hauptschulabschluss	45	*(46)*	91
2	Mittlere Reife	34	*(34)*	68
3	Abitur	55	*(55)*	110
		Σ 134	**Σ 135 *(135)***	**Σ 269**

Tabelle A-17 Bildungsgrade der Diagnosegruppen in der Stichprobe `parallel-cerebral`

Bildungsgrade der Diagnosegruppen					
Diagnose-gruppe	Bildungsgrad 0 # VP	Bildungsgrad 1 # VP	Bildungsgrad 2 # VP	Bildungsgrad 3 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	-	2	1	-	*(3)*
2	-	-	-	-	-
3	-	2	3	8	*(13)*
4	-	-	-	-	-
5	-	42	30	47	*(119)*
					Σ 135 *(135)*

Tabelle A-18 Kodierung der Altersgruppen in der Stichprobe `parallel-cerebral`

Kodierung der Altersgruppen		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	≤ 39. Lebensjahr	28	*(29)*	57
1	40. – 49. Lebensjahr	35	*(35)*	70
2	50. – 59. Lebensjahr	41	*(41)*	82
3	60. – 70. Lebensjahr	25	*(25)*	50
4	> 70. Lebensjahr	5	*(5)*	10
		Σ 134	**Σ 135 *(135)***	**Σ 269**

Tabelle A-19 Altersverteilung der Diagnosegruppen in der Stichprobe `parallel-cerebral`

Altersverteilung der Diagnosegruppen						
Diagnose-gruppe	Alters-gruppe 0 # VP	Alters-gruppe 1 # VP	Alters-gruppe 2 # VP	Alters-gruppe 3 # VP	Alters-gruppe 4 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	-	1	2	-	-	*(3)*
2	-	-	-	-	-	-
3	6	3	4	-	-	*(13)*
4	-	-	-	-	-	-
5	23	31	35	25	5	*(119)*
						Σ 135 *(135)*

Tabelle A-20 Geschlechterverteilung in der Stichprobe `parallel-cerebral`

Geschlecht		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	♀	49	*(50)*	99
1	♂	85	*(85)*	170
		Σ 134	**Σ 135 *(135)***	**Σ 269**

Tabelle A-21 Durchgeführte Tests in der Stichprobe `parallel-cerebral`

Durchgeführte Tests	#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
Raven	134	131	265
Cattell	134	131	265
Benton	134	131	265
Alle drei	134	126	260

Teilstichprobe „Testbatterie“(N=590):

Tabelle A-22 Kodierung der Diagnosen in der Teilstichprobe „Testbatterie“

Kodierung der Diagnosen		#VP *(davon* **mit** *cerebraler Beteiligung)*
0	Klinisch unauffällige Personen	325 *(0)*
1	F1 (psychotrope Substanzen)	64*(3)*
2	F2 (Schizophrenie)	25 *(0)*
3	F3 & F4 (Affektive & Neurotische Störungen)	47*(14)*
4	F5 & F6 & F8 & F9 (Persönlichkeitsstörungen etc.)	14 *(0)*
5	F0 & neurologische Patienten (Schlaganfall, SHT, Tumor)	115*(115)*
		Σ 590 *(132)*

Tabelle A-23 Kodierung der Bildungsgrade in der Teilstichprobe „Testbatterie“

Kodierung der Bildungsgrade		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	Kein Schulabschluss	1	11 *(0)*	12
1	Hauptschulabschluss	88	104*(47)*	192
2	Mittlere Reife	111	81*(31)*	192
3	Abitur	125	69*(54)*	194
		Σ 325	**Σ 265 *(132)***	**Σ 590**

Tabelle A-24 Bildungsgrade der Diagnosegruppen in der Teilstichprobe „Testbatterie“

Bildungsgrade der Diagnosegruppen					
Diagnose-gruppe	Bildungsgrad 0 # VP	Bildungsgrad 1 # VP	Bildungsgrad 2 # VP	Bildungsgrad 3 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	2	27	31	4	64*(3)*
2	1	14	5	5	25 *(0)*
3	3	16	14	14	47*(14)*
4	5	5	4	0	14 *(0)*
5	0	42	27	46	115*(115)*
					Σ 265 *(132)*

Tabelle A-25 Kodierung der Altersgruppen in der Teilstichprobe „Testbatterie“

Kodierung der Altersgruppen		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	≤ 39. Lebensjahr	153	103*(28)*	256
1	40. – 49. Lebensjahr	64	75*(37)*	139
2	50. – 59. Lebensjahr	63	56*(42)*	119
3	60. – 70. Lebensjahr	40	29*(23)*	69
4	> 70. Lebensjahr	5	2*(2)*	7
		Σ 325	**Σ 265 *(132)***	**Σ 590**

Tabelle A-26 Altersverteilung der Diagnosegruppen in der Teilstichprobe „Testbatterie“

Altersverteilung der Diagnosegruppen						
Diagnose-gruppe	Alters-gruppe 0 # VP	Alters-gruppe 1 # VP	Alters-gruppe 2 # VP	Alters-gruppe 3 # VP	Alters-gruppe 4 # VP	#VP *(davon* **mit** *cerebraler Beteiligung)*
1	31	21	12	0	0	64*(3)*
2	17	5	2	1	0	25 *(0)*
3	22	14	7	4	0	47*(14)*
4	11	2	0	1	0	14 *(0)*
5	22	33	35	23	2	115*(115)*
						Σ 265 *(132)*

Tabelle A-27 Geschlechterverteilung in der Teilstichprobe „Testbatterie“

Geschlecht		#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
0	♀	134	83*(46)*	217
1	♂	191	182*(86)*	373
		Σ 325	**Σ 265 *(132)***	**Σ 590**

Tabelle A-28 Durchgeführte Tests in der Teilstichprobe „Testbatterie“

Durchgeführte Tests	#klin. unauffällig	#klin. auffällig *(davon* **mit** *cerebraler Beteiligung)*	#VP
Raven	325	265*(132)*	590
Cattell	325	265*(132)*	590
Benton	325	265*(132)*	590
Alle drei	325	265*(132)*	590

Tabelle A-29 Kontraste in der Stichprobe „GPN parallel“

Kontraste der einzelnen Gruppen in der Stichprobe: „GPN parallel“ (N=210)			
Klinisch unauffällige Personen (G); psychiatr. Patienten ohne cerebr. Beteiligung (P); neurologische Patienten mit cerebr. Beteiligung (N)			
Kontrasttabelle	**G←→P**	**G←→N**	**P←→N**
t-Tests der Cattellscoremittelwerte der einzelnen Gruppen			
t Test	2,394	2,476	0,165
FG	136	137	135
Prob > |t|	0,0180	0,0145*	0,8693
t-Tests der Bentonscoremittelwerte der einzelnen Gruppen			
t Test	2,547	2,522	0,037
FG	137	136	135
Prob > |t|	0,0120*	0,0128*	0,9704

*p≤ 0,0166;**p≤ 0,0033 (Bonferoni-Korrektur)

Tabelle A-30 logarithmierte und summennormierte Itemparameter des Raventests

logarithmierte und summennormierte Itemparameter σ_i des Raventests sowie die Reihenfolge ihrer Schwierigkeiten für die jeweiligen Teilstichproben (1. am leichtesten)										
	Raven Gesamt		Raven klU		Raven klA		Raven NS		Raven HS	
Item1	-2,813	2.	-2,788	2.	-2,834	2.	-2,937	2.	-1,842	3.
Item2	-1,979	3.	-2,214	3.	-1,798	4.	-2,065	3.	-1,360	5.
Item3	-3,044	1.	-3,046	1.	-3,043	1.	-2,979	1.	-3,466	1.
Item4	-1,242	5.	-1,488	5.	-1,036	5.	-1,224	5.	-1,360	6.
Item5	-0,962	6.	-1,046	6.	-0,882	6.	-1,027	6.	-0,637	7.
Item6	-1,875	4.	-1,794	4.	-1,948	3.	-1,856	4.	-2,068	2.
Item7	0,504	12.	0,510	11.	0,506	12.	0,754	13.	-0,069	11.
Item8	0,732	14.	0,722	14.	0,752	14.	0,739	12.	0,721	13.
Item9	-0,354	9.	-0,395	8.	-0,309	9.	-0,382	9.	-0,246	10.
Item10	1,311	16.	1,285	16.	1,354	16.	1,347	16.	1,276	16.
Item11	0,632	13.	0,599	12.	0,676	13.	0,844	14.	0,224	12.
Item12	-0,770	7.	-0,854	7.	-0,688	8.	-0,664	7.	-1,498	4.
Item13	-0,521	8.	-0,222	9.	-0,832	7.	-0,507	8.	-0,572	8.
Item14	-0,052	10.	0,003	10.	-0,101	10.	0,022	10.	-0,295	9.
Item15	1,039	15.	1,065	15.	1,020	15.	1,000	15.	1,087	15.
Item16	0,467	11.	0,652	13.	0,276	11.	0,264	11.	0,836	14.
Item17	2,378	19.	2,318	19.	2,488	19.	3,156	20.	2,142	18.
Item18	1,791	17.	1,761	17.	1,846	17.	1,568	17.	1,949	17.
Item19	2,666	20.	2,782	20.	2,518	20.	2,071	19.	2,957	20.
Item20	2,094	18.	2,150	18.	2,034	18.	1,875	18.	2,222	19.

Tabelle A-31 Die Kontraste der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest

Kontraste der einzelnen Klassen des Raventests			
	Könner ←→ Durchschnittliche	**Könner ←→ Nichtkönner**	**Durchschnittliche ←→ Nichtkönner**
t-Tests der Scores der einzelnen Klassen			
t Test	12,904	27,960	14,083
FG	503	448	253
Prob > |t|	<,0001***	<,0001***	<,0001***
t-Tests des Lebensalters der einzelnen Klassen			
t Test	-2,920	-6,037	-3,053
FG	503	448	253
Prob > |t|	0,0037*	<,0001***	0,0025**
Wilcoxon-Tests der Altersgruppen der einzelnen Klassen			
χ^2 approx.	4,6979	30,1338	10,3232
FG	1	1	1
Prob> χ^2	0,0302	<,0001***	0,0013**
Wilcoxon-Tests der Bildungsgrade der einzelnen Klassen (Bildungsgrad 0 zugeordnet zu Bildungsgrad 1)			
χ^2 approx.	19,7925	42,4756	9,6920
FG	1	1	1
Prob> χ^2	<,0001***	<,0001***	0,0019**
Likelihood Ratio-Tests der Kontingenztafeln der einzelnen Klassen (Anzahl klinisch auffälliger Personen in den Klassen)			
χ^2	3,947	4,955	0,224
FG	1	1	1
Prob> χ^2	0,0469	0,0260	0,6359

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)

Tabelle A-32 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Raventest

Die Itemparameter der 3 Klassen des Mixed-Rasch-Modells beim Raventest (geordnet nach Schwierigkeit)						
Schwierigkeit (1.am leichtesten)	**Raven MRM3, 1.Kl.**		**Raven MRM3, 2.Kl.**		**Raven MRM3, 3.Kl.**	
1.	R-Item3	-3,038	R-Item1	-3,661	R-Item1	-2,970
2.	R-Item6	-2,652	R-Item3	-3,598	R-Item3	-2,889
3.	R-Item9	-2,584	R-Item2	-2,460	R-Item2	-2,008
4.	R-Item1	-2,015	R-Item6	-2,245	R-Item6	-1,346
5.	R-Item2	-1,559	R-Item5	-1,696	R-Item9	-1,202
6.	R-Item4	-1,474	R-Item4	-1,289	R-Item4	-0,967
7.	R-Item14	-1,322	R-Item12	-1,239	R-Item5	-0,843
8.	R-Item12	-1,132	R-Item13	-0,497	R-Item13	-0,580
9.	R-Item5	-0,479	R-Item16	-0,243	R-Item14	-0,312
10.	R-Item13	-0,407	R-Item7	0,405	R-Item18	-0,208
11.	R-Item11	0,480	R-Item8	0,534	R-Item15	-0,135
12.	R-Item7	0,499	R-Item11	1,045	R-Item16	0,121
13.	R-Item8	1,074	R-Item15	1,179	R-Item19	0,218
14.	R-Item16	1,118	R-Item14	1,388	R-Item10	0,373
15.	R-Item10	1,339	R-Item9	1,518	R-Item8	0,531
16.	R-Item15	1,391	R-Item20	1,794	R-Item12	0,723
17.	R-Item18	2,243	R-Item18	2,009	R-Item20	0,961
18.	R-Item17	2,526	R-Item10	2,019	R-Item11	1,212
19.	R-Item20	2,587	R-Item19	2,327	R-Item17	2,319
20.	R-Item19	3,407	R-Item17	2,711	R-Item7	7,000

Tabelle A-33 Korrelationen der Itemparameter der latenten Klassen beim Raventest

Klassenspezifische Itemparameter-Korrelationen beim Raventest (3-Klassen-Mixed-Rasch-Modell) (N=20)			
(Spearman Rho)	"Könner"	"Durchschnittliche"	"Nicht-Könner"
"Könner"		0,78***	0,74***
"Durchschnittliche"	0,78***		0,64***
"Nicht-Könner"	0,74***	0,64***	
(Produkt-Moment)	"Könner"	"Durchschnittliche"	"Nicht-Könner"
"Könner"		0,76***	0,52*
"Durchschnittliche"	0,76***		0,53*
"Nicht-Könner"	0,52*	0,53*	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle A-34 logarithmierte und summennormierte Itemparameter des Bentontests

logarithmierte und summennormierte Itemparameter σ_i des Bentontests sowie die Reihenfolge ihrer Schwierigkeiten für die jeweiligen Teilstichproben (1. am leichtesten)										
	Benton Gesamt		Benton klU		Benton klA		Benton NS		Benton HS	
item1	-1,475	2.	-1,436	1.	-1,532	3.	-1,355	3.	-2,383	1.
item2	-0,491	8.	-0,537	6.	-0,38	8.	-0,538	8.	-0,417	6.
item3	-1,462	3.	-1,35	2.	-1,559	2.	-1,470	2.	-1,508	2.
item4	-1,559	1.	-1,241	3.	-1,909	1.	-1,679	1.	-1,184	3.
item5	0,927	17.	0,939	17.	0,872	16.	0,886	16.	0,957	17.
item6	-0,064	11.	-0,048	11.	-0,107	11.	-0,170	10.	0,091	12.
item7	-0,25	9.	-0,266	9.	-0,242	9.	-0,320	9.	-0,142	9.
item8	-0,687	5.	-0,694	5.	-0,707	5.	-0,675	6.	-0,762	5.
item9	-0,126	10.	-0,097	10.	-0,184	10.	-0,170	11.	-0,068	10.
item10	0,142	13.	0,175	13.	0,103	12.	0,136	13.	0,134	13.
item11	-0,559	6.	-0,481	8.	-0,686	6.	-0,731	5.	-0,219	8.
item12	-0,52	7.	-0,518	7.	-0,541	7.	-0,648	7.	-0,273	7.
item13	0,843	16.	0,762	15.	0,952	17.	1,022	17.	0,688	15.
item14	0,642	14.	0,673	14.	0,598	14.	0,607	14.	0,671	14.
item15	0,074	12.	0,033	12.	0,141	13.	0,122	12.	-0,021	11.
item16	1,532	19.	1,378	19.	1,763	20.	1,564	18.	1,516	20.
item17	-0,842	4.	-0,905	4.	-0,814	4.	-0,872	4.	-0,802	4.
item18	1,485	18.	1,316	18.	1,712	19.	1,682	19.	1,385	18.
item19	0,801	15.	0,776	16.	0,832	15.	0,691	15.	0,895	16.
item20	1,59	20.	1,52	20.	1,686	18.	1,919	20.	1,443	19.

Tabelle A-35 Die Kontraste der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest

Kontraste der einzelnen Klassen des Bentontests			
	KönnerA ←→ KönnerB	**KönnerA ←→ Nichtkönner**	**KönnerB ←→ Nichtkönner**
t-Tests der Scores der einzelnen Klassen			
t Test	-0,832	-6,583	7,678
FG	384	308	522
Prob > |t|	0,4059	<,0001***	<,0001***
t-Tests der Fehler (BTF) der einzelnen Klassen			
t Test	1,387	6,404	-7,215
FG	384	308	522
Prob > |t|	0,1662	<,0001***	<,0001***
t-Tests des Lebensalters der einzelnen Klassen			
t Test	-0,197	2,574	-3,744
FG	384	308	522
Prob > |t|	0,8443	0,0105*	0,0002***
Wilcoxon-Tests der Altersgruppen der einzelnen Klassen			
χ^2 approx.	0,0191	4,5558	7,9887
FG	1	1	1
Prob> χ^2	0,8902	0,0328	0,0047*
Wilcoxon-Tests der Bildungsgrade der einzelnen Klassen (Bildungsgrad 0 zugeordnet zu Bildungsgrad 1)			
χ^2 approx.	0,9358	8,2940	32,1093
FG	1	1	1
Prob> χ^2	0,3334	0,0040*	<,0001***
Likelihood Ratio-Tests der Kontingenztafeln der einzelnen Klassen (Anzahl klinisch auffälliger Personen in den Klassen)			
χ^2	0,001	3,686	7,443
FG	1	1	1
Prob> χ^2	0,9816	0,0549	0,0064*

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)

Tabelle A-36 Itemparameter der 3-Klassen-Lösung des Mixed-Rasch-Modells beim Bentontest

Die Itemparameter der 3 Klassen des Mixed-Rasch-Modells beim Bentontest (geordnet nach Schwierigkeit)						
Schwierigkeit (1.am leichtesten)	**Benton MRM3, 1.Kl.**		**Benton MRM3, 2.Kl.**		**Benton MRM3, 3.Kl.**	
1.	B-Item4	-7,000	B-Item3	-1,883	B-Item14	-6,965
2.	B-Item1	-2,412	B-Item11	-1,862	B-Item2	-3,116
3.	B-Item3	-1,329	B-Item17	-1,548	B-Item8	-0,843
4.	B-Item17	-0,347	B-Item1	-1,306	B-Item3	-0,509
5.	B-Item7	-0,325	B-Item8	-1,182	B-Item1	-0,411
6.	B-Item8	-0,108	B-Item4	-1,168	B-Item9	-0,211
7.	B-Item12	-0,100	B-Item12	-1,129	B-Item11	-0,160
8.	B-Item2	0,127	B-Item9	-0,696	B-Item17	-0,119
9.	B-Item15	0,160	B-Item2	-0,682	B-Item4	0,081
10.	B-Item6	0,339	B-Item7	-0,567	B-Item6	0,140
11.	B-Item11	0,438	B-Item6	-0,420	B-Item12	0,189
12.	B-Item14	0,449	B-Item10	-0,264	B-Item15	0,323
13.	B-Item9	0,521	B-Item15	0,126	B-Item10	0,480
14.	B-Item10	0,544	B-Item13	0,186	B-Item19	0,910
15.	B-Item5	0,921	B-Item19	0,213	B-Item7	1,027
16.	B-Item13	1,295	B-Item5	0,691	B-Item20	1,364
17.	B-Item19	1,427	B-Item20	1,390	B-Item13	1,525
18.	B-Item18	1,523	B-Item16	1,459	B-Item16	2,046
19.	B-Item16	1,722	B-Item18	1,642	B-Item18	2,125
20.	B-Item20	2,154	B-Item14	7,000	B-Item5	2,126

Tabelle A-37 Korrelationen der Itemparameter der latenten Klassen beim Bentontest

Klassenspezifische Itemparameter-Korrelationen beim Bentontest (3-Klassen-Mixed-Rasch-Modell) (N=20)			
(Spearman Rho)	"Könner A"	"Könner B"	"Nicht-Könner"
"Könner A"		0,62**	0,58**
"Könner B"	0,62**		0,80***
"Nicht-Könner"	0,58**	0,80***	
(Produkt-Moment)	"Könner A"	"Könner B"	"Nicht-Könner"
"Könner A"		0,17	-0,45*
"Könner B"	0,17		0,35
"Nicht-Könner"	-0,45*	0,35	
ohne Ausreißer-Item 14 (N=19)			
(Produkt-Moment)	"Könner A"	"Könner B"	"Nicht-Könner"
"Könner A"		0,35	0,70***
"Könner B"	0,35		0,57*
"Nicht-Könner"	0,70***	0,57*	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle A-38 Die Kontraste der 4-Klassen-Lösung der LCA beim Cattelltest

Kontraste der einzelnen Klassen des Cattelltests						
	Könner ↔ Durch-schnitt A	**Könner ↔ Durch-schnitt B**	**Könner ↔ Nicht-könner**	**Durch-schnitt A ↔ Durch-schnitt B**	**Durch-schnitt A ↔ Nicht-könner**	**Durch-schnitt B ↔ Nicht-könner**
t Tests der Scores der einzelnen Klassen						
t Test	11,997	23,457	41,166	-3,671	-18,587	20,528
FG	289	440	316	287	163	314
Prob > |t|	<,0001***	<,0001***	<,0001***	0,0003**	<,0001***	<,0001***
t Tests bearbeiteter Items der einzelnen Klassen						
t Test	-3,947	8,721	4,690	-9,465	-5,877	-1,375
FG	289	440	316	287	163	314
Prob > |t|	<,0001***	<,0001***	<,0001***	<,0001***	<,0001***	0,1701
t Tests des Lebensalters der einzelnen Klassen						
t Test	-1,012	-6,470	-9,538	3,332	6,149	-3,950
FG	289	440	316	287	163	314
Prob > |t|	0,3126	<,0001***	<,0001***	0,0010*	<,0001***	<,0001***
Wilcoxon-Tests der Altersgruppen der einzelnen Klassen						
χ^2 approx.	0,0160	34,3273	63,7703	13,0748	31,1550	14,2279
FG	1	1	1	1	1	1
Prob> χ^2	0,8993	<,0001***	<,0001***	0,0003**	<,0001***	0,0002**
Wilcoxon-Tests der Bildungsgrade der einzelnen Klassen (Bildungsgrad 0 zugeordnet zu Bildungsgrad 1)						
χ^2 approx.	21,8351	11,8680	48,8182	4,7869	4,0156	20,8774
FG	1	1	1	1	1	1
Prob> χ^2	<,0001***	0,0006**	<,0001***	0,0287	0,0451	<,0001***
Likelihood Ratio-Tests der Kontingenztafeln der einzelnen Klassen (Anzahl klinisch auffälliger Personen in den Klassen)						
χ^2	2,109	13,349	27,749	1,163	8,031	5,941
FG	1	1	1	1	1	1
Prob> χ^2	0,1464	0,0003**	<,0001***	0,2807	0,0046*	0,0148

*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur)

Tabelle A-39 Itemparameter der 4-Klassen-Lösung der LCA beim Cattelltests

Die Itemparameter der 4 Klassen der Latenten-Klassen-Analyse beim Cattelltest (geordnet nach Schwierigkeit)								
Schwierigkeit (1.am leichtesten)	**Cattell LCA4, 1.Kl.**		**Cattell LCA4, 2.Kl.**		**Cattell LCA4, 3.Kl.**		**Cattell LCA4, 4.Kl.**	
1.	C-Item1	-6,000	C-Item1	-2,264	C-Item1	-0,660	C-Item8	-6,000
2.	C-Item2	-6,000	C-Item3	-1,545	C-Item3	-0,108	C-Item1	-1,837
3.	C-Item3	-4,078	C-Item6	-1,053	C-Item6	1,139	C-Item6	-1,061
4.	C-Item6	-2,495	C-Item2	-0,792	C-Item7	1,365	C-Item3	-0,811
5.	C-Item7	-1,095	C-Item5	-0,029	C-Item2	1,661	C-Item2	-0,059
6.	C-Item4	-0,467	C-Item4	0,243	C-Item4	2,018	C-Item7	0,069
7.	C-Item8	-0,215	C-Item7	1,196	C-Item9	2,157	C-Item5	0,685
8.	C-Item5	0,304	C-Item9	2,964	C-Item5	2,885	C-Item4	0,720
9.	C-Item9	2,247	C-Item13	4,924	C-Item10	2,962	C-Item9	2,173
10.	C-Item10	2,980	C-Item12	4,967	C-Item11	3,576	C-Item10	2,740
11.	C-Item12	3,792	C-Item8	5,137	C-Item13	3,914	C-Item11	2,888
12.	C-Item11	4,439	C-Item10	6,000	C-Item12	4,988	C-Item13	4,457
13.	C-Item13	6,000	C-Item11	6,000	C-Item8	6,000	C-Item12	6,000

Tabelle A-40 Korrelationen der Itemparameter der latenten Klassen des Cattelltests

klassenspezifische Itemparameter-Korrelationen beim Cattelltest (4-Klassen-LCA) (N=13)				
(Spearman Rho)	"Könner“	"Durchschnitt A"	"Durchschnitt B"	"Nicht-Könner"
"Könner“		0,83***	0,83***	0,82***
"Durchschnitt A"	0,83***		0,61*	0,54
"Durchschnitt B"	0,83***	0,61*		0,86***
"Nicht-Könner"	0,82***	0,54	0,86***	
(Produkt-Moment)	"Könner“	"Durchschnitt A"	"Durchschnitt B"	"Nicht-Könner"
"Könner“		0,67*	0,87***	0,71**
"Durchschnitt A"	0,67*		0,43	0,20
"Durchschnitt B"	0,87***	0,43		0,83***
"Nicht-Könner"	0,71**	0,20	0,83***	
ohne Ausreißer-Item 8 (N=12)				
(Produkt-Moment)	"Könner“	"Durchschnitt A"	"Durchschnitt B"	"Nicht-Könner"
"Könner“		0,88***	0,92***	0,86***
"Durchschnitt A"	0,88***		0,89***	0,93***
"Durchschnitt B"	0,92***	0,89***		0,84***
"Nicht-Könner"	0,86***	0,93***	0,84***	

*p≤ 0,05;**p≤ 0,01;***p≤ 0,001

Tabelle A-41 a-c) Klassischer Auswertungsmodus: Kontraste

(*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur))

a)

Kl.1„Mittleres Leistungsniveau“ & Kl.2„Niedriges Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	-5,502	413	<,0001***
Ravenscore	25,392	413	<,0001***
Cattellscore	11,031	413	<,0001***
bearbeitete Cattell-Items	2,736	413	0,0065*
Bentonscore	14,955	413	<,0001***
BTF	-14,418	413	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	24,0556	1	<,0001***
Bildungsgrad (0=1)	22,5570	1	<,0001***

b)

Kl.1„Mittleres Leistungsniveau“ & Kl.3„Hohes Leistungsniveau“			
	t –Test	FG	Prob > |t|
Alter	4,608	416	<,0001***
Ravenscore	-22,047	416	<,0001***
Cattellscore	-12,040	416	<,0001***
bearbeitete Cattell-Items	-1,539	416	0,1245
Bentonscore	-13,553	416	<,0001***
BTF	12,682	416	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	15,0640	1	0,0001***
Bildungsgrad (0=1)	30,1521	1	<,0001***

c)

Kl.2„Niedriges Leistungsniveau“ & Kl.3„Hohes Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	9,257	345	<,0001***
Ravenscore	-41,006	345	<,0001***
Cattellscore	-20,451	345	<,0001***
bearbeitete Cattell-Items	-3,922	345	0,0001***
Bentonscore	-25,412	345	<,0001***
BTF	21,523	345	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	54,5531	1	<,0001***
Bildungsgrad (0=1)	82,2703	1	<,0001***

Tabelle A-42 Klassischer Auswertungsmodus: Kontraste der Klasse 2 „Niedriges Leistungsniveau“

Kontraste der Klasse 2 „Niedriges Leistungsniveau“			
	klU⇔ Diag.o.cB	klU⇔ Diag.m.cB	Diag.o.cB ⇔ Diag.m.cB
t-Tests des Lebensalters			
t Test	-2,602	-2,697	-5,969
FG	124	122	92
Prob > |t|	0,0104*	0,0080*	<,0001***
t-Tests der Cattellscores			
t Test	2,033	3,330	1,138
FG	124	122	92
Prob > |t|	0,0442	0,0011**	0,2581
t-Tests der Bentonscores			
t Test	3,267	3,614	0,166
FG	124	122	92
Prob > |t|	0,0014**	0,0004**	0,8682
Wilcoxon-Tests der Bildungsgrade (Bildungsgrad 0=Bildungsgrad 1)			
χ^2 approx.	7,9233	0,3589	3,1152
FG	1	1	1
Prob> χ^2	0,0049*	0,5491	0,0776

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)

Tabelle A-43 a-o) Probabilistischer Auswertungsmodus: Kontraste

(*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur))

a)

Kl.1„jüngere Personen mit hohem Leistungsniveau“ & Kl.2„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Benton-Test“			
	t -Test	FG	Prob > |t|
Alter	-5,332	291	<,0001***
Ravenscore	15,580	291	<,0001***
Cattellscore	7,031	291	<,0001***
bearbeitete Cattell-Items	2,076	291	0,0387
Bentonscore	19,874	291	<,0001***
BTF	-18,981	291	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	23,1822	1	<,0001***
Bildungsgrad (0=1)	24,7286	1	<,0001***

b)

Kl.1„jüngere Personen mit hohem Leistungsniveau“ & Kl.3„jüngere Personen mit reduziertem Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	-2,038	302	0,0425
Ravenscore	18,223	302	<,0001***
Cattellscore	9,048	302	<,0001***
bearbeitete Cattell-Items	0,110	302	0,9123
Bentonscore	12,770	302	<,0001***
BTF	-13,864	302	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	0,9767	1	0,3230
Bildungsgrad (0=1)	39,7816	1	<,0001***

c)

Kl.1„jüngere Personen mit hohem Leistungsniveau“ & Kl.4„jüngere Personen mit leicht reduziertem Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	-1,343	307	0,1803
Ravenscore	14,817	307	<,0001***
Cattellscore	6,090	307	<,0001***
bearbeitete Cattell-Items	-0,286	307	0,7749
Bentonscore	8,151	307	<,0001***
BTF	-8,327	307	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	0,2654	1	0,6065
Bildungsgrad (0=1)	18,9642	1	<,0001***

d)

Kl.1„jüngere Personen mit hohem Leistungsniveau“ & Kl.5„ältere Personen mit Leistungseinbrüchen in allen drei Tests“			
	t -Test	FG	Prob > |t|
Alter	-7,646	292	<,0001***
Ravenscore	40,206	292	<,0001***
Cattellscore	18,748	292	<,0001***
bearbeitete Cattell-Items	2,165	292	0,0312
Bentonscore	27,482	292	<,0001***
BTF	-25,825	292	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	47,0286	1	<,0001***
Bildungsgrad (0=1)	61,9230	1	<,0001***

Fortsetzung nächste Seite

Tabellen A-43 a-o) Fortsetzung

(*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur))

e)

Kl.1„jüngere Personen mit hohem Leistungsniveau“ & Kl.6„ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“			
	t -Test	FG	Prob > |t|
Alter	-6,215	256	<,0001***
Ravenscore	18,278	256	<,0001***
Cattellscore	17,833	256	<,0001***
bearbeitete Cattell-Items	2,989	256	0,0031*
Bentonscore	14,810	256	<,0001***
BTF	-15,391	256	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	23,1945	1	<,0001***
Bildungsgrad (0=1)	22,7498	1	<,0001***

f)

Kl.2„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Benton-Test“ & Kl.3„jüngere Personen mit reduziertem Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	3,011	161	0,0030*
Ravenscore	0,563	161	0,5742
Cattellscore	1,347	161	0,1798
bearbeitete Cattell-Items	-1,579	161	0,1162
Bentonscore	-5,849	161	<,0001***
BTF	4,552	161	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	10,1884	1	0,0014**
Bildungsgrad (0=1)	0,3418	1	0,5588

g)

Kl.2„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Benton-Test“ & Kl.4„jüngere Personen mit leicht reduziertem Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	3,520	166	0,0006**
Ravenscore	-2,056	166	0,0414
Cattellscore	-0,939	166	0,3493
bearbeitete Cattell-Items	-1,896	166	0,0597
Bentonscore	-9,388	166	<,0001***
BTF	8,091	166	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	13,5475	1	0,0002**
Bildungsgrad (0=1)	2,1110	1	0,1462

h)

Kl.2„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Benton-Test“ & Kl.5„ältere Personen mit Leistungseinbrüchen in allen drei Tests“			
	t -Test	FG	Prob > |t|
Alter	-1,822	151	0,0704
Ravenscore	13,353	151	<,0001***
Cattellscore	9,286	151	<,0001***
bearbeitete Cattell-Items	0,104	151	0,9177
Bentonscore	5,638	151	<,0001***
BTF	-5,487	151	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	5,1767	1	0,0229
Bildungsgrad (0=1)	6,5769	1	0,0103

Fortsetzung nächste Seite

Tabellen A-43 a-o) Fortsetzung

(*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur))

i)

Kl.2„ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Benton-Test“ & Kl.6„ältere Personen mit stark reduziertem Leistungsniveau &relativem Leistungseinbruch im Cattelltest“			
	t -Test	FG	Prob > |t|
Alter	-1,729	115	0,0865
Ravenscore	3,044	115	0,0029*
Cattellscore	10,660	115	<,0001***
bearbeitete Cattell-Items	1,083	115	0,2809
Bentonscore	0,219	115	0,8273
BTF	-0,988	115	0,3253
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	2,0464	1	0,1526
Bildungsgrad (0=1)	0,1571	1	0,6918

j)

Kl.3„jüngere Personen mit reduziertem Leistungsniveau“ & Kl.4„jüngere Personen mit leicht reduziertem Leistungsniveau“			
	t -Test	FG	Prob > |t|
Alter	0,606	177	0,5451
Ravenscore	-2,961	177	0,0035*
Cattellscore	-2,317	177	0,0216
bearbeitete Cattell-Items	-0,136	177	0,7520
Bentonscore	-3,623	177	0,0004**
BTF	3,936	177	0,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	0,1759	1	0,6750
Bildungsgrad (0=1)	5,4973	1	0,0190

k)

Kl.3„jüngere Personen mit reduziertem Leistungsniveau“ & Kl.5„ältere Personen mit Leistungseinbrüchen in allen drei Tests“			
	t -Test	FG	Prob > |t|
Alter	-5,029	162	<,0001***
Ravenscore	14,376	162	<,0001***
Cattellscore	8,486	162	<,0001***
bearbeitete Cattell-Items	1,623	162	0,1066
Bentonscore	11,620	162	<,0001***
BTF	-10,051	162	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	25,6751	1	<,0001***
Bildungsgrad (0=1)	5,1130	1	0,0237

l)

Kl.3„jüngere Personen mit reduziertem Leistungsniveau“ & Kl.6„ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“			
	t -Test	FG	Prob > |t|
Alter	-4,445	126	<,0001***
Ravenscore	2,944	126	0,0039*
Cattellscore	9,689	126	<,0001***
bearbeitete Cattell-Items	2,384	126	0,0186
Bentonscore	4,482	126	<,0001***
BTF	-4,395	126	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	13,8298	1	0,0002**
Bildungsgrad (0=1)	0,0094	1	0,9229

Fortsetzung nächste Seite

Tabellen A-43 a-o) Fortsetzung:

(*p≤ 0,00833;**p≤ 0,00166;***p≤ 0,000166(Bonferoni-Korrektur))

m)

Kl.4„jüngere Personen mit leicht reduziertem Leistungsniveau“ & Kl.5„ältere Personen mit Leistungseinbrüchen in allen drei Tests“			
	t -Test	FG	Prob > |t|
Alter	-5,509	167	<,0001***
Ravenscore	18,509	167	<,0001***
Cattellscore	10,294	167	<,0001***
bearbeitete Cattell-Items	1,930	167	0,0553
Bentonscore	15,173	167	<,0001***
BTF	-13,249	167	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	30,1931	1	<,0001***
Bildungsgrad (0=1)	20,0178	1	<,0001***

n)

Kl.4„jüngere Personen mit leicht reduziertem Leistungsniveau“ & Kl.6„ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“			
	t -Test	FG	Prob > |t|
Alter	-4,778	131	<,0001***
Ravenscore	5,460	131	<,0001***
Cattellscore	11,053	131	<,0001***
bearbeitete Cattell-Items	2,640	131	0,0093
Bentonscore	7,061	131	<,0001***
BTF	-6,971	131	<,0001***
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	16,3625	1	<,0001***
Bildungsgrad (0=1)	3,2840	1	0,0700

o)

Kl.5„ältere Personen mit Leistungseinbrüchen in allen drei Tests“ & Kl.6„ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“			
	t -Test	FG	Prob > |t|
Alter	-0,215	116	0,8301
Ravenscore	-8,359	116	<,0001***
Cattellscore	1,195	116	0,2345
bearbeitete Cattell-Items	0,950	116	0,3441
Bentonscore	-3,886	116	0,0002**
BTF	3,102	116	0,0024*
(Wilcoxon)	CHI2aprox	FG	Prob>CHI2
Altersgruppe (4=3)	0,1501	1	0,6984
Bildungsgrad (0=1)	3,9542	1	0,0468

Tabelle A-44 Probabilistischer Auswertungsmodus: Kontraste der Klassen 1, 2 und 5

Kontraste der Klasse 1 „jüngere Personen mit hohem Leistungsniveau“			
	klU⇔ Diag.o.cB	klU⇔ Diag.m.cB	Diag.o.cB ⇔ Diag.m.cB
t-Tests der Ravenscores			
t Test	1,886	2,233	0,373
FG	183	175	70
Prob > |t|	0,0608	0,0268	0,7101
t-Tests der Cattellscores			
t Test	-0,723	2,346	2,233
FG	183	175	70
Prob > |t|	0,4703	0,0201	0,0287
Kontraste der Klasse 2 „ältere Personen mit reduziertem Leistungsniveau & relativem Leistungseinbruch im Bentontest“			
	klU⇔ Diag.o.cB	klU⇔ Diag.m.cB	Diag.o.cB ⇔ Diag.m.cB
t-Tests der Cattellscores			
t Test	-0,942	1,918	2,980
FG	48	59	39
Prob > |t|	0,3511	0,0599	0,0049*
t-Tests der Bentonscores			
t Test	-0,459	3,675	3,420
FG	48	59	39
Prob > |t|	0,6482	0,0005***	0,0015**
t-Tests des Lebensalters			
t Test	1,597	-0,011	-2,107
FG	48	59	39
Prob > |t|	0,1169	0,9909	0,0416
Wilcoxon-Tests der Bildungsgrade			
χ^2 approx.	0,1517	0,2940	0,7315
FG	1	1	1
Prob> χ^2	0,6969	0,5877	0,3924
Kontraste der Klasse 5 „ältere Personen mit Leistungseinbrüchen in allen drei Tests“			
	klU⇔ Diag.o.cB	klU⇔ Diag.m.cB	Diag.o.cB ⇔ Diag.m.cB
t-Tests der Ravenscores			
t Test	2,469	-0,254	-2,467
FG	54	51	43
Prob > |t|	0,0167	0,8005	0,0177

*p≤ 0,0166;**p≤ 0,0033;***p≤ 0,00033(Bonferoni-Korrektur)

Tabelle A-45 Probabilistischer Auswertungsmodus: F1-Diagnosen der Personen der Klasse 3

Diagnosen der Personen mit psychischen und Verhaltensstörungen durch psychotrope Substanzen der Klasse 3 „jüngere Personen mit reduziertem Leistungsniveau“ **im probabilistischen Auswertungsmodus**	**Alters gruppe**	**Bild-ungs-grad**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**2**	**2**
Störung durch Alkohol	**1**	**1**
Störung durch Alkohol	**1**	**2**
Störung durch Alkohol	**1**	**2**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**0**	**2**
Störung durch Alkohol	**0**	**1**
Störung durch Alkohol	**1**	**1**
Störung durch Alkohol	**2**	**2**
Störung durch Alkohol	**1**	**2**
Störung durch multiplen Substanzgebrauch	**0**	**2**

Tabelle A-46 Probabilistischer Auswertungsmodus: Diagnosen der Schlaganfall- und Schädelhirntraumapatienten der Klasse 6

Diagnosen der Schlaganfall- und Schädelhirntraumapatienten der Klasse 6 „ältere Personen mit stark reduziertem Leistungsniveau & relativem Leistungseinbruch im Cattelltest“ **im probabilistischen Auswertungsmodus**	**Alters gruppe**	**Bild-ungs-grad**
posterior Infarkt links	**3**	**1**
V.a. erneute linkszerebrale Ischämie bei Mediateilinfarkt links	**3**	**3**
Z.n. Mediainfarkt links , Z.n. Infarkt im Mediaversorgungsgebiet rechts	**3**	**2**
Z.n. schwerem SHT mit Schädelfraktur links parietal sowie Hirnkontusion rechts parietal	**2**	**1**
Z.n. linkshirnigem Insult im Mediabereich	**3**	**2**
Z.n. cerebraler Ischämie rechtshirnig im Bereich der Capsula interna	**3**	**1**
Ponsinfarkt und cerebellärer Infarkt rechts	**2**	**2**
nicht näher bezeichneter Schlaganfall	**3**	**1**
schweres SHT frontal links	**2**	**1**

Tabelle A-47 Raschhomogene Tests

Raschhomogene Tests (Angaben und Vollständigkeit ohne Gewähr)			
Test	Abkürzung	AUTOREN	Versuchspersonen-Pool der Normierung
Wiener Matrizen-Test	WMT	FORMANN & PISWANGER, 1979	2248VP[99]
Mannheimer Test zur Erfassung des physikalisch-technischen Problemlösens	MTP	CONRAD, BAUMANN & MOHR, 1980	VP?
Dreidimensionaler Würfeltest	3DW	GITTLER, 1990	VP?
Wortschatztest	WST	SCHMIDT & METZLER, 1992	573VP
Pilot's Spatial Test	PST	GRÖSSENBRUNNER, 1994	596VP
Gestaltwahrnehmungstest	GESTA	HERGOVICH & HÖRNDLER, 1994	VP?
Erkennen von Meta-Regeln	META	GATTERNIG & KUBINGER, 1994	257VP
Rechnen in Symbolen	RIS	SCHMOTZER, KUBINGER, & MARYSCHKA, 1994	236 VP & 165VP
Inventar komplexer Aufmerksamkeit	INKA	HEYDE, 1995	6950VP
Tachistoskopischer Verkehrsauffassungstest Mannheim Bildschirm-Vorgabe	TAVTMB	BIEHL, 1996	661VP, 3550VP, 91VP & 153VP
Raven`s Standard Progressive Matrices Plus	SPMPLS	STYLES & RAVEN, 1998	VP?
Bech-Rafaelsen-Melancholie-Skala	BRMS	STIEGLITZ, SMOLKA, BECH & HELMCHEN, 1998	VP?
Mechanisch-technisches Auffassungsvermögen	MTA	LIEDL, 1999	259VP & 556VP
Adaptiver Dreidimensionaler Würfeltest	A3DW	GITTLER, 2000	4064VP , 432VP & 161VP
Alkoholiker Trenn Verfahren	ATV	BIEHL, 2000	345VP
Managementarbeitsprobe	MAP	ETZEL & KÜPPERS, 2000	circa 500VP
Big Five Plus One	B5PO	HOLOCHER-ERTL, KUBINGER & MENGHIN, 2003	531VP
Trierer Integriertes Persönlichkeitsinventar	TIPI	BECKER, 2003	1026VP
Mentale Rotation	MR	BAUER, GUTTMANN, LEODOLTER & LEODOLTER, 2004	195VP
Inventar verkehrsrelevanter Persönlichkeitseigenschaften	IVPE	HERLE, SOMMER, WENZL & LITZENBERGER, 2004	360VP
HKS-Diagnostikum	HKSD	HÄUSLER, 2004	178VP

[99] VP=Versuchspersonen

Tabelle A-48 Adaptive Tests

Adaptive Tests (Angaben und Vollständigkeit ohne Gewähr)			
Test	Abkürzung	AUTOREN	Versuchspersonen-Pool der Normierung
Adaptives Intelligenz Diagnostikum	AID 1&2	KUBINGER & WURST, 1985; 1991	circa1318VP[100] 977VP (Neunormierung1995-97)
Begriffs-Bildungs-Test	BBT	KUBINGER, FISCHER & SCHUHFRIED, 1993	208VP
Syllogismen	SYL	SRP, 1994	290VP
Visueller Gedächtnis Test	VISGED	HORNKE & ETZEL, 1999a	590VP & 481VP
Verbaler Gedächtnis Test	VERGED	HORNKE & ETZEL, 1999b	588VP & 540VP
Adaptiver Matrizentest	AMT	HORNKE, ETZEL & RETTIG, 2000	1356VP & 400VP
Lexikon-Wissens-Test	LEWITE	WAGNER-MENGHIN, 2002a	294VP
Intelligenz-Struktur-Batterie	INSBAT	HORNKE et al. , 2004	341VP
Schlussfolgerndes Denken-Verbal	SDV	SPIEL, GÖßLER & GLÜCK, 2005	VP?

Tabelle A-49 Berechnung der Zugehörigkeitswahrscheinlichkeiten zu latenten Klassen

Die Wahrscheinlichkeit mit der eine Person mit dem Antwortmuster $\underline{x}$ einer Klasse g des Mixed-Rasch-Modells oder der Latenten-Klassen-Analyse (LCA) zugehört, beträgt:

$$p(g|\underline{x}) = \frac{\pi_g p(\underline{x}|g)}{\sum_{h=1}^{G} \pi_h p(\underline{x}|h)}$$; wobei π_g= Klassengröße der Klasse g; G=Klassenanzahl

vgl. ROST (1996, S.156 und S.172)

Tabelle A-50 Reliabilitätsmaß der probabilistischen Testtheorie

$$\text{Reliabilität (PTT)} = \text{Reliabilität } (\xi) = 1 - \frac{\sum_{\upsilon=1}^{N} \sigma^2(E_{\xi_\upsilon})}{N * \sigma^2(\hat{\xi})},$$

wobei: $\sigma^2(\hat{\xi}) = \frac{\sum_{\upsilon=1}^{N} (\xi_\upsilon - \bar{\xi})^2}{N-1}$, $\bar{\xi}$ = Mittelwert des Personenparameters, N=Personenanzahl

vgl. ROST (1996, S.354)

[100] VP=Versuchspersonen