Freie Universität Berlin
Fachbereich Erziehungswissenschaft und Psychologie
Wissenschaftsbereich Psychologie

Dissertation zum Thema:

Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT)

Berlin 2004

Tag der Disputation: 01. Juni 2004

Verfasserin:
Dipl.-Psych. Janine Becker

Anschrift:
Wolliner Str. 12a
10435 Berlin
☎ (privat): 030-44048599
☎ (beruflich): 030-450553123
email: janine.becker@charite.de

Erstgutachter:
Univ.-Prof. Dr. H. Westmeyer
Wissenschaftsbereich Psychologie
Freie Universität Berlin

Zweitgutachter:
PD Dr. med. habil. M. Rose
Med. Klinik m. S. Psychosomatik
der Charité Berlin

Danksagung

Die vorliegende Dissertation ist der Forschungsgruppe der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin gewidmet, welche mir die Möglichkeit eröffnet hat, einen Computergestützten Adaptiven Test zur Angstmessung (Angst-CAT) auf der Grundlage der Item Response Theorie (IRT) zu entwickeln und zu validieren.

Besonderer Dank gilt PD Dr. med. habil. M. Rose, der mich in die IRT-Methodik einführte, das Forschungsprojekt sicher navigierte, und mir als unersetzlicher, Betreuer stets hilfreich und engagiert zur Seite stand, sowie Prof. Dr. H. Westmeyer, welcher mich wohlwollend und mit außergewöhnlicher Sorgfalt begleitete und die Forschungsarbeit durch eine unermessliche Fülle von fachlichen Anregungen bereicherte.

Weiterhin bedanke ich mich bei Prof. Dr. med. B.F. Klapp, der den unschätzbaren Rahmen für das Gelingen der Forschungsarbeit bot, Dr. med. O. Walter, ohne dessen überragendes mathematisches und technisches Know-How die CAT-Methodik nicht realisiert worden wäre, und Dr. rer. nat. Dipl.-Psych. H. Fliege sowie Dipl.-Psych. R.-D. Kocalevent für ihre weitreichende fachliche und heiter zugewandte, kollegiale Unterstützung.

Mein besonderer Dank gilt auch den Diplomandinnen und Praktikant(inn)en der Klinik für ihre mehr als engagierte, fruchtbare Mitarbeit am Projekt. Zudem möchte ich dem gesamten Mitarbeiterteam für ihre große Kooperations-bereitschaft, sowie den Patienten der Klinik, die sich freiwillig bereit erklärten, am Projekt teilzunehmen, meinen warmen, herzlichsten Dank aussprechen.

Als gewinnbringend und erquicklich erlebte ich zudem eine institutionsübergreifende Forschungskooperation mit Dr. phil. habil. U. Ravens-Sieberer und Dipl.-Psych. M. Erhart des Robert-Koch-Instituts Berlin.

Zu guter Letzt’ - aber an meines Herzens erster Stelle - möchte ich meiner Familie, meinem Partner und meinen Freunden/innen dafür danken, dass Sie mich in dieser Zeit so warmherzig unterstützten, und mir die für das Gelingen einer solchen Arbeit nötige Geborgenheit in meinem Leben schenkten.

Zusammenfassung

Die vorliegende Dissertation, welche im Rahmen eines DFG-geförderten Forschungsprojekts der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin entstanden ist, hatte die Entwicklung und Validierung eines Computergestützten Adaptiven Tests (CAT) zur Angstmessung (Angst-CAT) zum Ziel.

Dies geschah aus zwei Gründen. Erstens haben Angststörungen in den westlichen Industrieländern eine sehr hohe Prävalenz und zweitens erschien es besonders lohnend zu überprüfen, ob sich die seit langem bekannten testtheoretischen Vorteile einer „modernen“ Testtheorie, namens Item Response Theorie (IRT), in der praktischen Anwendung der Testkonstruktion wieder finden lassen. Dies ist von Interesse, da bislang für die Messung der Zustandsangst zwar eine Vielzahl guter und etablierter Messinstrumente vorliegen, diese jedoch mehrheitlich auf der Grundlage einer „älteren“, der sogenannten Klassischen Test-Theorie konstruiert wurden, die einige messtheoretische Probleme aufwirft.

Einer der aus meiner Sicht wichtigsten Vorteile der IRT liegt in der Möglichkeit, sogenannte *Computergestützte Adaptive Tests (CAT)* zu konstruieren: CATs ermöglichen die individuelle Anpassung der Itemdarbietung an die Merkmalsausprägung einer Person. Dies geschieht, indem Testpersonen während der CAT-Bearbeitung nur die Items zur Beantwortung dargeboten werden, welche dem individuellem Ausprägungsniveau einer Person optimal entsprechen („adaptives Testen“). Welches Item jeweils während der CAT-Bearbeitung als „optimal“ gilt, hängt dabei sowohl von der individuellen Beantwortung vorangegangener Items, als auch von der vorher an einer Kalibrierungsstichprobe errechneten Iteminformation der einzelnen Items ab. Dadurch, dass einer Testperson nur die jeweils „passendsten“ bestmöglichen Items vorgelegt werden, kann eine deutliche *Itemreduktion* bei einem gleichzeitig *konstant hohen Messpräzisionsniveau* erreicht werden.

Die Reduktion in der Testlänge kann eine Entlastung für den Diagnostiker und die Testperson bedeuten. Während IRT-basierte CATs sich aufgrund dieser und weiterer testtheoretischer Vorteile in der *Leistungsdiagnostik* bereits vielfach mit gutem Erfolg bewährt haben, ist das Ziel vorliegender Dissertation zu untersuchen, ob auch die *klinisch-psychologische Diagnostik* von dieser

fortgeschrittenen Methodik profitieren kann. Dazu wurde die Entwicklung eines kurzen Screening-Instruments zur Erfassung von Zustands-Angst angestrebt, welches trotz einer kürzeren Testlänge eine Messung auf einem konstant hohen Messpräzisionsniveau erlauben soll.

Da das Forschungsfeld IRT-basierter CATs in der klinisch-psychologischen Diagnostik noch relativ jung ist, fehlt bislang ein allgemeiner Forschungskonsens über eine grundlegende methodische Strategie zur Konstruktion IRT-basierter CATs. Die hier realisierte Testentwicklung des Angst-CATs folgte daher verschiedenen Empfehlungen aus Lehrbüchern (z. B. Embretson & Reise, 2000; Hambleton et al., 1991; Wainer, 1990), Übersichtsartikeln (z. B. Hattie, 1984; Nandakumar, 1994; Meijer & Nering, 1999) und einer bereits im Rahmen der Lebensqualitätsforschung erprobten Testentwicklungsstrategie einer US-amerikanischen Forschungsgruppe (Ware et al., 2000, 2003). Sie gliedert sich in drei prinzipielle Schritte: (1.) die inhaltliche Auswahl angstrelevanter Items, (2.) die sequentielle statistische Itemanalyse und –selektion dieser Items mit dem Ziel, die Items mit der besten psychometrischen Qualität zur Konstruktion einer Itembank zu nutzen und (3.) die Implementierung der Itembank in einen computergestützten adaptiven Itemabfolge-Algorithmus, der die Präsentation der Items und die Schätzung der individuellen Angstausprägung (Theta-Schätzung) von Testpersonen ermöglicht.

In einem Dephi-Entscheidungsprozesses sind von 125 vorselektierten Items zur Angstmessung nach inhaltlichen Kriterien 81 angstrelevante Items (mit 2- bis 7-stufigen Likert-skalierten Antwortformaten) aus 13 etablierten psychometrischen Instrumenten (ADS, ALL, BDI, BSF, GBB, GT, NI-90, PGWI, PSQ, SF36, SKT, STAI, SWO) ausgewählt worden. Die statistische Itemanalyse und –selektion erfolgte an psychometrischen Daten von insgesamt N = 2.348 Patienten, die in der Medizinischen Klinik mit Schwerpunkt Psychosomatik im Rahmen ihrer poliklinischen, stationären bzw. konsiliarischen Behandlung zwischen 1995 und 2001 computergestützt erhoben wurden.

Da nicht alle Patienten der Gesamtstichprobe alle zu analysierenden Items beantwortet haben, erfolgte die statistische Itemanalyse und –selektion an drei Teilstichproben (N_1 = 1.010; N_2 = 834; N_3 = 775) der Gesamtstichprobe.

Die statistische Itemanalyse und –selektion verlief wiederum in drei sequentiellen Schritten: (a) der Untersuchung der Unidimensionalität mittels explorativer und konfirmatorischer Faktorenanalysen und der Analyse residualer Kovarianzen (Computerprogramme: SPSS, Mplus, AMOS), (b) der IRT-Analyse, welche die grafische Analyse der Item Response Curves (IRCs) und die Berechnung von Testinformationen, Standardmessfehlern und Reliabilitäten umschloss (Computerprogramm: TestGraf), und (c) der IRT-Modellierung, die der Schätzung der Itemparameter unter Anwendung des zweiparametrischen Generalized Partial Credit Modells (GPCM), der Untersuchung von Differential-Item-Functioning (DIF) und der Realisierung des Item-Link-Design diente (Computerprogramm: Parscale).

Die statistische Itemselektion führte zur Konstruktion einer 50 Items umfassenden Itembank, welche dem Angst-CAT zugrundegelegt wurde. Die Itembank besteht zu 70% aus Items, welche das Vorliegen von *Zustands-Angst* in Anlehnung an Spielbergers Definition (1972) in *positiver* Ausprägung (z. B. „ängstlich“ oder „besorgt“) und zu 30% aus Items, welche zur Angst *konträre* Zustände (z. B. „selbstsicher“ oder „entspannt“) erfassen. Mit der Itembank können gemäß der Konzeption des Angst-Konstruktes von Liebert und Morris (1967) sowohl *emotionale* (z.B. innere Unruhe) als auch *kognitive* Aspekte (z.B. Besorgtheit) erhoben werden.

Da das Angst-CAT eine möglichst objekt- und situationsübergreifende Messung von Zustands-Angst intendiert, wurden im Rahmen der inhaltlichen und statistischen Itemanalyse vor allem Items, welche gesundheitsspezifische Sorgen oder spezifische soziale Ängste sowie andere Konstrukte wie allgemeine Leistungseinbußen, Schlafstörungen und Depression erfassen, ausgeschlossen. Zudem wurden Items, welche vegetative Korrelate der Angst erheben aufgrund von Verletzungen der Unidimensionalitätsannahme aus der Itembank eliminiert.

Während Spielberger (1972) die vegetative Erregung als Teil der Zustands-Angst definierte, revidieren die Befunde der vorliegenden Forschungsarbeit im Einklang mit Forschungsbefunden neuerer Angstmodelle („Drei-Faktoren-Modell“, Barlowe et al., 1996; „Integratives hierarchisches Modell der Angst und Depression“, Mineka et al., 1998) diese Konzeption. Vegetative Angstkorrelate

wie z. B. Herzjagen, Zittern, Schwitzen, Schwindel etc. sind demnach vor allem spezifisch für *Panik*störungen und *nicht* für eine *allgemeine Zustands-Angst*.

In einer an die Testentwicklung anschließende Validierungsstudie an N = 102 psychosomatischen Patienten erwies sich das Angst-CAT als ein valides und reliables sowie ökonomisches psychometrisches Screening-Instrument zur Erfassung von Zustands-Angst.

Durchschnittlich genügte im Angst-CAT die Darbietung von 5-7 Items zur Schätzung der Angstausprägung (Theta-Schätzung) auf einem konstant hohen Messgenauigkeitsniveau mit einer Reliabilität von $\geq$ 0,9. Diese Höhe des Messpräzisionsniveaus wurde a priori als das Stoppkriterium festgesetzt, welches den CAT-Prozess beendet.

Die computergestützte adaptive Itemreduktion führte zu keinem nennenswerten Informationsverlust gegenüber der mit IRT-Methoden simulierten Darbietung aller Items (Walter et al., eingereicht). Jedoch nimmt bei der Messung von *extremen* Angstausprägungen die Anzahl der im Angst-CAT darzubietenden Items aufgrund eines etwas geringeren Informationsgehaltes dieser Items zu, wenn das a priori festgesetzte, konstante hohe Messpräzisionsniveau gewährleistet werden soll.

Die Itemreduktion erfolgte durch die adaptive Itemdarbietung mittels des *Maximum-Information-Verfahrens* (MI) auf der Basis der *Fisher Information* (Itemselektionsstrategie) und des *Bayes'schen Expected-A-Posteriori-Schätzverfahrens* (EAP), welches als Methode der Personenparameter-schätzung in der CAT-Forschung im Bereich der Leistungsdiagnostik bereits gut etabliert ist.

Die Validierungsstudie an N = 102 psychosomatischen Patienten belegte ferner eine mittelmäßige bis gute *konvergente Validität* des Angst-CATs in Form von Korrelationen zu anderen Angstinventaren (BAI, HADS-A; r = 0,51-0,76). Eine diagnosenspezifische konvergente Validität ist insofern gegeben, als Patienten mit der Diagnose einer Angststörung signifikant ($p \leq 0{,}001$) höhere Werte im Angst-CAT aufwiesen als Patienten ohne psychische Störung bzw. gesunde Personen (N = 35).

Die *diskriminante Validität* des Angst-CATs unterscheidet sich im Hinblick auf die untersuchten Konstrukte. Die psychometrische Diskrimination von Angst- und Depression (BDI, HADS) bzw. Neurotizismus (NEO-FFI) gestaltet sich –

wie theoretisch und empirisch in der Literatur bereits vielfach diskutiert – auch mit dem Angst-CAT schwierig.

Dagegen kann aufgrund geringer Korrelationen des Angst-CATs zu Skalen von zwei Persönlichkeitsinventaren (NEO-FFI, GT) auf eine *gute diskriminante Validität* bezüglich anderer Eigenschaftskonstrukte geschlossen werden. Eine diagnosenspezifische Diskrimination ist mit dem Angst-CAT bedingt gegeben, gestaltet sich jedoch aufgrund von Komorbidität nicht eindeutig, so dass die Angst-CAT-Werte stets im Zusammenhang weiterer klinischer Diagnostik interpretiert werden sollten.

Zusammenfassend lässt sich sagen, dass mit dem Angst-CAT ein kurzes, *messpräzises* und *valides Screening-Instrument* zur Messung einer objekt- und situationsübergreifenden aktuellen *Zustands-Angst* IRT-basiert entwickelt und validiert werden konnte, welches eine *mobile, ökonomische* (i. S. von Item- und Zeitersparnissen) und damit eine *patientenfreundliche* Erfassung der Angstausprägung ermöglicht.

Dieser Befund, welcher im Einklang mit positiven Forschungsergebnissen IRT-basiert-entwickelter CATs im Bereich der Leistungsdiagnostik steht, deutet darauf hin, dass auch die klinisch-psychologische Diagnostik von der Entwicklung und dem Einsatz IRT-basierter CATs profitieren kann.

Angesichts des jungen Forschungsstandes auf diesem Gebiet werden mit vorliegender Dissertation jedoch auch eine Reihe von Fragen aufgeworfen. So steht nicht nur der hier erprobte *methodische* Weg der IRT-basierten Testkonstruktion des Angst-CATs, sondern auch die dargestellten *Ergebnisse* und deren Bewertung zur Diskussion. Besondere Schwierigkeiten bestehen dabei aus meiner Sicht in *fehlenden Bewertungsmaßstäben* zur Beurteilung der Güte IRT-basierter Itemparameterwerte, sowie der Etablierung allgemeiner Kriterien für die Bewertung von Gesamttests und den Vergleich der Datenanpassung verschiedener IRT-Modelle. Solange hier kein Konsens zwischen den Anwendern der IRT-Methoden erreicht werden kann, wird die Resonanz bzw. der Verbreitungsgrad IRT-basierter CATs in der klinisch-psychologischen Diagnostik letztendlich wohl maßgeblich von der Einstellung der Anwender zur *IRT* im Speziellen und zur *Computerdiagnostik* im Allgemeinen abhängen.

Inhaltsverzeichnis

1. Einleitung

1.1. Zielsetzung

In der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin werden eine Vielzahl von psychometrischen Fragebögen zur Eingangs- und Verlaufsdiagnostik im poliklinischen, konsiliarischen und stationären Setting eingesetzt.

Aufgrund einer hohen Prävalenz von Angststörungen allgemein (9,2 – 28,3% Lebenszeitprävalenz; Neumer, 2000) und im psychosomatischen Bereich im Speziellen (24,4 – 29,4% Punktprävalenz; Fliege, Rose, Bronner & Klapp, 2002) ist man im psychosomatischen Bereich an einer informationsreichen, ökonomischen und patientenfreundlichen Erfassung von Angst besonders interessiert. Angst gilt hier seit jeher als „das Symptom im Grenzland zwischen körperlicher und psychischer Störung“ (Sims & Snaith, 1993, S. 46), da sie sowohl vegetativ wie seelisch erlebt und durch körperliche Krankheit sowie psychische Konflikte verursacht wird.

Die in der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin nach klinikinternen Anforderungen zusammengestellten Testbatterien beinhalten psychometrische Verfahren, welche auf der Grundlage der „Klassischen Test-Theorie“ (KTT) konstruiert sind. Die Zusammenstellung verschiedener Testverfahren ermöglicht eine breite und differenzierte Psychodiagnostik, ist jedoch aufgrund des großen Umfangs konventioneller Papier-und-Bleistift-Testformen von einer Mehrbelastung der Patienten und Diagnostiker begleitet. Diese Mehrbelastung durch die Darbietung großer Mengen von Items (aus unterschiedlichen Testverfahren) äußert sich in Ermüdungserscheinungen sowie Motivationsproblemen der Patienten und mindert dadurch nicht zuletzt die Qualität der erhobenen Daten. Zudem ist die konventionelle Papier-und-Bleistift-Diagnostik kosten- und zeitaufwendig und somit sowohl für Patienten wie auch für Diagnostiker ressourcenintensiv.

Das Ziel vorliegender Dissertation ist es zu erproben, ob durch die Entwicklung eines „Computergestützten Adaptiven Tests“ (CAT) auf der Grundlage einer anderen (moderneren) Testtheorie - der sogenannten „Item Response Theorie“ (IRT) - ein messpräzises Verfahren zur Angsterfassung entwickelt werden kann, welches durch einen geringeren Itemumfang Patienten und Diagnostiker weniger belastet.

Während in konventionellen Testverfahren zur Angstmessung ein Standardset an Items allen Patienten gleichermaßen präsentiert wird, und somit auch Patienten Items dargeboten werden, die keine oder nur eine geringe individuelle Relevanz für sie haben, bietet das Computergestützte Adaptive Testen (CAT) die Möglichkeit, die Items an das Ausprägungsniveau der Angst eines Patienten „adaptiv“ anzupassen. Dies führt dazu, dass Patienten nur diejenigen Items vorgelegt werden, welche für sie auch wirklich aussagekräftig bzw. informationsreich sind, womit die für eine präzise Angstmessung benötigte Itemanzahl verringert wird. Die Psychodiagnostik wird damit effizienter und ökonomischer.

Die Anwendung der „Item Response Theorie“ (IRT) stellt im Bereich klinisch-psychologischer Testentwicklung ein wenig erforschtes Gebiet dar (siehe Kapitel 3.5.2.). Die Dissertation wird exemplarisch zeigen, inwieweit die praktische Anwendung der IRT im Rahmen einer Testentwicklung und -validierung die zu erwartenden praktischen, ökonomischen und testtheoretischen Vorteile bietet.

1.2. Aufbau der Dissertation

Die vorliegende Dissertation gliedert sich in zwei Teile. Der erste - theoretische - Teil umfasst allgemeine Einführungen zum Konstrukt der Angst (Kapitel 2), zur Item Response Theorie (IRT; Kapitel 3) und zum Computergestützten Adaptiven Testen (CAT; Kapitel 4), welches auf der Grundlage der IRT realisiert werden kann.

Im ersten dieser theoretischen Kapitel (Kapitel 2) wird ein Überblick über das Konstrukt der Angst gegeben, welcher der Einordnung der Entwicklung eines Messinstruments zur Angsterfassung in die umfangreiche psychologische Forschungstradition der Angst dienen soll. Besonders zentral erscheinen hier die Definition der Angst als normales Phänomen und als Störung sowie die Einbettung in die differentialpsychologische Theorienlandschaft, auf deren Grundlage die Messung von Angst erfolgt.

Da der entwickelte Computergestützte Adaptive Test zur Angstmessung („Angst-CAT“) auf der Basis der „Item Response Theorie“ (IRT) konstruiert wurde, wird darauffolgend dieser testtheoretische Ansatz in Abgrenzung zur konventionellen „Klassischen Test-Theorie“ (KTT) erörtert (Kapitel 3).

Der Theorieteil wird schließlich durch ein Kapitel, welches die Grundzüge des Computergestützten Adaptiven Testens erläutert, den aktuellen Forschungsstand in diesem Bereich vorstellt und die Vor- und Nachteile dieser Form des Testens zusammenfasst, abgeschlossen (Kapitel 4).

Der zweite - empirische - Teil befasst sich mit der Darstellung der Entwicklung (Kapitel 5) und Validierung des Angst-CATs (Kapitel 6), indem jeweils zuerst die untersuchten Stichproben und die angewandten Methoden vorgestellt werden, auf deren Grundlage die Präsentation der Ergebnisse der Testentwicklung und -validierung erfolgt. Abschließend ist das letzte Kapitel der Diskussion der Ergebnisse gewidmet (Kapitel 7).

2. Angst

2.1. Einleitung

Angst als eine fundamentale Erlebensform menschlicher Existenz (Krohne, 1996) beschäftigt die Menschen seit jeher. So intuitiv verstehbar wie der Begriff auf den ersten Blick erscheint, so verschiedenartig sind die Perspektiven aus denen Menschen diesen Begriff erforschen. So befassen sich nicht nur Philosophen seit Jahrhunderten mit dem Thema, sondern auch Dichter und Künstler, Laien und Wissenschaftler verschiedener Disziplinen (Psychologie, Medizin, Philosophie, Theologie, Soziologie, Biologie, Politologie, Medienwissenschaften, Wirtschaftswissenschaften etc.).

Während singuläre Ereignisse (Katastrophen wie der Terroranschlag auf das World Trade Center, in New York am 11.09.2002) episodenhaft eine Zunahme des öffentlichen Interesses an dem Phänomen der Angst auslösen, räumen manche Autoren der Angst gar den Stellenwert eines „allgemeinen Zeitgeists“ ein und bezeichnen das 20. Jahrhundert als das „Zeitalter der Angst“ (May, 1950; Spielberger, 1980). Als Argumente für ein „Zeitalter der Angst“ können für unsere Gesellschaft charakteristische angstauslösende Bedingungen, wie die globalen Verunsicherungen ausgelöst durch den exponentiellen technischen Fortschritt (z. B. Reproduktionsmedizin, ABC-Waffenentwicklung), durch den raschen wirtschaftlichen (z. B. Globalisierung und Liberalisierung) und sozialen Wandel (z. B. Entwurzelung durch Arbeitsmarktveränderungen und Vereinsamung durch Urbanisierung) sowie Glaubwürdigkeitsverluste bezüglich politischer Autoritäten (z. B. Spendenaffären) und Instanzen (z. B. WTO[1], IWF[2]) ins Feld geführt werden.

Eine *wissenschaftliche* Annäherung an das Thema „Angst“ ist in der Psychologie seit Beginn des 20. Jahrhunderts zu verzeichnen. Seither herrscht eine rege Forschungs- und Publikationstätigkeit zum Thema Angst, die besonders nach dem zweiten Weltkrieg (wahrscheinlich nicht ohne Grund) Auftrieb gewonnen hat. So ergibt eine Literaturrecherche in den Datenbank „PsyInfo“ und „PsycArticles“, den zwei größten Datenbanken (u. a. der American Psychological Association; APA), welche die wichtigsten psychologischen Fachzeitschriften weltweit auswerten, dass in den letzten

[1] WTO = World Trade Organisation (Welthandelsorganisation).
[2] IWF = Internationale Währungs-Fonds.

10 Jahren (1993-2003) 20.870 Publikationen zum Thema „Angst“ verfasst wurden. Angesichts dieser kaum noch zu überblickenden Fülle an Forschungsarbeiten beschränkt sich vorliegende Arbeit auf einige meines Erachtens zentrale Aspekte der Angst.

2.2. Historische Begriffsentwicklung

Die Faszination des Themas „Angst“ lässt sich bis in die Antike zurückverfolgen. Schon im 4. Jahrhundert v. Chr. wurde dieser Gefühlszustand von Philosophen wie Hippokrates und Aristoteles beschrieben, welche sich vor allem auch mit der Beziehung zwischen einem gestörten Affekt und körperlicher Krankheit befassten (Sims & Snaith, 1993). Auch die Begriffe „Angst“ („anxo“, gr.: niedergedrückt, beengt), „Panik“ und „Phobien“ sind griechischen Ursprungs. So waren „Pan“ und „Phobos“ griechische Götter, denen als personifizierte Verursacher von Angst die Aufgabe zuteil wurde, Feinde in die Flucht zu schlagen. Während in Griechenland somit die Furcht vor Göttern objektbezogen war, vermutet Finzen (1988) in der an die griechische Antike anschließenden Epoche des (zerfallenden) römischen Reiches ein von mangelnder Ordnung und Geborgenheit geprägtes Gesellschaftsgefühl, welches eine unbestimmte gegenstandslose „Weltangststimmung“ vor Dämonischem provozierte. Diese solle den Weg für das Aufkommen des Christentums, welche die Weltangst im Jetzt zu überwinden versprach, geebnet haben. Seit dem 16. Jahrhundert begann schließlich ein verstärktes literarisches Interesse an dem Thema „Angst“. So kann der Gebrauch des Wortes „Angst“ bis zu einem Bericht von Lovell über „den Schmerz und die Angst des Ventrikels“ (1661) zurückverfolgt werden, der ähnlich wie die Schriftsteller Burton, Taylor und Flecknoe (Sims & Snaith, 1993) sehr akurat Angstzustände jedoch zunächst nur im Zusammenhang mit Depression, Schmerz und körperlichen Erkrankungen (v. a. Koronarerkrankungen) schilderte. Im 19. und 20. Jahrhundert befassten sich die Philosophen Kierkegaard (1844) und Klages (1926; später auch: Heidegger, 1979, Sartre, 1962, und Jaspers, 1973) mit dem Phänomen der Angst. Erste Klassifikationsbemühungen *klinischer* Angst reichen bis ins Jahr 1798 zurück, in dem Rush eine erste Liste verschiedener Formen der Phobie abfasste (Sims & Snaith, 1993, S. 38). Die Agoraphobie wurde bereits 1871 von Westphal als eigenes psychiatrisches Syndrom eingeführt. Die Verbreitung einer

grundsätzlichen Unterscheidung zwischen normaler und pathologischer Angst geht auf Freud (Breuer & Freud, 1895), die Popularisierung der systematischen Abgrenzung von „Zwangserkrankungen“ und „Phobien“ auf Kraepelin (1918) zurück. Obwohl der Begriff der Panik seit dem Ende des 18. Jahrhunderts (Freud, 1940) bekannt ist, wurde ihm der Status einer eigenständigen nosologischen Einheit erst 1980 durch das DSM-III zuerkannt.

2.3. Definitionen

2.3.1. Was ist Angst?

Angst ist ein elementarer Affekt und ein zentrales Symptom seelischer Störungen. Obwohl der Begriff „Angst“ aufgrund seiner Alltagsnähe intuitiv leicht verständlich erscheint, existieren in der Psychologie Hunderte von verschiedenen Definitionen zu diesem Konstrukt. Im Folgenden seien exemplarisch drei aufgeführt.

> „Unter Angst versteht man allgemein eine Stimmung oder ein Gefühl der Beengtheit, Beklemmung oder Bedrohung, einen unangenehmen, spannungsreichen, oft quälenden Zustand.“ (Hogen, 2001, S. 38)

> „...ein mit Beengung, Erregung, Verzweiflung verknüpftes Lebensgefühl, dessen besonderes Kennzeichen die Aufhebung der willensmäßigen und verstandesmäßigen »Steuerung« der Persönlichkeit ist.“
> (Häcker & Stapf,1998, S. 40)

> „...ein affektiver Zustand des Organismus, der durch erhöhte Aktivität des autonomen Nervensystems sowie durch die Selbstwahrnehmung von Erregung, das Gefühl des Angespanntseins, ein Erlebnis des Bedrohtwerdens und verstärkte Besorgnis gekennzeichnet ist.“
> (Stöber & Schwarzer, 2000, S. 189; Krohne, 1996, S. 5; Spielberger, 1972)

Vorangegangenen Definitionen ist gemein, dass Angst grundsätzlich als ein (Lebens-) *Gefühl*, eine *Stimmung* bzw. ein *affektiver Zustand* angesehen wird, der zumindest in der ersten Definition explizit als „*unangenehm*“ beschrieben wird.

Das Gefühl der *Beengtheit* wird nur in den ersten beiden, das der *Erregung* nur in den letzten beiden Definitionen expliziert. Die zweite Definition fokussiert zusätzlich Auswirkungen der Angst auf der Verhaltensebene (Kontrollverlust).

Die von Freud (1940) vorgezeichnete und von Spielberger (1972) formulierte dritte Definition umfasst sowohl emotionale (Bedrohungserleben), kognitive

(Besorgnis) und physiologische Aspekte (erhöhte Aktivität des autonomen Nervensystems) der Angst.

In diesen Definitionen deutet sich schon ein Dilemma des Angstbegriffs an. Es ist trotz umfangreicher Forschungsbemühungen umstritten, wie viele Komponenten der Angst zugerechnet werden, ob die Angst als ein eindimensionales oder mehrdimensionales Konstrukt zu konzipieren ist (siehe Kapitel 2.7.3.4.), oder ob es spezifische Aspekte gibt, welche *nur* kennzeichnend für das Phänomen der Angst sind, da eine Abgrenzung zu benachbarten Konstrukten oft schwer fällt (siehe Kapitel 2.5.).

2.3.2. Auslöser der Angst

Nach Benesch (1995, S. 91) können prinzipiell drei Quellen der Angst unterschieden werden: a) äußere Angstreize, b) innere Angstgründe und c) äußerlich-innerliche Interdependenzen, welche sich in einem „Vorgang der Aufschaukelung“ verstärken können. Zu a) rechnet man Objekte / Situationen und Personen, die Angst auslösen, verstärken und aufrechterhalten können, worauf sich speziell der lerntheoretische Ansatz und die Verhaltenstherapie fokussiert. Zu b) zählen innere Konflikte, die vor allem im psychodynamischen Ansatz eine wesentliche Rolle spielen, sowie Kognitionen, mit denen sich kognitive Ansätze auseinandersetzen. Die äußerlich-innerlichen Interdependenzen werden heute vor allem im Rahmen der kognitiven Verhaltenstherapie der Angst (Margraf, 2000) als therapeutische Ansatzpunkte genommen.

Laut Stöber und Schwarzer (2000) können grundsätzlich zwei Angstthemen unterschieden werden: die Angst vor der *körperlichen* Bedrohung und der Angst vor der *Selbstwert*bedrohung, zu der die soziale Angst und die Leistungsangst gerechnet werden können. Eine ähnliche themenspezifische Ordnung von Ängsten nehmen auch Tewes und Wildgrube (1999) vor. Sie unterscheiden in einer hierarchischen Taxonomie nach dem Allgemeinheitsgrad zwischen 1. *Existenzangst*, zu der sie Todes-, Krankheits-, Verletzungs-, Flug-, Höhen-, Gewitter-, Dunkel- und Kriegsangst rechnen, 2. *sozialer Angst*, welche Scham, Verlegen- und Schüchternheit, Angst vor dem anderen Geschlecht, Sexualität, Publikum und dem Vorgesetzten umfasst, und 3. *Leistungsangst*, zu der sie Bewertungs-, Prüfungs-, Schul- und Berufsangst zählen.

Da grundsätzlich alle äußeren Objekte[3], Personen und Situationen sowie auch innere Reize (Schuldgefühle, Triebimpulse etc.) Angst auslösen können, ist die Zahl möglicher bereichsspezifischer Ängste und Phobien[4] unbegrenzt.
Prinzipiell ist in diesem Zusammenhang darauf hinzuweisen, dass viele Autoren die von Kierkegaard (1844) postulierte begriffliche Unterscheidung zwischen *Furcht*, die als auf einen Gegenstand gerichtet definiert wird, und *Angst*, welche als ungerichtet, objektlos und frei flottierend angesehen wird, treffen (Peters, 2000).

2.3.3. Funktionen der Angst

> Angst ist ... „eine emotionale Reaktion auf das Erkennen oder vermeintliche Erkennen einer Gefahr, unabhängig davon, ob diese Gefahr auch objektiv gegeben ist.“ (Spielberger, 1972, S. 482)

Bezüglich der Funktionen von Angst sind sich Angstforscher aller theoretischer Ansätze erstaunlich einig. Angst dient dem Schutz vor Gefahren und ist damit lebenserhaltend. In einer Bedrohungssituation hat sie die Funktion eines „Warnsignals“ (Hogen, 2001) oder „Gefahrenschutzinstinkts“ (Häcker & Stapf, 1998, S. 40), welcher den Organismus durch eine Steigerung der Aktivität des sympathikotonen Nevensystems im Sinne der Cannon’schen Notfallreaktion (Spielberger, 1980) mobilisiert, um drohende Gefahr abzuwenden. Mit der Initiierung einer Reihe von lebenserhaltenden physiologischen Reaktionen (siehe Kapitel 2.3.4.), welche vom Hirnstamm aus gesteuert werden, und eine allgemeine Aktivierungs- und Leistungssteigerung bewirken, geht auch eine Erhöhung der Aufmerksamkeit und Handlungsmotivation einher. Die rasche Aufmerksamkeitsfokussierung auf das bedrohende Moment führt dabei zur Handlungsunterbrechung weniger wichtiger Aufgaben (Kazdin, 2000, S. 209). Somit stellt Angst aus evolutionstheoretischer Perspektive eine evolutionsgeschichtlich früh entwickelte Anpassungsleistung dar (Darwin, 1965).
Adaptiv ist Angst jedoch nicht nur im Hinblick auf den Schutz vor objektiven, realen Gefahren, sondern auch bezüglich des Schutzes des eigenen Selbstwertes oder Selbstbildes.

---

[3] Im weitesten Sinne ist hier auch die Angst vor dem eigenen Körper bzw. den eigenen Körpergrenzen (in pathologischer Ausprägung > Boderline Störung) und Körperausmaßen (> Essstörungen) und seiner Gesundheit (> Hypochondrie) aufzuführen.
[4] Phobien sind pathologische Formen übersteigerter objektbezogener Furcht (siehe Kapitel 2.6.1.).

Battegay (1970) weist ferner darauf hin, dass Angst auch soziale Funktionen erfüllt. So kann sie dazu führen, dass der angsterlebende Mensch die Aufmerksamkeit Anderer auf sich zieht, es wird ein Appell an die Mitwelt gesendet, der Hilfeleistung zu initiieren vermag, und im neurotischen Sinn kann Angst auch der Machtausübung über andere Personen dienen, bzw. die Funktion einer Sicherungstendenz starrer Ordnung bzw. des Stillstands haben.

2.3.4. Reaktionen der Angst und ihre Bewältigung / Abwehr

Konzeptuell können drei verschiedene Reaktionsebenen der Angst unterschieden werden: die physiologische / körperliche, die verhaltensmäßig-expressive / motorische und die subjektive Ebene (Emotionen und Kognitionen) (Krohne, 1996, S. 5; Benesch, 1995, S. 92).

Als körperliche Begleiterscheinungen der Angst treten Puls- und Herzfrequenz-steigerungen, Palpitationen, Tachykardie, Druckschmerzen oder Kloßgefühle in der Brust- und Herzgegend, ein erhöhter Blutdruck und Adrenalinspiegel, eine gesteigerte Atemfrequenz (Erstickungsgefühle), erhöhte Muskelspannungen, Zittern, Schwitzen, Pupillenerweiterung, Errötung, Mundtrockenheit, abdominelle Beschwerden, Beschleunigung der Darmtätigkeit (Diarrhoe), Harndrang, Übelkeit, Erbrechen, Schwindel, Kribbel- und Taubheitsgefühle (Parästhesien), Depersonalisations- und Derealisationsempfindungen sowie Ohnmachtsgefühle auf. Diese möglichen Begleiterscheinungen der Angst können als „Angstäquivalente“ das subjektive Angsterleben sogar in den Hintergrund treten lassen.

Auf der Verhaltensebene können sich klassische Kampf- oder Fluchtreaktionen („fight-or-flight reaction“; Cannon, 1975) mit Aktivitätssteigerung bis hin zu aggressiven Handlungen und Vermeidungsverhalten oder Verhaltens-hemmungen, zeigen, die bis zur Erstarrung oder Lähmung reichen können. Häufig wird die Angst dabei von einem spezifischen Gesichtsausdruck sowie Störungen im Sprachfluss begleitet. Langanhaltende Angst- bzw. Stress-reaktionen wurden von Selye (1957) in einem Prozessmodell konzipiert.

Auf subjektiver Ebene können die verschiedenartigsten Gefühlsempfindungen wie Beengungs-, Bedrohungs- und Schuldgefühle, sowie Ärger, Traurigkeit, Scham, Aggression auftreten, welche kombiniert das „Angstgefühl“ ausmachen (Kazdin, 2000, S. 210). Um der emotionalen Vielschichtigkeit gerecht zu werden, konzipierten Watson und Clark (1984) die Angst als ein

Emotionsmuster, welches sich aus einer Reihe von negativen Emotionen wie Wut, Trauer, Zorn, Schuld, Frustration, Nervosität, Selbstunsicherheit zusammensetzt, und fassten sie unter dem Sammelbegriff der „Negativen Affektivität“ zusammen (zu Weiterentwicklungen dieser Modellvorstellungen siehe Kapitel 2.5.).

Die Bewältigung der Angst kann auf zweierlei Weise erfolgen: durch den Einsatz von Copingstrategien oder von Abwehrmechanismen. Copingstrategien sind meist aktive Bewältigungsformen, welche der bewussten und flexiblen Anpassung der Person an die Situation dienen. In Tabelle 1 sind drei verschiedene Konzeptionen von Copingstrategien zusammengefasst (Stöber & Schwarzer, 2000).

Tabelle 1: Coping – Modelle.

Autoren	Jahr	Modell	Copingstrategien
Billings & Moos	1984	Dreidimensionales Coping Modell	bewertungs-,[5] problem-,[6] emotionszentriertes Coping[7]
Byrne	1961	Repression-Sensitization-Modell	Informationssuche Informationsabwehr
Krohne	1993	Zweidimensionales Modell der dispositionellen Angstbewältigungsstile	vigilant-kognitives, kognitiv-vermeidendes Coping[8]

Abwehrstrategien sind unbewusste, oft rigide, nicht altersentsprechende Mechanismen der Angstabwehr, welche von Psychoanalytikern konzipiert wurden (Freud, 1936). Zu ihnen zählen z. B. die Regression, Realitätsleugnung, Verdrängung, Projektion, Verschiebung, Sublimation und Überkompensation. Eine Reihe sozialpsychologischer Forschungsarbeiten belegen, dass Verunsicherung den Wunsch vergrößert, mit Anderen zusammen zu sein, und dass die Gesellschaft anderer Menschen angstmindernd sein kann, da sie soziale Vergleiche, Neubewertung, emotionale, informative und instrumentelle Unterstützung verspricht (Ströbe, Hewstone & Stephenson, 1996).

[5] z. B. Neubewertung der Situation.
[6] z. B. aktive Informationssuche oder Suche nach sozialer Unterstützung.
[7] z. B. emotionale Regulationsmechanismen („Tief durchatmen“, Beruhigungsstrategien, Musik hören.
[8] z. B. Ablenkung, Vermeidung, Bagatellisierung.

2.4. Theorien der Angst

In der psychologischen Literatur finden sich eine Vielzahl von Theorien zur Erklärung des Phänomens der Angst. Diese Vielfältigkeit der Theorien deutet bereits auf ein globales Dilemma hin. Angst ist als Phänomen zu vielschichtig, um sie erschöpfend in einer Theorie zu behandeln. Daher können einzelne Theorien auch immer nur einzelne Aspekte der Angstentstehung und des Angsterlebens hervorheben, was zwangsläufig eine Kritik der Einseitigkeit der meisten Theorien berechtigt. Da die meisten Theorien sich nicht widersprechen, plädiere ich bei der Betrachtung der Angst für eine eklektizistische Sichtweise, in der jeder Theorien ihr spezifischer Erklärungsstellenwert zukommt. Um den Rahmen vorliegender Arbeit nicht zu sprengen, wird im Folgenden nur differentialpsychologische Theorien der Angst fokussiert, da die Messung der Angst als ein nomothetisches Persönlichkeitskonstrukt am ehesten der differentialpsychologischen Forschungstradition zuzuordnen ist.

2.4.1. Differentialpsychologische Theorien der Angst

Im differentialpsychologischen Ansatz erforscht man interindividuelle Unterschiede verschiedener Merkmalsausprägungen. Diesem Bemühen liegt die Annahme zugrunde, dass es stabile Merkmale (Trait-Ansatz) gibt, in dessen Ausprägung sich Individuen über eine längere Zeitspanne und über verschiedene Situationen hinweg stabil unterscheiden (intraindividuelles Kohärenzprinzip und transsituative Konsistenzannahme; Laux & Glanzmann, 1996, S. 121). Demnach richtet sich ein Forschungsfokus auf die abstrakte Erfassung dieser interindividuellen Unterschiede. Obgleich eigentlich eine interindividuelle Unterscheidbarkeit der Ängstlichkeit als Persönlichkeits-konstrukt hinsichtlich der erlebten *Häufigkeit* und *Intensität* der Angst angenommen werden kann, wurden diese beiden Aspekte der Angst stets gemeinsam erforscht. Der systematischen empirischen Erforschung interindividueller Unterschiede in der Disposition zur Angstreaktion (auch: Angstneigung / Ängst-lichkeit / Angstbereitschaft genannt), welche mit Cattell und Scheier in den 60er Jahren begann, gingen bereits in den 50er Jahren einige Bemühungen der Operationalisierung „manifester Angst“ (Triebtheorie von Taylor, 1958, fußend auf der Lerntheorie von Hull, 1943) voraus.

Cattell und Scheier (1960) können als Väter der empirischen, faktorenanalytischen Angstforschung angesehen werden. Sie ergründeten,

inwiefern Ängstlichkeit als eine eigene faktorenanalytische Dimension identifizierbar ist, und schlussfolgerten aus ihren Studien, dass Ängstlichkeit allen Kriterien einer „trait definition“ und einer „type definition“ entspreche und damit als eine allgemeine Persönlichkeitsdimension betrachtet werden könne. Die „trait definition“ sei erfüllt, wenn die von Klinikern der Angst zugeschriebenen Variablen möglichst „rein“ auf einem „Angstfaktor“ laden (d. h. auf allen anderen Faktoren einer Faktorenlösung möglichst gering laden); die „type definition“ sei erfüllt, wenn der Angstfaktor mit anderen Angstindikatoren (z. B. Diagnosen der Angst oder anderen Angsttests) hoch korreliere. Cattell und Scheier (1960) fanden in ihren faktorenanalytischen Untersuchungen, dass Ängstlichkeit als ein stabiler Faktor zweiter Ordnung („FQII“) als gut gesichert gelten kann. Dieser setzt sich aus den folgenden sechs Cattell’schen Primärfaktoren zusammen: Triebspannung, Neigung zu Schuldgefühlen, fehlende Willenskontrolle, fehlende Ichstärke, Misstrauen und Furchtsamkeit. Das Cattell’sche faktorenanalytisch fundierte Persönlichkeitskonstrukt „Ängstlichkeit“ konnte in zahlreichen Replikationsstudien durch enge Korrelationen zu Faktoren des Guilfordschen Persönlichkeitssystems (E-Faktor: fehlende emotionale Stabilität; O-Faktor: Hypersensitivität / Misstrauen; P-Faktor: überkritische Einstellung gegenüber Menschen; Guilford, Zimmermann & Guilford, 1976) und zum Faktor Neurotizismus des Eysenckschen Persönlichkeitssystems (Fünf-Faktoren-Modell; Eysenck, 1947) sowie zur Repression-Sensitization-Skala von Byrne (1961) empirisch gesichert werden. Über den genauen Zusammenhang von „Neurotizismus“ und „Ängstlichkeit“ herrscht jedoch noch Uneinigkeit unter den Forschern. Obgleich Eysenck und Eysenck (1985) und Gray (1981) Ängstlichkeit als eine Kombination aus Neurotizismus und niedriger Extraversion konzipierten, weisen Costa und McCraes (1985) faktorenanalytische Untersuchungen mit hohen Korrelationen von vier der sechs Facetten der Neurotizismusskala des NEO-PIs (Neurotizismus-Extraversion-Offenheit-Psychotizismus-Introversion-Persönlich-keits-Inventar) mit der Neurotizismus-Skala des EPQ (Eysenck Personality Questionnaire) deutlich darauf hin, dass Ängstlichkeit und Neurotizismus sehr ähnliche, wenn nicht sogar identische Persönlichkeitskonstrukte auf einem hohen, allgemeinen Abstraktionsniveaus darstellen (Amelang & Bartussek, 2001, S. 450ff; siehe Kapitel 2.7.3.1.).

2.4.1.1. Exemplarisch: Das State-Trait-Modell der Angst

Das State-Trait-Modell der Angst erfreut sich seit seiner Konzeption im Jahre 1972 durch Spielberger einer großen Beliebtheit. Die Popularität und breite Rezeption dieses Modells liegt wahrscheinlich darin begründet, dass es alltagspsychologische Überzeugungen reflektiert, auf empirischen, faktoren-analytischen Ansätzen von Cattell und Scheier (1960) aufbaut, seine Wurzeln bereits bei Freud (1940) zu finden sind, und in der Entwicklung eines modellkonformen, ökonomischen Messinstruments mündete, welches im Folgenden intensive Forschungstätigkeit anregte.

Das Modell konzipiert Angst zweidimensional in Form einer Zustands-Angst (State) und einer Eigenschafts- (Trait). Diese beiden „Grundpfeiler“ der Angst werden nach Spielberger (1972) folgendermaßen definiert:

> „State-Angst ist ein emotionaler Zustand, welcher durch Anspannung, Besorgtheit, Nervosität, innere Unruhe und Furcht vor zukünftigen Ereignissen gekennzeichnet ist. Physiologisches Korrelat ist eine erhöhte Aktivität des autonomen Nervensystems.“

> „Trait-Angst ist eine erworbene, zeitstabile Verhaltensdisposition, welche bei einem Individuum zu Erlebens- und Verhaltensweisen führt, eine Vielzahl von objektiv wenig gefährlichen Situationen als Bedrohung wahrzunehmen.“

Die Annahme, dass Individuen sich konsistent und kohärent in einer dispositionellen Ängstlichkeit (Trait) unterscheiden, ist grundlegende Voraussetzung für eine sinnvolle Angstmessung. In dem von Spielberger entwickelten State-Trait-Anxiety-Inventory (STAI) schlägt sich die konzeptionelle Unterscheidung zwischen einer State- und einer Trait-Angst in der Konstruktion zweier getrennter Skalen nieder. Da intensive Forschungstätigkeiten mit dem STAI mehrfach zwar die empirische Unterscheidbarkeit aber nicht die statistische Unabhängigkeit der beiden Konstrukte belegen konnten (siehe Kapitel 2.7.3.4.), diskutiert man heute einen prinzipiell stufenlosen Übergang der Zustands- zur Eigenschafts-Angst im Sinne eines State-Trait-Kontinuums (Hermann, Scholz & Kreuzer, 1991).

Der Zusammenhang von State- und Trait-Angst wurde 1972 von Spielberger folgendermaßen beschrieben. Aussagen über zukünftiges Angsterleben ließen sich auf der Grundlage der Feststellung einer Angstdisposition (Trait), welche sich aus der Häufigkeit und Intensität vergangener Angstzustände (State)

ableite, treffen. Insofern nahm bereits Spielberger einen engen (induktiven) Zusammenhang zwischen den beiden Konstrukten an.

Die zentrale Aussage des State-Trait-Modells lautet, dass hochängstliche Menschen Situationen, die mit einer Bedrohung des Selbstwerts verknüpft sind, bedrohlicher als Niedrigängstliche wahrnehmen, d. h. in solchen Situationen einen höheren Anstieg der Zustands-Angst aufweisen (Spielberger, 1972). Somit trifft das Modell nicht nur Annahmen über eine dispositionelle und situative Angst, sondern behandelt auch deren Interaktion (> interaktionistischer Ansatz). Desweiteren entwarf Spielberger ein Prozessmodell zur Angst-entstehung, welches hier jedoch aus Platzgründen nicht näher erörtert werden kann (siehe Laux & Glanzmann, 1996, S. 110).

Obgleich das State-Trait-Modell einen großen Einfluss auf die Entwicklung der Angstforschung hatte, wird es angesichts der Komplexität und Differenziertheit heutiger Forschungsergebnisse oftmals für seine konzeptuellen und empirischen Schwächen kritisiert; Stöber und Schwarzer (2000, S. 191) schreiben gar, dass Modell und Instrument mittlerweile als überholt gelten.

2.4.1.2. Kritik am State-Trait-Modell der Angst

Zu den Hauptkritikpunkten gehört: a) die Fragwürdigkeit, ob State- und Trait-Angst tatsächlich qualitativ unterscheidbare Konstrukte sind, oder ob es nicht vielmehr - wie es sich auf der Testebene andeutet - einen stufenlosen Übergang eines State-Trait-Kontinuums gibt (Laux & Glanzmann, 1996), b) die Annahme, dass Hoch- und Niedrig-Ängstliche in nicht bedrohlichen Situationen die gleiche State-Angst aufweisen, konnte empirisch nicht bestätigt werden, für Hoch-Ängstliche ist eine generell höherer State-Angst belegt (Krohne, 1996) und c) die Einschränkung des Ansatzes auf selbstwertbedrohliche Situationen.

Desweiteren wird bezüglich einzelner Annahmen, die im Rahmen des Modells getroffen werden, Kritik geübt. So wurde z. B. das Modell erst 1985 um eine kognitive Angst-Komponente („Besorgtheit“) erweitert, das Postulat der „Proportionalität“ der erlebten Gefährdung und Intensität der Angst ist wie einige andere Annahmen empirisch nicht überprüfbar, da es zu unkonkret formuliert ist (Ist Propotionalität als eine lineare oder nicht-lineare Beziehung zu verstehen? Wie spezifisch sind die einzelnen Aspekte? Sind autonome Reaktionen konkordant mit dem subjektiven Erleben? etc.). Aus diesen Gründen ist eine unkritische Übernahme des State-Trait-Ansatzes sicher nicht zu empfehlen;

dennoch gibt es kein anderes Modell der Angst, welches dieses bisher abgelöst hat. Obgleich es zahlreiche Forschungsbemühungen gibt, das Konstrukt der Angst systematisch in verschiedene Dimensionen zu unterteilen (siehe Kapitel 2.7.3.4.), ist eine Einigung in diesem diffizilen Forschungsfeld noch nicht gelungen. Dies gründet sich möglicherweise in der faktorenanalytischen Methodik, welche zwar die Differenzierung zwischen einzelnen Angst-komponenten erlaubt, die Auswahl der als bedeutsam eingestuften Faktoren jedoch oft willkürlich erscheinen lässt.

2.5. Angst und Depression

Angstgefühle und depressive Verstimmungsgefühle kommen sehr häufig gemeinsam vor. *Epidemiologische* Studien weisen auf eine *hohe Komorbidität* zwischen Angststörungen und depressiven Störungen hin (14,6 – 45,9%, Neumer, 2000, S. 53; 50-65%; DSM-IV, Saß, Wittchen & Zaudig, 1996). Möller, Laux und Deister (1996, S. 112) beteuern, dass eine genaue *klinische* Trennung der beiden Emotionen auf Syndromebene nicht möglich sei und es noch unklar sei, ob beide Phänomene als Ausdrucksformen *einer* zugrunde liegenden psychischen Störung gelten können, oder ob beide als „Symptome“ aufeinander bezogen seien. *Klassifikatorisch* werden beide Syndrome zunächst als getrennte nosologische Einheiten betrachtet, jedoch wird derzeit diskutiert, ob - ähnlich im ICD-10 eine Störungskategorie „Angst und depressive Störung gemischt“ existiert - die sich auch im DSM-IV etablieren sollte (Neumer, 2000). Dafür sprechen klinische Studien, die zeigen, dass sich Patienten und Psychiater in der Art und Weise des Umgangs mit den Begriffen Angst und Depression unterscheiden. Während sich bei Psychiatern keine Korrelation zwischen den Begriffen Angst und Depression finden ließ, überlappten sich diese Konzepte bei den Patienten erheblich (Sims & Snaith, 1993, S. 31). Helmchen und Linden (1986) fanden, dass die Differenzierung von Angst und Depression im diagnostischen Gespräch oft nicht zuverlässig gelingt, da unter anderem „Depression von Patienten oft als Bezeichnung für jedwelchen unangenehmen Gefühlszustand verwendet wird, dessen Verursachung sie sich nicht erklären können“ (Rauchfleisch, 1992, S. 38).

Angst und Depression sind spezifische Muster von Emotionen, Kognitionen, Verhaltensweisen und physiologischen Merkmalen, welche sich teilweise überschneiden. Garber, Miller und Abramson (1980) versuchen

Gemeinsamkeiten und Unterschiede zwischen Angst und Depression auf unterschiedlichen *Symptom*-Ebenen herauszustellen. Das Konstrukt der Hilflosigkeit[9], welches nach Garber und Mitarbeiter (1980) sowohl für die Angst als auch für die Depression charakteristisch sei, vermag eine konzeptionelle Brücke zwischen der Angst und der Depression zu schlagen. Schon in den Klassifikationssystemen zeigt sich die Überschneidung zwischen Angst und Depression bezüglich eines verstärkten *Hilflosigkeitserlebens* (Agoraphobie / Panikstörung, DSM-IV, Saß el., 1996, S. 456f.). Die zentrale Rolle der Hilflosigkeit im Rahmen der *Depressionsentstehung* führt Seligman in seiner „Theorie der erlernten Hilflosigkeit“ (1975) aus. So generiere die wiederholte Erfahrung mangelnder Kontrolle eine generalisierbare Erwartung von Unkontrollierbarkeit, was nach Abramson, Seligman und Teasdale (1978) wiederum zu einem pessimistischen Attributionsstil führe (lerntheoretischer Erklärungsansatz). „Pessimistische, negative Zukunftsperspektiven“ zählen auch zu den Hauptkriterien einer *Depressiven* Episode (ICD-10, F.32; Dilling, Mombour & Schmidt, 2000, S. 139).

Zur Differenzierung zwischen Angst und Depression schlugen Garber und Mitarbeiter (1980) das Konstrukt der *Hoffnungslosigkeit* vor („Theorie der Hoffnungslosigkeit“; Stotland, 1969). Dies griff Ulich (1989, S. 208) auf, der im Falle eines Überschreitens der Grenze zwischen Hilflosigkeit zu Hoffnungslosigkeit von einer Einmündung der Angst in die Depression spricht. Das von Garber und Mitarbeitern (1980) erdachte *Kontinuum* der Angst und Depression steht jedoch zunächst dem Verständnis eines *simultanen* Angst- und Depressionserlebens entgegen. Die Autoren erklären das gleichzeitige Erleben von Angst und Depression auf zwei Weisen. Entweder oszilliere ein Individuum ständig zwischen verschiedenen Graden der Kontrollierbarkeitseinschätzung (sicher / nicht kontrollierbar: Depression; unsicher / kontrollierbar: Angst) oder es treffe unterschiedliche Kontrollierbarkeitseinschätzungen gleichzeitig in Abhängigkeit von spezifischen Situationen. Auch eine *sequentielle* Beziehung des Angsterlebens zum depressiven Erleben ist denkbar, wenn man sich vor Augen führt, dass Angst jeweils *vor* einem Objektverlust / Selbstwertverlust entsteht, während sich Traurigkeit bzw. Depression *als Folge* davon einstellen kann (Helmchen & Linden, 1986).

[9] Hilflosigkeit wird definiert als die „Unabhängigkeit eines Outcomes von dem Verhalten des Individuums“ (Garber, Miller & Abramson, 1980).

Neben diesen *epidemiologischen* und *klinischen* Überlegungen zur Differenzierung von Angst und Depression widmen sich seit den 80er Jahren auch *psychometrische* Studien dem Thema. Sie belegen einen engen Zusammenhang zwischen Angst- und Depressionsphänomenen (Korrelationen zwischen Angst- und Depressionsmaßen liegen nach Finney, 1985, bei r = 0,5; nach Laux, Glanzmann, Schaffner & Spielberger, 1981, bei r = 0,68 / 0,72 und nach Mineka, Watson & Clark, 1998, bei r = 0,61-0,78, $\bar{r}$ = 0,69).

Faktorenanalytische Modellierungen von Angst und Depression entstammen vor allem zwei Forschergruppen. Die Forschergruppe um Watson und Clark (1984) schlug in Ermangelung einer deutlichen empirischen Trennung (d. h. aufgrund einer hohen gemeinsamen Varianz) von Angst und Depression in einem *„Tripartite Modell*“ ein allgemeines Konzept der „*negativen Affektivität*“ vor, welches Emotionen wie Angst, Traurigkeit, Nervosität, Wut, Enttäuschung integriert. Neben diesem *globalen* Faktor konzipieren die Autoren weiterhin *zwei spezifische* Sekundärfaktoren, von denen der eine als *angst*spezifisch angesehen wird, da er sich aus einem Symptommuster der somatischen Anspannung und *vegetativen* Überregbarkeit zusammensetzt, und der zweite *depressions*spezifisch im Sinne eines Mangels an positiven Affekt (Anhedonie) sei (Clark & Watson, 1991; Watson & Clark, 1984; Watson et al., 1995). In diesem frühen Modell, gilt – in Abgrenzung zur Depression – die *vegetative Übererregbarkeit* bzw. die somatische Anspannung als *spezifisches* Charakteristikum der Angst.

Diese Vorstellung wurde im *„Drei-Faktoren-Modell“* von Barlow und Mitarbeitern weiterentwickelt (Balow et al., 1996; Zinbarg & Barlow, 1996; Chorpita et al., 1998). Die Forschergruppe unterscheidet ebenfalls *drei* Grundemotionen: a) die negative Affektivität als Ausdruck von Angst, b) die autonome Erregung als Ausdruck von Furcht bzw. Panik und c) Anhedonie und Hoffnungslosigkeit als Indikatoren für Depression (siehe Garber et al., 1980).

Die grundlegende Weiterentwicklung, welche in diesem Modell formuliert ist, liegt in der Ablösung von der bislang typischen Vorstellung, dass das Konstrukt der Angst vor allem durch Symptome vegetativer Erregung abzubilden sei. Für Barlow und Mitarbeiter (1996) sind diese *vegetativen* Symptome *nicht* für das Konstrukt der Angst im Allgemeinen charakteristisch, sondern *spezifisch* für *akute Panik*- bzw. Furchtzuständen.

Die Konzeption eines spezifischen *separaten* vegetativen Indikators, welcher für *Panik*zustände kennzeichnend, und *nicht* im Sinne eines *globalen*, breiten Angstfaktors zusammen mit allen anderen Angstsymptomen zu interpretieren sei, setzte sich gestützt durch zahlreiche richtungsweisende Befunde aus Strukturgleichungsanalysen (Brown et al., 1997; Chorpita et al., 1998) schließlich durch. Für diese Modellierung spricht auch die nosologische Einordnung von attackenweise auftretenden vegetativen Angstsymptomen bei Panikstörungen (F41.0; ICD-10; Dilling et al., 2000; DSM-IV; Saß et al., 1996; siehe Kapitel 2.6.1.). Während in einem neuen Modell, welches von der anfangs erörterten Forschungsgruppe um Watson und Clark (Mineka et al., 1998)[10] stammt, eine differenzierte Integration *verschiedener* Komponenten der Angst im Rahmen einer übergeordneten *hierarchischen* Struktur angestrebt wird, fokussiert eine IRT-basierte Studie von Krüger und Finger (2001) nun in jüngster Zeit wieder eine *eindimensionale* Modellierung von Angst und Depression durch einen beiden Konstrukten gemeinsamen *„Internalisierungs-faktor“*.

Es lässt sich resümieren, dass der skizzierte Forschungsdiskurs seit Jahrzehnten verschiedene Modellierungen von Angst (und Depression) erbrachte und zum jetzigen Zeitpunkt noch nicht abgeschlossen erscheint. Offensichtlich gestaltet sich die psychometrische Diskimination zwischen verschiedenen Komponenten der Angst und der Depression schwierig und stellt eine hohe Herausforderung an die psychometrische Forschung dar.

[10] Zum „Integrativen Hierarchischen Modell von Angst und Depression“ (Mineka et al., 1998): Dieses Modell erklärt jedes klinische Syndrom (d. h. spezifische Angst- oder depressive Störungen), durch einen allen Syndromen gemeinsamen Faktor höherer Ebene („negative Affektivität“) und durch eine spezifische Komponente. Die Syndrome unterscheiden sich in ihrem Verhältnis der Varianz, welche von einem gemeinsamen Faktor, und der Varianz, welche von einem spezifischen Faktor aufgeklärt werden kann. Zudem differieren die spezifischen Komponenten - je nach Syndrom - in der Anzahl und Gewichtung verschiedener Symptome.

2.6. Angst als Störung

Angst kann in einen Primäraffekt (im Sinne einer Zustands-Angst), ein Persönlichkeitsmerkmal (Eigenschafts-Angst) und eine pathologische Angst unterschieden werden.

Pathologische Merkmale der Angst sind nach Lieb und Wittchen (1998, S. 882): a) eine unbegründete, unangemessen starke (Intensität) und häufige (Frequenz) Angst, b) welche konsistent und überdauernd sei, c) Vermeidungsverhalten begründet und Angst vor dem Kontrollverlust mit einschließt, sowie d) zu Beeinträchtigungen der Lebensqualität (> sozialer und beruflicher Leidensdruck) führe. Möller und Mitarbeiter (1996) heben darüber hinaus hervor, dass auch das Fehlen von Angst krankheitswertig sein könne (z. B. im Rahmen soziopathischer Persönlichkeitsstörungen).

Angst kann als eigene psychische Erkrankung (siehe Kapitel 2.6.1.) oder als Symptom(komplex) im Rahmen anderer psychischer Störungen (Depression, Schizophrenie, Zwangsstörung, Persönlichkeitsstörung, Anpassungsstörung), körperlicher Erkrankungen (z. B. internistische Erkrankungen wie Schilddrüsenüber- bzw. -unterfunktion, Nebennierenrindenüberfunktion, Hypoglykämie, koronare Erkrankungen, Atemwegserkrankungen, Vitamin B_{12}-Mangel und neurologische Erkankungen wie Multiple Sklerose, hirnoganische Anfallsleiden, Chorea Huntington etc.) sowie substanzinduziert (Entzug von Alkohol, Opiaten, Anxiolytika etc., Intoxikation von Halluzinogenen, Alkohol, Nikotin, Koffein, Amphetaminen, Kokain etc.) auftreten.

Für Angststörungen charakteristische Symptome manifestieren sich dabei auf der subjektiv-emotionalen, kognitiven, behavioral-motorischen und / oder physiologischen Ebene (Möller et al., 1996; Freyberger & Stieglitz, 1996).

Zu den subjektiv-emotionalen Merkmalen der Angst gehören die Angst vor der Angst („Erwartungsangst“), Ängste, die Kontrolle zu verlieren, verrückt zu werden, zu ersticken oder zu sterben; typische kognitive Symptome sind anhaltende Sorgen, Grübeln, kognitive Einschränkungen (auf gefährliche Stimuli), Desorganisiertheit und Konzentrationsschwierigkeiten, behavioral-motorisch ist vor allem das Vermeidungsverhalten zentral sowie körperliche Unruhe, Zittern, Spannungskopfschmerz und die Unfähigkeit, sich zu entspannen (zu typischen körperlichen Angstsymptomen siehe Kapitel 2.3.4.).

2.6.1. Klassifikation (ICD-10 und DSM-IV)

Unter dem Oberbegriff „*Angststörungen*“ werden mehrere Störungsgruppen zusammengefasst, die durch unterschiedliche Erscheinungsweisen der Angst geprägt sind. Die wesentlichen Formen sind die *Phobien*, *Panikstörungen* und *Generalisierte Angststörungen* („Neurotische Belastungs- und somatoforme Störungen“, F.4 des ICD-10, Dilling et al., 2000). Das DSM-IV (Saß et al., 1996) fasst in einer umfassenderen Definition auch *Zwangsstörungen*, *akute Belastungsstörungen* und die *Posttraumatische Belastungsstörung* unter dem Sammelbegriff der „Angststörungen“ zusammen.

Im Folgenden werden die drei klassischen *Angststörungen*, welche im ICD-10 (F.4) beschrieben werden, erläutert. Unter *Phobien* (F.40) versteht man unbegründet starke Ängste, welche ausschließlich oder überwiegend durch eindeutig definierte, im Allgemeinen ungefährliche Situationen oder (außerhalb der Person liegende) Objekte hervorgerufen werden und Vermeidungsverhalten provozieren. Nach einer Unterteilung, welche auf Marks (1970) zurückgeht, unterscheidet man zwischen *Agoraphobien*, *sozialen Phobien* und sogenannten *spezifischen Phobien*.

Die *Agoraphobie* wird gekennzeichnet durch eine deutliche und anhaltende Furcht vor und / oder dem Vermeiden von mindestens zwei der folgenden Situationen: Menschenmengen, öffentliche Plätze, alleine Reisen, Reisen mit weiter Entfernung von Zuhause; das Schlüsselsymptom ist das Fehlen eines nutzbaren Fluchtweges (Stumm & Pritz, 2000, S. 34). Bei *sozialen Phobien* steht die Furcht, im Zentrum der Aufmerksamkeit zu stehen, sich peinlich oder erniedrigend zu verhalten und gegebenenfalls das Vermeidungsverhalten solcher Situationen im Vordergrund. Die Angst wird als übertrieben oder unvernünftig empfunden. Zu den *spezifischen Phobien* werden anhaltende Ängste vor einem umschriebenen Objekt oder einer umschriebenen Situation verstanden, welche ebenfalls Vermeidungsverhalten provozieren können, und ein Individuum in seinem Leben oder seinen alltäglichen Aktivitäten beeinträchtigen (z. B. Zoophobie, Akrophobie, Klaustrophobie, Verletzungsphobie etc.).

Als *Panikstörung* (F.41.0) wird Angst klassifiziert, wenn ohne sichtbaren Anlass ausgeprägte Angst oder Panik wiederholt in Form von spontanen, unerwarteten Panikattacken auftritt (4 mal pro Monat) mit einer spezifischen Erwartungsangst

verknüpft ist und regelmäßig zu intensiven vegetativen Symptomen führt, welche Leidensdruck hervorrufen. Eine Komorbidität von Panikstörung und Agoraphobie ist häufig (21,6%; Neumer, 2000).

Unter einer *Generalisierten Angststörung* (F.41.1) versteht man mindestens sechs Monate lang anhaltende Sorgen und Befürchtungen bzgl. alltäglicher Ereignisse und Probleme, welche nicht nur auf bestimmte Situationen und Objekte begrenzt sind. „Es bestehen unrealistische Befürchtungen, motorische Spannung und vegetative Übererregbarkeit“ (Möller et al., 1996, S. 110).

2.6.2. Epidemiologie

Angst stellt laut Möller und Mitarbeitern (1996, S. 98) eine der häufigsten psychopathologischen Symptome dar. In Allgemeinarztpraxen geben mehr als die Hälfte der Patienten Angst als subjektive Beschwerde an; davon wird in 20% der Fälle die Angst als behandlungsbedürftig angesehen. In der Allgemeinbevölkerung findet sich Angst als behandlungsbedürftiges Symptom bei 10% aller Menschen. Die Angaben zur Lebenszeitprävalenz schwanken in sechs aktuellen epidemiologischen Studien zwischen 9,2 und 28,3% (Neumer, 2000, S. 57). Phobische Störungen sind am häufigsten (mit 13% Lebenszeit-prävalenz; LP). Die soziale Phobie steht an erster Stelle dieser Störungsgruppe mit einer Monatsprävalenz (MP) von 6-8% (Kessler, McGonagle, Zhao, Nelson, Hughes, Eshelman, Wittchen, & Kendler, 1994), gefolgt von den Generalisierten Angststörungen (MP: 2-3%) und Panikstörungen (LP: 2-3%; Möller et al., 1996, S. 99). Frauen sind wesentlich häufiger betroffen als Männer (bis zu zweifach höheres Erkrankungsrisiko je nach Störungsgruppe). Hinsichtlich weiterer soziodemografischer Faktoren lassen sich nur geringfügige Unterschiede finden. Nach dem 45. Lebensjahr nimmt die Inzidenz von Angststörungen deutlich ab.

2.6.3. Verlauf und Therapie

Angststörungen neigen zu Chronifizierung aufgrund a) eines häufigen Vermeidungsverhaltens und einer ständigen Erwartungsangst, welche die Angst verstärkt (zum Teufelskreismodell der Angst siehe Abbildung 3 in Kapitel 2.7.3.4.), und oftmals zu sozialer Isolierung führt, b) einer häufigen Komorbidität mit anderen Erkrankungen und c) einer häufig ungünstigen (Selbst-)Medikation (Missbrauch von Anxiolytika, Alkohol und anderen Drogen). Zur Angstreduktion bieten sich unterschiedliche Therapieansätze an.

Grundsätzlich kann zwischen psychopharmakologischen (Benzodiazepine, Antidepressiva, Betablocker etc.) und nichtpharmakologischen Therapieansätzen unterschieden werden. Zu den nichtpharmakologischen Ansätzen zählen stützende ärztliche Gespräche, Entspannungsverfahren (Autogenes Training, Progressive Muskelrelaxation, Bio-Feedback etc.), sozio-therapeutische Strategien (berufliche Reintegration, Alltagsbewältigung etc.) und psychotherapeutische Therapieverfahren.

Auf der Grundlage verschiedener Theorien zur Entstehung und Aufrechterhaltung von Angststörungen versucht man Angststörungen mit verschiedenen psychotherapeutischen Ansätzen zu beheben. Während tiefenpsychologisch orientierte Verfahren aufdeckend arbeiten, indem sie den der Angst zugrundeliegenden Konflikt behandeln, wird in Ansätzen der kognitiven Verhaltenstherapie durch kognitive Umstrukturierung (Neubewertungen) und systematische Verhaltensübungen (graduierte Angstexposition: „systematische Desensibilisierung“; massive Reizüberflutung: „flooding“) versucht, dem „Teufelskreis“ der Angst entgegenzuwirken. Humanistische Therapieansätze stellen die Persönlichkeitsentfaltung im Sinne einer Förderung des Kongruenzerlebens durch die Akzeptanz inkongruenter, potentiell angstauslösender (abgespaltener) Persönlichkeitsanteile in den Vordergrund.

2.7. Messung von Angst

Ängstlichkeit ist ein differentiell-psychologisches Konstrukt, welches sich der direkten Beobachtung entzieht. Es lässt sich zusammen mit anderen Konstrukten (z. B. Depression) in einem (nomologischen) Netzwerk von Beziehungen verorten und mit Hilfe empirischer Indikatoren beschreiben.

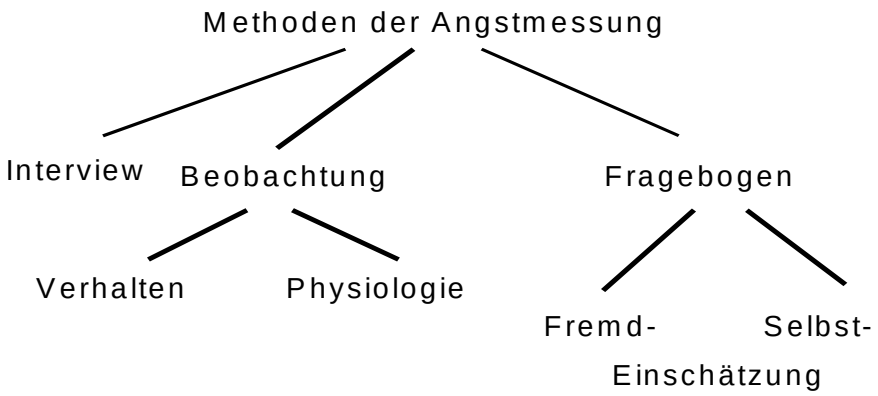


Abbildung 1: Methoden der Angstmessung – ein Überblick.

Es gibt drei verschiedene Gruppen von Methoden zur Erhebung unterschiedlicher empirischer Indikatoren der Angst: die Interviewtechnik, die Beobachtung und die Fragebogenerhebung (Selbst-/Fremdbeschreibung). Einen Überblick über diese verschiedene Methoden und deren Untergruppen, welche in diesem Kapitel vorgestellt werden, illustriert vorangegangene Abbildung 1.

2.7.1. Interviewtechniken

Im Folgenden wird ein kurzer Überblick über Interviewtechniken und Beobachtungverfahren zur Angstmessung gegeben, um dann schließlich den Hauptfokus auf die Erörterung verschiedener Fragebogenverfahren zu legen.

Die verschiedenen Interviewtechniken, die zur Angsterhebung genutzt werden, stammen aus dem klinischen Bereich und dienen vor allem der strukturierten Diagnostik der Angst als Störung (siehe Kapitel 2.6.1.). Zu diesen Interviews, die, um eine möglichst hohe Erhebungsobjektivität zu gewährleisten, zumeist vollstrukturiert sind, gehören das DIA-X[11] (Computerversion: M-CIDI[12]; Diagnostik nach DSM-IV und ICD-10, siehe Kapitel 6.5.3.), das SKID[13] (DSM-IV), das DIPS[14] sowie z.B. der semistrukturierte Leitfaden der AMDP[15].

2.7.2. Beobachtung von Angst

Beobachtungstechniken zur Angstmessung dienen entweder der Erhebung von Verhalten oder von physiologischen Parametern. Die *Verhaltensbeobachtung* kann mit Hilfe sogenannter „Kategoriensysteme“ direkt im „natürlichen“ Feld (in vivo) oder im Labor (in vitro) erfolgen, wobei letzteres in der Psychologie häufiger ist. In der Psychologie werden vor allem Verhaltenskorrelate der Angst (Mimik, Gestik, Vokalisation, Motorik) erfasst. Speziell die experimentelle Angstinduktion zur Erfassung von Angst wirft allerdings eine Reihe von ethischen Problemen auf. Die Verhaltenserfassung im klinischen Bereich geschieht im Rahmen verhaltenstherapeutischer Ansätze mit Hilfe sogenannter

[11] DIA-X: Diagnostisches Expertensystem für Psychische Störungen (Wittchen & Pfister, 1996).
[12] M-CIDI: Munich Composite International Diagnostic Interview (Wittchen & Pfister, 1996).
[13] SKID I und II: Strukturiertes Klinisches Interview nach DSM-IV (Wittchen, Wunderlich, Guschwitz & Zaudig, 1997).
[14] DIPS: Diagnostisches Interview bei psychischen Störungen (Magraf, Ehlers & Schneider, 1994).
[15] AMDP: Arbeitsgemeinschaft von Methodik und Dokumentation in der Psychiatrie (1997).

Angst-Tagebücher, in denen der Patient sein selbst beobachtetes Verhalten (und andere Erlebensaspekte der Angst) systematisch dokumentiert.[16]

Welche Möglichkeiten bei der *Messung von physiologischen Parametern* als Korrelate des Angsterlebens existieren, erörtert Krohne (1996, S. 42ff) ausführlich. Physiologische Parameter können auf allen biologischen Reaktionsebenen abgeleitet werden (siehe Kapitel 2.3.4.). So werden in Laborstudien *zentralnervöse* Angstkorrelate (z. B. eine erhöhte kortikale Aktivierung im EEG), Parameter des *peripheren* Nervensystems (z. B. eine erhöhte Aktivierung im EKG, EDA, EMG), des *neuroendokrinologischen* (z. B. eine erhöhte Konzentration von Adrenalin, Noradrenalin, ACTH, Kortisol sowie der Wachstumshormone und Endorphine) und des immunologischen Systems (z. B. die Reduktion von T-Zellen) untersucht. Da dies die Anwendung von apparativen Einrichtungen erfordert, ist diese Erfassung meist kompliziert, kostspielig und erfordert viel Erfahrung. Die Hoffnung, dass es bestimmte *angstspezifische physiologische* autonome Aktivierungsmuster gibt, konnte bisher *nicht* bestätigt werden (Fahrenberg, 1967). Da „der Intensität nach kein genaues psychophysiologisches Korrelat zur subjektiv erlebten Erregung“ (Tewes & Wildgrube, 1999, S. 29), welche mit der Angst einhergeht, existiert, gilt bisher die Selbsteinschätzung per Fragebogen als die zuverlässigste Quelle zur Differenzierung zwischen Emotionen (Krohne, 1996).

2.7.3. Fragebögen

Fragebögen - auch Skalen oder Inventare genannt - gehören zu den populärsten psychologischen Methoden zur Erfassung psychischer Erlebens- und Verhaltensweisen. Sie erheben über die Darbietung einzelner Items (Fragen/Aussagesätze/Wörter), welche Gefühle und Meinungen von sich selbst und der Umgebung beinhalten, den Grad der Zustimmung oder Ablehnung einer Person. Ein Gesamtpunktwert („score“) wird aus den einzelnen Itembeantwortungen der Testperson ermittelt, von dem aus auf das Ausmaß einer bestimmten Merkmalsausprägung (hier: Angst) geschlossen wird (zum Zusammenhang zwischen einer Messung und einem latenten Merkmal siehe Kapitel 3 zur Item Response Theorie, IRT).

[16] Beispiele für Angst-Tagebücher sind das Marbuger Angst-Tagebuch (Margraf & Schneider, 1990), das Generalisierte Angsttagebuch, (Wittchen, Schuster & Vossen, 1997) und das Angsttagebuch für Panikstörungen (Börner, Gülsdorff, Margraf, Osterheider, Philipp & Wittchen, 1997).

Fragebögen haben gegenüber aufwendigen Interview- und Beobachtungstechniken den Vorteil, dass sie in der Durchführung und Auswertung schnelle, leichte und einfache Verfahren sind, welche Merkmale objektiv, reliabel und valide messen können.

Nachteilig ist an Fragebögen allgemein, dass ihre Aussagekraft durch spezifische Antworttendenzen[17] verfälscht werden kann. Desweiteren setzen sie kognitive Fähigkeiten, wie z. B. eine gewisse Selbstreflexion sowie die Motivation bzw. den Willen voraus, Aussagen über sich oder Andere bzw. spezifische Konstrukte zu treffen. Die Motivation zur Selbstauskunft ist speziell im klinischen Bereich aufgrund eines hohen Leidensdrucks der Patienten oftmals gegeben, in der Arbeits- und Organisationspsychologie ist jedoch z. B. im Rahmen von (Stellen-)Bewerbungen mit verstärkten Verfälschungstendenzen zu rechnen.

Fragebogenverfahren können in Selbst- und Fremdbeurteilungsverfahren eingeteilt werden. Das im deutschsprachigen Raum am weitesten verbreitete Fremdeinschätzungsverfahren zur Angsterfassung ist laut Swinson, Cox und Fergus (1993) die Hamilton-Angst-Skala (HAMA; Hamilton, 1959, 1977). Einen Überblick über verschiedene Klassen von Selbstbeurteilungsverfahren gibt Abbildung 2.

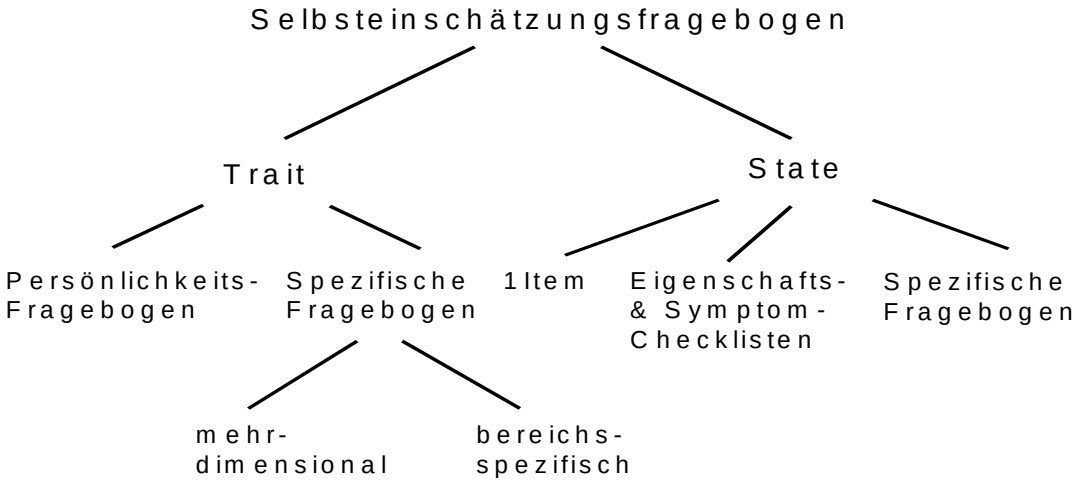


Abbildung 2: Sebsteinschätzungsfragebögen zur Angstmessung – ein Überblick.

[17] Beispiele für formale Antworttendenzen: Zustimmungs- / Ablehnungstendenz; Tendenz zur Mitte oder den Extremen; Beispiele für inhaltlich-begründete Antworttendenzen: soziale Erwünschtheit).

Laut einer Psyndex[18] Testrecherche (1945-2003) liegen allein im deutschsprachigen Raum bisher über 58 verschiedener Fragebögen zur Selbsteinschätzung der Angst vor. Am ehesten lassen sich diese Fragebögen in solche unterscheiden, welche sich der Erfassung der Ängstlichkeit als ein Persönlichkeitsmerkmal entweder in Form von Persönlichkeitsfragebögen (siehe Kapitel 2.7.3.1.) oder in Form sogenannter Trait-Angst-Verfahren (siehe Kapitel 2.7.3.2.) widmen, und solche, welche der ausschließlichen Erhebung der Zustands-Angst dienen (siehe Kapitel 2.7.3.3.).

2.7.3.1. Persönlichkeitsfragebögen

Ängstlichkeit kann als ein Persönlichkeitsmerkmal in Form von Subskalen im Rahmen verschiedener Persönlichkeitsinventare erfasst werden (sehe Tabelle 2) Diese Persönlichkeitsinventare basieren auf unterschiedlichen Persönlichkeitstheorien (z.B. EPI: Hierarchisches Persönlichkeitsmodell von Eysenck, 1947; NEO-FFI: „Big Five"-Ansatz von Tupes & Christal, 1961; 16 PF: Hierarchisches Persönlichkeitsmodell nach Cattell, 1974). Insbesondere das Konstrukt „Neurotizismus" steht im engen Zusammenhang mit Angst, wie sie von klassischen Angstinventaren wie dem STAI gemessen werden (Korrelation: $r_{EPI\text{-}N\ /\ STAI\text{-}T} = 0{,}77$; Laux et al., 1981).

Tabelle 2: Überblick über Persönlichkeitsinventare, mit denen Ängstlichkeit erfasst werden kann.

Inventar	Abkürzung	Autoren	Jahr	Angst-Subskala
Eysenck-Persönlichkeits-Inventar	EPI	Eggert	1983	Neurotizismus
NEO-Fünf-Faktoren-Inventar	NEO-FFI	Borkenau & Ostendorf	1993	Neurotizismus
16-Persönlichkeits-Faktoren-Test	16 PF-R	Schneewind & Graf	1998	Primärfaktor: Besorgtheit; Globalfaktor: Ängstlichkeit
Freiburger-Persönlichkeits-Inventar	FPI	Fahrenberg, Hampel & Selg	1989	Gesundheitssorgen
Minnesota Mulitphasic Personality Inventory	MMPI-2	Hathaway, McKinley & Engel	2001	Hypochondrie, Neurotizimus

Weitere Persönlichkeitsinventare, mit denen Angst erfasst werden kann, sind analytische projektive Verfahren wie der *Rorschach-Test* (Rorschach, 1954), der *Thematische-Apperzeptions-Test* (TAT; Murray, 1991) und die *Holtzman*

[18] Psyndex: Datenbank der Zentralstelle für Psychologische Information und Dokumentation der Universität Trier. Sie enthält Nachweise und Abstracts zu deutschsprachigen Publikationen aus der Psychologie und ihren Randgebieten. Hier sind Artikel aus 250 Zeitschriften, Monographien, Beiträge aus Sammelwerken sowie Dissertationen und Reportliteratur aus Deutschland, Österreich und der Schweiz sowie Beschreibungen von in deutschsprachigen Ländern seit 1945 gebräuchlichen psychologischen Testverfahren dokumentiert.

Inkblot Technik (HIT; Holtzman, Thorper & Swartz, 1961). Mit diesen Verfahren ist Angst allerdings weniger gut quantifizierbar als mit den oben genannten Persönlichkeitsinventaren, da ein hoch idiosynkratisches Vorgehen - wie es bei der Anwendung projektiver Verfahren erfolgt - oftmals zu Lasten der Testgütekriterien geht. Daher werden diese Verfahren eher selten zur Angsterfassung genutzt.

2.7.3.2. Trait-Angst-Verfahren

Verfahren, welche Ängstlichkeit als Persönlichkeitsmerkmal gesondert erfassen, können in Verfahren zur Erfassung einer allgemeinen Trait-Angst sowie in mehrdimensionale und bereichsspezifische Verfahren gegliedert werden. Diese Instrumente basieren jeweils mehr oder minder eng auf verschiedenen Theorien der Angst.

Die ersten allgemeinen Trait-Angst-Skalen wurden aus den Items des MMPI's entwickelt. Dazu zählen die *Skala zur Erfassung Manifester Angst* (MAS, Lück & Timaeus, 1969), welche auf einer triebtheoretischen Angstvorstellung von Taylor (1953) beruht, und damals einen Boom in der Angstforschung auslöste, die *Welsh Scale* (Welsh, 1952) und die *Finney Scales* (Finney, 1962). Sie wurden mit Hilfe faktorenanalytischer Untersuchungen konstruiert, sind jedoch nur noch von historischem Wert. In den 70er Jahren wurde das unter Psychometrikern sehr verbreitete State-Trait-Anxiety-Inventory (STAI; Spielberger, Gorsuch & Lushene, 1970) entwickelt. Es initiierte eine Vielzahl von empirischen Forschungsarbeiten, welche die strukturelle Unterscheidung einer Zustands- und Eigenschafts-Angst weitestgehend belegen. Auch wenn das STAI den Gipfel der Popularität überschritten zu haben scheint und trotz einer Reihe von gerechtfertigten Kritikpunkten an diesem Instrument (Begrenzung auf die Erfassung von Bewertungsängstlichkeit, umstrittener Gebrauch angstnegativer Items, Kritik, dass ein „Amalgam" aus Angst, negativer Affektivität und Depression gemessen werde; Krohne, 1996, S. 31), ist es der MAS-Skala als Omnibusverfahren überlegen.

Zwei weitere faktorenanalytische Verfahren zur Messung einer allgemeinen Trait-Angst sind die „Institute for Personality and Ability Testing Anxiety Scale" (IPAT; Cattell & Scheier, 1963) sowie die Objektive Testbatterie (OA-TB-75; Häcker, Schmidt, Schwenkmezger & Lutz, 1975).

Desweiteren werden *mehrdimensionale* Trait-Angst-Tests - auch stimulus-orientierte Methoden genannt - unterschieden. Diese erfassen habituelle Angstbereitschaften in Bezug auf verschiedene potentiell angstäuslösende Situationen oder Objekte (siehe Kapitel 2.3.2.). Die wohl bekanntesten drei sind die Endler Multidimensional Anxiety Scale (EMAS; Endler, Edwards & Vitelli, 1991), welche Angst vor sozialer Bewertung, physischer Gefahr, mehrdeutigen Situationen und Alltäglichem getrennt erfasst, der Interaktions-Angst-Fragebogen (IAF; Becker, 1997), der auf gleichnamigen Theorienansatz basiert, und eine Angsterfassung in Bezug auf acht unterschiedliche situative Bedingungen ermöglicht (vor physischer Verletzung, öffentlichen Auftritten, Selbstbehauptungs- und Abwertungssituationen, physischen und psychischen Angriffen sowie Bewährungssituationen) und das S-R-Inventar zur Erfassung von Angst (Walter, Leifert & Linster, 1975), welches der Messung der Angstbereitschaft in Abhängigkeit von elf angstauslösenden Situationen dienen soll.

Schließlich existieren eine Reihe von *bereichsspezifischen* Trait-Angst-Tests, welche Angst ausschließlich in Bezug auf einzelne Situationen / Objekte erheben. Eine Vielzahl solcher Tests sind zur Erfassung *sozialer Angst* (z. B. der SPAIK; Melfin, Florin & Warnke, 2001), *physischer / Verletzungsangst* (z. B. die Geburtsangstskala, GAS; Ettrich, Krauss & Sandau, 1992; der Fragebogen zur Erfassung der Angst vor einem Herzinfarkt, AF-HI; Mrazek, 1985), *Sportängstlichkeit* (z. B. Bilder-Angst-Test für Bewegungs-situationen, BAT; Bös & Mechling, 1985) und *Prüfungs*ängstlichkeit (z. B. Test Anxiety Scale, TAS; Sarason, 1978) entwickelt worden. Bereichsspezifische Angst-Tests sind globalen Trait-Angst-Skalen dann vorzuziehen, wenn sie die interessierenden Situationen / Objekte hinreichend erfragen. Allgemeine Trait-Angst-Skalen sind vor allem dann günstiger, wenn die Erfassung der Angst in selbstwertrelevanten Situationen intendiert ist, und keine bereichsspezifischen Angst-Skalen vorliegen, die zu der zu erhebenden Situation „passen“ (Laux & Glanzmann, 1996, S. 119). Spezifische Trait-Angst-Tests für Kinder liegen ebenfalls vor (z. B. Kinder-Angst-Test, KAT-II; Thurner & Tewes, 2000).

2.7.3.3. State-Angst-Verfahren

Um eine möglichst einfache State-Angsterfassung in Laboruntersuchungen oder bei Feldbeobachtungen zu ermöglichen, wurden zu Beginn der Angstforschung sogenannte „Ein-Item-Skalen“ entwickelt, welche verbal mittels eines vertikalen „Furcht-Thermometers“ (Walk, 1958)[19] oder nonverbal mittels „Fingerspannen-Skalierung“ (Birbaumer, Tunner, Hölzl & Mittelstaedt, 1973) eine ereignissimultane Angsterfassung ermöglichen.

Umfangreicher sind die in der Forschung zur Erfassung der State-Angst beliebten Eigenschaftswörterlisten wie z. B. die von Janke und Debus (1978), welche Angst (neben weiteren Befindlichkeitsvariablen) in Form von (antithetischen) Adjektivlisten, erhebt. Im klinischen Bereich werden Symptomchecklisten bevorzugt, welche der Erfassung pathologischer Angstausprägungen (neben anderen Symptomen) dienen können (z. B. Symptom-Checkliste von Derogatis, SCL-90-R; Franke, 1995). Einen systematischen Überblick über Selbst- und Fremdbeurteilungsverfahren zur Erfassung der Angst im klinischen Bereich, welche zur Diagnostik der Angst als Störung in enger Anlehnung an die Klassifikationssysteme ICD-10 und DSM-IV dienen, geben Margraf und Bandelow (1997).

Zu den klinischen Tests, welche ausschließlich zur Angstdiagnostik genutzt werden, zählen unter anderem das State Trait Anxiety Inventory (Laux et al., 1981), das Beck-Angst-Inventar (BAI; Margraf & Ehlers, 1995) und die Hospital Anxiety and Depression Scale (HADS; Hermann, Buss & Snaith, 1995; siehe Kapitel 6.3.). Die klassische State-Angst-Skala des STAIs (Laux et al., 1981) umfasst eine Liste von Zustandsbeschreibungen („ich bin...“, „ich fühle mich...“), welche zeitlich fluktuierende Angstzustände erfassen sollen. Sie weist eine hohe Änderungssensitivität und eine hohe interne Konsistenz auf, wurde jedoch suboptimal konstruiert (Querschnitts- statt Längsschnittdesign, hohe interne Konsistenz führt zur Maximierung interindividueller Unterschiede auf Kosten der Messung intraindividueller Veränderungen).

2.7.3.4. Unidimensionale versus multidimensionale Angstmessung

Inwiefern ist es gerechtfertigt zwischen einer State-Angst und einer Trait-Angstmessung zu unterscheiden? Infolge der Entwicklung des STAIs wurde mehrfach die faktorielle Differenzierung zwischen einer State- und einer Trait-

[19] Zitiert nach Krohne (1996).

Angst empirisch belegt (Steyer, Schmidt & Eid, 1999). Jedoch existiert konzeptionell wie empirisch ein enger Zusammenhang zwischen diesen beiden Konstrukten der Angst. Spielberger (1972) formuliert den Zusammenhang folgendermaßen: je stärker die Trait-Ausprägung, desto wahrscheinlicher wird ein Individuum den emotionalen Zustand, der zu dem Trait passt, erfahren. Wenn also eine hohe Trait-Ausprägung als eine hohe Wahrscheinlichkeit einer hohen State-Ausprägung definiert wird, so erscheint es nach Uhlenhuth (1985, S. 676) möglich, aus der Berechnung des Mittelwertes wiederholter State-Messungen die Trait-Ausprägung abzuleiten (zum State-Trait-Kontinuum siehe Kapitel 2.4.1.1.). Die Möglichkeit einer solchen „indirekten“ Trait-Angstmessung wirft dann die Frage auf, inwiefern eine separate Trait-Angstmessung bei Verlaufstestungen der State-Angst überhaupt noch gerechtfertigt ist. Usala und Hertzog (1991) begründen die Notwendigkeit einer eigenständigen Erhebung der Trait-Angst mit der Retest-Reliabilität. Sie fanden, dass Trait-Angst-Skalen eine höhere Stabilität (r = 0,9) als State-Angst-Aggregate (r = 0,72) und State-Angst-Skalen (r = 0,66) aufweisen. Ist jedoch aufgrund einer höheren Retest-Reliabilität eine Unterscheidung dieser beiden Konstrukte der Angst sinnvoll? Wie eng ist denn der Zusammenhang zwischen der State- und der Trait-Angst? Endler, Magnusson, Ekehammar und Okada (1976) untersuchten den statistischen Zusammenhang von State- und Trait-Angst-Skalen und zeigten in faktorenanalytischen Studien mit dem STAI, dass diese beiden Skalen höher miteinander korrelierten als verschiedene State-Angst-Skalen untereinander. Auch im Testhandbuch des STAIs werden Korrelationen zwischen den beiden Skalen von r = 0,56 bis r = 0,75 berichtet.

Sprechen diese Ergebnisse nicht doch für ein allgemeines State-Trait-Angst-Kontinuum? Im Zuge der Forderung nach verstärkt idiografischer Angstforschung konstatiert Tunner (1978), dass eine „allgemeingültige Angstdimension von universeller Gültigkeit für alle Individuen heute nicht mehr unterstellt werden kann“ (S. 209). Sollte es eine allgemeingültige State-Trait-Angstdimension nicht geben, und ist die Sinnhaftigkeit der Trennung zwischen einer State- und einer Trait-Angst umstritten, so stellt sich die Frage nach anderen und gegebenenfalls besseren Differenzierungen unterschiedlicher Aspekte des Angsterlebens. Seit Liebert und Morris (1967), welche erstmals das Angsterleben in zwei Komponenten unterteilten: in die Aufgeregtheit

(„*emotionality*“), d. h. das subjektive Empfinden der Angst und ihrer einhergehenden Wahrnehmung körperlicher Erregung, sowie die Besorgnis („*worry*“), d. h. die unter Bedrohung auftretenden Gedanken (Sorgen, Zweifel, Misserfolgserwartungen, negative Selbstbewertungen), haben sich Angstforscher mit möglichen Differenzierungen des Angsterlebens befasst. Vor allem auf der Basis faktorenanalytischer Studien, wurden unterschiedliche Aspekte der Angst voneinander unterschieden.

Tabelle 3: Verschiedene faktorenanalytische Studien zur Differenzierung des Angst-Konstrukts.

Autoren	Jahr	Inventar	Faktoren
Mandler & Sarason	1952	Test Anxiety Qestionnaire (TAQ)	1. Zuversicht, 2. Autonome Reaktionen, 3. Vermeidungstendenzen.
Endler, Hunt & Rosenstein	1962	S-R-Inventory[20]	1. Angstgefühle, 2. Vegetative Reaktionen, 3. Muskelspannung.
Fenz & Epstein	1965	Manifest Anxiety Scale (MAS)	1. Angstgefühle, 2. Autonome Übererregbarkeit, 3. Symptome der Anspannung der Muskulatur.
Liebert & Morris	1967	Worry-Emotionality-Questionnaire (WEQ)	1. Emotionalität („emotionality”), 2. Besorgnis („worry“)
Lushene	1970	WEQ	1. Autonome, 2. Kognitive, 3. Motorische Komponenten.
Newmark, Faschingbauer, Finch & Kendall[21]	1979	STAI, MMPI	1. Adjustment, 2. Passivity, 3. Somatic concern, 4. Anxiety proneness.
Sedlmayer	1980	Unklar	1. Emotional-kognitive, 2. Physiologische, 3. Motorische Komponenten.
Sarason	1984	Test Anxiety Inventory (TAI)	1. Wahrnehmung körperlicher Reaktionen, 2. Besorgtheit, 3. Aufgabenirrelevante Kognitionen, 4. Anspannung.
Rost & Schermer	1987	TAI	1. Wahrnehmung körperlicher Reaktionen, 2. Selbstwertbedrohliche Kognitionen.
Krohne & Hindel	1990	Sportlicher Wettkampf	1. Emotionale Anspannung, 2. Selbstzweifel, 3. Hilflosigkeit.
Endler, Edwards & Vitelli	1991	Endler Multidimensional Anxiety Scale (EMAS)	1. Autonome Aufgeregtheit, 2. Kognitive Besorgnis.
Hodapp	1991	TAI	1. Aufgeregtheit, 2. Besorgtheit, 3. Kognitive Interferenz, 4. Mangel an Zuversicht.
Slangen, Kleemann & Krohne	1993	Operative Angst	1. Affektive, 2. Kognitive, 3. Vegetative Symptome.

[20] Stimulus-Response Inventory of Anxiousness (Walter, Leifert & Linster, 1975).
[21] Zitiert nach Krohne (1996).

Tabelle 3 fasst die unterschiedlichen Bemühungen um Differenzierung verschiedener Komponenten der Angst in unterschiedlichen Bereichen (allgemeine Ängstlichkeit, Testangst, sportbezogene und operative Angst) von 13 Forschern bzw. Forschergruppen seit 1952 zusammen.

Es fällt auf, dass die verschiedenen Forscher aufgrund faktorenanalytischer Studien zu einer Reihe von Vorschlägen gelangen, die sich teilweise überschneiden, jedoch bisher noch keine einheitliche theoretische Konzeption verschiedener Angstkomponenten gefunden werden konnte. Dies mag im Umstand der Methodik (Faktorenanalyse) begründet sein, welche oft beliebige, instabile, faktorielle Strukturen offenbart, die im nachhinein von den einzelnen Forschern mit „Inhalt“ gefüllt werden müssen. Desweiteren könnte es unterschiedliche Angstkomponenten für unterschiedliche Angstbereiche (allgemeine Angst, Prüfungs- oder Sportangst) geben.

Am häufigsten - da wahrscheinlich am augenscheinlichsten - wird die körperliche Symptomebene als eigener Faktor benannt (vegetativ, physiologisch, autonom: 10 Nennungen), gefolgt von emotionalen (Emotion / Gefühl / affektiv: 7 Nennungen), kognitiven (Kognitionen / kognitive Interferenz / aufgabenirrelevante Kognitionen: 7 Nennungen) und motorischen (Muskelspannung: 5 Nennungen) Faktoren bzw. Komponenten der Angst. Schließlich halten drei Autoren den Mangel an Zuversicht / Selbstzweifel bzw. Hilflosigkeit für eine separate Angstkomponente; Andere ergänzen ihre Konzeptionen um Faktoren der behavioralen Ebene (Passivität, Vermeidung, „Adjustment“). Obgleich die faktorenanalytischen Bemühungen um eine Strukturierung der Angst erstrebenswert erscheinen, konnte bisher selbst für die historisch früheste vorgeschlagene grundlegende Differenzierung zwischen einer emotionalen und einer kognitiven Komponente der Angst (Liebert & Morris, 1967) keine eindeutige empirische Trennung im Sinne einer statistischen Unabhängigkeit der Komponenten belegt werden. So schreibt Krohne (1996), dass von vornherein für die beiden Komponenten „kein voneinander unabhängiges Variieren angenommen werden“ (S. 32) könne. Korrelationen zwischen den beiden Komponenten erfasst durch die eigens für diese konzeptuelle Trennung entwickelten Inventare WEQ und TAI liegen zwischen $r = 0,4$ und $r = 0,65$ (WEQ; Morris et al., 1970, 1981, 1983) bzw. $r = 0,5$ und $r = 0,8$ (TAI; Krohne, 1996, S. 66). Benson und Mitarbeiter (1992) vermuten,

dass letztere Korrelationen aufgrund von Messfehlern, die jeder Messung manifester Variablen anhaften, unterschätzt sind, und führten Analysen mit latenten (messfehlerfreien) Variablen durch, die zu einer „Bereinigung“ der Korrelation (r = 0,82 / 0,92) führten. Krohne (1996) folgert, dass bei so hohen Korrelationen „nicht ernsthaft von einer gelungenen Differenzierung zweier Komponenten gesprochen werden“ (S. 66) kann. Drei mögliche Gründe für diesen mangelnden Fortschritt um strukturelle Differenzierung der Angstkomponenten werden vermutet. Erstens sei die Zuordnung der einzelnen Items zu den beiden Komponenten uneindeutig (siehe Tabelle 4).

Tabelle 4: Die Zuordnung der Items des WEQ zur Emotionalitäts (E)- bzw. Besorgnis (B)-Skala (Morris, Davis & Hutchings, 1981).

Itemtext	Skala
Das Herz schlägt mir bis zum Hals.	E
Ich bin bekümmert.	B
Ich bin so angespannt, dass mir fast schlecht ist.	E
Ich habe Angst, dass ich für die Prüfung nicht genug gelernt habe.	B
Ich habe ein beklemmendes Gefühl.	E
Ich glaube, dass andere über mich enttäuscht sein werden.	B
Ich bin aufgeregt.	E
Ich glaube, das ich in der Prüfung nicht das leiste, was ich eigentlich leisten könnte.	B
Ich bin übernervös.	E
Ich glaube nicht, dass ich in dieser Prüfung besonders gut abschneiden werde.	B

Zweitens sei in der Testkonstruktion bereits ein „Fehlschlag“ dadurch angelegt, dass insbesondere Items mit einer hohen Trennschärfe, d. h. einer hohen Korrelation mit *einem* Gesamtscore (Emotion und Kognition) zur Testkonstruktion ausgewählt wurden, was zu einer unnötigen, ja im oben ausgeführten Sinne sogar kontraproduktiven Homogenisierung der Gesamtskala geführt habe, und drittens existierten komplexe Auslösungs- und Rückmeldungsbeziehungen zwischen den verschiedenen Manifestationen der Angst, welche eine Differenzierung derselben erschwerten, wenn nicht sogar verhinderten (Krohne, 1996). Um die auch von anderen Autoren verschiedener theoretischer Richtungen vermutete enge Beziehung zwischen verschiedenen Ebenen des Angsterlebens zu verdeutlichen, sei an dieser Stelle das „Teufelskreismodell der Angst“ von Margraf (2000) angeführt (Abbildung 3).

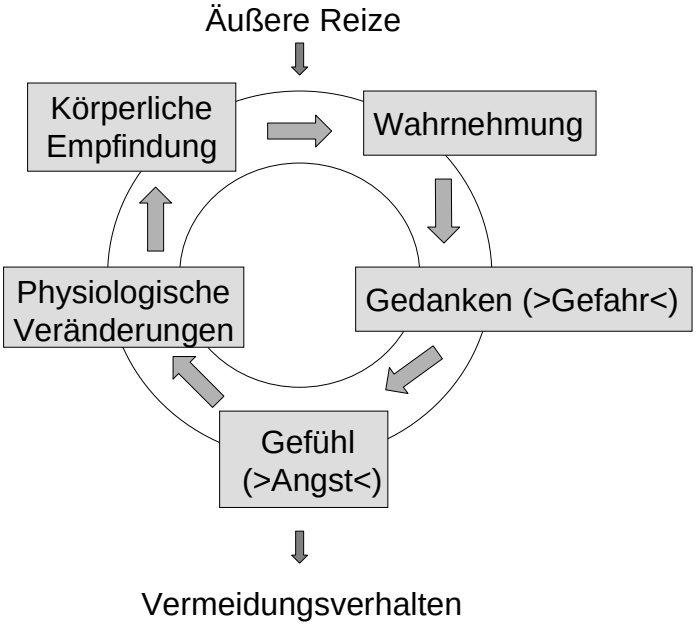


Abbildung 3: Teufelskreismodell der Angst (Margraf, 2000) zur Verdeutlichung des Zusammenhangs verschiedener Aspekte des Angsterlebens.

3. Die Item Response Theorie

3.1. Einleitung

Psychologische Tests verfolgen das Ziel, mit wissenschaftlichen Methoden quantitative Aussage über den relativen Grad der individuellen Ausprägung eines Merkmals (hier z. B. Angst) zu treffen (Lienert & Raatz, 1994). Um eine solche Aussage fundiert zu ermöglichen, basieren psychologische Tests auf einer *Testtheorie*. Sie beschreibt den Zusammenhang zwischen dem zu erfassenden Merkmal und dem Testverhalten (Rost, 1996). Zwei Testtheorien können unterschieden werden:

1. die Klassische Test-Theorie (KTT) und
2. die Item Response Theorie (IRT).

Die KTT ist die ältere Testtheorie, deren jahrzehntelange Tradition bis zum Anfang des letzten Jahrhunderts zurückreicht (Spearman, 1907) und seit dem Testboom in den 30er Jahren als Standard in der Testentwicklung galt und heute noch gilt. Zu den namhaften frühen Vertretern der KTT zählen Gulliksen (1950) und Novick (1966). Letzterer gab der KTT die grundlegende axiomatische Endform (siehe Kapitel 3.2.1.).

Die Wurzeln der IRT liegen bei Rasch (1960) und Birnbaum (1968), welche erstmals mathematische, stochastische Modelle in die psychologische Forschung einführten. In einem wegbereitenden Textbuch von Lord und Novick (1968), in denen Rasch ein und Birnbaum vier Kapitel publizierten[22], wurde die IRT, welche seither auch den Namen „probabilistische“ Testtheorie trägt (Rost & Spada, 1982), Ende der 60er Jahre einem breiten Fachpublikum zugänglich gemacht. Zur Rezeption der Geschichte der IRT, welche durch zwei Entwicklungslinien (eine US-amerikanische um Lord & Novick, 1968, und eine Europäische um Rasch, 1960) gekennzeichnet ist, werden Embretson und Reise (2000) empfohlen.

Lange Zeit glaubte man, dass aufgrund der zahlreichen Potentiale der IRT, welche einige im Rahmen der KTT aufgeworfenen messtheoretischen Probleme zu lösen verspricht (siehe Kapitel 3.3.3.), die jüngere / modernere Testtheorie (IRT) die ältere Testtheorie (KTT) ablöst.

[22] Textbuch von Lord & Novick (1968): Kapitel 17-20 von Birnbaum; Kapitel 21 von Rasch.

Eine Abkehr von der KTT fand jedoch nicht in dem Maße wie von vielen erwartet statt. Obgleich seit Beginn der Entstehung der IRT das Interesse an ihrer Anwendung im wissenschaftlichen Forschungskontext wuchs und seither unvermindert floriert (siehe Kapitel 3.5.), setzte sich dieser Trend - abgesehen von einigen umfangreichen Testprogrammen größerer Institutionen (wie z. B. des ETS, 1996, oder der Bundeswehr, Hornke, Küppers & Etzel, 2000; siehe Kapitel 3.5.1.) - nicht im Alltag der Testpraxis durch, d. h. die meisten der üblicherweise im klinischen Alltag eingesetzten Testverfahren, welche über Testverlage erhältlich sind, sind KTT-basiert entwickelte Instrumente.

Gründe für dieses „Schattendasein“ der IRT knapp ein halbes Jahrzehnt nach ihrer Entstehung versucht Rost (1999) zu eruieren. Sie liegen wahrscheinlich in der ungünstigerweise entstandenen polarisierenden Konkurrenzsituation der beiden Testtheorien zueinander. In diesem Konkurrenzverhältnis offenbarte sich im Rahmen von Forschungsarbeiten bereits früh, dass sich die Anwendung der IRT - trotz ihrer vielen messtheoretischen Vorteile (siehe Kapitel 3.3.3.) - bei der realen Testkonstruktion schwieriger gestaltet als die Anwendung von Methoden der KTT (mögliche Gründe: Komplexität der IRT-Modelle, benutzer*un*freundliche IRT-Software etc.; zu den Nachteilen der IRT siehe Kapitel 3.3.4.). Weiterhin zeigte sich in einer Reihe von wissenschaftlichen Studien in den 70er Jahren vielfach eine mangelnde Datenanpassung der IRT-Modelle an klinisch-psychologische Daten (mündliche Mitteilung von Prof. Dr. Westmeyer). Als Konsequenz werden seither IRT-Konzepte und Methoden bei der Entwicklung der im Testalltag gängigen Instrumente, welche von Testverlagen vertrieben werden, vernachlässigt.

Im Gegensatz zum Alltag der Testpraxis erfuhr die IRT jedoch im *wissenschaftlichen* Forschungskontext seit ihrer Entstehung großes Interesse (siehe Kapitel 3.5.). Die anfängliche Wahrnehmung einer polarisierenden Konkurrenzsituation der beiden Testtheorien zueinander weicht hier langsam der Vorstellung, die beiden Testtheorien als komplementär zueinander zu betrachten. Rost (1999) zum Beispiel argumentiert durch das Aufzeigen messtheoretischer Brückenschläge zwischen den Theorien, dass eine die Testtheorien kontrastierende, polarisierende Darstellung messtheoretisch nicht gerechtfertigt sei. Im Einklang mit Embretson und Hershberger (1997) hält er eine Integration beider Testtheorien für wünschenswert (Rost, 1996).

Die formale Umsetzung einer solchen Integration der Testtheorien findet sich bereits bei Steyer und Eid (1993); ein Beispiel für den Versuch einer konzeptionellen und anwendungsbezogenen Kombination beider Testtheorien geben Verstralen, Bechger und Maris (2001).

Im Folgenden werden zunächst die Grundzüge der KTT mitsamt ihren messtheoretischen Unzulänglichkeiten erörtert, um auf dieser Grundlage ein besseres Verständnis für die Unterschiede und Möglichkeiten der IRT zu entwickeln.

3.2. Die Klassische Test-Theorie (KTT)

Die KTT bietet „ein Arsenal pragmatisch orientierter Prinzipien oder Regeln zur Konstruktion, Erprobung und Evaluation psychometrischer Tests und zur Interpretation von Testergebnissen“ (Stumpf, 1996, S. 411). Im engeren Sinn ist sie eine „*Messfehlertheorie*“ (Rost, 1999), auf deren Grundlage sich Messinstrumente auf der Ebene der *Tests* – die IRT bietet Methoden zur *item*bezogenen Analyse – analysieren und bewerten lassen (Hambleton, Swaminathan & Rogers, 1991).

Erstmals wurde die KTT, deren theoretische Grundlagen im Beginn des letzten Jahrhunderts (Spearman, 1904) liegen, von Gulliksen (1950) zusammenfassend dargestellt, und in Form von rein formallogisch gesetzten Annahmen systematisch entwickelt und ausgebaut. Spätere Arbeiten von Novick (1966) und Zimmermann (1975) zeigen, dass die KTT auch von schwächeren Annahmen als den von Gulliksen (1950) Konstatierten abgeleitet werden kann.

Obgleich die KTT im Gegensatz zur IRT kein empirisch überprüfbares mathematisches Modell darstellt (Embretson & Reise, 2000), ist sie der älteste und bis heute am weitesten verbreitete Ansatz innerhalb der Psychometrie, dem eine lange Tradition an Konstruktionen von Messinstrumente, die gute Reliabilitäten aufweisen und sich pragmatisch bewährt haben, zu verdanken ist.

3.2.1. Axiome der KTT

Die KTT trifft keine Aussagen über ein latentes Merkmal wie die IRT (Rost & Spada, 1982), sondern bietet ein Set von Axiomen, welches die Beziehungen *zwischen* und die messtheoretischen Charakteristika *von* einem beobachteten Messwert (Testverhalten = „x“), einem wahren Wert („w“) und einem Fehlerwert (error = „e“) einer Person j in einem Test t festlegt. Dieses Set von Axiomen, stellt die Grundlage der Reliabilitätstheorie in der KTT dar.

Die wichtigsten Axiome der KTT sind:

1. $x_{tj} = w_{tj} + e_{tj}$,
2. $\sum_{j=1}^{\infty} (e_{tj}) = 0$; $r(e_{tj}, w_{tj}) = 0$; $r(e_{tj}, w_{uj}) = 0$; $r(e_{ti}, e_{uj}) = 0$,
3. x_{tj}, w_{tj} und e_{tj} sind normalverteilt.

Die Postulate definieren, dass (1.) sich jeder beobachtete Wert x_{tj} einer Person j in Test t additiv aus einem wahren Wert w_{tj} und einem Fehlerwert e_{tj} zusammensetzt, (2.) der Fehlerwert e_{tj} eine Zufallsvariable mit einem Erwartungswert ($\sum$) von 0 ist und unabhängig vom wahren Wert eines Tests (w_{tj}) oder eines anderen Tests u (w_{uj}), sowie vom Fehlerwert eines anderen Tests (e_{uj}) ist (Kranz, 1979; Steyer & Eid, 1993) und es wird angenommen, dass (3.) der beobachtete Wert x_{tj}, der wahre Wert w_{tj} und der Fehlerwert e_{tj} normal verteilt sind. Sind die aufgeführten Axiome realisiert, und setzt man voraus, dass die zu messende Variable in der Messsituation einen konstanten Wert besitzt, so ist es möglich, den wahren Wert w durch Messwiederholungen zu approximieren (Lehmann, 1983; Kristof, 1983). Eine indirekte Annäherung an den wahren Wert w ist somit durch eine unendliche Anzahl von Messungen, welche entweder in Form wiederholter Messungen an ein und derselben Testperson oder einer einmaligen Messung an vielen Testpersonen realisiert werden kann, möglich (Amelang & Zielinski, 1996). Problematisch ist hier jedoch die Realisierung einer Messsituation mit einer *konstanten* Variable, da besonders im psychologischen Bereich unter Einwirkung der Messung und erst recht der Messwiederholung eine *Variation* der zu messenden Variablen zu erwarten ist. Auf der Grundlage oben genannter Axiome, werden im Rahmen der KTT weitere für die Messung zentrale theoretische Ableitungen (Theoreme) formuliert, welche die Zerlegung der Varianz eines Testwertes (s_{xt}^2; siehe 4.) und die Berechnung der (Retest-) Reliabilität (r_{tt}, siehe 5.) behandeln, woraus sich der Standardmessfehler (s_{et}; siehe 6.) herleiten lässt.

4. $s_{xt}^2 = s_{wt}^2 + s_{et}^2$,
5. $Rel_{tt} = \frac{s_{wt}^2}{s_{xt}^2}$,
6. $s_{et} = s_{xt} * \sqrt{1 - r_{tt}}$.

Die Erfassung der Reliabilität in Form einer wiederholten Messung (Retest-Reliabiliät, siehe 5.) ist ein pragmatischer Versuch der Realisierung des idealen theoretischen Konzepts „paralleler Messungen“. Dieses in der KTT wichtige Konzept, welches jede Art der Reliabilitätsmessung begründet, ist wie folgt definiert: Eine parallele Messung ist gegeben, wenn bei zwei Messungen x und x’ angenommen werden kann, dass sie die gleichen wahren Werte (w; siehe 7.) und die gleichen Messfehlervarianzen (s_e^2; siehe 8.) aufweisen (Novick, 1966). Die Reliabilität (Rel_x) kann dann durch die Korrelation der beiden Messungen bestimmt werden (siehe 9.).

7. $w_x = w_{x'}$,
8. $s_{ex}^2 = s_{ex'}^2$,
9. $Rel_x = r(x, x')$.

Problematisch ist hier jedoch, dass sich parallele Messungen in der Realität nur schwer realisieren lassen. Für eine umfassende Darstellung der KTT sei Steyer und Eid (1993) empfohlen.

3.2.2. Grenzen der KTT

Die Schwächen der KTT sind seit den 70er Jahren allgemein bekannt (Lumsden, 1976; Fischer, 1983; Kristof, 1983). Die Wichtigsten dieser können - ohne Anspruch auf Vollständigkeit - wie folgt zusammengefasst werden (Embretson & Reise, 2000):

1. die Axiome der KTT sind empirisch *nicht* überprüfbar,
2. das postulierte Skalenniveau (ISK)[23] ist fragwürdig,
3. die KTT-basiert berechenbaren Item-, Test- und Personenstatistiken sind stichproben*abhängig*,
4. die Annahme der Gleichheit des Messfehlers über alle Merkmals-ausprägungen ist empirisch *nicht begründet*,
5. die Reliabilität ist abhängig von der Test*länge*,
6. die Annahme der intraindividuellen *Invarianz der wahren Werte* ist nur bedingt vertretbar (Amelang & Zielinski, 1996, S. 61)[24] und
7. die *norm*bezogene Interpretation der Testwerte ist inhaltlich wenig aussagekräftig.

[23] ISK: Intervallskalenniveau.
[24] Die Annahme einer intraindividuellen Invarianz der wahren Werte einer Person erscheint nur bezüglich kurzer Zeiträume und nur für bestimmte Merkmalsbereiche vertretbar.

Eine der bedeutsamsten Unzulänglichkeit der KTT liegt wohl in der Stichproben*abhängigkeit* (Punkt 3) der auf ihrer Grundlage berechenbaren

(a) Item- bzw. Teststatistiken und

(b) Testwerte von Personen.

Sowohl die Schwierigkeit und die Trennschärfe von *Items*, als auch die interne Konsistenz, der Standardmessfehler, die Reliabilität und die Validität von *Tests* hängen von der jeweils untersuchten Personenstichprobe ab (Embretson, 1996; Embretson & Hershberger, 1997; Embretson & Reise, 2000; Hambleton et al., 1991; Hambleton & Slater, 1997; Suen, 1990). Dies ist ungünstig, weil die an einer Basisstichprobe errechneten Item- und Teststatistiken somit nicht ohne weiteres auf andere Stichproben übertragbar sind. Eine Generalisierung ist strenggenommen nur erlaubt, wenn parallele Messungen angenommen werden, und die Merkmalsausprägung in der Population normalverteilt ist. Beides ist so meistens nicht voraussetzbar.

Die Abhängigkeit des individuellen *Testwerts* von dem jeweils beantworteten Set von Items ist aus psychometrischer Sicht nicht erwünscht, da ein Messergebnis über eine spezifische Testsituation hinausgehende generalisierbare Schlussfolgerungen über eine Merkmalsausprägung einer Person erlauben sollte. So können Testwerte aus unterschiedlichen Tests, welche die Erfassung des gleichen Konstrukts intendieren, in der Regel nicht direkt miteinander verglichen werden (Ausnahme: parallele Messungen), da den Testwerten keine testübergreifende gemeinsame Skalierung zugrunde liegt.

Die Interpretation von KTT-basierten Testwerten erfolgt über komparative Aussagen zu anderen Messwerten, d. h. zumeist werden Testwerte *normbezogen* interpretiert (Punkt 7). Eine normbezogene Interpretation sagt jedoch wenig über die inhaltliche Bedeutung des Merkmalsausprägungsgrades aus, da die Testwerte nicht in direktem Bezug zu den Iteminhalten gesetzt werden (wie bei der IRT, siehe Kapitel 3.3.3.).

Weiterhin ist hervorzuheben, dass die in der KTT formulierte Annahme, dass der *Standardmessfehler* über alle Merkmalsausprägungen hinweg *konstant* ist, nicht der empirischen Realität entspricht (Punkt 4). Vielmehr besteht eine nicht-lineare Beziehung zwischen der Merkmalsausprägung von Personen und dem Standardmessfehler in der Form, dass dieser im mittleren Merkmals-

ausprägungsbereich am geringsten ausfällt und zu den extremen Ausprägungsbereichen hin zunimmt (Embretson & Reise, 2000).

Zudem sind in der KTT mit dem Konzept der *Reliabilität* einige methodische Schwierigkeiten verknüpft. *Parallele Messungen*, deren Realisierung in der KTT theoretisch idealerweise zur Erfassung der Reliabilität angestrebt werden (siehe Kapitel 3.2.1.), sind in Reinform in der Praxis nicht herstellbar. Desweiteren hängt die Reliabilität in der KTT von der Testlänge ab (Punkt 5), was eine Korrektur (mittels der Spearman Formel) notwendig macht.

Zusammenfassend lässt sich resümieren, dass die KTT eine Reihe von Grundannahmen postuliert, welche theoretisch wie empirisch nicht begründet und unangemessen sind. Es werden messtheoretische Probleme aufgeworfen, deren Lösungsversuche im Rahmen der KTT als nicht ideal bewertet werden müssen.

3.3. Die Item Response Theorie (IRT)

Die IRT wird häufig als „moderne“ Testtheorie bezeichnet, da sie sich vor allem in den letzten beiden Jahrzehnten bei der Konstruktion und Evaluation von psychometrischen Tests (v.a. in der Leistungsdiagnostik) als nützlich erwiesen hat (Hambleton et al., 1991). Ein zentraler Vorteil der IRT liegt in der Möglichkeit Computergestützte Adaptive Tests entwickeln zu können (CAT; siehe Kapitel 4.3). Weiterhin verspricht die IRT eine Reihe von Messproblemen, welche bei der Anwendung der KTT aufgetreten sind (siehe Kapitel 3.2.2.), zu lösen.

Genaugenommen ist die IRT nicht eine *einzelne* Theorie, sondern umfasst eine *Familie von* formalen, mathematischen, probabilistischen Mess*modellen*, welche postulieren, dass dem beobachtbaren Testverhalten (*manifeste V*ariable) eine Fähigkeit / Eigenschaft bzw. Disposition (*latente* Variable) zugrunde liegt, die das Testverhalten „steuert“ (Rost & Spada, 1982, S. 60). Während die Messung in der KTT als eine *direkte* Messung zu verstehen ist, konzipiert die IRT die Messung als *indirekt*. Das beobachtbare Verhalten stellt also lediglich einen *Indikator* für ein - in IRT Begrifflichkeiten ausgedrückt - *latentes Trait* dar, auf dessen Ausprägung es zu schließen gilt (Müller, 1999). Die IRT beinhaltet theoretisch wie empirisch gerechtfertigtere Messprinzipien als die KTT (Embretson & Reise, 2000), welche indirekt empirisch überprüfbar sind (Rost, 1999). Somit sind IRT-Modelle im Gegensatz zur KTT prinzipiell

falsifizierbar (Hambleton et al., 1991), da eine Reihe von Annahmen über die Daten expliziert werden, welche auf einen Datensatz zutreffen können, d. h. eine modellbasierte Vorhersage des Testverhaltens erlauben, oder nicht.

3.3.1. Kernannahmen der IRT

Das „Herzstück" der IRT stellt die Modellierung des Itemantwortverhaltens durch eine mathematische non-lineare Funktion, welche *Item Response Function* (IRF) genannt wird (Suen, 1990), dar. Die IRF kann als *Item Response Curve* (IRC) grafisch visualisiert werden.

> (1.) Die IRF bzw. IRC beschreibt die non-lineare Beziehung zwischen der Wahrscheinlichkeit eines manifesten Antwortverhaltens in Abhängigkeit von der Ausprägung einer Person auf dem zugrundeliegenden latenten Trait. (Embretson & Reise, 2000, S. 46f)

Je nach Art des IRT-Modells werden zur besten Modellierung des Antwortverhaltens unterschiedliche Funktionstypen (Normale Ogivenfunktion, logistische Funktion etc.) angenommen. Abbildung 4 (links) zeigt IRCs von zwei dichotomen Items (Rasch-Modell), Abbildung 4 (rechts) veranschaulicht die IRCs eines polytomen Items (Generalized Partial Credit Modell, GPCM; Muraki, 1992; zu den unterschiedlichen Modellen siehe Kapitel 3.4.). Auf der Abzisse ist die Ausprägung des latenten Traits (in z-Werten) und auf der Ordinate die Antwortwahrscheinlichkeit (von 0 bis 1) abgetragen (zu IRCs bei der Itemanalyse siehe Kapitel 5.4.2.1.).

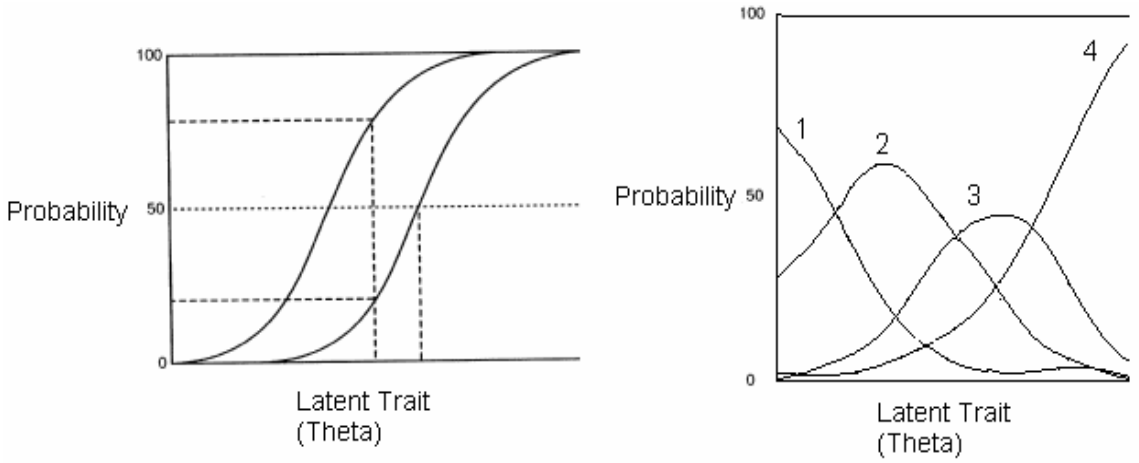


Abbildung 4: Item Response Curves (IRCs). Links: IRCs modelliert mit dem einparametrischen Rasch-Modell. Rechts: IRCs modelliert mit dem zweiparametrischen Generalized Partial Credit Modell (GPCM).

Die IRCs, welche auf der Grundlage der Familie dichotomer Rasch-Modelle (siehe Kapitel 3.4.2.) modelliert werden, unterscheiden sich nicht in ihrem Kurven*verlauf* (logistische Kurven, welche asymptotisch gegen 0 und 1 konvergieren) sondern lediglich in ihrer *Positionierung* auf der Abszisse (> parallele Kurvenverläufe). Abbildung 4 (links) zeigt zwei Items, für welche jeweils nur eine Antwortkategorienkurve (p) abgetragen ist, da die zweite Antwortkategorienkurve (q = 1-p) bei dichotomen Items genau umgekehrt verläuft.

Die IRCs, welche auf der Grundlage polytomer IRT-Modelle modelliert werden - wie hier am Beispiel des GPCMs (siehe Kapitel 3.4.3.) - verlaufen bei Modellkonformität eingipflig und glockenförmig, jedoch nicht unbedingt symmetrisch (siehe Abbildung 4, rechts, IRCs Nr. 2 & 3). Die IRC der ersten Antwortkategorie verhält sich stets stetig monoton fallend (siehe IRC Nr. 1), die IRC der letzten Antwortkategorie stetig monoton steigend (siehe IRC Nr. 4). Abbildung 4 (rechts) zeigt die Antwortkategorienkurvenverläufe für vier Antwortkategorien eines Items. Die Kurvenverläufe unterscheiden sich in der *Positionierung* auf dem latenten Trait und der *Steigung* (innerhalb *und* zwischen Items).

Die IRC kann - wie erwähnt - mittels einer mathematischen Formel (IRF) - beschrieben werden, welche durch Item- und Personenparameter (zu den unterschiedlichen Itemparametern siehe Kapitel 3.4.1.) spezifiziert wird. Der Begriff Parameter deutet daraufhin, dass es sich hier um zunächst unbekannte Kenngrößen handelt, welche es im Rahmen der IRT-basierten Messung zu schätzen gilt (zu den Schätzalgorithmen siehe Kapitel 4.3.3.3. / 4.3.3.4.).[25]

Die Parametrisierung von Itemeigenschaften (Itemparameter) und der Merkmalsausprägung (Personenparameter) in *einer* Gleichung deutet auf die zweite grundlegende Besonderheit der IRT-Modelle hin:

> (2.) Item- und Personenparameter werden auf *einer gemeinsamen Skala* liegend konzipiert. (Hambleton & Slater, 1997, S. 32)

Dies hat vorteilhafte Implikationen für die Interpretation der Personen- und Itemparameter (siehe Kapitel 3.3.3.). Der Personenparameter wird in der IRT

[25] Da die Schätzalgorithmen einen hohen Rechenaufwand erfordern und nur computergestützt realisierbar sind, ist die Darstellung derselben aus Kapitel 3 (IRT) in Kapitel 4 (Computerdiagnostik) ausgelagert worden.

mit dem griechischen Buchstaben „θ“ (= Theta) gekennzeichnet und entspricht dem in der KTT üblichen Summenscore eines Tests. Die Theta-Skala hat per se keinen natürlichen Referenzpunkt (Suen, 1990), sondern wird üblicherweise in z-Werten dargestellt (M = 0; SD = 1). Die Theta-Werte sind wie folgt zu interpretieren: je *größer* die Theta-Werte, desto *stärker* ist das zu messende Merkmal ausgeprägt bzw. desto *schwieriger* ist ein Item und umgekehrt: je *geringer* der Theta-Wert, desto weniger ist das zu messende Merkmal ausgeprägt bzw. desto *leichter* ist ein Item.

Obgleich beide Parameter auf einer gemeinsamen Skala positioniert werden, können sie unabhängig voneinander geschätzt werden („Separierbarkeit von Item- und Personenparametern“; Rasch, 1960). Diese dritte zentrale Charakteristik der IRT wird auch „*Invarianz Eigenschaft*“ genannt:

> (3.) Itemparameter und Personenparameter sind *stichprobenunabhängig.*
> (Hambleton, Swaminathan & Rogers, 1991, S. 18)

Es bedeutet, dass die in der IRT geschätzten Itemstatistiken von der untersuchten Personenstichprobe unabhängig sind, d. h. im Falle, dass die Daten den vom IRT-Modell spezifizierten Annahmen entsprechen, die berechneten Itemstatistiken wie z. B. die Schwierigkeit oder Diskriminationsfähigkeit von einzelnen Items über verschiedene Stichproben von Personen generalisierbar sind.

Umgekehrt hängt die Schätzung der individuellen Merkmalsausprägung Theta *nicht* von dem spezifischen Set dargebotener Items ab. Dies erlaubt die Vergleichbarkeit von Theta-Werten von Personen, denen z. B. im Rahmen eines individuellen Itemselektionsprozesses beim adaptiven Testen unterschiedliche Items zur Beantwortung vorgelegt werden (siehe Kapitel 4.3.3.3.).

Die Eigenschaft der Stichprobenunabhängigkeit von Parameterschätzungen stellt *die* zentrale Voraussetzung für das *adaptive Testen* dar.

Nicht nur Theta-Werte von Personen, welche unterschiedliche Itemsets beantwortet haben, können verglichen werden, da sie auf einer gemeinsamen Skala abgebildet werden, sondern auch ein Vergleich von *individuellen* Standardmessfehlern, welche bei der Erhebung von Personen mit

unterschiedlichen Merkmalsausprägungen eingegangen werden, ist im Rahmen der IRT möglich, da ein weiteres zentrales Messprinzip wie folgt lautet:

> (4.) Der Standardmessfehler variiert in Abhängigkeit von der Ausprägung auf dem latenten Trait θ. (Embretson, 1996, S. 342)

Während bei der praktischen Anwendung der KTT unterstellt wird, dass der Standardmessfehler für einen Gesamttest über alle Merkmalsausprägungen konstant ist, ermöglicht die IRT eine individuelle Erfassung desselben. Dies erlaubt beim adaptiven Testen die Kontrolle des Standardmessfehlers einer Messung und ermöglicht eine konstant hohe Messung über das gesamte Kontinuum der Merkmalsausprägung (zum Stoppkriterium, siehe Kapitel 4.3.3.6.).

Eng verschwistert mit dem Konzept des Standardmessfehlers ist die *Reliabilität*. Die IRT eröffnet Möglichkeiten der Reliabilitätsbestimmung, welche sich von der in der KTT üblichen unterscheiden. Es gilt folgendes:

> (5.a) Die Berechnung der Reliabilität macht keine parallelen Messungen nötig.
>
> (5.b) Die Reliabilität hängt nicht von der Testlänge ab.

Beide Aussagen zur Reliabilität zeigen, dass die IRT hier KTT-spezifische Probleme (Schwierigkeit der Herstellung genuin paralleler Messungen und die Abhängigkeit der Reliabilität von der Testlänge) zu lösen vermag.

An dieser Stelle konnten nur die wichtigsten Grundzüge der IRT vorgestellt werden. Für einen systematischen Überblick der Unterschiede zwischen Messprinzipien der KTT versus der IRT seien Embretson (1996), Embretson und Hershberger (1997) und Embretson und Reise (2000) empfohlen.[26]

3.3.2. Voraussetzungen der IRT

IRT-Modelle unterscheiden sich in ihren jeweils postulierten mathematischen Annahmen (siehe Kapitel 3.4.). Insbesondere das Rasch-Modell impliziert einige spezifische testtheoretische Besonderheiten, welche in Kapitel 3.4.2. separat erläutert werden. Eine zentrale Voraussetzung, welche von *allen* IRT-Modellen gleichermaßen postuliert wird, ist die *lokale stochastische*

[26] Embretson und Reise (2000) bieten den vollständigsten Überblick mit zehn voneinander abgrenzbaren Messregeln. In Embretson und Hershberger (1997) sowie Embretson (1996) fehlen noch einige der Abgrenzungen, welche in dem zuletzt erschienenen Buch publiziert sind.

Unabhängigkeit. Sie wird definiert als die Unabhängigkeit der Antwortwahrscheinlichkeit eines Items von der Antwortwahrscheinlichkeit eines vorangegangenen Items bei konstanter Merkmalsausprägung. Das heißt, die Wahrscheinlichkeit, ein Item richtig zu beantworten, hängt *nicht* davon ab, ob das vorangegangene Item richtig oder falsch beantwortet wurde, wenn die Merkmalsausprägung von Personen gleich ist (Rost & Spada, 1982). Oder anders ausgedrückt, es wird vorausgesetzt, dass das latente Trait der einzige Faktor ist, welcher das Antwortverhalten beeinflusst (Hambleton et al.,, 1991). Methodisch kann dies überprüft werden, indem beispielsweise in einer Faktorenanalyse nach der Herauspartialisierung des dominanten Faktors keine Restkorrelationen zwischen den Items verbleiben. Aus dieser Eigenschaft kann auf die *Homogenität* von Items geschlossen werden (Amelang & Zielinski, 1996). Wobei die Homogenität als die Eigenschaft von Items definiert wird, dieselbe Fähigkeit bzw. dasselbe Merkmal zu erfassen (Rost & Spada, 1982). Die *Unidimensionalität*, ist eng mit diesen beiden Konzepten verwandt. Sie ist gegeben, wenn dem Antwortverhalten nur ein *einziges* latentes Trait zugrunde liegt. Untersucht wird sie meist durch die Suche nach einem dominanten Faktor (mittels Faktorenanalysen, siehe Kapitel 5.3.2.1.; Hambleton et al., 1991). Ist die Forderung der meisten IRT-Modelle nach Unidimensionalität erfüllt, so ist auch die lokale stochastische Unabhängigkeit gegeben. Jedoch kann die lokale stochastische Unabhängigkeit auch erreicht werden, wenn die Daten nicht eindimensional sind (Hambleton et al., 1991, S. 11). Die lokale stochastische Unabhängigkeit und die Homogenität sind notwendige Bedingungen bei der Anwendung jeglicher IRT-Modelle, da sie die zentrale Voraussetzungen für die Stichprobenunabhängigkeitsannahme (siehe Kapitel 3.3.1.) darstellen. Unidimensionalität wird nicht von allen IRT-Modellen verlangt, sondern nur von eindimensional konzipierten Modellen gefordert.

3.3.3. Potentiale der IRT

Die IRT bietet einige psychometrische Vorteile, um eine Reihe von Messproblemen zu lösen. Diese gründen sich auf den in Kapitel 3.3.1. eingeführten messtheoretischen Prinzipien. Die Vorzüge der IRT liegen vor allem in neuen / alternativen bzw. erweiterten Möglichkeiten der statistischen Analyse von Items, die weitreichende Implikationen für die Skalenanalyse, -entwicklung und -bewertung haben. So ist z. B. die *lokale stochastische*

Unabhängigkeit die Voraussetzung für die *Stichprobenunabhängigkeit* der Item- und Personenparameterschätzung, welche wiederum die methodische Grundlage für das *adaptive Testen* darstellt.

Vorteilhaft für das adaptive Testen ist außerdem eine statistische Kenngröße, welche von der IRT eingeführt wird, und die mit dem Standardmessfehler und der Reliabilität (siehe Kapitel 5.4.2.2./3.) eng verwandt ist. Es ist die *Iteminformationsfunktion* $I(\theta, i)$. Sie beschreibt die Information, welche ein Item i zur Diskrimination zwischen verschiedenen Merkmalsausprägungen bei der Theta-Schätzung beiträgt, in Abhängigkeit von Theta (Suen, 1990). Obgleich sie mathematisch auf unterschiedliche Weise abgeleitet werden kann, stellt sie konzeptuell das Verhältnis der Steigung der ICC (1. Ableitung der ICC: $P'_i(\theta)^2$) zum erwarteten Standardmessfehler auf der jeweiligen Ausprägung des Theta-Kontinuums dar. Sie berechnet sich durch folgende Formel:

Gleichung G.1.:
$$I(\theta, i) = \frac{P'_i(\theta)^2}{P_i(\theta)\ Q_i(\theta)}$$

$P_i(\theta)$ = Wahrscheinlichkeit einer richtigen Antwort; $Q_i(\theta)$ = Wahrscheinlichkeit einer falschen Antwort ($Q_i(\theta) = 1 - P_i(\theta)$).

Die Iteminformation ist der Kennwert, welcher zur Itemselektion, d. h. zur Auswahl des „passendsten“ Items für ein Individuum, im Rahmen des IRT-basierten adaptiven Testens genutzt werden kann (siehe Kapitel 4.3.3.3.). Ferner ist sie bei der Itembankentwicklung von Tests interessant, da sie erlaubt, Items mit einem geringen Informationsgehalt bei der Testkonstruktion auszuschliessen. Auch zur Bewertung der Indikation verschiedener Tests kann sie aufschlussreich sein. Durch die pure Summierung der *Item*informationen aller Items kann nämlich die *Testinformation* berechnet werden, welche genutzt werden kann, um zu bewerten, welcher Test in welchen Bereichen der Merkmalsausprägung den höchsten Informationswert bietet (Embretson & Reise, 2000).

Neben diesen beiden für das *adaptive* Testen bedeutsamen Vorzügen der IRT und den bereits in Kapitel 3.3.1. eingeführten Vorteilen, die sich aus den alternativen messtheoretischen Annahmen ergeben, bietet die IRT weiterhin durch die Annahme der Stichprobenunabhängigkeit der Parameterschätzung „elegante“ Möglichkeiten...

1. des Inbezugsetzens unterschiedlicher Skalen („Equating“),
2. des metrischen Verbindens der Items von verschiedenen Skalen („Linking“ z. B. durch sogenannte „Anker-Test-Designs“),
3. der Analyse von systematischen Itemantwortverzerrungstendenzen („Differential-Item-Functioning“, DIF) und
4. der Analyse der Anpassung der Itemantworten einer Person an das Modell („Personen-Fit-Statistiken“).

Während in der KTT aufwendige Prozeduren des Inbezugsetzens verschiedener Skalen, welche die Messung derselben Merkmalsausprägung intendieren, nötig sind (z. B. „Equipercentile or linear equating“; Kolen, 1986), bietet die IRT spezifische „Linking-Designs“, welche ein direktes Inbezugsetzen von Skalen, über mehrere Itemparameter erlauben (Vale, 1986), so dass die Entwicklung einer gemeinsamen, instrumentenübergreifenden Metrik möglich ist. Exemplarisch sei hier das „Anker-Test-Design“ hervorgehoben, welches es erlaubt, die Itemparameter verschiedener Items, welche an verschiedenen Personenstichproben kalibriert wurden, auf einer *gemeinsamen* Metrik zu positionieren (in IRT-Begrifflichkeiten: kalibrieren), wenn ein Set von gemeinsamen Items („Anker-Items“) beiden Personenstichproben dargeboten wurde (siehe Kapitel 5.3.2.3.3.).

Die Analyse von DIF ist speziell im Hinblick auf die häufig diskutierte Testfairness im Rahmen der Testkonstruktion und –evaluation ein wichtiger Aspekt. Während in der KTT „Item bias“ (systematische Antwortverzerrungen) üblicherweise durch die Invarianz des Faktorenladungsmusters von Items eines Tests, welcher an verschiedener Stichproben oder zu unterschiedlichen Messzeitpunkten erhoben wurde, mittels konfirmatorischer Faktorenanalysen untersucht wird (Reise, Widaman & Pugh, 1993), bietet die IRT detailliertere Möglichkeiten der DIF-Analyse (Thissen Steinberg & Gerrard, 1986). So können Itemantwortverzerrungstendenzen spezifisch auf der Grundlage der IRFs untersucht werden, d. h. z. B. in Bezug auf *einzelne* Antwortkategorien oder in Abhängigkeit von *verschiedenen* Itemstatistiken (Schwierigkeit, Diskriminations-fähigkeit etc.).

Die Erfassung von Personen-Fit (Meijer, 1996), also der Konsistenz des Antwortverhaltens einer Testperson zu den IRT-Modellannahmen, ist nicht nur ein methodisches Spezifikum der IRT, sondern von allgemein psychometrischer

Relevanz, wenn eine Identifizierung von Personen, welche formale (zur Mitte oder zu den Extremen) oder inhaltliche Antworttendenzen (aufgrund von sozialer Erwünschtheit, etc.) aufweisen, gewünscht ist.

Abschließend seien noch zwei Vorzüge der IRT hervorgehoben, welche den Anwendern von IRT-basierten Tests sofort auffallen dürften, und daher von direkter praktischer Relevanz sind. Zum einen ermöglichen einige IRT-Modelle (z. B. das GPCM, siehe Kapitel 3.4.3.) die Verwendung *verschiedener* Antwortformate (dichotome und verschiedene polytome Formate) zwischen Items innerhalb *eines* IRT-basierten CATs, zum anderen unterscheidet sich eine IRT-basierte Testscore – *Interpretation* von Theta von der in der KTT üblichen *norm*bezogenen Interpretation (Embretson & Reise, 2000).

Während in KTT-basierten Verfahren ein Messergebnis in der Regel in Bezug auf eine Normstichprobe interpretiert wird (sogenannte komparative Messung), kann in der IRT – aufgrund der Positionierung der Item- und Personenparameter auf einer *gemeinsamen* Skala (siehe Kapitel 3.3.1.) – zusätzlich zur normbezogenen Interpretation auch eine Interpretation der Theta-Schätzung bezogen auf Iteminhalte erfolgen. Während in der KTT also eine Aussage getroffen wird, die beispielsweise wie folgt lautet: „Person j hat ein Messergebnis auf der Skala „Angst“, welches größer ist als bei 85% aller Personen einer Normstichprobe“, kann in einem IRT-basierten Test die geschätzte Merkmalsausprägung mit Hilfe des Inhalts der Items beschrieben werden, die durch ihre Itemparameter in der Nähe der geschätzten Merkmalsausprägung lokalisiert sind. Ein Beispiel für eine solche *inhalts*bezogene direkte Interpretation wäre: „Die Merkmalsausprägung der Angst von Person j kann behaftet mit einem Vorhersagefehler v durch die Items „häufige Angstattacken“ (Item i_1), „starke Unsicherheit“ (Item i_2) und „Zittern“ (Item i_3) am besten beschrieben werden“. Eine solche Beschreibung der Merkmalsausprägung kann eine informationsreiche Ergänzung zur üblichen normbezogenen Interpretation von Testwerten sein.

3.3.4. Nachteile der IRT

Obgleich die bisherigen Erläuterungen zeigen, dass die IRT neue Wege bei der Lösung vielfältiger Messprobleme eröffnet, ist sie kein psychometrisches „Allheilmittel“. Ihre Anwendung wirft ebenfalls eine Reihe von Schwierigkeiten auf, die im Folgenden zusammengefasst werden sollen.

Zunächst stellt die Anwendung der IRT höhere Anforderungen an personelle, technische und finanzielle Ressourcen als die KTT. Psychodiagnostisches und statistisches Fachwissen zur richtigen Anwendung der Methoden sowie technische Expertise bei dem Gebrauch der - leider meist eher benutzer*un*freundlichen - IRT-Software und gegebenenfalls bei der eigenständigen Entwicklung von IRT-basierten computergestützten Schätzalgorithmen (bei CAT-Anwendungen) sind erforderlich. Die Anschaffungskosten für Hard- und Software, welche aufgrund aufwendiger Rechenleistungen im Rahmen von IRT-Modellierungen unabdingbar ist, müssen kalkuliert werden, und es bleibt abzuwägen, ob dieser insgesamt hohe Initialaufwand lohnt. In der Praxis zeigt sich, dass vor allem Organisationen, welche routinemäßig breitangelegte Testuntersuchungen an großen Personenkollektiven durchführen (wie z. B. der Educational Testing Service, ETS, 1996), von IRT-Anwendungen im Allgemeinen (siehe Kapitel 3.5.1.) und von auf dieser Basis implementierten Computer Adaptiven Testungen (CAT; siehe Kapitel 4.6.1.) im Besonderen profitieren. Die über die letzten Jahrzehnte zunehmende Forschungsaktivität im Hinblick auf IRT-und CAT-Anwendungen zeigt, dass die angeführten Hindernisse überwindbar sind.

Trotz der zunehmenden Forschungsarbeiten bestehen noch eine Reihe von methodischen Unsicherheiten, welche auf einen großen Forschungsbedarf hindeuten. Schwierig gestaltet sich bei der Anwendung der IRT, dass...

a) methodische *Standards* zur Entwicklung IRT-basierter Tests bislang fehlen,
b) die erforderliche *Größe der Kalibrierungsstichprobe* zur *robusten* Parameterschätzung unsicher ist: je nach IRT-Modell und Forscher werden unterschiedliche Personenstichprobengrößen (n) empfohlen:
    - Rasch-Modelle: Linacre (1994), Wright (1996): n > 150;
    - GRM-Modell:
        - Embretson und Reise (2000): n > 350; Reise und Yu (1990) : n > 500;

- GPCM-Modell:
  - Cella und Chang (2000): bei dichotomen Items: n > 1.000, bei polytomen Items: n > 1.000;
  - Muraki und Bock (1999): n = 500 – 1.000;

c) die Robustheit der Parameterschätzungen bei *Verletzungen* der IRT-Modellannahmen umstritten sind,[27]

d) die *Wahl des* angemessenen *IRT-Modells* schwierig ist, sowie die Auswirkungen einer unpassenden Modellwahl auf die Parameterschätzung nicht bekannt sind,

e) *Modell-Fit-Statistiken* vor allem bei zweiparametrischen Modellen unzulänglich erforscht sind (Van der Linden & Hambleton, 1997; siehe Kapitel 5.3.2.3.4.),

f) *mehrdimensionale* IRT-Modelle bislang (zumindest in der Persönlichkeitsdiagnostik) vernachlässigt werden und

g) eine pragmatische Anwendungsforschung zur Erprobung *iteminhaltsbezogener* Interpretationen weitgehend fehlt (siehe Kapitel 3.3.3.).

3.4. IRT-Modelle

3.4.1. Ein Überblick

Die Entwicklung von IRT-Modellen begann in den 40er / 50er Jahren mit Vertretern wie Lord (1952), der als Vater des „Normal Ogive Modells“ (NOM) angesehen werden kann, sowie Rasch (1960) und Birnbaum (1968), welche alternativ zum mathematisch komplexen NOM die logistische Funktion einführten.

Damit war die Familie der „Rasch-Modelle“ geboren, welche eine rege Forschungs- und Modellentwicklungstätigkeit anstieß. Die meisten Modelle, die in dieser Anfangsphase der IRT-Geschichte entstanden, sind *eindimensional* konzipierte Modelle, welche für die Modellierung des Antwortverhaltens von Items mit *dichotomem* Antwortformat entwickelt wurden. Erst in den 80er Jahren gelang es einer Reihe von Forschern (Samejima, 1969, 1996; Andrich, 1978; Masters, 1982) IRT-Modelle zu entwickeln, die auch auf Items mit *polytomem* Antwortformat anwendbar sind, und seither vielfach erprobt wurden. Etwas später entstanden IRT-Modelle, welche für die Modellierung

[27] Studien von Dorans und Kingston (1985), Forsyth, Saisangjan und Gillmer (1981) sowie Rentz und Barshaw (1977) ergaben die relative Robustheit der Parameterschätzungen bei Modellverletzungen; Studien von Cook, Eignor und Taft (1984), Loyd und Hoover (1980) sowie Slinde und Linn (1978) konnten dies nicht belegen.

multidimensionaler Daten entwickelt wurden (Bock, Gibbons & Muraki, 1988; Carstensen, 2000; Keldermann, 1997; McKinley & Way, 1992; Reckase, 1997; Rost & Carstensen, 2002; Segall, 1996, 2001).

Mittlerweile existieren eine Fülle von unterschiedlichen IRT-Modellen, welche sich nach verschiedenen Aspekten taxonomisch ordnen lassen, wie z. B. der Art der IRF (Moosbrugger, 1984), der Art der Variablen (Rost, 1996), der *Anzahl* der Itemparameter (Weiss & Davison, 1981) und der *Separierbarkeit* von Itemparametern (Müller, 1997). Die Klassifikation der verschiedenen Modelle erfolgt am häufigsten nach der Zahl der in der IRF spezifizierten Itemparameter (siehe Abbildung 5).

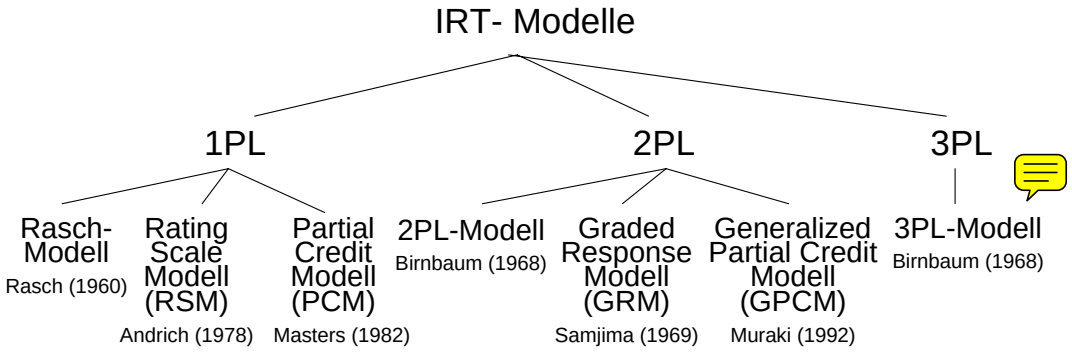


Abbildung 5: Überblick über die wichtigsten IRT-Modelle.

Es werden Modelle, welche einen, zwei bzw. drei Itemparameter postulieren, unterschieden. Die einparametrischen Modelle (1PLM) beschreiben das Antwortverhalten mit Hilfe von einem einzigen Itemparameter, dem „Location Parameter“ („b“; Lokationsparameter), welcher die Positionierung eines Items auf dem latenten Trait bestimmt. Zu den 1PL-Modellen gehört das eindimensionale Rasch-Modell (Rasch, 1960; siehe Kapitel 3.4.2.), das Rating Scale Modell (RSM; Andrich, 1978) sowie das Partial Credit Modell (PCM; Masters, 1982). Zur Form der IRCs dieser Modelle sei auf Abbildung 4 in Kapitel 3.3.1. verwiesen. Zweiparametrische Modelle (2PLM) sind komplexere Modelle und nutzen neben dem Lokationsparameter einen zweiten Itemparameter, den „Slope Parameter“ („a“; Steigungsparameter), zur Spezifizierung der Beziehung zwischen dem beobachtbaren Antwortverhalten und der latenten Variable (zur Form der IRC von 2PL-Modellen siehe Abbildung 4 in Kapitel 3.3.1.). Und schließlich wird in dreiparametrischen

Modellen (3PLM, z. B. Birnbaum, 1968) zusätzlich zu den beiden genannten Itemparametern ein „Guessing Parameter“ („c“; Rateparameter) konzipiert, welcher besonders bei der Modellierung des Antwortverhaltens in Tests, in denen Testpersonen möglicherweise die richtige Antwort „raten“ können (z. B. Leistungstest), eine Rolle spielt. Grafisch zeigen sich solche „Rateeffekte“ in Antwortkategorienkurven (IRCs), die ihren Ursprung dann nicht bei Null haben, sondern in einem Wert größer Null, dem sie sich asymptotisch annähern. Modelle, welche sowohl zwei- als auch dreiparametrisch spezifiziert werden können, sind z. B. das Graded Response Modell (GRM; Samejima, 1969, 1996) und das Generalized Partial Credit Modell (GPCM; Muraki, 1992). Letzteres Modell wurde zur Itemparameterschätzung des hier entwickelten Angst-CATs genutzt (siehe Kapitel 3.4.3.).

Die verschiedenen Modelle unterscheiden sich in vielfältigen Aspekten, jedoch können manche auch als Generalisierungen oder Spezialfälle von anderen angesehen werden (Levine et al., 1992).

Im Folgenden werden exemplarisch zwei Modelle vorgestellt, das Rasch-Modell in seiner Ursprungsform (siehe Kapitel 3.4.2.), welches als „Mutter“ aller IRT-Modelle angesehen werden kann, und das GPCM als Beispiel für ein polytomes Modell (siehe Kapitel 3.4.3.). Im Anschluss daran werden auf der Grundlage dieser spezifischen Modellausführungen einige Vor- und Nachteile unterschiedlicher Modelle diskutiert und gegeneinander abgewogen (siehe Kapitel 3.4.4.). Den Abschluss dieses Unterkapitels (siehe Kapitel 3.4.) bildet schließlich ein Kapitel über die Wahl des adäquaten IRT-Modells und die Bestimmung seiner Gültigkeit (siehe Kapitel 3.4.5.).

3.4.2. Das Rasch-Modell

Der dänische Mathematiker Rasch (1960) entwickelte eine Familie von einparametrischen IRT-Modellen für dichotome Items, die nach ihm benannt wurden. In dieser Modellfamilie wird die Lösungswahrscheinlichkeit als (nicht lineare) logistische Funktion, welche durch die Personenfähigkeit (Personenparameter) und Aufgabenschwierigkeit (Itemparameter: Lokations~) spezifiziert wird, modelliert (siehe Gleichung G.2.).

Gleichung G.2.: $$p\left(x_{ij}\right) = \frac{\exp\left(x_{ji}\left(\theta_j - b_i\right)\right)}{1 + \exp\left(\theta_j - b_i\right)}$$

$p\left(x_{ji}\right)$ = Wahrscheinlichkeit für das Antwortverhalten x einer Person j auf das Item i. Xij kann entweder den Wert 1 annehmen (für die Lösung des Items) oder den Wert 0 (für die Nicht-Lösung des Items). Die Gleichung G.2. lässt offen, ob das Item gelöst wird oder nicht.. θ_j = Personenfähigkeit (Personenparameter) einer Person j; b_i = Aufgabenschwierigkeit (Lokationsparameter) eines Items i.

Das dichotome Rasch-Modell ist - verglichen mit anderen 2- bzw. 3PL-Modellen - in seinen zugrundeliegenden Annahmen recht restriktiv, da Items nur in ihrem Lokationsparameter b_i divergieren dürfen. Dies drückt sich in den IRCs verschiedener Items so aus, dass das Modell postuliert, dass sich diese in ihrem Kurven*verlauf* nicht unterscheiden, sondern nur in der *Positionierung* auf dem latenten Merkmalskontinuum variieren, d. h. die IRCs verschiedener Items gleichen sich (es gibt keine Überschneidungen zwischen den IRC) und sind lediglich auf der Abszisse parallel verschoben (siehe Abbildung 4, links, in Kapitel 3.3.1.). Weitere zentrale Modellcharakteristiken sind - neben den bereits in Kapitel 3.3.1. erläuterten IRT-Modelleigenschaften der Stichprobenunabhängigkeit der Parameterschätzung und der lokalen stochastischen Unabhängigkeit - das Postulat der Summenwerte als erschöpfende Statistik und das der spezifischen Objektivität. Dass Summenwerte als erschöpfende Statistik genügen, bedeutet, dass durch die reine Addition der Itemantworten die latente Merkmalsausprägung schätzbar ist. Dies ist insofern vorteilhaft, als eine Gewichtung verschiedener Items nicht erfolgen muss, und damit der Aufwand zur Berechnung des Testwerts einer Person relativ gering ist. Die Eigenschaft der erschöpfenden Statistik bezieht sich nicht nur auf die Personenparameterschätzung, sondern auch auf die Itemparameterschätzung. So lässt sich durch die Antworten von Personen einer Stichprobe auf ein spezifisches Item auch der Lokationsparameter durch pure Addition schätzen. Nach erfolgreicher Überprüfung der Modellkonformität wird zudem angenommen, dass die Eigenschaft der spezifischen Objektivität gegeben ist. Diese ist erfüllt, wenn der Schwierigkeitsunterschied zweier Items unabhängig davon festgestellt werden kann, ob Personen mit niedrigen oder hohen Merkmalsausprägungen untersucht wurden, d. h. in der Umkehrung, dass Unterschiede zwischen Personenparametern unabhängig von den verwendeten Items festgestellt werden können

3.4.3. Das Generalized Partial Credit Modell (GPCM)

Das Generalized Partial Credit Modell (GPCM) wurde ursprünglich von Muraki (1990) entwickelt. Es stellt eine erweiterte Form des Partial Credit Modells (PCM) von Masters (1982) für polytome Items dar. Masters (1982) PCM erlangt seinen Namen durch die Besonderheit, dass es die abgestufte Bewertung der Antworten (Partial Credit) konzipiert (siehe Kapitel 3.4.4.).

Das GPCM gründet sich auf der Annahme, dass die Wahrscheinlichkeit $P_{ih}(\theta)$, die Antwortkategorie *h* eines Items *i* zu wählen, in Form der in Gleichung G.3. (Muraki, 1997) dargestellten logistischen „Item Category Response Function“ (ICRF, Itemantwortfunktion) beschrieben werden kann.

$$Pih(\theta j) = \frac{\exp\left[\sum_{j=1}^{h} Z_{ij}(\theta j)\right]}{\sum_{c=1}^{mi} \exp\left[\sum_{j=1}^{c} Zij(\theta j)\right]}$$

Gleichung G.3.:

$$Z_{ih}(\theta j) = \left[\sum_{j=1}^{h} Zij(\theta j)\right] = Dai(\theta j - bih) = Dai(\theta j - bi + d_{ih})$$

θ = Personenparameter (Merkmalsausprägung); Indizes: *i* = Item; *h* = Antwortkategorie; j = Person; D = „Skalierungskonstante“ (= 1,7) hat die Funktion, die logistische Funktion an die „Normal Ogive Function“ anzugleichen (Lord, 1952).

In der ICRF werden folgende Itemparameter[28] berücksichtigt:

***a**i*: „Slope Parameter“ (Steigungsparameter). Er spezifiziert die gemittelte Steigung über alle Antwortkategorienkurven (IRCs) eines Items und stellt einen Indikator für die Diskriminationsfähigkeit eines Items auf einer bestimmten Merkmalsausprägungsstufe (Theta-Wert) dar. Er steht in enger Beziehung zum KTT-basierten Reliabilitätsindex.

b**i*: „Location Parameter“ (Lokationsparameter). Bei Leistungstests ist er der Parameter, der analog zum in der KTT berechenbaren Schwierigkeitswert steht. Er drückt die Positionierung eines Items auf dem latenten Merkmalskontinnum (Theta) aus, und liegt mit dem Personenparameter auf einer gemeinsamen Skala (siehe Kapitel 3.3.1.). Bei dichotomen Modellen (z. B. dem dichotomen Rasch-Modell) ist er das Lot des Wendepunktes der IRC auf dem latenten Merkmalskontinuum (Theta), bei polytomen Modellen wird er über den Mittelwert der Antwortkategoriengrenzen (d**ih*) berechnet.

***b**ih*: „Item Threshold Parameter“ (Schwellenparameter). Er spezifiziert die absolute Lokalisation der Antwortkategoriengrenzen von Items auf dem latenten

[28] Zur Erläuterung der Bedeutung der Itemparameter siehe Kapitel 3.3.1. und zur Taxonomie von IRT- Modellen nach der Anzahl der berücksichtigten Itemparameter siehe Kapitel 3.4.1.

Trait (Theta). Grafisch ist er als Lotpunkt auf der Abszisse zu verorten, an dem zwei Itemantwortkategorienkurven (IRCs) sich schneiden.

d_{ih}: „Item Category Parameter“ (Parameter der Antwortkategoriengrenzen). Dieser Parameter spezifiziert die Lokalisation der Antwortkategoriengrenzen von Items auf dem latenten Trait (Theta) in Relation zum Lokationsparameter.

Die Besonderheit des GPCM (Muraki, 1990, 1992, 1997) liegt - verglichen mit dem anfänglich erwähnten PCM von Masters (1982) - in (a) der Lockerung der Annahme der gleichen Diskriminationsfähigkeit von Items, und (b) der Möglichkeit, das Antwortverhalten auf Items mit unterschiedlichen Antwortformaten zu modellieren. Die Lockerung der Annahme der gleichen Diskriminationsfähigkeit von Items zeigt sich in der Berücksichtigung eines Steigungsparameters, welcher für jedes Item einzeln geschätzt wird. Grafisch drückt sich dies in zwischen verschiedenen Items in ihrer Steigung variierenden Kurvenverläufen (IRCs) aus. Die Möglichkeit der Berücksichtigung von Items mit unterschiedlichen Antwortformaten bei der Konstruktion einer gemeinsamen Skala ist insofern sinnvoll, als abhängig vom Inhalt der Fragen oft unterschiedliche Antwortformate nötig erscheinen, und zudem bei der Kalibrierung großer Itembanken Itemparameter von Items aus unterschiedlichen Instrumenten (welche oft verschiedene Antwortformate aufweisen) gemeinsam kalibriert werden können.[29] Für eine detailliertere Erörterung des Modells verweise ich den interessierten Leser auf Muraki (1990, 1992, 1997).

3.4.4. IRT-Modelle im Vergleich

Da eine ausführliche Darstellung aller IRT-Modelle den hier gegebenen Rahmen sprengen würde, werden im Folgenden nur einige wichtige Unterschiede zwischen den bekanntesten *unidimensionalen* Modellen hervorgehoben (Überblick siehe Kapitel 3.4.1.). Für eine detaillierte Einführung in die gebräuchlichsten IRT-Modelle empfehle ich das Handbuch von van der Linden und Hambleton (1997).

Zunächst werden Besonderheiten von zwei *einparametrischen* Modellen (RSM, PCM) herausgestellt, gefolgt von der Abgrenzung zu mehreren *zweiparametrischen* Modellen (GRM, M-GRM, GPCM).

Das Rating Scale Modell (RSM) von Andrich (1978) sowie das Partial Credit Modell (PCM) von Masters (1982) sind *einparametrische* Modelle, die der

[29] Siehe Kapitel 5.3.1.

Familie der Rasch-Modelle zugehörig sind, und mit ihr die Eigenschaft der erschöpfenden Statistik sowie der einheitlichen Fixierung des Steigungsparameters auf einen Wert von $a_i = 1$ gemeinsam haben. Das RSM kann vom PCM abgeleitet werden (Embretson & Reise, 2000) und stellt ein restriktiveres Modell für ordinale, d. h. strikt geordnete (polytome) Daten dar, welches für alle Items dieselben konstanten Schwellenparameter annimmt („Äquidistanz zwischen Antwortkategorien“).

Das PCM (Masters, 1982) kann als ein Spezialfall des „Normal Ogive Modell“ (NOM) angesehen werden (Thissen & Steinberg, 1986). Es erlangte seinen Namen durch die Besonderheit, dass es eine abgestufte Bewertung (Partial Credit) der Antworten konzipiert. Bei seiner Anwendung werden polytome Antwortformate in „m-1“ hypothetische, dichotome Subitems zerlegt. Während das RSM ordinal geordnete Antwortkategorien verlangt, können mit dem PCM dagegen auch Items, deren Antwortkategorienparameter nicht geordneten sind, analysiert werden.

Sowohl das RSM als auch das PCM erlauben Analysen von Items mit unterschiedlichen Antwortformaten nur in isolierten Gruppen (Blöcken). Die isolierte Itemanalyse von Items verschiedener Antwortformate kennzeichnet auch die Anwendung von zwei *zweiparametrischen* Modellen: dem Graded Response Modell (GRM) von Samejima (1969) und dem Modified Graded Response Modell (M-GRM) von Muraki (1990). Das GRM postuliert einheitliche Steigungen der Antwortkategorienkurven innerhalb eines Items und nutzt eine über die Antwortkategorien kumulierende Schätzfunktion zur Parameterschätzung (Embretson & Reise, 2000). Das M-GRM (Muraki, 1990) ist eine Modifikation des GRMs. Im Unterschied zum GRM, welches eine Variation der Kategorienschwellenparameterwerte zwischen Items erlaubt, liegt die Besonderheit des M-GRMs in der Zerlegung der Antwortkategorienparameter in einen für jedes Item spezifischen Lokationsparameter und in für alle Items einer Skala geltende *einheitliche* Kategorienparameterwerte.

Das Generalized Partial Credit Model (GPCM, Muraki, 1992) ist verglichen mit den vorangestellten Modellen dasjenige mit den geringsten Restriktionen in den Modellannahmen. Es erlaubt die gemeinsame Analyse von Items mit unterschiedlichen Antwortformaten, frei variierende Steigungen der Antwort-

kategorienkurven (IRCs) innerhalb eines Items sowie frei zwischen Items variierende Steigungs-, Kategorienschwellen- und Lokationsparameterwerte.

Für alle *zweiparametrischen Modelle* (GRM, M-GRM und GPCM) gilt die für die Rasch-Modelle charakteristische Eigenschaft der *erschöpfenden Statistik nicht*, da mehr als ein Itemparameter in die Schätzung des Personenparameters eingeht und damit eine Gewichtung der Itemantworten erfolgt, welche die Anwendung dieser Modelle mathematisch (rechen-) aufwendiger macht.

3.4.5. Zur Wahl eines IRT-Modells und Bestimmung des Modell-Fits

Die Diskussion um das „*beste*“ IRT-Modell währt bereits drei Jahrzehnte. Der Standpunkt, je *mehr* Parameter ein Modell berücksichtigt, desto besser kann es die empirische Realität modellieren, läuft dem „Prinzip der Sparsamkeit“ („principle of parsimony“, Embretson & Hershberger, 1997, S. 246) zuwider. In der Tat erscheinen in manchen Anwendungsfällen komplexe (mehr-parametrische) IRT-Modelle jedoch besser zu den empirischen Daten zu passen, da sie weniger restriktive Annahmen setzen. Allerdings unterliegen sie im Falle *geringer* Personenstichprobengrößen in ihrer Datenanpassung IRT-Modellen mit wenigen Parametern. Dies äußert sich dann in *instabilen* Parameterschätzungen. Mitunter kann ein Mangel an identifizierbaren Parametern auch der Anwendung komplexerer IRT-Modelle im Wege stehen (Van der Linden & Hambleton, 1997).

Die Wahl eines IRT-Modells kann von den folgenden Aspekten abhängen:

1. der Art des theoretischen Konstruktes:
   - ist es unidimensional oder multidimensional?
   - sind Rateparameter sinnvoll?[30]
2. dem Ziel der Parameterschätzung (präzise Schätzungen werden eher über 2/3 PL-Modelle erreicht; Embretson & Reise, 2000),
3. der Gewichtung von Itemantworten (müssen diese aus inhaltlichen Gründen gewichtet werden, so bieten sich 2/3 PL-Modelle an, ist dies nicht der Fall, so kann mit Rasch-Modellen gearbeitet werden),
4. der Praktikabilität (die Parameterschätzungen mit Rasch-Modellen gestaltet sich einfacher als diejenige von 2/3 PL-Modellen) und
5. der Datenanpassung an das Modell (Modell-Fit).

[30] Rateparameter sind v.a. bei IRT-Modellierungen von Leistungstests, weniger bei Persönlichkeitsskalen sinnvoll (Suen, 1990).

Insbesondere der letzte Punkt: die Frage, ob die Daten konsistent mit dem gewählten Modell sind, erregt häufig Aufmerksamkeit und Kopfzerbrechen. Ziel ist es, ein Modell zu wählen, welches möglichst gut zu den empirischen Daten passt, bzw. die Daten (z. B. mittels Itemselektion) oder die Konstrukte (z. B. durch Re-Konzeptualisierungen) so zu verändern, dass sie zu dem Modell passen. Hierbei ist es wichtig, sich vor Augen zu führen, dass Modelle stets Idealisierungen darstellen, die nie gänzlich der Realität entsprechen (Van der Linden & Hambleton, 1997). Die Tatsache, dass die Passung zwischen Daten und Modellen empirisch untersucht werden kann, ist eine Besonderheit der IRT (in der KTT nicht gegeben, siehe Kapitel 3.2.). Die empirische Überprüfung der Modellkonformität ist insofern zentral, als von ihr das Inkrafttreten zentraler Modelleigenschaften wie z. B. der Stichprobeninvarianz (siehe Kapitel 3.3.1.) abhängt, und damit die Güte der Parameterschätzung beeinflusst wird. Empirische Modellgeltungstests können auf zweierlei Wegen erfolgen: mittels grafischer Kontrollen der Residuen und / oder durch eine numerische Erfassung. Für letzteres werden häufig χ^2-Tests durchgeführt, welche jedoch durch ihre Sensitivität gegenüber der Stichprobengröße in Kritik geraten sind. Während statistische Modellgeltungstests für Rasch-Modelle weitgehend erforscht und etabliert sind (Andersen, 1973; Glas, 1988; Keldermann, 1984; Molenaar, 1974), gilt dies nicht für die Modellgeltungstests von 2/3 PL-Modellen (Van der Linden & Hambleton, 1997, S. 16). Gut etablierte statistische Tests existieren für diese nicht, und selbst wenn sie existieren würden, zögen Van der Linden und Hambleton (1997) deren Nützlichkeit in Zweifel. Denn unabhängig davon, ob ein Modell tatsächlich zu den Daten passt oder nicht, wird - lässt man sich von χ^2-Statistiken leiten - bei genügend großen Personenstichproben jedes Modell verworfen. Überspitzt formulierte dies McDonald bereits 1989 so: „[the] failure to reject an IRT model is simply a sign that sample size was too small“ (S. 212). Als Alternativen zu den χ^2-Fit-Statistiken werden drei Wege vorgeschlagen (Van der Linden & Hambleton, 1997):

1. die Überprüfung der Gültigkeit der IRT-Modellvoraussetzungen, z. B. durch die gezielte Untersuchung der Unidimensionalität und der Modellkonformität der IRCs,
2. die Überprüfung der Invarianz von Itemparametern zwischen verschiedenen IRT-Modellen und Personenstichproben und
3. die Überprüfung der Modellvorhersage im Rahmen von simulierten und realen Validierungsuntersuchungen.

Abgesehen von diesen drei Alternativstrategien zur Überprüfung der Modellgültigkeit stellt sich dennoch die Frage, wie mit einem potentiellen Ergebnis eines numerischen “Modell-Misfits”, also der Tatsache, dass statistische Modellgeltungstests nahe legen, dass es keine Passung zwischen Daten und Modell gibt, bei der Anwendung von χ^2-Fit-Statistiken umgegangen werden soll. Prinzipiell sind zwei Konsequenzen zur gezielten Verbesserung der Fit-Statistiken denkbar: eine gezielte Itemselektion oder eine Lockerung der Restriktionen eines Modells (oder die Wahl eines weniger restriktiven Modells). Diese Strategien sind jedoch nur sinnvoll, wenn man diese Fit-Statistiken für gültig und damit handlungsleitend hält. Generell halten sich die meisten der IRT-Forscher bezüglich der Nennung spezifischer Richtlinien zum Umgang mit ungenügenden Ergebnissen in der Fit-Statistik bedeckt. Allgemein empfehlen Van der Linden und Hambleton (1997), dass der Umgang mit Misfits von folgenden Faktoren abhängig sei:

1. der Art des Misfits,
2. der Verfügbarkeit von Ersatzitems,
3. dem mit dem Neuschreiben von Items verbundenen Aufwand,
4. der Verfügbarkeit von Kalibrierungsstichproben und
5. dem Testziel.

Da drei dieser Punkte (2.-4.) Praktikabilitätserwägungen beinhalten, deutet sich hier an, dass oftmals praktische Einschränkungen zur (vorläufigen) Akzeptanz von Misfits, von denen vermutet wird, dass sie lediglich statistische „Artefakte“ darstellen, führen.

3.5. Aktueller Forschungsstand zur IRT

3.5.1. IRT Anwendungen in der Leistungsdiagnostik

Die IRT erfuhr seit den 80er Jahren mit der Verfügbarkeit von Software zur computergestützten Anwendung von IRT-basierten Methoden, welche sich in der Regel als sehr rechenaufwändig erweisen, in der Leistungs- und Eignungsdiagnostik eine weite Verbreitung. IRT-Anwendungen finden sich mittlerweile weltweit in Australien, Belgien, China, England, Indonesien, Israel, Japan, Kanada, Kora, den Niederlanden, Schweden, Spanien, Taiwan, der Türkei und den U.S.A. (Hambleton & Slater, 1997). Vor allem größere Testorganisationen, welche umfangreiche Routinetestungen durchführen, wie der Educational Testing Service (ETS), das American College Test (ACT) Board, das National Board of Medical Examiners (NBME), das College Board, die Psychological Corporation und der Law School Admissions Council (LSAC) nutzen die Potentiale der IRT zur Entwicklung und Evaluation von psychometrischen Tests (Embretson & Reise, 2000). Da eine umfassende Darstellung der internationalen anwendungsbezogenen Forschungsarbeiten zur IRT in der Leistungsdiagnostik an dieser Stelle nicht möglich ist, sei exemplarisch nur auf einzelne IRT-basiert konstruierte Tests wie die Graduate Record Examination (GRE; ETS, 1996), die Woodcock-Johnson-Psycho-Educational-Battery (Woodcock, 1989) sowie den Computerized Placement Test (CPT; College Board, 1993) hingewiesen. Die genannten Tests deuten auf den Trend zur Computerisierung von umfangreichen Testbatterien vor allem im Bereich der *Leistungsdiagnostik* im *anglo-amerikanischen* Sprachraum hin. In diesem Bereich wurden auch die ersten IRT-basierten Computergestützten Adaptiven Tests (CATs) entwickelt (siehe Kapitel 4.6.). Weiterhin finden sich hier auch erste Ansätze zur Anwendung mehrdimensionaler IRT-Modelle (Carstensen, 2000; McKinley & Way, 1992; Reckase, 1997; Rost & Carstensen, 2002; Segall, 1996, 2001). Verglichen mit der Anwendung der IRT im Bereich der *Persönlichkeitsforschung* lässt sich zusammenfassen, dass im Bereich der *Leistungsdiagnostik* die Geschichte der IRT begann und hier bislang auch das „Gros“ der Forschungsarbeiten zu verorten ist. Für einen Einstieg in die IRT-basierte Forschung im Bereich der Leistungsdiagnostik im *deutsch*sprachigen Raum sei auf drei Forschungskreise verwiesen, welche sich um Vertreter wie Hornke (1981, 1989, 1993, 1994, 1996, 1999; Hornke & Habon, 1984; Hornke

& Etzel, 1999a,b; Hornke, Küppers & Etzel, 2000), Kubinger (1986, 1993, 1996, 1999; Kubinger & Wurst, 2000) und Rost (1996, 1999; Rost & Carstensen, 2002; Rost & Spada, 1982) zentrieren.

3.5.2. IRT Anwendungen in der klinischen und Persönlichkeitsdiagnostik

Trotz ihrer Potentiale wurde die IRT - verglichen mit ihrer weiten Verbreitung im Bereich der *Leistungs*diagnostik - in der *Persönlichkeits*diagnostik bisher eher wenig genutzt (Steinberg & Thissen, 1995). In jüngster Zeit wird jedoch ein Trend zu einer zunehmenden Nutzung von IRT-Modellen zur Untersuchung der psychometrischen Eigenschaften von bereits etablierten Persönlichkeits-inventaren deutlich (Ozer & Reise, 1994). Es finden sich allerdings nur wenige Persönlichkeitsinventare (Thissen, Steinberg, Pyszczynski & Greenberg, 1983), welche *gänzlich* IRT-basiert entwickelt wurden (Embretson & Reise, 2000). Die meisten IRT-Anwendungen im Bereich der Persönlichkeitsforschung beziehen sich auf die Untersuchung bereits *existierender* psychometrischer Instrumente mit IRT-Methoden.

Mögliche Ursachen für die relativ geringe Verbreitung der IRT-Methodik bei der *Entwicklung* von *Persönlichkeits*inventaren mögen darin liegen, dass in den 70er Jahren IRT-Analysen von Persönlichkeitsinventaren durchgeführt wurden, welche wenig erfolgreich waren (persönliche Mitteilung von Prof. Dr. Westmeyer). Weiterhin kann der Mangel an genuin IRT-basiert entwickelten Persönlichkeitsinstrumenten auch - neben den in Kapitel 3.3.4. aufgeführten Nachteilen der IRT (z. B. benutzerunfreundliche Software, Erfordernis großer Kalibrierungsstichproben, hoher Rechenaufwand) - in einer ungenügenden Vermittlung von IRT-Kenntnissen und einer daraus resultierenden Unsicherheit bezüglich des Nutzens dieser Methodik im Rahmen der Persönlichkeitsforschung begründet sein (Childs, Dahlstrom, Kemp & Panter, 2000). Spezifisch für die Persönlichkeitsforschung ist außerdem, dass in ihr oftmals Konstrukte beforscht werden, deren Erfassung mit Daten konfrontiert, welche nicht so einfach wie diejenigen in der Leistungsdiagnostik den der IRT zugrundeliegenden messtheoretischen Annahmen entsprechen. So ist z. B. der Anspruch der Unidimensionalität bei vielen persönlichkeits-theoretischen Konstrukten schwierig realisierbar oder gar nicht intendiert

(Waller & Reise, 1989); und obgleich es multidimensionale IRT-Modelle gibt, gestaltet sich deren Anwendung komplizierter und ist noch weit weniger erforscht als die eindimensionaler IRT-Modelle. Weiterhin zweifeln manche Autoren (z.B. Reise, 2000), ob die Annahme monoton verlaufender Itemcharakteristiken bei *Persönlichkeits*items überhaupt gerechtfertigt sei. Dem entgegnen Rost, Carstensen und Davier (1999), dass fast alle konventionellen Persönlichkeitsfragebögen auf der Annahme basierten, dass ein höherer Ausprägungsgrad des zu messenden Traits auch zu einer stärkeren Zustimmung zum jeweiligen Iteminhalt führe; eine Annahme nicht-monotoner Itemfunktionen müsse zu *gänzlich* anderen Auswertungsformen führen, so dass auch die sonst in der KTT übliche Interpretation von Summenscores sich verbiete (Rost & Luo, 1997).

Wenn bislang *allein* auf der Grundlage der IRT kaum Persönlichkeitsinventare *entwickelt* wurden, stellt sich die Frage, welche Anwendungen die IRT im Bereich der Persönlichkeitsforschung denn erfährt.

Eine Sichtung der aktuellen Literatur zeigt, dass hier die IRT vor allem zur detaillierten Analyse der psychometrischen Eigenschaften von Antwortkategorien, Items und Skalen genutzt wird (u. a. Analyse der Skalenstruktur, Bewertung der Informationsfunktionen und Betrachtungen der Item Response Curves (IRCs) im Hinblick auf die Modellkonformität und Diskriminationsfähigkeit von Items und Antwortkategorien). Weiterhin werden mit IRT-Methoden Antworttendenzen, Antwortinkonsistenzen sowie Itempositionseffekte exploriert, sowie Differential-Item-Functioning (DIF) zwischen verschiedenen Subpopulationen (Geschlechtsunterschiede, kulturelle / sprachliche Unterschiede zwischen verschiedenen Testversionen etc.) erforscht.

Im Folgenden werden eine Reihe von Forschungsarbeiten zur Anwendung der IRT-Methodik im Bereich der Persönlichkeitsforschung zusammengefasst (Tabelle 5).

Tabelle 5: Überblick über IRT-Anwendungen im Bereich der Persönlichkeits- und klinischen Diagnostik.

Autoren	Jahr	Inventar	IRT-Modell
Gibbons, Clark, Cavanaugh & Davis	1985	Beck Depression Inventory (BDI)	Rasch-Modell
Bouman & Kok	1987	BDI	Rasch-Modell
Waller & Reise	1989	Absorption Scale	2 PL-Modell (Birnbaum, 1968)
Reise & Waller	1990	Multidimensional Personality Questionnaire (MPQ)	2 PL-Modell
King, King, Fairbank & Schlenger	1993	Mississippi Scale for Combat-Related Posttraumatic Stress Disorder	unklar
Ellis, Becker & Kimmel	1993	Trier Personality Inventory (TPI)	3 PL-Modell
Santor, Ramsay & Zuroff	1994	BDI	Nonparametrisches Modell (Ramsay, 1995)
Harvey, Murry & Markham	1994	Meyer-Briggs Type Indicator	unklar
Steinberg	1994	State Trait Anxiety Inventory (STAI-Trait)	Nonparametrisches Modell
Santor, Zuroff, Ramsay, Cervantes & Palacious	1995	BDI, Center of Epidemiological Studies-Depression Scale (CES-D), NEO-PI (N)	Nonparametrisches Modell
Waller, Tellegen, McDonald & Lykken	1996	Negative Emotionality Scale	2 PL-Modell
Gray-Little, Williams & Hancock	1997	Rosenberg Self-Esteem Scale	GRM (Samejima,1969)
Cooke & Michie	1997	Hare Psychopathy Checklist – Revised	GRM
Schmit & Ryan	1997	NEO-PI Conscientiousness Scale	GRM
Rost, Carstensen & Davier	1999	NEO-FFI	Eindim. Rasch-Modell & Mixed Rasch-Modell
Cooke, Michie, Hart & Hare	1999	Screening Version of the Hare Psychopathy Checklist (PCL:SV)	GRM
Rouse, Finger & Butcher	1999	MMPI-Psy-5 Scale	2 PL-Modell
Reise & Henson	2000	NEO PI-R	GRM
Orlando, Sherbourne & Thissen	2000	CES-D	GRM
Santor & Coyne	2000	Hamilton Rating Scale for Depression	Nonparametrisches Modell
Childs, Dahlstrom, Kemp & Panter	2000	MMPI-Depression Scale	2 PL-Modell
Chernyshenko, Stark, Chan, Drasgow & Williams	2001	16 Personality Factor Questionnaire (16 PF), Big Five Personality Measure	2/3 PL-Modell: GRM, Maximum likelihood formula scoring (MFS, Levine, 1974)
Ferrando	2001	Neuroticism Scales of Maudsley Medical Questionnaire (MMQ), Maudsley Personality Inventory (MPI), Eysenck Personality Inventory (EPI), Eysenck Personality Questionnaire (EPQ)	2 PL-Modell
Cooke, Kosson & Michie	2001	Psychopathy Checklist-Revised (PCL-R)	GRM
Marshall, Orlando, Jaycox, Foy & Belzberg	2002	Modified Version of the Peritraumatic Dissociative Experience Questionnaire (PDEQ)	GRM
Orlando & Marshall	2002	Post Traumatic Stress Disorder Checklist (PTSD-C)	GRM

Gemeinsam ist den in Tabelle 5 angeführten Forschungsarbeiten, dass sie ihren Schwerpunkt auf die Analyse *bereits existierender* psychometrischer Instrumente legen.

Die Anwendung von IRT-Methoden in der Persönlichkeitsforschung begann in den 80er Jahren durch die zunehmende Verbreitung von IRT-Software. Während zunächst Skalen zur Erfassung von Depressivität mit IRT-Methoden reanalysiert wurden, widmeten sich in den folgenden Jahren Persönlichkeitsforscher sowohl der Untersuchung einzelner weiterer Konstrukte (Neurotizismus, Selbstwirksamkeit etc.), psychopathologischer Checklisten (PCL, PDEQ, PTSD), sowie ganzer Persönlichkeitsinventare (TPI, NEO-FFI, 16PF, MMQ, MPI, EPI, EPQ und MMPI; siehe Tabelle 5).

Auffällig ist, dass in den Anfängen verstärkt ein- und zweiparametrische logistische Modelle (1PLM: Rasch, 1960; 2PLM: Birnbaum, 1968; Software: Bilog); später dann vor allem das Graded Response Modell (GRM; Software: Multilog, Thissen, 1991) und nonparametrische Modellierungen (Software: TestGraf, Ramsay, 1995) genutzt wurden. Eine Sichtung dieser Forschungsarbeiten (die Stichprobengrößen der Studien variieren bis zu N_{max} = 13.059 Personen; Chernyshenko et al., 2001) erlaubt das Fazit, dass - obgleich bezüglich zweiparametrischer Modelle wie z. B. dem Graded Response Modells keine Fit-Statistiken existieren und daher eine Bewertung schwer fällt - die Anwendung von IRT-Modellen im Bereich der Persönlichkeitsdiagnostik möglich und gewinnbringend ist (Embretson & Reise, 2000; Ferrando, 2001; Hambleton & Slater, 1997; Santor & Ramsay, 1998; Steinberg & Thissen, 1995). Durch die IRT-basierte differenzierte Analyse auf der Itemebene konnten für spezifische Instrumente Empfehlungen zur Optimierung der Tests durch Verbesserungen der Antwortformate, Elimination von wenig informativen Items oder von Items mit DIF ausgesprochen sowie verschiedene Testformen verglichen und unter Umständen einander angeglichen werden (mittels IRT-basierter „Equating“-Methoden; Orlando, Sherbourne & Thissen, 2000). Die angeführten Forschungsarbeiten legen nahe, dass die IRT-Methodik genauere Aussagen über die Beziehung zwischen dem Antwortverhalten und den zugrundeliegenden Konstrukten sowie eine Verbesserung des inhaltlichen Verständnisses des Messbereiches ermöglicht (z. B. Chernyshenko et al., 2001).

4. Computerdiagnostik

4.1. Einleitung

Unter Computerdiagnostik im psychologischen Bereich versteht Jäger (1990):

> „eine strategische Variante innerhalb der Diagnostik [...], um psychologisch relevante Variablen zu erfassen, deren Auswahl zu steuern, die erhaltenen Informationen zu einem Urteil zu verdichten und gegebenenfalls schriftlich und / oder bildlich darzustellen.“ (S. 91)

Nach ihm ist kein Abschnitt des psychologischen diagnostischen Prozesses ungeeignet, um ihn innerhalb der Computerdiagnostik zu realisieren (Jäger, 1990, S. 93). Die Geschichte der computergestützten psychologischen Diagnostik begann in den 20er Jahren, als erstmals automatisierte Testscorerechenmaschinen zur Berufseignungsdiagnostik in den U.S.A. eingesetzt wurden (SVIB: Strong Vocational Interest Blanks; Moreland, 1992). Seither trägt die zunehmende weltweite Verbreitung von Computern aufgrund stetiger technischer Fortschritte in der Hard- und Software-Entwicklung begleitet von einer allgemeinen Kostenreduktion dazu bei, dass in vielen psychologischen Feldern Computer als technische Hilfsmittel zur Diagnostik eingesetzt werden. Der Höhepunkt in der Computerdiagnostik ist aufgrund der fortschreitenden Soft- und Hardware-Entwicklung noch nicht abzusehen (Kubinger, 1993). Dies trifft vor allem auf den klinisch-psychologischen Bereich zu, in dem Computerdiagnostik bislang eher vernachlässigt wurde (Jäger & Krieger, 1994; Hänsgen & Bernascomi, 2000).

Die erste computerdiagnostische Anwendung im klinisch-psychologischen Bereich lässt sich in die 60er Jahre zurückdatieren, als in der Mayo-Klinik in Minnesota (U.S.A.) das international weit verbreitete Minnesota Multiphasic Personality Inventory (MMPI), ein umfangreicher klinischer Persönlichkeits-fragebogen, erstmals computergestützt erhoben wurde (Swenson, Rome, Pearson & Brannick, 1965). Inzwischen existieren weltweit Hunderte von psychodiagnostischen Computeranwendungen, welche grob in die folgenden Einsatzbereiche eingeteilt werden können:[31]

[31] Der dokumentarische und organisatorische Einsatz von Computern in der psychologischen Praxis und Forschung wurde hier nicht extra aufgeführt, da dieser mittlerweile selbstverständlich erscheint (Farrell konstatierte z.B. bereits 1989, dass jeder vierte klinische Psychologe regelmäßig zu dokumentarischen Zwecken einen Computer nutzt). Und klassische

1. Computergestütztes Testen:
   a) Testentwicklung,
   b) Testdurchführung,
   c) Testauswertung,
   d) Testevaluation,
      - Computergestütztes Adaptives Testen (CAT)
2. Computergestützte Interviews,
3. Computer Basierte Test Interpretationsprogramme (CBTI),
4. Computergestützte Expertensysteme.

Um einen Überblick über die genannten Computeranwendungen zu erleichtern, entspricht die formale Aufzählungsreihenfolge (1.-4) ihrem Verbreitungsgrad. Der internationale „Markt“ computergestützter Tests, die von Psychologen / Medizinern / Informatikern und auch fachfremden (!) Anbietern entwickelt werden, ist mittlerweile so groß, dass er kaum noch überschaubar erscheint. In einem über 10 Jahre alten Kompendium wurden bereits mehr als 1.000 computergestützte Tests weltweit aufgelistet (Sweetland & Keyser, 1991), dennoch ist deren Einsatz im Rahmen klinisch-psychologischer Diagnostik im europäischen Raum noch relativ selten (Hänsgen & Bernascomi, 2000).

Als ein Spezialfall computergestützter Tests können Computergestützte Adaptive Testverfahren (CAT) in den Kanon der Computerdiagnostik eingegliedert werden. Deren Verbreitungsgrad ist bislang noch sehr begrenzt (siehe Kapitel 4.6.). Spezifisch für CATs ist, dass sie sich die enorme Rechen- und Speicherkapazitäten von Computern zunutze machen, um Testungen möglichst individuell an die jeweilige Testperson „anzupassen“ (adaptiv). Die „Anpassung“ der Testung erfolgt, indem einzelnen Testpersonen jeweils nur die Fragen gestellt werden, welche für die Messung am informativsten sind (siehe Kapitel 3.3.). An zweiter Stelle der Popularität computergestützter Diagnostik stehen meines Wissens computerdiagnostische klinische Interviews (z. B.

Testentwicklungen – wie die des NEO-FFIs sind heutzutage ohne computergestützt berechnete Faktorenanalysen nur noch schwer vorstellbar). Der Einsatz von Computern im klinisch-therapeutischen Bereich stößt dagegen schnell an seine Grenzen. Für einen (leider veralteten) Überblick wird Bloom (1992) empfohlen. Hier werden Software Programme aus den 80er Jahren zur Unterstützung der Beck’schen kognitiven Therapie (Selmi, Klein, Greist, Johnson & Harris, 1982), der systematischen Desensibilisierung zur Behandlung von Phobien (Ghosh, Marks & Carr, 1984) und ein PC-Therapieprogramm mit zirkulären Fragen (Colby, Watt & Gilbert, 1966) erwähnt.

Psyndex[32] Recherche zwischen 1977 und 2003: 151 Artikel zur Nutzung des computergestützten Interviews CIDI; Wittchen & Pfister, 1996), welche in der Regel hoch strukturiert sind, und entweder vom Diagnostiker während des Gesprächs genutzt oder vom Patienten alleine interaktiv mit dem Computer bearbeitet werden. Der klinische Nutzen und die Validität solcher Interviews ist derzeit jedoch noch umstritten (Wetzler & Marlowe, 1994). Ebenso umstritten, aber noch weniger etabliert sind Computer Basierte Test Interpretationsprogramme (CBTI), die aufgrund häufig fehlender Validierungsuntersuchungen in die Kritik gerieten (Wetzler & Marlowe, 1994; Hornke, 1993; Garb, 2000).

Am wenigsten verbreitet, obgleich erwiesen wurde, dass allgemein die statistische Modellierung des diagnostischen Prozesses einer rein intuitiven klinischen Diagnostik überlegen ist (Wiggins, 1981), sind computergestützte diagnostische Expertensysteme. Sie wurden im deutsprachigen Raum bislang vor allem für den schulpsychologischen Bereich entwickelt, wo sie in der Einzelfalldiagnostik einerseits als wissensbasierte, interaktive Systeme den Diagnostiker in seinen Entscheidungen (bzgl. Hypothesenauswahl, Testindikationsentscheidungen und Testbewertungen) während des gesamten diagnostischen Prozesses regelgeleitet unterstützen (z. B. DIASYS; Hageböck, 1994, Westmeyer & Hageböck, 1992) oder auch „nur“ der statistischen Analyse und Interpretation von einzelnen psychometrischen Testbefunden dienen (z. B. PSYMEDIA, Hageböck, 1990).

[32] Psyndex: Datenbank der Zentralstelle für Psychologische Information und Dokumentation der Universität Trier. Sie enthält Nachweise und Abstracts zu deutschsprachigen Publikationen aus der Psychologie und ihren Randgebieten. Hier werden Artikel aus 250 Zeitschriften, Monographien, Beiträge aus Sammelwerken sowie Dissertationen und Reportliteratur aus Deutschland, Österreich und der Schweiz sowie Beschreibungen von in deutschsprachigen Ländern seit 1945 gebräuchlichen psychologischen Testverfahren dokumentiert.

4.2. Computergestütztes Testen

4.2.1. Vorteile

Viele Wissenschaftler sind sich einig, dass computergestützte Tests die folgenden Vorteile bieten:

1. Verbesserung der Datenqualität durch eine Erhöhung der Gütekriterien:
    - Objektivität,
    - Reliabilität,
    - Validität;
2. Ökonomische Vorteile:
    - Zeitersparnis,
    - Arbeitserleichterung,
    - Kostenreduktion,
    - Nützlichkeit;
3. Nutzung von Potentialen durch:
    - Multimedia,
    - Interaktive und Adaptive Strategien (z. B. durch CAT).

Einer der drei aus meiner Sicht wesentlichsten Vorteile computergestützter Tests ist die Verbesserung der klassischen Gütekriterien (Lienert & Raatz, 1994). Indem der Testleiter, welcher konventionell Papier-und-Bleistift-Testdarbietungen leitete, durch einen Computer ersetzt wird, entfallen mögliche Testleitereffekte (Schötzau-Fürwentsches & Grubitzsch, 1991; Kubinger, 1993). Dies bedeutet, dass mögliche Faktoren, welche die soziale Interaktion beeinflussen können, als „Störvariablen“ wegfallen, da z. B. ein Computer niemals müde, gelangweilt oder frustriert ist, sich jeder (moralischen) Wertung enthält und darüber hinaus über ein „konsistentes, perfektes Gedächtnis“ verfügt (Wetzler & Marlowe, 1994, S. 56ff). So wird die Testerhebung maximal standardisiert und die *Objektivität* steigt.

Indirekt wird dadurch auch die *Reliabilität* günstig beeinflusst (Retest-/ Interrater-Reliabilität). Einen direkten Einfluss auf die Reliabilität hat die Reduktion von (menschlichen) routinebedingten Auswertungs- bzw. Messfehlern (Butcher, 1987, S.17, schätzt, dass Auswertungsfehler aufgrund menschlichen Versagens in durchschnittlich 10% der Fälle vorkommen), d. h. der Computer bietet eine hohe Verrechnungs- bzw. Auswertungssicherheit (Kubinger, 1993; Gregory, 1996; Garb, 2000). Direkte Validitätsverbesserungen

haben sich einige Wissenschaftler (Johnson & Johnson, 1981; Lucas, Mullin, Luna & McInroy, 1977) zeitweise dadurch erhofft, dass „anonyme“ Computerbearbeitungen die Bereitschaft erhöhen könnten, offener intime / persönliche Fragen zu beantworten. Dies konnten Menghin und Kubinger (1996) jedoch empirisch nicht bestätigen. Weiterhin wird vermutet, dass die „hohe face validity“ (Kubinger, 1993) von computergestützten Tests sowie deren ansprechende mobile Darbietung (z. B. per Taschencomputer, siehe Rose et al., 1999, 2003) aufgrund des impliziten spielerischen Moments motivationsfördernd sein kann, und sich somit die Datenqualität und indirekt auch die *Validität* verbessert. Aufgrund eines diesbezüglichen Forschungs-defizits lassen sich darüber jedoch noch keine empirischen Aussagen treffen.

Zu den möglichen erheblichen *ökonomischen* Vorteilen zählt die *Zeitersparnis* bei der Testdurchführung und –auswertung für den Diagnostiker (Rose et al., 1999: Zeiteinsparungen von 2/3) und die Testpersonen (Butcher, 1987, S. 19: Zeiteinsparungen von 15-50%). Desweiteren können computergestützte Tests insofern zu einer massiven *Arbeitserleichterung* des Diagnostikers führen, als sie von gleichförmigen (organisatorischen und administrativen) Routine-tätigkeiten befreien (Schötzau-Fürwentsches & Grubitzsch, 1991; Jäger & Krieger, 1994) und durch die Standortflexibilität des Computers bzw. mobilen Taschencomputers die Arbeitskapazität des Diagnostikers von der Fragebogenbearbeitungszeit der Testperson(en) entkoppeln (Kleinmuntz & McLean, 1968). Eine Arbeitserleichterung stellt auch die schnelle Berechnung komplizierter Auswertungsalgorithmen, die einfache Doku-mentation (Speicherung), Verwaltung (Organisation in Datenbanken) und Verknüpfung großer Testdatenmengen (z. B. zur „online“-Aktualisierungen von Testnormen) sowie deren schnelle Abrufbarkeit dar. In diesem Zusammenhang ist die Vermeidung von „missing data“ durch computergestütztes Testen interessant. Rose und Mitarbeiter (1999) berichten beispielsweise über eine Zunahme der Vollständigkeit von Testdatensätzen von 15% (Papier-und-Bleistift-Tests: 80%; computergestützte Tests: 95%). Sie kann evoziert werden, indem der Computer so eingestellt wird, dass die nächste Frage nur erscheint, wenn die vorherige beantwortet wurde (Itemdarbietungskontrolle).

Verglichen mit umfangreichen Papier-und-Bleistift-Testheften weist Butcher (1987) auch darauf hin, dass bei der computergestützten Testvorgabe einzelner Items ein „Verrutschen“ auf dem herkömmlichen Antwortbogen vermieden wird. Schließlich führen Einsparungen von Testmaterial und Personalkosten zu Test*kosten*reduktionen von bis zu 50% (Gregory, 1996; Hornke, 1993, 1999; Rose et al., 1999; Weiss & Vale, 1987; zu den Nachteilen computergestützter Tests, siehe Kapitel 4.2.2.). Dies kann sich nach Hornke (1993, S. 115) bei 200.000 Testuntersuchungen jährlich in Kosteneinsparungen von 1,1 Mio. DM (pro Jahr) niederschlagen.[33]

Hieraus mag man leicht auf die *Nützlichkeit* von computergestützten Tests allgemein schließen. Kubinger (1993) merkt dazu jedoch an, dass die bloße Computerisierung von Papier-und-Bleistift-Tests einen Test als solchen nicht „nützlicher“ mache (S. 133). Ebenso wenig ist es nützlich, denselben Test mehrfach zu computerisieren (z. B. von verschiedenen Anbietern). Ein Test wird computergestützt dann nützlich, wenn anfangs erläuterte *Vorteile* genutzt werden können oder Potentiale genutzt werden, welche sich aus den Möglichkeiten des Computers ergeben. Dazu zählt z. B. die Nutzung von Multimedia (Gregory, 1996) durch die Ausschöpfung visueller (Tabellen, Grafiken, Video, Animationen), akustischer (Geräusche, Töne, Sprache, Musik), taktiler (z. B. Messung des Tastendrucks, z. B. mit „touchpads“), zeitlicher (Messung von Antwortlatenz bzw. Festlegung verschiedener Bearbeitungs-geschwindigkeiten z. B. bei der Leistungsdiagnostik), interaktiver und adaptiver Potentiale (zu den Vorteilen von CAT siehe Kapitel 4.4.). Dadurch kann Diagnostik realitätsgerechter - z. B. durch (Arbeitsalltags-) Simulationen im Rahmen der Berufseignungsdiagnostik - und individueller - z. B. durch adaptives Messen - werden.

[33] Rechenbeispiel zu Einspareffekten nach Hornke (1993): Eine Einsparung von 5 Items bei 200.000 Probanden macht einen Gewinn von 200.000[Pbn]*5[eingesparte Items]* 20sek.[Testzeit pro Item] = 5555 eingesparte Teststunden (z. B. beim Graduate Record of Examination pro Jahr mühelos erreicht). Wird ein Organisationsstundensatz von 200 DM zugrunde gelegt, so sind das Einsparungen von 1,1 Mio. DM pro Jahr.

4.2.2. Nachteile

Neben den genannten Vorteilen computergestützter Tests wird in der Literatur auch auf eine Reihe von möglichen Nachteilen hingewiesen.

Diese können in Kategorien negativer Auswirkungen in Bezug auf a) den Diagnostiker, b) die Testpersonen und c) die Datenqualität gegliedert werden.

Computerdiagnostik setzt eine gewisse technische Kompetenz im Umgang mit Computern voraus. Ist der *Diagnostiker* wenig vertraut mit Computern, so kann allein der Umstand, dass ein Computer eingesetzt wird, zu (technokratischer) Angst, Zurückhaltung, Skepsis, Vorbehalten und schließlich Ablehnung führen (Butcher, 1987; Hornke, 1993; Jäger & Krieger, 1994). Wird der Einsatz von spezifischer Software als „undurchschaubar“ erlebt, so entsteht Angst vor Kontrollverlust (Butcher, 1987). Da zunehmend auch „Fachfremde“ (Mediziner, Informatiker, Mathematiker, Laien aus der Privatwirtschaft etc.) computergestützte Tests entwickeln, ist die Gefahr einer Entprofessionalisierung (Schötzau-Fürwentsches & Grubitzsch, 1991) nicht von der Hand zu weisen. Auch eine gewisse Selbstwertbedrohung (Garb, 2000) scheint verständlich, wenn die Sorge entsteht, durch einen Computer ersetzt zu werden (Butcher, 1987; Gregory, 1996) und in der jeweiligen Institution nicht darauf fokussiert wird, die durch den Computereinsatz frei gewordenen Personalressourcen für wichtigere, interessantere und kreativere (z. B. therapeutische) als rein administrative Aufgaben zu nutzen (siehe Kapitel 4.2.3.).

Neben diesen potentiellen negativen Auswirkungen auf (a) den Diagnostiker müssen auch mögliche Nachteile für (b) die *Testpersonen* diskutiert werden. Kubinger (1993) weist beispielsweise auf die Möglichkeit einer ungewollten psychischen Stressinduktion hin, räumt aber ein, dass bislang empirisch nicht belegt werden konnte, dass Testpersonen sich subjektiv durch den Computereinsatz überfordert fühlen. Weiterhin beklagen einige Autoren (Butcher, Keller & Bacon, 1985; Kubinger, 1999), dass Variablen der sozialen Interaktion (z. B. durch Verhaltensbeobachtungen) bei der Anwendung von computergestützten Tests nicht erfasst werden. Dem ist entgegen zu halten, dass bei den klassischen Papier-und-Bleistift-Tests (ausgenommen projektiven Verfahren) Verhaltensbeobachtungen der Testpersonen ebenfalls nicht standardisiert gesammelt werden, sondern höchstens ein subjektiver Eindruck der Testbearbeitung beim Diagnostiker entsteht.

Ein wichtiger Faktor, den es in diesem Zusammenhang zu berücksichtigen gilt und der häufig befürchtet wird, ist eine mögliche Abhängigkeit zwischen Testergebnis und Computer*erfahrung*. Erste Untersuchungen weisen darauf hin, dass nach der vorangegangenen Applikation eines entsprechenden Lernprogramms zum Gebrauch der Software keine signifikanten Testniveauunterschiede zwischen Personen mit und ohne Computererfahrung resultieren (Hergovich, 1992). Hier ist jedoch besonders im Leistungsbereich weitere Forschung nötig.

Potentielle Gefahren im Hinblick auf die Testfairness sollten stets reflektiert werden. So gibt Kubinger (1993) zu bedenken, dass ethische, kulturelle, geschlechtsspezifische und sensorische Faktoren ein Testergebnis verzerren können. Interessant ist die These, dass durch die rein visuelle Darbietung der Testinstruktion beim computergestützten Testen möglicherweise „auditive" Wahrnehmungstypen diskriminiert werden könnten, da die Instruktion computergestützter Tests nur visuell, Papier-und-Bleistift-Testinstruktionen jedoch in der Regel auditiv *und* visuell erfolgen.

Schließlich mag der Computereinsatz, wie Kubinger (1999) vermutet, dazu führen, dass Items weniger sorgfältig bearbeitet werden als in Papier-und-Bleistift-Testversionen, d. h. der Computereinsatz per se zu vorschnellen Antworten und Überlesen verleiten kann. Dies führt zur dritten groben Klasse der Nachteile: die Gefahr der Verringerung der Datenqualität (c).

Diese droht, wenn...

1. entwickelte computergestützte Tests nicht ausreichend validiert werden (Gregory, 1996),
2. unkritisch Normen von Papier-und-Bleistift-Tests auf die vermeintlich äquivalente Computerversion übertragen werden (zur Äquivalenz-forschung siehe u. a. Mead & Drasgow, 1993; Kubinger, 1993, Jäger & Krieger, 1994; Rose et al., 1999, 2003; Schwenkmezger & Hank, 1993),
3. sich durch den Einsatz eines fehlerhaften Computer-Programms wieder-holt Fehler reproduzieren (Schötzau-Fürwentsches & Grubitzsch, 1991),
4. ein Computerausdruck gerade bei Kenntnismangel und unter Zeitdruck dazu verleitet, „blind" der Technik zu vertrauen, da er autorisiert (auch

ohne Unterschrift > Diffusion der Verantwortlichkeit; Butcher, 1987; Gregory, 1996; Schötzau-Fürwentsches & Grubitzsch, 1991) erscheint.

Insbesondere der letzte Punkt ist eng mit der Gefahr eines Testmissbrauchs verknüpft, der im medizinischen Bereich dadurch provoziert werden kann, dass Mediziner Psychodiagnostik als einen Gebührenposten kassenärztlich „abrechnen“ können (Computerausdrucke werden hier also im doppelten Sinne als „bare Münze“ genommen; Schötzau-Fürwentsches & Grubitzsch, 1991, S. 309).

Da keine strikten berufspolitischen juristischen Grenzen zum Gebrauch von computergestützten Tests existieren, ist auch die Gefahr des Missbrauchs gegeben. Diese ist jedoch nicht nur auf computergestützte Tests beschränkt, sondern gilt gleichermaßen auch für Papier-und-Bleistift-Tests.

Ein Aspekt, der jüngst im Zeitalter der Computer-Hacker und Wireless Local Area Networks (LAN) psychometrischer Daten bei computergestützten Tests in den Vordergrund gerückt wird, ist der der Datensicherheit (Gregory, 1996). Allgemein muss speziell bei der Benutzung von institutionseigenen Netzwerken diese weitestgehend durch Datenverschlüsselungen und Zugriffsbegrenzungen (Passwords) gewährleistet sein.

4.2.3. Zum Umgang mit computergestützten Tests

Da für die Entwicklung von computergestützten Tests oftmals nicht nur psychologisches Fachwissen, sondern auch Fachwissen aus der Medizin, Mathematik und Informatik benötigt wird, implizieren Gedanken über computergestützte Tests auch berufspolitische Überlegungen. Schötzau-Fürwentsches und Grubitzsch (1991) betonen in Übereinstimmung mit einem Großteil von Psychodiagnostikern, dass unabdingbare Voraussetzung für die Anwendung psychodiagnostischer Verfahren (hier speziell computergestützter Tests) eine qualifizierte psychologische Ausbildung sei. Auf eine wissenschaftlich abgesicherte und fundierte computergestützte Psycho-diagnostik wurde schon vor mehr als 30 Jahren großer Wert gelegt. So formulierten 1986 das Testkuratorium und die American Psychological Association (APA) zeitgleich Richtlinien zur computergestützten Diagnostik (APA, 1986; Testkuratorium, 1986). In ihnen wird auf die Bedeutung eines wohlüberlegten, verantwortungsbewussten, nachvollziehbaren, transparenten und reflektierten Umgangs mit computergestützten Tests hingewiesen und

Empfehlungen in Bezug auf die Kontrolle und Bewertung von Ergebnissen ausgesprochen.

Mehrere Autoren (Jäger & Krieger, 1994; Wetzler & Marlowe, 1994) betonen in diesem Zusammenhang, dass der Computer lediglich ein technisches Hilfsmittel im Rahmen des diagnostischen Prozesses darstelle, welches bei begründeter Indikation als Ausgangspunkt der diagnostischen Hypothesenbildung fungieren könne. Der Computereinsatz solle einseitig abhängig vom Urteil des Psychodiagnostikers sein und keinen Selbstzweck erfüllen, sondern im Interesse der Testperson(en) stattfinden. Ergebnisse sind persönlich, gruppiert nach Konstrukten, verständlich auf Item- und Skalenniveau mit der Angabe von Vergleichsgruppen/-werten ökonomisch und für den Laien verständlich rückzumelden. Die unreflektierte Anwendung undurchschaubarer von Laien entwickelter computergestützter Tests, die einer „black box“ ähneln, verbiete sich, und die Verwendung automatisierter nicht valider Interpretationsprogramme sei zu vermeiden (Jäger & Krieger, 1994). Letztendlich ist jeder Testentwickler von computergestützten Tests (bzw. CATs) herausgefordert, qualitativ hochwertige Tests nach wissenschaftlichen Kriterien in transparenter Weise zu konstruieren und zu validieren, sowie die Soft- und Hardware leicht verständlich und benutzerfreundlich zu gestalten. Der wissenschaftlichen Fundierung computergestützter Psychodiagnostik kommt in jedem Fall das Primat über technische Überlegungen zu.

4.2.4. Computergestützte Tests zur Angstmessung

Im deutschen Sprachraum existieren bereits eine Reihe von computergestützten Testverfahren zur Angstmessung, welche auf den Prinzipien der KTT entwickelt wurden. Im Rahmen des Computerbasierten Ratingsystems zur Psychopathologie (CORA, Hänsgen & Merten, 1994) liegen computergestützte Versionen der folgenden fünf Fragebögen vor:

- Hamilton-Angst-Skala (HAMA; Hamilton, 1959, 1977),
- Selbstbeurteilungs-Angst-Skala (SAS; Collegium-Internationale-Psychiatriae-Scalarum (CIPS), 1996),
- Interaktions-Angst-Fragebogen (IAF; Becker, 1997),
- State-Trait-Angst-Inventar (STAI-State; Laux et al., 1981),
- Fragebogen zur Angst vor körperlichen Symptomen.

4.3. Computergestütztes Adaptives Testen (CAT)

4.3.1. Einleitung

Das allgemeine Prinzip einer Adaptivität / Adaptation (= Anpassung) findet sich in der psychologischen Diagnostik auf zwei verschiedenen Ebenen realisiert. So kommen nach Kisser (1995) adaptive Strategien auf einer „Makroebene“ zum Einsatz, wenn die Auswahl der Untersuchungsbereiche (z. B. Fähigkeiten, Einstellungen) und die Art und Reihenfolge einzusetzender Erhebungsinstrumente (Fragebogen, Verhaltensbeobachtung, Interview,...) von spezifischen diagnostischen Fragestellungen abhängig gemacht wird. Ein Diagnostiker sollte demnach im Idealfall sein diagnostisches (und damit treatmententscheidendes) Vorgehen dem individuellen Fall „anpassen“.

Auf der „Mikroebene“ ist Adaptivität gegeben, wenn die Darbietung einzelner Fragen, Experimente und Testaufgaben an den Einzelfall angepasst wird. Die Grundidee des adaptiven Testens besteht in der Annahme, dass ein Test am besten misst, wenn der Testperson im Laufe eines Tests genau diejenigen Fragen (= Items) dargeboten werden, welche über die Testleistung der Testperson das meiste aussagen, welche also am „informativsten“ für die Diagnostik sind.

Daraus ergibt sich die Frage, welche Items am „informativsten“ (und übrigens auch am subjektiv interessantesten / motivierendsten) für eine Person sind.

Nach Birnbaum (1968) sind es diejenigen Fragen / Aufgaben, welche einen mittleren Schwierigkeitsgrad für eine spezifische Person aufweisen. Da die Einschätzung der mittleren Schwierigkeit einer Testaufgabe von der individuellen Fähigkeit abhängt, wird die mittlere Schwierigkeit allgemein in Abhängigkeit von der Lösungswahrscheinlichkeit einer Testaufgabe definiert. So besitzt ein Item i für eine bestimmte Person j eine mittlere Schwierigkeit, wenn die Wahrscheinlichkeit einer Person j dieses Item i zu lösen p_{ij} (richtig) = 0,5 entspricht, d. h. wenn es gleich wahrscheinlich ist, dass die Testperson das Item löst bzw. nicht löst (p_{ij} (richtig) = p_{ij} (falsch) = 0,5; Birnbaum, 1968). Hier zeigt sich bereits, dass die Wahrscheinlichkeitstheorie eine wesentliche Grundlage des adaptiven Testens darstellt, weshalb IRT-basierte Tests von manchen Autoren auch als Realisierungen eines „stochastischen Testdesigns“ (Wainer, 1990, S. 130) bezeichnet werden.

Es lässt sich zusammenfassen, dass beim adaptiven Testen eine Anpassung der Itemdarbietung an das Fähigkeitsniveau einer Testperson wie folgt geschieht:

> „Adaptives Testen ist interaktiv, indem Testpersonen diejenigen Items dargeboten werden, von denen man auf der Grundlage des Wissens um die Beantwortung bereits beantworteter Items annimmt, dass sie für die zu testende Person am informativsten sind." (Freie Übersetzung nach Embretson, 1992, S. 129)

Konkret folgt daraus folgendes strategisches Vorgehen:

> Wenn die Testperson ein Item „falsch" beantwortet, wird ihr als nächstes ein „einfacheres" Item gestellt, antwortet die Testperson auf das Item hingegen „richtig" wird ein „schwierigeres" Item dargeboten.

Die Anfänge des adaptiven Testens finden sich zu Beginn des letzten Jahrhunderts in Frankreich, wo Binet 1909 einen adaptiven Papier-und-Bleistift-Test zur Messung von Intelligenz im Rahmen der Schuleignungsdiagnostik (Pädagogik) entwickelte. Er realisierte eine sogenannte „upward / downward"-Strategie (Gregory, 1996, S. 589), bei der für jede Testperson eine „obere" und „untere" Fähigkeitsgrenze erhoben wurde, indem jeder Testperson einerseits so lange immer schwierigere Items gestellt wurden, bis sie eine bestimmte Anzahl von Testaufgaben mit gleicher Schwierigkeit immer falsch beantwortete („upward"), und andererseits einer Testperson so lange immer leichtere Items gestellt wurden, bis sie eine bestimmte Anzahl von Testaufgaben mit gleicher Schwierigkeit immer richtig beantwortete („downward"; zu unterschiedlichen Formen adaptiven Testens siehe Kapitel 4.3.2.). Dieser Intelligenztest blieb lange Zeit der einzige adaptive Test seiner Art, bis in den 60er Jahren durch das Aufkommen der Item Response Theorie (IRT, siehe Kapitel 3) und der rapiden technischen Entwicklung von Computern ein idealer Nährboden für die weitere Erforschung von Computergestützten Adaptiven Tests (CATs) entstand. Im Rahmen eines umfangreichen Forschungsprogramms verfolgte als erster Forscher Lord (1980) in den 60er Jahren die Entwicklung von IRT-basierten CATs in der Schuleignungsdiagnostik in den U.S.A. (Educational Testing Service). Dies initiierte unterstützt von dem U.S. Armed Services und der U.S.

Civil Service Commission (Hambleton & Zaal, 1990) die Entwicklung einer Reihe weiterer IRT-basierter computergestützter adaptiver Leistungs- und Eignungstests (Scholastic Aptitude Test, SAT; California Achievement Tests, CAT; Stanford Achievement Tests and the Woodcock-Johnson-Psycho-Educational-Battery).

Dabei impliziert adaptives Testen per se nicht den Einsatz eines Computers. So wurde der erste adaptive Test in der Leistungsdiagnostik wie eingangs erwähnt als Papier-und-Bleistift-Verfahren entwickelt (IQ-Test von Binet, 1909).

Computer erleichtern jedoch aufgrund ihrer hohen Rechen- und Speicherkapazität (besonders bei der Anwendung von IRT-basierten Tests ist diese aufgrund der hohen Rechenanforderungen beinahe unabdingbar) das adaptive Testen ungemein.

Dabei dient der Computer folgenden Aufgaben (Weiss & Vale, 1987):

- Selektion der Items,
- Präsentation der Items,
- Registrierung der Itemantwort,
- Berechnung eines Fähigkeitsscores (während der Testdarbietung),
- Beenden des Tests.

4.3.2. Varianten des Adaptiven Testens

Seit den 70er Jahren entwickelten sich eine Reihe von verschiedenen Formen adaptiver Tests, denen gemein ist, dass sie den „Spagat“ zwischen Individual- und Gruppendiagnostik zu lösen versuchen, indem sie über eine große Itemzahl verfügen (Itembank), welche alle Schwierigkeitsgrade abdecken sollten und aus deren Menge jeweils die Items ausgewählt und dargeboten werden, welche dem Fähigkeitsniveau einer Person optimal entsprechen („tailored testing“: maßgeschneidertes Testen; Weiss, 1985).

Die bislang entwickelten adaptiven Tests, welche teilweise in Papier-und-Bleistift-Format und teilweise in Form von CATs vorliegen, können in verschiedene Gruppen klassifiziert werden, welche sich in ihrer Art der Realisierung der Adaptivität unterscheiden. Die folgende Abbildung 6 gibt einen groben Überblick über die verschiedenen Formen adaptiver Tests.

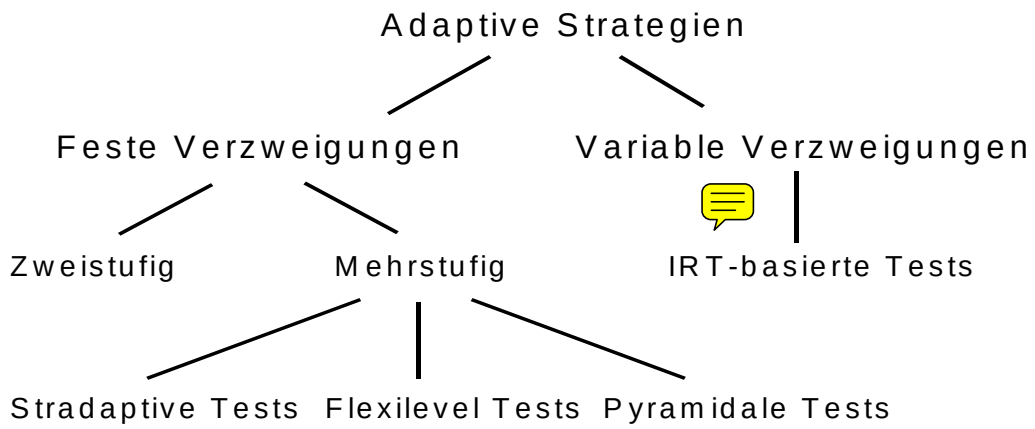


Abbildung 6: Überblick über verschiedene Formen von adaptiven Testsstrategien.

Allgemein lassen sich zwei grundlegende adaptive Teststrategien unterscheiden: Tests beruhend auf festen (vorher fixierten) Verzweigungsstrukturen, welche die Itemauswahl bestimmen, und Tests mit variablen Verzweigungswegen, die auf der Grundlage der Item Response Theorie (IRT) berechnet werden.

Im Folgenden wird zunächst das Grundprinzip von Tests mit *festen* Verzweigungsstrukturen vorgestellt, bevor der Schwerpunkt auf die Testform mit *variablen* Verzweigungswegen gelegt wird, welche in vorliegender Dissertation realisiert wurde: ein IRT-basierter CAT (zur IRT siehe Kapitel 3).

Adaptive Tests, welche sich feste Verzweigungsstrategien zunutze machen („branching tests“; Thissen & Mislevy, 1990), beruhen auf einer durch die Schwierigkeit von Items festgelegten Struktur und Hierarchisierung des Itempools, d. h. diesen Tests liegt ein statisches Verzweigungsschema, zugrunde, welches während der Testkonstruktion entwickelt wurde. Adaptive Tests mit festen Verzweigungen können in Zweistufige und Mehrstufige unterschieden werden. *Zweistufige* fest verzweigte adaptive Tests sind minimal adaptiv („two stage procedure“; Lord, 1980; Hambleton & Zaal, 1990).

Sie bestehen meist aus einem anfänglichen Set von Screening-Aufgaben, welche alle Schwierigkeitsgrade grob abdecken („routing test“), und einem in Abhängigkeit von den Antworten auf diese Anfangsaufgaben nachgeschalteten für die Testperson optimalen Subset von Fragen, das am besten dem (vor-) ermitteltem Fähigkeitsniveau entspricht, und damit eine differenziertere (End-) Testung erlaubt.

Unter *mehrstufigen* adaptiven Tests mit festen Verzweigungsregeln versteht man klassischerweise Tests, welche sich durch Verzweigungen auf der Itemebene auszeichnen (denkbar sind aber auch Verzweigungen auf der Skalenebene). Hier kann entweder anhand inhaltlicher Gesichtspunkte die Itemmenge so strukturiert sein, dass eine Gruppe von Items einem spezifischen Inhaltsbereich („testlet“) angehört, so dass der Itempool in verschiedene Subsets von Items geordnet werden kann („stratified / stradaptive Tests“; Lord, 1980), welche je nach „Anpassung“ bearbeitet werden, oder die Strukturierung der Items erfolgt in Abhängigkeit von der Schwierigkeit. Letzteres ist das grundlegende Prinzip der „flexilevel Tests“ und der „pyramidalen Tests“ (Lord, 1980). *Flexilevel* Tests verfügen über jeweils *ein* Item auf jeder Schwierigkeitsstufe. Die Itempräsentation beginnt mit einem mittelschwierigen Item und vollzieht sich entweder in Richtung schwierigere („downward“) oder leichtere („upward“) Items (Binet, 1909). Durch dieses Vorgehen kann ein Test in seiner Testlänge halbiert werden. *Pyramidalen* Tests liegt eine pyramidenartige Strukturierung des Itempools zugrunde, da sie über *mehrere* Items pro Schwierigkeitsstufe verfügen, und damit die Itemauswahl in Form eines „Entscheidungsbaumes“ mit multiplen Verzweigungen die rein binäre Itemauswahlstrategie der Flexilevel Tests übertreffen (z. B. Adaptives-Intelligenz-Diagnostikum, AID; Kubinger & Wurst, 2000).

Natürlich wurden in der Vergangenheit noch eine Reihe weiterer Formen adaptiver Tests („Robbins-Monro branching method“; „Implied orders tailored testing“ etc.) erprobt. In jüngster Vergangenheit seien hier interessante Ansätze, bei denen die Itembankstrukturierung theoriegeleitet nach Prinzipien der strukturellen Informationstheorie erfolgte (Guthke, Räder, Caruso & Schmidt, 1991), sowie ein Ansatz erwähnt, der sich das methodische Prinzip des „Cluster-Branchings“ als Grundlage der Itembankstrukturierung zunutze machte (Laatsch & Choca, 1994). Abgesehen von diesen Publikationen finden sich jedoch in diesem Forschungsfeld vor allem eher veraltete adaptive Ansätze, welche zum Teil verworfen wurden bzw. heute nur noch von historischem Wert sind. Daher wird hier auf eine ausführliche Darstellung dieser verzichtet (für einen historischen Überblick wird Lord, 1980, empfohlen).[34]

[34] Desweiteren finden bei Butcher und Mitarbeiter (1985) allgemeine adaptive Teststrategien Erwähnung, welche vor allem das Ziel verfolgen, Testpersonen zu klassifizieren, so z. B. die „Countdown Strategie“, welche eine Testung von Personen impliziert bis ein „Cut Score“

Zusammenfassend ist die grundlegende Gemeinsamkeit adaptiver Tests mit festen Verzweigungen ein nach der Item*schwierigkeit* (andere Itemparameter wie z. B. die Iteminformation bei einer IRT-basierten CAT-Anwendung, siehe Kapitel 4.3.3.3., werden nicht genutzt) vorstrukturierter Itempool, der die Grundlage der Itemauswahl bildet. Meist ist die Testlänge auf eine bestimmte dargebotene Itemanzahl fixiert und nicht durch eine logische Stoppfunktion (wie z. B. durch ein bestimmtes Messgenauigkeitskriterium wie bei IRT-basierten CATs siehe Kapitel 3.3.3.) begründet. Weiterhin nachteilig erscheint, dass dem adaptiven Testprozess keine gemeinsame Metrik (wie bei IRT-basierten CATs) zugrunde liegt, was die Vergleichbarkeit der Testergebnisse im strengen Sinn unmöglich macht.

Die IRT vermag diese drei „Mängel“ der fixierten adaptiven Tests zu beheben, da sie folgende Möglichkeiten eröffnet:

1. die Berechnung mehrerer Itemparameter:
   - Implikation: Nutzung derselben zur gezielten Itemauswahl;
2. die Berechnung von Messgenauigkeiten (bzw. Reliabilitäten) in Abhängigkeit zur Merkmalsausprägung:
   - Implikation: Nutzung dieser als Stoppfunktion;
3. die Positionierung von Items und Personen auf einer gemeinsamen Metrik:
   - Implikation: Vergleichbarkeit von Testergebnissen.

Obgleich im folgenden Kapitel zunächst auf die methodischen Grundzüge IRT-basierter CATs fokussiert wird, sei schon anhand der drei beschriebenen Potentiale der IRT hervorgehoben, dass diese „neue“ Testtheorie seit ihrer Entstehung als die eleganteste (und aufwendigste) Methodologie bei der Realisierung von CATs gilt (zur IRT siehe Kapitel 3).

---

erreicht wird, und der „adaptiv typologische Ansatz“, bei dem aufgrund von vorausgehenden Item-Diskriminationsanalysen Items in Form von Stufentests vorgeschlagen werden. Beide adaptive Formen finden kaum Anwendung.

4.3.3. Grundzüge IRT-basierter CATs

Die Wurzeln IRT-basierter CATs finden sich bei Lord und Novick (1968), welche durch ein bahnbrechendes Textbuch, mit einem Kapitel von Rasch und vier Kapiteln[35] von Birnbaum (1968), die statistischen Grundlagen der stochastischen Testtheorie in die psychologische Forschung einführten und damit den Grundstein der IRT legten (Wainer, 1990). Die IRT bietet als eine „Familie“ mathematischer Modelle eine kohärente Methodologie, welche das Testverhalten einer Person zu beschreiben versucht, und die Berechnung von Itemcharakteristiken ermöglicht, die über die konventionellen Statistiken bei der Testkonstruktion auf der Basis der Klassischen Test-Theorie (KTT) hinausgehen (siehe Kapitel 3.).

Durch die Anwendung der IRT zur Testkonstruktion können - verglichen mit den in Kapitel 4.3.2. erörterten verschiedenen Formen adaptiver Strategien - mögliche Gewinne von computergestützten adaptiven Tests maximiert werden.

Charakteristisch für CATs ist, dass eine spezifische Interaktionsregel zwischen Computer und Testperson eingehalten wird, die lautet: „Präsentiere dem Pbn nur solche Items, die geeignet für ihn sind!“ (Hornke, 1994, S. 321). Um die Itemeignung bei *IRT*-basierten CATs zu bestimmen, sind in der Regel umfangreiche (Vor-) Kalibrierungsuntersuchungen an den später zu präsentierenden Items nötig (Kubinger, 1996). Sie dienen der Berechnung von Itemcharakteristiken, welche folgendermaßen genutzt werden können:

1. zur Selektion der „besten“ Items für die Itembank,
2. zur Programmierung des Itemselektionsalgorithmus und
3. zur Berechnung des Skalenwertes einer Person (Personenparameterschätzung).

Der Veranschaulichung eines IRT-basierten computergestützten adaptiven Testablaufs dient Abbildung 7, welche im Folgenden erläutert wird.

Die Numm[illegible]n im Text beziehen sich auf die Nummern in der Abbildung:

(1.) Die initiale Skalenberechnung geht z.B. von dem Mittelwert der klinischen Population aus ($\theta_0 = 0$; zur Startfunktion siehe Kapitel 4.3.3.2.). (2.) Die Wahl des ersten Items in der Regel auf ein Item, welches mit seinen Antwortalternativen in diesem Bereich die höchste Information verspricht

[35] Textbuch von Lord & Novick (1968): Kapitel 17-20 von Birnbaum, Kapitel 21 von Rasch.

(z. B Fisher-Information, zur Itemselektionsstrategie siehe Kapitel 4.3.3.3.). Nach (3.) der Auswahl einer Antwortalternative auf das erste Items durch die Testperson, wird (4.) der aktuelle Skalenwert anhand eines Personenparameter-Schätzalgorithmus (siehe Kapitel 4.3.3.4.) und das Messgenauigkeitsniveau der jeweiligen Schätzung berechnet. Eine dementsprechende Itemdarbietung und Neuschätzung des Skalenwertes geschieht iterativ und sukzessiv bis (5.) eine bestimmte Stoppfunktion, wie z.B. die maximale Anzahl von Items dargeboten wurde und / oder die Messpräzision hinreichend erfüllt ist. Dann wird (6.) der CAT-Prozess beendet (siehe Kapitel 4.3.3.6.). (7.) Ist die Skala Teil einer Testbatterie so wird (8.) die nächste Skala zur Messung eines weiteren Konstruktes ausgewählt. Wird nur eine Skala in einem CAT-Prozess angewandt, so wird (9.) der CAT-Prozess nach Erfüllung des Stoppkriteriums beendet.

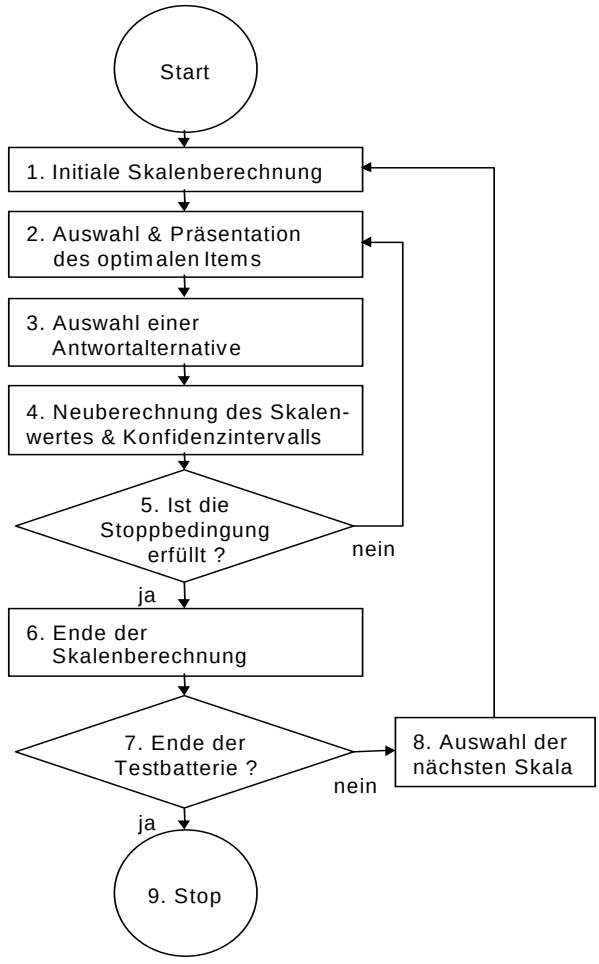


Abbildung 7: Flussdiagramm eines IRT-basierten computergestützten adaptiven Testprozesses (Wainer, 1990, S. 108).

Zusammenfassend lässt sich sagen, dass für IRT-basierte CATs folgende Aspekte charakteristisch sind:

1. die *sofortige* Registrierung jeder einzelnen Itemantwort,
2. die *iterative* Neuschätzung des Personenparameters mit Hilfe der Itemantwort(en) und der Itemcharakteristiken,
3. die iterative Auswahl des informativsten Items der erzielten Schätzung,
4. die iterative Bestimmung des Konfidenzintervalls der erzielten Schätzung,
5. die regelgeleitete Entscheidung über Fortsetzung oder Abbruch der Testung,
6. die finale modellbasierte Personenparameterschätzung stellt das Testergebnis dar.

Im Folgenden werden einige der bereits eingeführten Themen IRT-basierter CATs näher fokussiert.

4.3.3.1. Itembank

Der Güte der Itembank kommt bei der Entwicklung eines CATs eine zentrale Rolle zu. So kann nach Embretson und Reise (2000) ein CAT nur so gut sein wie seine Itembank, d. h. die Güte der Itembank entscheidet letztendlich über die Effektivität des CATs.

Leider existieren in der Psychologie wenig einheitliche Regeln, nach denen bei der Testkonstruktion vorgegangen werden sollte. Embretson und Reise (2000) unterscheiden drei Testkonstruktionsansätze: a) den „empirical keying approach“, welcher sich auf die Vorhersage von Verhalten von Probanden fokussiert, jedoch ohne einen unidimensionalen Messanspruch zu stellen; b) den „construct approach“, darunter wird der traditionelle Testkonstruktions-ansatz - wie er im Rahmen der Klassischen Test-Theorie (KTT) favorisiert wird - verstanden (bestehend aus der Berechnung von Faktorenanalysen, Inter-Item- und Item-Test-Korrelationen etc.), und c) eine IRT-basierte Skalenkonstruktion, welche eine umfangreiche Kalibrierung von IRT-Parametern an einer zuvor erhobenen *Kalibrierungsstichprobe* umfasst.

Ein Vorteil IRT-basierter Itembanken gegenüber KTT-basierten Itempools liegt in dem Potential, Items mit unterschiedlichen Antwortformaten auf einer Skala zu integrieren. Ein Nachteil ist mit dem Umstand der Kalibrierung verknüpft.

Da eine der Anforderungen an eine „gute“ Itembank ihre Größe ist, ist das eigentlich ideale Vorgehen, speziell für den CAT neue Items zu entwickeln, oft aufgrund des damit verknüpften großen Erhebungsaufwandes nicht realisierbar. In der Praxis folgt man der Annahme, dass in der Regel schon ein potentiell guter Itempool für die Erfassung bestimmter Konstrukte (d. h. gute Indikatoren für das latente Trait) geschrieben wurde (z. B. Items aus KTT-basierten Fragebögen; Weiss, 1985; Embretson & Reise, 2000), der - falls er bereits an einer ausreichend großen Kalibrierungsstichprobe erhoben wurde - zur Berechnung IRT-basierter Parameter genutzt werden kann. Dabei sind die Anforderungen, welche an eine *Kalibrierungsstichprobe* gestellt werden, nach Embretson und Reise (2000) nicht sehr hoch. Die Kalibrierungsstichprobe (von Personen) muss nicht repräsentativ sein (aufgrund der in der IRT formulierten Unabhängigkeit der Item- und Personenparameterschätzung) und darf bzw. sollte möglichst heterogen in Bezug auf das zu messende Merkmal sein.

Während die Anforderungen an die Kalibrierungsstichprobe gering erscheinen, existieren eine Reihe von strengen psychometrischen Anforderungen an eine „gute“ Itemstichprobe (*Itembank)*, welche nach folgenden Aspekten zusammengefasst werden (Hambleton & Zaal, 1990; Wainer, 1990; Weiss, 1985; Embretson & Reise, 2000):

1. Größe der Itembank,
2. Homogenität der Itembank,
3. Erfassung eines *weiten* Bereichs des Merkmalsausprägungskontinuums,
4. Hohe Diskriminationsfähigkeit der Items,
5. Ausschluss „schlechter“ Items,
6. Validität der Itembank.

Für die erwünschte *Größe* der Itembank liegen bisher nur Erfahrungswerte aus der Leistungsdiagnostik vor. Hier rät Weiss (1985) zu Itemmengen von N_{items} = 100-200, Hornke (1993) zu Itemmengen von N_{Items} = 70-200, während Embretson und Reise (2000) N_{Items} = 100 empfehlen, jedoch darauf hinweisen, dass für den Bereich der Persönlichkeitsdiagnostik weniger Items nötig seien, da diese in der Regel ein polytomes Antwortformat haben (Dodd, De Ayala & Koch, 1995; Master & Evans, 1986).

Weiterhin ist die *Homogenität* einer Itembank speziell bei der Entwicklung eines unidimensionalen CATs zentral. Diese kann durch die Selektion anhand von

inhaltlichen Itemtext-Kriterien (durch Expertenurteile), sowie mittels Unidimensionalitätsüberprüfungen (Faktorenanalysen, Analysen residualer Kovarianzen) gewährleistet werden. Schließlich ist die Erfassung eines *weiten* Bereichs des *Merkmalsausprägungsspektrums* vor allem dann erwünscht, wenn es sich um die Konstruktion eines sogenannten „equal precise“ Tests handelt, also ein Test entwickelt werden soll, der anstrebt, die Merkmalsausprägung von Personen unterschiedlicher Ausprägungsniveaus gleich gut zu messen. Diese Anforderung muss nicht erfüllt werden im Falle sogenannter „peaked“ Tests (kriteriumsbasierter Tests), welche das Ziel verfolgen, Personen anhand eines bestimmten computergestützten Testscores (Kriteriumswertes) in zwei Gruppen zu klassifizieren. In diesem Fall wären nur Items mit einer hohen Information um den Kriteriumstestwert nötig (Embretson & Reise, 2000).

Die Anforderung einer hohen *Diskriminationsfähigkeit* versteht sich vor diesem Hintergrund von selbst. Schwieriger gestaltet sich schon der Ausschluss „*schlechter*“ Items. Denn es gibt in der IRT-Entwicklung von Itembanken bisher noch keine einheitlichen Bewertungsstandards der Qualität von Items. So können sich Selektionskriterien einerseits auf die Überprüfung der Unidimensionalität, die Kontrolle der Diskriminationsfähigkeit, die „Passung“ an das ausgewählte IRT-Modell („Modell-Fit“) oder ähnliches beziehen. Weitere Forschung ist in diesem Feld dringend erforderlich.

Einig sind sich die meisten Forscher, dass die Itembank eines CATs einer umfangreichen *Validierung* unterzogen werden sollte, um sicher zu stellen, dass das CAT wirklich das misst, was es zu messen vorgibt (siehe Kapitel 6.).

Zusammenfassend ist hervorzuheben, dass speziell bei CATs hohe Anforderungen an die Items gestellt werden, da durch die adaptive Reduktion der Testlänge „schlechte“ Items vor allem zu Beginn der Testung den Testverlauf stärker negativ beeinflussen können als bei konventionellen Tests (Wainer, 1990). Allerdings bieten IRT-basierte CATs die Möglichkeit, ihre bestehenden Itembanken kontinuierlich über das Hinzufügen speziell „gezüchteter guter“ Items (durch sogenannte Item-Link-Designs; siehe Kapitel 3.3.3. und 5.3.2.3.3.) und den Ausschluss „schlechter“ Items zu verbessern.

4.3.3.2. Startfunktion

Je kürzer ein adaptiver Test ist, desto mehr Einfluss hat das erste dargebotene Item auf das Messergebnis (Lord, 1980). Aus diesem Grund wird der Startfunktion an dieser Stelle ein eigenes Unterkapitel gewidmet. Nach Embretson und Reise (2000) existieren drei Möglichkeiten, wie ein CAT begründeterweise beginnen kann:

a) mit der Darbietung eines leichten Items,
b) mit der Darbietung eines Items in Abhängigkeit vom Vorwissen,
c) mit der Darbietung eines Items mit mittlerer Schwierigkeit.

Die Darbietung eines leichten Items als „Start-Item“ bei Leistungstests wird von Wainer und Kiely (1987) empfohlen. Indem Frustrationen durch die anfängliche Vermeidung der Darbietung schwerer Items vermieden werden, solle sich die initiale Testangst reduzieren. Zudem sollte bei Leistungstests darauf geachtet werden, dass das erste Item keinem Lerneffekt unterliegen kann, so dass es bei Retests nicht in seiner Aussagekraft reduziert ist.

Eine Präsentation des ersten Items in Abhängigkeit vom Vorwissen aus einer vorangegangenen Testung erscheint sinnvoll, um Redundanz in Mehrfachmessungen zu vermeiden. Da jedoch in den meisten Fällen kein Vorwissen um die Merkmalsausprägung einer Testperson besteht, werden CATs in den meisten Fällen mit der Darbietung eines Items mittlerer Schwierigkeit begonnen. Dies ist vor dem Hintergrund der Annahme einer Normalverteilung der Merkmalsausprägung insofern sinnvoll, da ein Item mittlerer Schwierigkeit initial die beste Schätzung der Merkmalsausprägung erlaubt (Thissen & Mislevy, 1990).

4.3.3.3. Itemselektion

Der Itemselektion liegt in der Regel einer von mehreren möglichen Algorithmen zugrunde, welche speziell für IRT-basierte CATs als Software entwickelt (programmiert) werden müssen. Nach Thissen und Mislevy (1990, S. 103) werden derartige Algorithmen als Regelwerk definiert, welches festlegt, welche Fragen in welcher Reihenfolge von Probanden beantworten werden sollen.

Es lassen sich bei IRT-basierten CATs zwei grundlegende Algorithmen / Verfahren der Itemselektion unterscheiden:[36] [37]

[36] Für einen Überblick über verschiedene Itemselektionsverfahren siehe Thissen und Mislevy (1990) sowie Schnipke und Green (1995).

1. das Maximum-Information-Verfahren (MI) und
2. das Bayes'sche Sequentialverfahren (BE).

Die Idee des Maximum-Information-Verfahrens (MI) stammt wahrscheinlich ursprünglich von Urry (1977)[38], der vorschlug, immer diejenigen Items zu präsentieren, welche für die jeweilige Schätzung der Merkmalsausprägung die höchste Iteminformation aufweisen (d. h. p_{ij} (richtig) = 0,5; entspricht einer mittleren Itemschwierigkeit). Die Iteminformation (meist: Fisher-Information, möglich ist aber auch die Kullback-Leibler Information o. ä.) entnimmt der Computer entweder einer vorher an einer Kalibrierungsstichprobe berechneten Iteminformationstabelle oder er errechnet die Iteminformation simultan während des computergestützten adaptiven Prozesses. Die erste Realisierung des MI-Verfahrens erfolgte im Jahre 1977 durch Brown und Weiss, welche diese Itemselektionsstrategie (mittels eines Rückgriffs auf eine Iteminformations-tabelle durch den Testadministrator) in Papier-und-Bleistift-Format umsetzten. Um zu vermeiden, dass ein Item mehrfach dargeboten wird, da es in mehreren Bereichen die höchste Iteminformation besitzt, kann dieses Verfahren so abgewandelt werden, dass eine Zufallsauswahl des „besten" Items pro Schwierigkeitsbereich realisiert wird. Dies setzt jedoch voraus, dass mehrere Items mit einem ähnlich hohen Informationsgehalt pro Schwierigkeitsbereich in der Itembank vorliegen. Veerkamp und Berger (1997) schlagen eine Abwandlung des Selektionsalgorithmus vor, in dem die Items mit jeweils der höchsten mittleren Information innerhalb eines bestimmten Konfidenzintervalls des Merkmalsausprägungskontinuums ausgewählt werden.

Das Bayes'sche Sequentialverfahren (Bayesian Estimation, BE) wurde erstmals 1969 von Owen publiziert. Es basiert auf der Annahme einer bestimmten Form und Verortung der Merkmalsausprägungsverteilung („a priori"-Verteilung; Weiss & Vale, 1987) - in der Regel einer Normalverteilung (Thissen & Milevy, 1990) - und kombiniert diese in einem komplizierten Rechenalgorithmus mit den bekannten Itemcharakteristiken und dem Antwortverhalten einer Person. Die Itemauswahl verfolgt hierbei das Ziel, die „a posteriori belief distribution"

[37] Neben diesen beiden am häufigsten zur Anwendung kommenden Itemselektionsverfahren (1. & 2.) sei der Vollständigkeit halber darauf verwiesen, dass es auch die Möglichkeit gibt, die Itemselektion gänzlich in Abhängigkeit von Inhalts- und Zeitkriterien zu gestalten (Eggen, van der Linden, Scrams & Schnipke, 1999 zitiert nach Meijer und Nering, 1999).

[38] Urry (1977) selbst nutzte jedoch auch das Bayes'sche Sequentialverfahren und nicht die MI-Itemselektionsstrategie.

(Hambleton & Zaal, 1990, S. 350) so weit wie möglich einzuengen. Dazu wird jeweils das Item mit der kleinsten erwarteten „a posteriori"-Varianz gewählt, so dass der Standardmessfehler minimiert (Thissen & Mislevy, 1990) und eine möglichst genaue Schätzung ermöglicht wird. Diese Art der Itemselektion hängt logischerweise stark von der Adäquatheit der Vorannahme über die „a priori"-Verteilung ab. Van der Linden und Hambleton (1997) schlagen in diesem Zusammenhang vor, Wissen um bereits bekannte „a priori"-Verteilungen zu nutzen. Vergleicht man MI und BE miteinander, so heben Meijer und Nering (1999) hervor, dass beide Itemselektionsverfahren als stabil gelten und sich insbesondere, wenn sich die „Start-Items" gleichen, bei längeren Tests (N = 20 Items; Thissen & Mislevy, 1990) kaum unterscheiden. In kürzeren CATs sei jedoch eine Anwendung des BEs dem MI vorzuziehen (Hambleton et al., 1991). Abschließend sei eingeräumt, dass die Güte der beiden Itemselektionsstrategien in starkem Maße davon abhängt, inwiefern das Antwortverhalten den IRT-Modellannahmen entspricht.

Im Umgang mit diesen ausgefeilten mathematischen Itemselektionsverfahren weisen Thissen und Mislevy (1990) darauf hin, dass die Itemselektion sich nie gänzlich unreflektiert auf mathematische Berechnungen gründen sollte, sondern Forscher den Itemselektionsalgorithmus inhaltlich reflektieren und gegebenenfalls durch eine Iteminhaltsbalancierung[39] die Itemdarbietung kontrollieren sollten.

4.3.3.4. Personenparameterschätzung

Zur Schätzung der Merkmalsausprägung einer Person, in der IRT auch „Personenparameterschätzung" oder „θ (=Theta)"-Schätzung genannt, kommen in der adaptiven Forschung zur Zeit die folgenden vier verschiedenen Verfahren zum Einsatz:

1. **die Maximum-Likelihood-Schätzung (MLE),**
2. **die Weighted-Maximum-Likelihood-Schätzung (WLE),**
3. **die Expected-A-Posteriori-Schätzung (EAP) und**
4. **die Maximum-A-Posteriori-Schätzung (MAP).**

[39] Iteminhaltsbalancierung ist eine freie Übersetzung des Begriffs „Content Balancing" (Wainer, 1990, S. 122). Bei adaptiven Tests mit heterogenem Iteminhalt besteht die Gefahr, dass der Itemselektionsalgorithmus allein aufgrund statistischer Kennwerte die Itemselektion gestaltet und damit unter Umständen der gesamte Inhaltsbereichs des zu messenden Konstrukts nicht hinreichend erfasst wird. Um dem vorzubeugen, können Strategien zur Iteminhaltsbalancierung - wie z. B. die Strukturierung des Itempools in homogene Testlets, aus denen adaptiv Items gewählt werden - angewandt werden.

Die ersten beiden Ansätze (MLE und WLE) basieren auf dem Likelihood-Schätzverfahren und gehen auf ein von Lord (1980) formuliertes Grundprinzip zurück, der vorschlug, die Wahrscheinlichkeit einer bestimmten Merkmalsausprägung aus einer mathematischen „Kombination“ („joined likelihood function“) der Wahrscheinlichkeit des individuellen Antwortmusters einer Person und des Wissens um die Itemcharakteristiken der dargebotenen Items zu schätzen. Es wird jeweils der Merkmalsausprägungswert auf dem Theta-Kontinuum als beste Schätzung angenommen, an dem die Likelihood Funktion ihr Maximum aufweist.

Der dritte und vierte Ansatz (EAP und MAP) hat seine Wurzeln bei Owen (1969). Ihm liegt das Bayes’sche Schätzverfahren der Merkmalsausprägung auf der Grundlage einer „a priori“-Verteilung zugrunde. Beide Ansätze greifen bei der Theta-Schätzung auf Maße der zentralen Tendenz (EAP: Arithmetischer Mittelwert; MAP: Modalwert) der angenommenen „a priori“-Verteilung (Normalverteilung) zurück. Was wiederum kritisch ist, wenn die vermutete „a priori“-Verteilung nicht der tatsächlichen empirischen Merkmalsausprägungs-verteilung entspricht. Allerdings nimmt mit steigender Testlänge der potentiell verzerrende Einfluss der „a priori“-Verteilungsannahme ab und die „Likelihood“-Verteilung gewinnt an Einfluss.

Alle Ansätze gelten als konsistent und effektiv in ihrer Anwendung (Chen, 1997), ihre Robustheit ist jedoch sowohl von der (IRT-) Modellkonformität des Antwortverhaltens als auch der dargebotenen Itemanzahl abhängig. So weisen eine Reihe von Autoren (Thissen & Mislevy, 1990; Wang, 1995, 1999) darauf hin, dass mit zunehmender Itemdarbietungszahl die Robustheit der Schätzung steigt und die Unterschiede zwischen den einzelnen Algorithmen abnehmen.

Vergleicht man die verschiedenen Ansätze, so tendiert der MLE-Ansatz allgemein zu einer Schätztendenz zu den Extremen (Lord, 1983). Desweiteren funktioniert seine Anwendung in folgenden drei Spezialfällen nicht: a) wenn nur ein Item dargeboten wird (also als Anfangsschätzalgorithmus; Voraussetzung für das Funktionieren des MLE-Algorithmus ist mindestens eine richtige und eine falsche Antwort auf jeweils ein Item), b) wenn alle Items richtig, und c) wenn alle Items falsch beantwortet werden (da in diesen Fällen die Schätzung gegen unendlich läuft).

Die Weighted-Likelihood-Schätzung (WLE; Warm, 1989) gilt als eine Weiterentwicklung des MLE-Ansatzes, der die Wurzel der Testinformationsfunktion als Gewichtung in die Schätzung (bei ein- bzw. zweiparametrischen Modellanwendungen) einfließen lässt, so dass seine Anwendung auch in den oben genannten drei „Spezialfällen“ möglich ist. Nach Meijer und Nering (1999) produziert dieser Ansatz weniger „bias“ (Testergebnisverzerrung).

Auch die EAP- (Bock & Mislevy, 1982) und MAP-Algorithmen können bereits nach der ersten Antwort auf ein Start-Item genutzt werden, da sie auf die vermutete „a priori“-Verteilung zurückgreifen. Dies kann zu einer Verbesserung der Theta-Schätzung führen (Meijer & Nering, 1999). Zudem kommt es zu keinen „Unendlichkeitsschätzungen“. Der Nachteil dieser Verfahren liegt jedoch, im Falle der Darbietung nur weniger Items und einer starken Abweichung des Mittelwerts der „a priori“-Verteilung von der geschätzten Likelihood, in einer „Schätztendenz zur Mitte“. Vergleicht man EAP- und MAP-Algorithmus, so ist der MAP- dem EAP- Algorithmus durch eine geringere Verzerrungstendenz überlegen, während umgekehrt der MAP- den EAP-Algorithmus durch einen etwas geringeren Standardmessfehler übertrifft (Meijer & Nering, 1999).

Möchte man Vorteile beider Ansätze (EAP / MAP und MLE / WLE) nutzen, so kann unter Umständen eine „Step-size-procedure“ (Embretson & Reise, 2000, S. 266f) empfehlenswert erscheinen, bei der die Anfangsschätzung auf der Basis von EAP bzw. MAP erfolgt, bis eine Schätzung auf Basis des MLE- bzw. WLE-Algorithmus möglich wird.

4.3.3.5. Itemdarbietung

Bislang findet sich wenig Forschung zur Itemdarbietung und deren Kontrolle (Thissen & Mislevy, 1990). Meijer und Nering (1999) sowie Embretson und Reise (2000) regen bei der Erforschung dieses Feldes folgende Fragestellungen an:

1. Dürfen bekannte Items mehrmals in einem CAT-Prozess dargeboten werden?
2. Welchen Einfluss haben Vorwissen bzw. Lerneffekte auf das CAT?
3. Sollen alle Items im Laufe eines bestimmten Zeitintervalls dargeboten werden, z. B. durch ein Itembankrotationssystem?
4. Welchen Einfluss hat die Darbietungszeit auf die Itemantwort?

5. Kann während der Itemdarbietung inkonsistentes Antwortverhalten identifiziert und eventuell beeinflusst werden?
6. Welchen Einfluss haben Itempositions-/-reihenfolgeeffekte?

Es bleibt zu hoffen, dass dieses spannende Forschungsfeld in naher Zukunft weitere Forschungsarbeiten motiviert.

4.3.3.6. Stoppfunktion

Um einen computergestützten adaptiven Algorithmus zu beenden, bieten sich prinzipiell drei Stoppkriterien an (Hambleton & Zaal, 1990):

1. ein festgelegtes Messfehlerkriterium (> Reliabilitätskriterium),
2. eine bestimmte Testlänge (minimale bzw. maximale Itemanzahl),
3. ein bestimmtes Klassifikationskriterium („Cut-Off-Wert“).

IRT-basierte CATs bieten gegenüber KTT-basierten Verfahren den großen Vorteil der Berechnung des *individuellen Messfehlers* einer Personenparameterschätzung. Dies ermöglicht die Realisierung eines „equal precise“ Tests, d. h. eines Tests, der empirisch gesichert auf allen Merkmalsausprägungsstufen gleich gut misst. Um dieses zu gewährleisten, kann die eigentliche Testlänge eines IRT-basierten CATs variabel gehalten werden. Konventionelle „fixed-length“ Testverfahren bieten diese Möglichkeit nicht.

Ein IRT-basierter CAT kann aber natürlich genauso in seiner *Testlänge* auf eine bestimmte maximale und / oder minimale Itemdarbietungszahl festgelegt werden, was vor allem bei großen Forschungserhebungen aus ökonomischen Gründen wünschenswert sein kann. Aufgrund der relativen Kürze eines CATs ist eine maximale Begrenzung jedoch häufig nicht nötig. Im Gegenteil merken Hambleton und Zaal (1990) an, dass für Laien die extreme Kürze von CATs mitunter unglaubwürdig oder sogar suspekt wirken könne, so dass eventuell eher (auch im Sinne des Vorbeugens eines „bias“) eine Limitierung im Hinblick auf die minimale Anzahl dargebotener Items angezeigt sei, um die „face validity“ (Augenscheinvalidität) zu erhöhen.

Als Stoppkriterium kann ebenfalls eine Kombination aus minimaler Testlänge und einem bestimmten Messfehlerkriterium gewählt werden.

Und schließlich können im Rahmen kriteriumsorientierter Tests auch sogenannte *„Cut-Off-Werte“* (bezüglich eines Testwertes oder Konfidenzintervalls) als Abbruchkriterien fungieren, welche der reinen

Klassifikation von Personen in zwei (oder mehr) Gruppen dienen. Bei der Nutzung solcher „Cut-Off-Werte“ als Stoppfunktion, erhöht sich meist die Testlänge / -zeit, je näher die Schätzung der Merkmalsausprägung einer Person dem vorher festgelegten „Cut-Off-Wert“ kommt (Weiss & Vale, 1987).

4.3.3.7. Wahl der Soft- und Hardware

Hornke (1996) hebt hervor, dass die Anforderungen, welche CATs an die Hardware stellen, weniger problematisch sind als diejenigen, die CATs an die Software-Programmierung stellen. Er fasst zusammen, dass die Hardware langlebig sein, und sich ihre Benutzeroberflächen für Laien (Testpersonen) handhabbar gestalten sollte (ergonomische Erwägungen, Benutzer-freundlichkeit, gute Lesbarkeit von Itemtexten, einfache Tastenbedienung etc.; Wainer, 1990).

Die einzigen universellen Software-Pakete, welche der Umsetzung des CAT-Prozesses nach bereits stattgefundener Itemkalibrierung dienen können, sind meines Wissens das „Micro-CAT“ (Hambleton & Zaal, 1990), welches 1988 von der Assessment Systems Corporation entwickelt wurde, und der „ADTEST“, der 1994 von Ponsoda, Olea und Revuelta vorgestellt wurde. Mitunter wird von den einzelnen Forschergruppen computergestützte adaptive Testsoftware auch selbst entwickelt (Ware et al., 2000, 2003).

Allgemein gilt die Empfehlung, Software zu entwickeln, welche nicht als „Inselprodukt“ oder „Exot“ auf dem Markt wahrgenommen wird, sondern Software (sowie auch Hardware) so zu standardisieren, dass sie über Schnittstellen zu anderen Komponenten und zu unterschiedlichen Zeitversionen (z. B. von Computersystemen: Windows; Linux etc.) kompatibel ist. In diesem Sinne sollten CATs wie „Haushaltsgeräte mit Bedienungsanleitung“ für ausgebildete Psychodiagnostiker leicht zu handhaben sein, jedoch stets einer professionellen Pflege und einer ernsthaften, verantwortungsbewussten Administration unterliegen.

4.4. Vorteile IRT-basierter CATs

Die zwei Hauptvorteile, welche von vielen Autoren (Weiss & Vale, 1987; Kubinger, 1993; Kisser, 1995; Hornke, 1999; Gregory, 1996; Amelang und Zielinski, 1996; Embretson & Reise, 2000) für IRT-basierte CATs ins Feld geführt werden, sind die Verbesserung a) der *Testökonomie* bzw. –effizienz und b) der *Messgenauigkeit*.

Wie im Kapitel zu computergestützten Tests bereits angedeutet, können diese zu *Zeit*- und *Kosteneinsparungen* von bis zu 50%, IRT-basierte CATs sogar zu Einsparungen von 50-80% führen (Weiss & Vale, 1987; Hornke, 1993, 1996; Gregory, 1996), da durch adaptives Testen die Zeit der Testadministration sowie der Testauswertung und –dokumentation erheblich verringert werden kann und die laufenden Materialkosten (Papier, Bleistifte etc.) gegen eine einmalige Anschaffungsgebühr der Software und Hardware entfallen. Dies ist in großen Forschungsprogrammen von Belang, aber auch für den unmittelbaren klinischen Alltag relevant. Denn durch adaptives Testen (und die adaptive Auswahl von Testverfahren) wird Testen auf Nachfrage möglich, und dies kann zu einer Erleichterung der klinischen Fokusbildung führen. Embretson und Reise (2000) beschreiben exemplarisch einen solchen Nutzen an dem Beispiel eines kognitiven Screening-Instruments, dem bei diagnostischen Hinweisen auf kognitive Defizite ein Gedächtnistest adaptiv nachgeschaltet werden kann.

Desweiteren wird ähnlich wie in Kapitel 4.2.1. darauf verwiesen, dass der eingesparte Aufwand an Routineadministration Zeit für weitere Diagnostik oder Therapie bietet. Eine erhöhte Testökonomie kann nicht nur dem Diagnostiker, sondern auch der Testperson zugute kommen, da durch die alleinige Darbietung derjenigen Items, die für die individuelle Testperson am informativsten sind, die Testperson durch die Psychodiagnostik zeitlich wie emotional weniger belastet wird. D. h. Über- und Unterforderung und damit einhergehende Frustration und Verwirrung bei der Darbietung zu schwieriger Items, sowie Ärger und Langeweile bei der Präsentation zu leichter Items (sowie potentiell resultierende Verminderungen der Datenqualität z. B. durch Flüchtigkeitsfehler oder Motivationseffekte) können durch ein adaptives Testvorgehen vermieden werden (Wainer, 1990). Im Idealfall fühle sich - so Hornke (1993) - die Testperson optimal gefordert und schreibe der CAT-

Testung bedingt durch eine hohe Standardisierung und Augenscheinvalidität eine hohe Testfairness zu.

Die Bestimmung und Kontrolle der *Messgenauigkeit* (Reliabilität) resultiert aus den Möglichkeiten der IRT (siehe Kapitel 3.3.3.). Sie wird durch eine Reihe von Autoren (Weiss, 1985; Weiss & Vale, 1987; Kisser, 1995; Amelang & Zielinski, 1996; Gregory, 1996; Meijer & Nering, 1999; Embretson & Reise, 2000) als der zweite Hauptvorteil adaptiven Testens genannt.

Während des adaptiven Testens ist eine Erhöhung der Messgenauigkeit durch einzelne Items kumulativ abschätzbar, so dass sowohl Aussagen darüber getroffen werden können, wie stark einzelne Items die Messgenauigkeit beeinflussen, als auch mit welcher Messgenauigkeit der gesamte individuelle CAT-Prozess einhergeht. Ersteres erlaubt die Auswahl der Items, welche für ein bestimmtes Merkmalsausprägungsniveau die höchste Messgenauigkeit aufweisen, woraus die eingangs erwähnte Testökonomie resultiert (Amelang & Zielinski, 1996). Eine solche Erhöhung der Messgenauigkeit kann sich auch positiv auf die Validität auswirken (Weiss & Vale, 1987). Desweiteren erlaubt die Kontrolle einer konstanten Messgenauigkeit über verschiedene Merkmalsausprägungsniveaus hinweg den interindividuellen Vergleich von einzelnen Testpersonen sowie den Vergleich von Gruppenkollektiven (trotz unterschiedlicher Art und Anzahl dargebotener Items, d. h. trotz variabler Testlänge; Kisser, 1995). Durch die dadurch bedingte Vermeidung von Decken- oder Bodeneffekten können z. B. Gruppenvergleiche wie in der Lebensspannenforschung, der (Therapie-) Evaluationsforschung und bei Wachstums- /Veränderungsmessungen verbessert werden (Embretson, 1992; Embretson & Reise, 2000).

Neben der Messgenauigkeitsberechnung eröffnet ein IRT-basiertes Vorgehen auch die Möglichkeit der Berechnung weiterer Parameter, wie z. B. der Iteminformationsfunktion, die zur Itemselektion genutzt wird, sowie der Testinformationsfunktion (siehe Kapitel 3.3.3.), welche die Vergleichbarkeit der Messgenauigkeit unterschiedlicher Tests in Bezug auf unterschiedliche Merkmalsausprägungsbereiche oder Personenkollektive und somit gezielte Test-Indikationsentscheidungen ermöglicht (dies übersteigt die Möglichkeiten der KTT). Desweiteren wird darauf hingewiesen, dass IRT-basierte Testscores die empirische Wirklichkeit adäquater als KTT-basierte Testscores abzubilden

vermögen (Kubinger, 1993), was u. a. aus dem Einbezug einer größeren Anzahl von Parametern (z. B. des Rateparameters bei dreiparametrischen IRT-Modellen; Wainer, 1990) resultiere.

Neben den zwei genannten Hauptvorteilen (a) der Testökonomie und (b) der Messgenauigkeit und deren positiven Implikationen (Entlastung des Diagnostikers und der Testperson, Vergleichbarkeit von Messwerten) sowie (c) den zuletzt genannten Vorteilen, welche mit der Berechnung zusätzlicher *IRT-spezifischer Parameter* verbunden sind, werden IRT-basierten CATs in der Literatur eine Reihe von weiteren Vorteilen zugeschrieben.

Diese können grob in Vorteile unterteilt werden, welche sich auf: d) Unterschiede in der *Testform* (Testlänge, Antwortformate und Testinstruktionen) und e) Unterschiede in der *Durchführung* und *Auswertung* von IRT-basierten CATs beziehen.

Vergleicht man konventionelle Verfahren (KTT-basierte Papier-und-Bleistift-Versionen) mit IRT-basierten CATs so stechen die Vorteile einer variablen, adaptiven und damit kürzeren Testlänge,[40] eines variablen Antwortformats (Hambleton & Zaal, 1990; Hambleton et al., 1991) und einer möglichen „maßgeschneiderten“ adaptiven Instruktion (Kisser, 1995) ins Auge.

IRT-basierte CATs unterscheiden sich weiterhin von konventionellen Verfahren, indem umfangreiche Antwortbögen, welche die Gefahr des „Verrutschens“ in der Itemtext- / Antwortzeile mit sich bringen können (Embretson & Reise, 2000) durch eine „Item-by-Item“ Präsentation ersetzt werden. Simultan zur Darbietung einzelner Items vollzieht sich die Schätzung der Merkmalsausprägung der Testperson, so dass eine schnelle / sofortige Testergebnisberechnung, -dokumentation und –rückmeldung (Feedback) ermöglicht wird (Hambleton & Zaal, 1990; Hambleton et al., 1991; Embretson & Reise, 2000).

Aus der kontinuierlichen Verrechnung der Itemantworten einer Testperson ergeben sich zwei weitere Vorteile: zum einen lässt sich dadurch inkonsistentes Antwortverhalten einer Testperson bereits während des CAT-Prozesses identifizieren und eventuell korrigieren (Meijer & Nering, 1999) und zum anderen resultieren daraus (potentielle) Vorteile bezüglich der Itembankentwicklung. So machten bereits 1985 Butcher, Keller und Bacon darauf aufmerksam, dass im Rahmen IRT-basierter CATs eine kontinuierliche

[40] Hornke (1999) zeigt an der Entwicklung von drei CATs eine adaptive Itemreduktion um 2/3 der vorherigen Testlänge (durchschnittliche Anzahl dargebotener Items: 7).

Aktualisierung der Itembank möglich sei, z. B. durch die Einspeisung von „neuen“ Testitems und deren simultaner Kalibrierung im Rahmen des CAT-Prozesses. Durch eine „Züchtung“ „guter“ Items und die Identifikation und den Ausschluss „schlechter“ Items (z. B. durch die Berechnung von Item Response Curves, IRCs; siehe Kapitel 3.3.1.) kann die einem CAT zugrunde liegende Itembank ständig verbessert werden (Thissen & Mislevy, 1990). Meines Wissens wurde das Potential einer simultan zum CAT-Prozess möglichen Aktualisierung der Itembank jedoch in der Praxis noch nicht erprobt.

4.5. Nachteile IRT-basierter CATs

Der größte Nachteil IRT-basierter CATs liegt in den hohen Anfangskosten, welche die Entwicklung und Implementierung solcher Verfahren begleiten (Meijer & Nering, 1999). Diese sind sowohl finanzieller (Kosten von Soft- und Hardware) wie auch personeller (psychodiagnostische, statistische, technische Qualifikationen) Art. Am aufwendigsten ist wohl die umfangreiche Itembankkalibrierung, welche die Erhebung einer Vielzahl von Items an einem großen Personenkollektiv voraussetzt. So wird im individuellen Fall mit detaillierten Kosten-Nutzen-Analysen (Thissen & Mislevy, 1990) abzuwägen sein, ob sich die Entwicklung und Implementierung von IRT-basierten CATs in der jeweiligen Institution bzw. Organisation lohnt.

Während vor einigen Jahrzehnten die technischen Möglichkeiten (begrenzte Rechnerkapazitäten) noch die Grenzen IRT-basierter CAT-Entwicklungen steckten, stellen Hardware-Begrenzungen heutzutage aufgrund des raschen technischen Fortschritts und der ubiquitären Verbreitung von Computern kein ernsthaftes Hindernis mehr dar.

Problematisch ist in diesem Zusammenhang wohl eher die relative Benutzer*un*freundlichkeit der Software, mit der IRT-basiert Itembanken kalibriert werden, sowie der relative Unbekanntheitsgrad der - verglichen mit der KTT - eher komplizierten IRT. Diese beiden Umstände führten bislang zumindest im klinischen Bereich nur zu einer geringen Verbreitung der Methodik (Rost, 1999; siehe Kapitel 3.3.4. und 3.5.). Das damit verbundene Forschungsdefizit lässt viele Fragen offen.

So zweifelt beispielsweise Kisser (1995), ob die erhoffte Zeitersparnis bei IRT-basierten CATs sich bei deren Anwendung in der Realität tatsächlich zeigt. Es existieren zwar einige Belege für eine kürzere Bearbeitungszeit von CATs

(Hornke, 1993, 1996, 1999), allerdings vermutet Kisser (1995), dass eine geringe Anzahl von Items (wie beim CAT) nicht unweigerlich zu einer Testzeitverkürzung führe, wenn die Bearbeitung von Items mit *unterschiedlichen* Antwortformaten mehr Zeit als die Beantwortung von Items mit dem *gleichen* Antwortformat (wie bei konventionellen Verfahren) in Anspruch nähme. Bezüglich der Untersuchung der Zeitersparnis fand Hornke (1996) in einer seiner Studien, dass im Laufe des CAT-Prozesses die Itembearbeitungszeit abnähme, gleichzeitig verringerte sich jedoch auch die Konstanz der Messergebnisse, was Hornke auf einen Sorgfalts-, Aufmerksamkeits- und / oder Motivationsverlust (> Flüchtigkeitsfehler) der Testpersonen in der Interaktion mit dem Computer zurückführte. Auch dies ist kritisch zu bewerten.

Neben der Überprüfung der tatsächlichen Zeitersparnis, sind bislang weitere grundlegende Aspekte von IRT-basierten CATs unerforscht. So bestehen beispielsweise große Forschungsdefizite im Hinblick auf...

1. die methodischen Standards der IRT-basierten Itembankentwicklung (Selektionskriterien),
2. die Itembanksicherheit (v. a. bei „wireless LAN-Applikationen“[41]),
3. die Robustheit von Item- und Personenparameterschätzungen…
    a) über verschiedene Zeiten, Kontexte und Stichproben (Gefahr des „Parameterdrifts“; Bock & Mislevy, 1988),
    b) bei unterschiedlichen Itembankgrößen,
    c) bei unterschiedlichen Antwortformaten (Kisser, 1995),
    d) auf der Grundlage unterschiedlicher Itemselektionsalgorithmen,
    e) bei unterschiedlichen Itemreihenfolgen und –positionen,
    f) bei unterschiedlichen Itemdarbietungskontrollen (Zeitrestriktionen, Unmöglichkeit des Zurückblätterns / Korrigierens, Iteminhalts-balancierung),
    g) im Falle von Vorwissen um Items (Lerneffekte),
    h) im Falle von (Computer-) Testangst,
    i) bzgl. der Verletzung von IRT-Modellannahmen wie z. B.:
        - Unidimensionalitätsverletzungen,
        - Item-Misfits,
        - Personen-Misfits (Wainer, 1990; Kubinger, 1999).

[41] Wireless LAN (Local Area Network) = kabellose Datenübertragung in lokalen Computernetzwerken.

4. die Anwendung von IRT-Modellen auf polytome Items (Dodd et al., 1995),
5. die beste Kommunizierbarkeit IRT-basierter Testscores (Theta), welche in Einheiten der Standardnormalverteilung (z-Werte) ausgegeben werden und für den Laien (Testpersonen) nicht intuitiv verständlich erscheinen (Embretson & Reise, 2000),
6. die Äquivalenzprüfung (Kubinger, 1999) von Papier-und-Bleistift-Verfahren, computergestützten Tests und CATs,
7. die prospektive Validität von IRT-basierten CATs und
8. die allgemeine Qualitätssicherung IRT-basierter CATs.

Wie die aufgeführten Forschungsdefizite (siehe auch Kapitel 3.3.4.) verdeutlichen, stecken IRT-basierte CATs (v. a. im klinisch-psychologischen Bereich) noch weitgehend in den Kinderschuhen (siehe Kapitel 4.6.). Das junge Forschungsfeld IRT-basierter CATs ist durch ein Mosaik technischer Artikel (Embretson & Hershberger, 1999) gekennzeichnet. Empirische Befunde zur Anwendung IRT-basierter CATs im psychologischen Bereich beschränken sich größtenteils auf Simulationsstudien (Kisser, 1996; Hornke, 1999; Gardner et al., 2002). Daher kommt der Entwicklung und Erprobung „echter“ CATs in der Praxis bei der Beantwortung oben genannten Forschungsfragen ein großer Stellenwert zu.

4.6. Aktueller Forschungsstand zu IRT-basierten CATs

Die Sichtung der Literatur zum Thema IRT-basierter CATs gestaltet sich etwas verwirrend, da CATs entwickelt wurden, welche sich andere adaptive Strategien zunutze machen als die IRT (siehe Kapitel 4.3.2.). Zum Beispiel wandten Ben-Porath, Slutske und Butcher (1989) die „Countdown“-Methode an, um ein CAT des Minnesota Multiphasic Personality Inventory (MMPI) zu entwickeln (Roper, Ben-Porath & Butcher, 1991; Handel, Ben-Porath & Watt, 1999).

Zudem existieren eine Reihe von Forschungsarbeiten zur Anwendung der IRT bei der Itembankentwicklung, die in der Entwicklung von CATs mündeten, bei denen jedoch der Itemselektionsalgorithmus und die Personenparameter-schätzung nicht IRT-basiert erfolgen, sondern „konventionell“ programmiert sind. Tabelle 6 gibt einen Überblick über solche CATs im deutschsprachigen Raum. Sie begrenzt sich auf einen Überblick der im internationalen Raum

aktuell eingesetzten CATs, welche *gänzlich* IRT-basiert sind. Das heißt, es werden nur CATs aufgeführt, bei denen die IRT sowohl bei der Itemparameterkalibrierung, als auch bei der Itemselektion und Personenparameterschätzung im Rahmen des CAT-Prozesses angewandt wurde.

Tabelle 6: Überblick über CATs im deutschen Sprachraum, bei denen die Itembankentwicklung IRT-basiert erfolgte (die Itemselektion und Testergebnisberechnung jedoch nicht IRT-basiert sind).

Inventar	Bereich	Autoren	Jahr	Ort
Verbal Memory Test	Gedächtnistest	Hornke & Etzel	1999a	Institut für Psychologie, Rheinisch-Westfälische TH Aachen, Deutschland.
Visueller Gedächtnis Test	Gedächtnistest	Hornke & Etzel	1999b	Schuhfried-Testverlag, Mödling, Österreich.
Adaptive Three-Dimensional Cube Comparison Test (A3DW)	Eindimensionaler Intelligenztest	Gittler	1999	Institut für Psychologie der Universität Wien, Österreich.
Adaptiver Matrizentest (AMT)	Wehrpsychologische Eignungsdiagnostik	Hornke & Habon	1984	Institut für Psychologie, Rheinisch-Westfälische Technische Hochschule, TH Aachen, Deutschland.
Adaptiver Zahlenfolgen-Lerntest (AZAFO)	Lernfähigkeitstest	Vahle & Rittner	1995	Schuhfried-Testverlag, Mödling, Österreich.
Computergest. Intelligenz-Lerntest-Batterie (ACIL)	Intelligenztest	Beckmann & Guthke	1999	Schuhfried-Testverlag, Mödling, Österreich.
Adaptiver Analogien-Lerntest (ADANA)	Lernfähigkeitstest	Stein	1995	Schuhfried-Testverlag, Mödling, Österreich.
Syllogismen	Eindimensionaler Intelligenztest	Srp & Hörndler	1994	Swets Test Services, Frankfurt am Main, Deutschland.
Begriffs-Bildungs-Test (BBT).	Informations-verarbeitungtest	Kubinger, Fischer & Schuhfried	1993	Schuhfried-Testverlag, Mödling, Österreich.

4.6.1. IRT-basierte CATs in der Leistungs- und Eignungsdiagnostik

IRT-basierte CATs sind vor allem im Bereich der Fähigkeitseinschätzung zur *Eignungsdiagnostik* auf internationaler Ebene mittlerweile gut etabliert. Die zwei größten Anwendungsgebiete liegen im Bereich der *Schul*diagnostik und der *militärischen* Eignungsdiagnostik.

In der *Schuleignungsdiagnostik* sind eine Reihe IRT-basierter CATs in der Anwendung wie z. B. in den U.S.A. der „Scholastic Aptitude Test“ (SAT), die „Graduate Record Examination“ (GRE, 1996; Educational Testing Service, ETS, 2001), der „Computerized Placement Test“ (College Board, 1993) und verschiedene Mathematik, Lese- und Schreibtests innerhalb des „COMPASS“-Programms (American College Testing, 1993; Dodd et al., 1995), in Südafrika der „Learning Potential Computerized Adaptive Test“ (LPCAT; de Beer, 2000) sowie in den Niederlanden zwei Mathematikleistungstests (National Institute for Educational Measurement; Verschoor & Straetmans, 1999).

Im Rahmen von *wehrpsychologischen Untersuchungen* werden sowohl in Deutschland als auch in den U.S.A. IRT-basierte CATs eingesetzt. In Deutschland zählen Hornke und seine Mitarbeiter zu den Hauptvertretern dieser Richtung, welche IRT-basierte CATs zur Diagnostik „Verbaler Analogien“ (Hornke, 1989), zur Messung der Gedächtnis- und Orientierungsleistung (Hornke, 1999) und zur Intelligenz („Matrizentest“; Hornke, 1999) entwickelten. In der U.S. Armee wird zur Eingangsdiagnostik ein IRT-basierter CAT namens „Armed Services Vocational Aptitude Battery“ (ASVAB; Curran & Wise, 1994; Sands, Waters & McBride, 1997) angewandt. Da der initiale Entwicklungsaufwand IRT-basierter CATs recht hoch ist, finden sich die meisten Anwendungen IRT-basierter CATs in *größeren Organisationen* und Institutionen, welche regelmäßig umfangreiche psychodiagnostische Testungen durchführen. So machen sich neben amerikanischen Schulbehörden (z. B. Portland Public School District, Kingsbury & Houser, 1993) und Militäreinrichtungen (z. B. U.S. Department of Defense) auch Prüfungsbüros von medizinischen Ausbildungseinrichtungen die IRT-basierte CAT-Methodik bei der Durchführung von Examina zunutze (z. B. American Society of Clinical Pathologists; Lunz, Bergstrom & Wright, 1992; National Council of State Boards of Nursing; Zara, 1988; American Board of Internal Medicine; Reshetar, Norcini & Shea, 1993).

4.6.2. IRT-basierte CATs in der klinischen und Persönlichkeitsdiagnostik

Während die Anwendungsbeispiele von IRT-basierten CATs im Bereich der Leistungs- und Eignungsdiagnostik zeigen, dass diese Methodik in diesem Bereich bereits verbreitet ist, gilt dies nicht für den Bereich der klinischen Diagnostik sowie der Messung von Einstellung und Persönlichkeitseigenschaften.

Im Bereich der *klinisch-medizinischen Diagnostik* existieren meines Wissens (neben der Forschungsgruppe an der Charité Berlin, in dessen Rahmen vorliegende Arbeit geschrieben wurde), nur drei Forschungsgruppen, welche folgende IRT-basierte CATs entwickelt haben:

- Ware, Bjorner und Kosinski (2000):
  Dynamic Health Assessment (DynHA):
  - Headache Impact Test (HIT)[42],
  - Dynamic SF-36 Health Survey,
  - Depression Impact Test etc;
- Simms und Clark (submitted):
  Schedule for Nonadaptive and Adaptive Personality;
- Gardner, Kelleher und Pajer (2002):
  Pediatric Symptom Checklist (PSC).

Im Bereich der *Einstellungsdiagnostik* (sowie Leistungsmessung) findet sich neben Reise und Waller (1990), welche die Absorption Scale des MPQ (Tellegen, 1982) IRT-basiert computer-adaptiv erprobten, und Andrich (1978) eine rege Forschungstätigkeit nur in der Forschungsgruppe um Dodd, Ayala und Koch (1995). Diese sticht jedoch dafür durch eine hohe Publikationsfreudigkeit hervor (De Ayala, 1989, 1992; Dodd, 1990; Dodd et al., 1988, 1989, 1993; Koch & Dodd, 1985, 1989; Koch et al., 1990), indem sie gezielt die Anwendung verschiedener IRT-Modelle auf polytome Items fokussiert (bislang werden in der IRT-basierten CAT-Eignungsdiagnostik fast ausschließlich dichotome Items genutzt).

Polytome Items werden auch vielfach in der *Persönlichkeitsdiagnostik* verwandt. Jedoch hinkt die IRT-basierte CAT-Forschung in diesem Bereich dem Forschungsstand, wie er z. B. bereits in der Eignungsdiagnostik gediehen ist,

[42] Ware, Kosinski, Bjorner, Bayliss, Batenhorst, Dahlöt, Tepper & Dowson (2003).

stark hinterher. Ursächlich hierfür könnte u. a. die Diskussion um die (Uni-) Dimensionalität von Persönlichkeitskonstrukten (die meisten IRT-basierten CATs sind bislang unidimensional konstruiert; komplexe multidimensionale IRT-Ansätze finden sich meines Wissens nur bei Gardner et al., 2002) sowie das allgemein geringe wirtschaftliche Interesse an der Persönlichkeitsdiagnostik sein (Persönlichkeitsdiagnostik ist im Rahmen von Eignungsdiagnostik umstritten; die psychologische Diagnostik wird im chronisch unterfinanzierten öffentlichen Gesundheitswesen eher vernachlässigt).

Genuine IRT-basiert entwickelte CATs zur Messung von Persönlichkeits-variablen existieren meines Wissens weder im deutschen noch im inter-nationalen Sprachraum.

Jedoch publizierten kürzlich Reise und Henson (2000) in einer Simulationsstudie eine computergestützte adaptive Version des bereits etablierten NEO-PIs, dessen Itemselektion und Personenparameterschätzung IRT-basiert anhand des Graded Response Modells erfolgt, dessen Itembank jedoch nicht mit IRT-Methoden entwickelt wurde. Zudem bereiten Simms und Clark eine Publikation vor, in der ein Persönlichkeitsfragebogen („Schedule for Nonadaptive and Adaptive Personality“, SNAP; Clark, 1993) als IRT-basierte CAT-Version anhand des 2PL-Modells von Birnbaum entwickelt und an N=413 Studenten erfolgreich validiert wurde. Die meisten Forschungsarbeiten in diesem Gebiet sind noch nicht so weit fortgeschritten und beschränken sich größtenteils auf die Erprobung von IRT-Methoden im Rahmen der (Re-)Analyse bzw. Bewertung bereits etablierter KTT-basierter Verfahren (zum aktuellen Forschungsstand bzgl. IRT-Anwendungen und zu möglichen Gründen dieses Forschungsdefizits siehe Kapitel 3.5.2.).

Zusammenfassend lässt sich resümieren, dass sich das in vorliegender Arbeit entwickelte IRT-basierte CAT zur Angstmessung (Angst-CAT) als eine *klinisch-psychologische* Pionierarbeit in die oben genannten US-amerikanischen Forschungsarbeiten zur klinisch-medizinischen Diagnostik, Einstellungs- und Persönlichkeitsdiagnostik einreihen lässt.

5. Die Entwicklung des Computergestützten Adaptiven Tests zur Angstmessung (Angst-CAT)

5.1. Ziel

Wie in Kapitel 4.4. erörtert, bietet IRT-basiertes Computergestütztes Adaptives Testen (CAT) eine Vielzahl von Vorteilen. Die Nutzung dieser Vorteile hinkt jedoch im klinisch-diagnostischen Alltag dem theoretischen Wissen um die Vorzüge IRT-basierten Computergestützten Adaptiven Testens hinterher (zum Forschungsstand siehe Kapitel 4.6.).

Da Patienten mit Angststörungen in psychosomatischen Kliniken gehäuft auftreten (24,4% – 29,4%; Fliege et al., 2002), ist hier das Interesse an einer zuverlässigen, messgenauen, patientenfreundlichen und ökonomischen Diagnostik besonders groß.

Um die psychometrische Angstmessung in diesem Feld zu verbessern, wurde ein IRT-basierter Computergestützter Adaptiver Test zu Angstmessung (Angst-CAT) entwickelt. Angestrebt wurde die Konstruktion eines kurzen Screening-Instruments, welches als *eindimensionales* Breitbandverfahren bei gesunden Testpersonen einsetzbar sein sowie im klinisch-therapeutischen Bereich seine Anwendung finden soll. Wenngleich man mit einer *mehrdimensionalen* Testkonstruktion intuitiv sicher eher den vielfältigen Facetten des Phänomens der Angst (siehe Kapitel 2.4.) gerecht würde, so scheint aus wissenschaftlicher Sicht nach jahrzehntelangen Forschungsbemühungen um eine empirische Differenzierung verschiedener statistisch unabhängiger Angstkomponenten (siehe Kapitel 2.7.3.4.) eine solche nicht zu gelingen. Für eine unidimensionale Testung sprechen zusätzlich ökonomische Gründe sowie der aktuelle, junge methodische Forschungsstand (siehe Kapitel 3.5.). Bislang wurden meines Wissens nur zwei gänzlich IRT-basierte CAT-Versionen im Bereich der Persönlichkeitsdiagnostik (Reise & Henson, 2000; Simms & Clark, in Vorbereitung) und einige wenige in der klinischen Diagnostik (Gardner, Kelleher & Pajer, 2002; Ware et al., 2000, 2003) entwickelt. Dieser Forschungsrückstand deutet darauf hin, dass sich hier die Anwendung schwierig gestaltet (zu den Gründen siehe Kapitel 3.5.2.). Unsere Forschungsgruppe hat sich entschieden, zunächst die Entwicklung eines eindimensionalen CATs (Angst-CAT) zu erproben, bevor sie den nächsten Schritt zur Entwicklung mehrdimensionaler CATs geht.

Da das Angst-CAT im klinisch-therapeutischen Bereich zur Eingangs- und Verlaufsdiagnostik genutzt werden sollte, wurde es als Verfahren zur Erfassung der Zustands-Angst (State-Anxiety, siehe Kapitel 2.4., 2.7.3.3) entwickelt. Dies bietet den Vorteil, durch eine angestrebte hohe Veränderungssensitivität auch Therapieverlaufsevaluationen zu ermöglichen.

Dass im Bereich der State-Angst-Messung laut Amelang und Zielinski (1996) „fraglos ein gewisser Mangel an Verfahren zur Abschätzung *aktueller Zustände*" (S. 287) herrscht, begründet dieses Vorhaben desweiteren. Auf eine Messung der Angst als stabiler Persönlichkeitseigenschaft wurde verzichtet, da - wie in Kapitel 2.7.3.4. erörtert - Zustands- und Eigenschafts-Angst so eng miteinander korrelieren, dass aus meiner Sicht die separate Erfassung von Eigenschafts-Angst im klinischen Alltag nicht zwingend notwendig ist, da Eigenschafts-Angst ggf. durch eine Mittelung intraindividueller Zustands-Angstscores zu verschiedenen Messzeitpunkten abgeleitet werden kann (Uhlenhuth, 1985).

5.2. Stichprobe der Testkonstruktion

5.2.1. Gesamtstichprobe

Die statistische Itemanalyse und -selektion zur Entwicklung des Computergestützten Adaptiven Tests zur Angsterfassung (Angst-CAT) erfolgte an insgesamt N = 2.348 Patienten, die sich in der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin zur Diagnostik oder Therapie in den Jahren 1995 bis 2001 vorstellten. Tabelle 7 fasst die wesentlichen soziodemografischen, Tabelle 8 die klinischen Charakteristika dieser Stichprobe zusammen.

Tabelle 7: Soziodemografische Charakteristika der zur Testkonstruktion des Angst-CATs genutzten Gesamtstichprobe.

Charakteristika	Kategorie / Parameter	Angaben
Geschlecht	Weiblich	68,5%
	Männlich	31,5%
Alter	Arithmetischer Mittelwert ($\overline{X}$)	41,31 Jahre
	Standardabweichung (SD)	14,31 Jahre
Familienstand	verheiratet (mit Partner zusammen lebend)	38,7%
	verheiratet (ohne Partner zusammen lebend)	5,3%
	unverheiratet (mit Partner)	14,3%
	ledig (ohne Partner)	23,7%
	geschieden / verwitwet	16,0%
	fehlende Angaben	2, 0%

Tabelle 8: Klinische Charakteristika der zur Testkonstruktion des Angst-CATs genutzten Gesamtstichprobe.

Charakteristika	Kategorie	Angaben
Erhebungsbereich	Stationär	55,3%
	Ambulant	33,4%
	Konsiliarisch	11,3%
Diagnosen[43]	Angststörungen (F.40-41)	13%
	Depressive Störungen (F.32-34)	30%
	Essstörungen (F.50)	18%
	Somatoforme Störungen (F.45)	24%
	Primär somatische Erkrankungen (nicht F)	10%

Im Rahmen der klinisch-psychologischen Routinediagnostik (Testbatterien) wurden an diesen Patienten 13 psychometrische Verfahren angewandt, welche sich im psychosomatischen Bereich bewährt haben (ADS[44], ALL[45], BDI[46], BSF[47], GBB[48], GT[49], NI-90[50], PGWI[51], PSQ[52], SF36[53], SKT[54], STAI[55], SWO[56]). Der Einsatz der Instrumente erfolgte computergestützt mittels Handcomputer, sogenannter „PDA's" (Personal Digital Assistants der Firma Psion), deren Einsatz bereits erprobt ist (Rose, Hess, Hörhold, Brähler & Klapp, 1999; Rose, Walter, Fliege, Becker, Hess & Klapp, 2003). In der medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité werden seit 1995 zur psychologischen Routinediagnostik oben genannte Handcomputer (16,5 x 8,8 x 2,3 cm, 280g) eingesetzt, welche eine mobile, d. h. standortunabhängige, selbstständige Beantwortung der Fragen durch die Patienten ermöglichen. Dazu werden *vor* der computergestützten Fragebogenerhebung (Routinetestbatterien) vom Klinikpersonal die Patienten-Identifikationsdaten in die jeweiligen Hand-

[43] Die Diagnosestellung erfolgte durch klinisch erfahrene Diagnostiker nach den Kriterien des ICD-10 (Dilling et al., 2000). Die Prozentwerte der Diagnosen summieren sich nicht zu 100%, da Komorbidität zwischen einzelnen Störungen häufig ist.
[44] ADS: Allgemeine-Depressions-Skala (Hautzinger & Bailer, 1993).
[45] ALL: Fragebogen zum Alltagsleben (Bullinger, Kirchberger & Steinbüchel, 1993).
[46] BDI: Beck-Depressions-Inventar (Hautzinger, Bailer, Worall & Keller, 1994).
[47] BSF: Berliner-Stimmungs-Fragebogen (Hörhold & Klapp, 1993; Rose et al., in Druck).
[48] GBB: Gießener-Beschwerde-Bogen (Brähler & Scheer, 1995).
[49] GT: Gießen-Test Selbst & Idealselbst (Beckmann, Brähler & Richter, 1991).
[50] NI: Narzissmus-Inventar (NI: Deneke & Hilgenstock, 1989; NI-90: Schöneich, Rose, Danzer, Thier, Weber & Klapp, 2000).
[51] PGWI: Psychological General Wellbeing Index (Ludwig, Geier & Bullinger, 1990).
[52] PSQ: Perceived Stress Questionnaire (Levenstein, Prantera, Varvo, Scribano, Berto, Luzi & Andreoli, 1993).
[53] SF36: Fragebogen zum Gesundheitszustand (Bullinger & Kirchberger, 1998).
[54] SKT: Subjektive-Krankheitstheorien-Ursachenvorstellung (Faller, 1997).
[55] STAI: State Trait Anxiety Inventory (Laux, Glanzmann, Schaffner & Spielberger, 1981).
[56] SWO: Fragebogen zu Selbstwirksamkeit, Optimismus und Pessimismus (Scholler, Fliege & Klapp, 1999)

computer eingegeben. *Nach* der Datenerhebung wird der Handcomputer an einen Computer angeschlossen. Die psychodiagnostischen Daten werden so auf eine klinikinterne Datenbank übertragen und automatisch (grafisch) ausgewertet. In fortlaufenden Studien werden die Reliabilität und Validität sowie die Datenstruktur der eingesetzten Instrumente überprüft und (Test-)Normen mittels gesammelter Daten an psychosomatischen Patientenkollektiven aktualisiert. Eine umfangreiche Studie zu den Auswirkungen der vollständigen Umstellung der psychometrischen Routinediagnostik auf die oben beschriebene mobile, computergestützte Erhebungsmethode an N = 1.400 (Papier-und-Bleistift-Version) bzw. N = 9.000 (Computerversion) psychosomatischen Patienten erbrachte drei zentrale Ergebnisse (Rose et al., 1999, 2003). Erstens werde, so Rose und Mitarbeiter (1999), die Datenorganisation verbessert, wodurch ein schnellerer Zugriff für klinische und wissenschaftliche Zwecke gewährleistet sei, zweitens führten die mobilen computergestützten Erhebungen zu Einsparungen von 2/3 des gesamten Dokumentations-aufwandes und drittens ließen sich hinsichtlich der Datenstruktur keine grundlegenden Stabilitäts- oder Verteilungsunterschiede zwischen der Papier- und der Computerversion feststellen[57] (siehe auch Kubinger, 1999).

5.2.2. Teilstichproben

Nicht alle Patienten der Gesamtstichprobe konnten aus ökonomischen Gründen und aufgrund einer psychodiagnostischen Mehrbelastung *alle* Items (N_{Items} = 81) beantworten, welche im Rahmen der theoretischen Itempoolerstellung (siehe Kapitel 5.3.1.) als inhaltlich relevant für die Angstmessung angesehenen wurden. Daher erfolgte die statistische Itemanalyse und –selektion (siehe Kapitel 5.3.2.) an drei Teilstichproben (N_1 = 1.010; N_2 = 834; N_3 = 775), welche gebildet wurden, um einen möglichst großen initialen Itempool untersuchen zu können. Die Teilstichproben überlappen sich sowohl bezüglich einzelner Items (bis zu N = 28 Items), als auch bezüglich einer Gruppe von Patienten (bis zu N = 275).

[57] Die Äquivalenzprüfung von einem Instrument zur Erfassung von Trait-Merkmalen (GT) zeigte keine Unterschiede zwischen der Papier- und Computerversion, die Äquivalenzprüfung an Instrumenten zur Erfassung von State-Merkmalen zeigte bzgl. eines Verfahrens (BSF) keine Unterschiede und bzgl. eines Verfahrens (GBB) eine Tendenz zu etwas höheren Skalen-mittelwerten in der Computerversion, so dass hier eine Normierungsaktualisierung notwendig wurde.

Die Itemüberlappung ermöglicht das Zusammenfassen der Teilstichproben mittels eines „Item-Link-Designs“ (siehe Kapitel 5.3.2.3.3.) auf einer gemeinsamen Skala. Es wird vermutet, dass die Personenüberlappung zu einer stabileren Itemparameterschätzung zwischen den Teilstichproben beiträgt. Negative Auswirkungen der Personenüberschneidung auf die Itemanalyse und -selektion werden zunächst nicht angenommen, da eine der zentralen messtheoretischen Annahmen der IRT (siehe Kapitel 3.3.1. „Invarianz Eigenschaft“) lautet, dass die Item- und Personenparameterschätzung bei Modellkonformität stichprobenunabhängig ist (Embretson, 1996; Embretson & Reise, 2000). Diese Stichprobenunabhängigkeit bezieht sich sowohl auf die Schätzung der Itemstatistiken, d. h. die berechneten Schwierigkeits- und Diskriminationsparameter von Items sind von der untersuchten Personenstichprobe unabhängig und damit generalisierbar, als auch auf die Schätzung individueller Merkmalsausprägungen (Theta), von der im Rahmen der IRT angenommen wird, dass sie von dem spezifischen Set dargebotener Items unabhängig ist. Dies erlaubt die Vergleichbarkeit von Theta-Werten von Personen, denen unterschiedliche Itemsets zur Beantwortung vorgelegt werden und ermöglicht überhaupt erst das adaptive Testen.
Abbildung 8 gibt einen Überblick über die drei Teilstichproben, welche der statistischen Itemanalyse und -selektion zugrunde liegen.

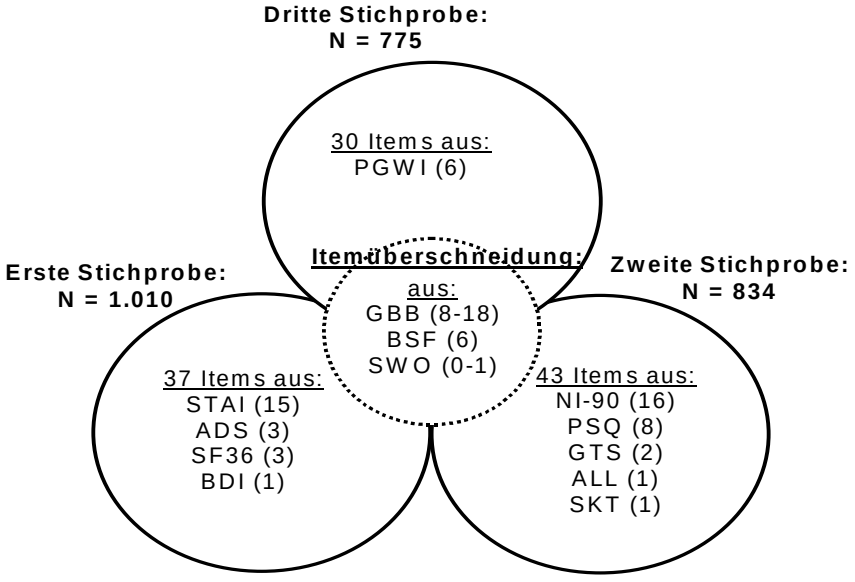


Abbildung 8: Überblick über die drei Teilstichproben, an denen die statistische Itemanalyse und -selektion erfolgte (Testabkürzungen siehe Fußnoten S. 106).

Die jeweilige gesamte Itemmenge der Teilstichproben in Abbildung 8 ergibt sich aus den in dem jeweiligen Kreis dargestellten Items plus einer Anzahl von Items, welche gemeinsam in mehreren Stichproben von Patienten erhoben wurden. So setzen sich die 37 Items aus der ersten Teilstichprobe aus den im Kreis dargestellten 22 Items (STAI:15; ADS: 3; SF36: 3; BDI: 1 Item) plus weiteren 15 Items aus der Itemüberschneidungsmenge (hier: GBB: 8; BSF: 6; SWO: 1) zusammen; die Itemmenge von 43 Items der zweiten Teilstichprobe resultiert aus 28 Items (NI: 16; PSQ: 8; GT: 2; ALL: 1; SKT: 1) plus 15 Items aus der Itemüberschneidungsmenge (hier: GBB: 8; BSF: 6, SWO: 1); und die 30 Items umfassende Itemmenge der dritten Teilstichprobe entstammt dem PGWI (6), GBB (18) und BSF (6).

Die analysierten Items der drei Teilstichproben wurden im Anschluss an eine umfangreiche Itemanalyse und –selektion miteinander verbunden (zum „Item-Link-Design“, siehe Kapitel 5.3.2.3.3.), um einen Computergestützten Adaptiven Test (CAT) mit möglichst vielen psychometrisch hochwertigen Items zu generieren. Das methodische Vorgehen der theoretischen Erstellung der Itembank und der statistischen Itemanalyse und -selektion wird in Kapitel 5.3. erläutert, die Ergebnisse der Untersuchung der drei Teilstichproben in Kapitel 5.4. dargestellt, und die gesamte Itembank, in der alle selektierten Items der drei Teilstichproben zusammengefasst wurden, wird in Kapitel 5.4.4. beschrieben.

5.3. Methoden der Entwicklung der Itembank

Das Vorgehen bei der Testentwicklung lässt sich in drei prinzipielle Schritte gliedern (Abbildung 9). Im ersten Schritt wurde ein Itempool zur Messung von „Angst“ theoriegeleitet erstellt. Der zweite Schritt besteht aus der statistischen Itemanalyse und –selektion. Im dritten Schritt wurden die Items, welche sich in den vorangegangenen Schritten bewährt haben, als Itembank einem computergestützten, adaptiven Itemabfolge-Algorithmus zugrundegelegt, welcher die Schätzung des sogenannten Theta-Wertes ermöglicht, was der sonst üblichen Testwertberechnung („Summenscore“) der Angstausprägung entspricht.

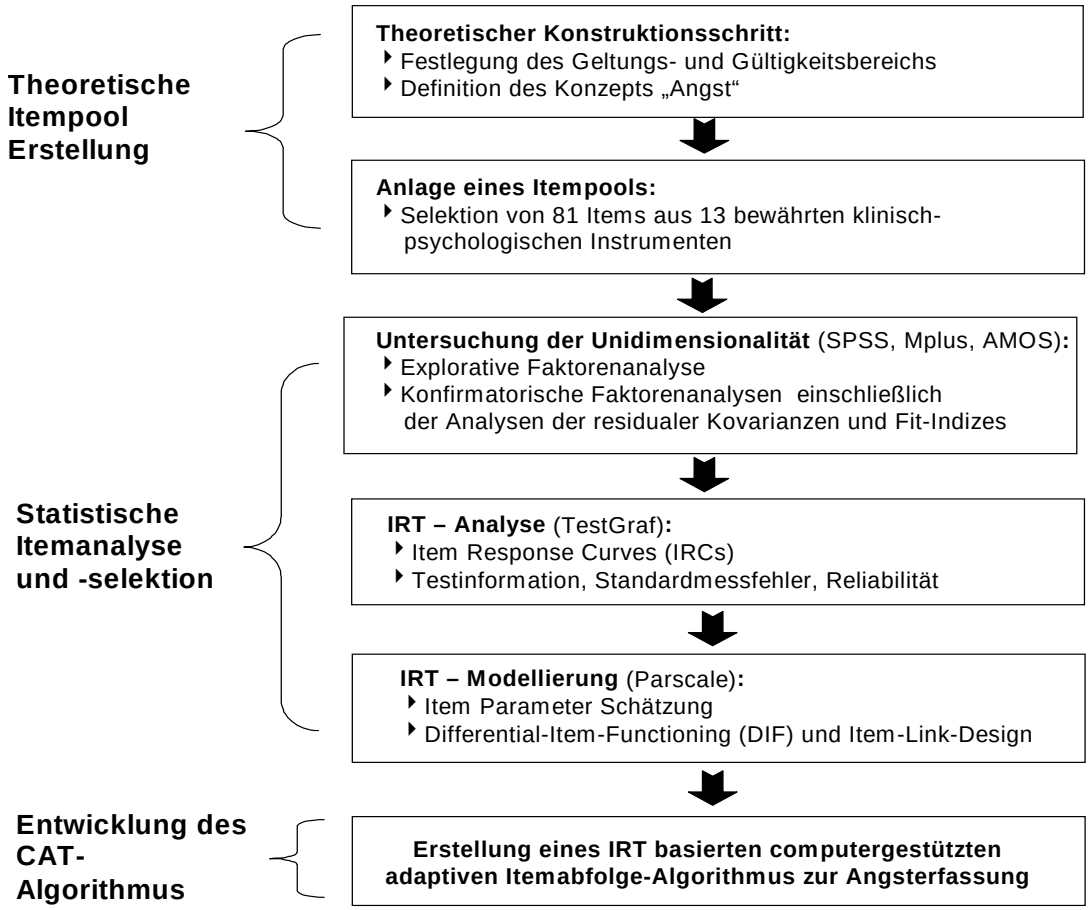


Abbildung 9: Ablaufschema der Entwicklung des IRT-basierten Angst-CATs.

5.3.1. Theoretische Erstellung der Itembank

Die Testkonstruktion begann mit einem theoriegeleiteten Teil, in dem zunächst der Geltungs- und Gültigkeitsbereich des zu entwickelnden Instruments festgelegt wurde. Wie bereits in Kapitel 5.1. ausgeführt und begründet, intendiert das Angst-CAT die eindimensionale Erfassung der Zustands-Angst in der Allgemeinbevölkerung, bei Patienten mit chronischen somatischen Erkrankungen und bei psychosomatischen bzw. psychiatrischen Patienten. Um die Messung einer globalen Ausprägung der Angst mit dem Instrument zu gewährleisten und eine abstrakte, situationsübergreifende Messung der Angst zu ermöglichen, wurde auf den Einbezug situations- bzw. objektspezifischer Aspekte der Angst (siehe Kapitel 2.3.2., 2.6.1. und 2.7.3.4.) weitgehend verzichtet.

Weiterhin wurde in dem theoriegeleiteten Teil, das Konstrukt „Angst“ theoretisch reflektiert und konzeptionell definiert (siehe auch Kapitel 2). Die Autorin schließt sich bei der Definition der Angst Spielberger (1972) an, der Zustands-Angst als

einen „emotionalen Zustand, der durch Anspannung, Besorgtheit, Nervosität, innere Unruhe und Furcht vor zukünftigen Ereignissen gekennzeichnet ist“ (S. 482) definiert (siehe Kapitel 2.4.3.1.). Die Definition entspricht damit weitgehend den Kriterien, die in der ICD-10 (Dilling et al., 2000) für eine generalisierte Angststörung (F41.1) genannt sind. Hier werden für die Angststörung „Befürchtungen, motorische Spannungen und vegetative Übererregbarkeit“ als charakteristisch angesehen.

Um die verschiedenen Ausprägungsgrade der Angst darstellen zu können, wurden im Rahmen der Itemkonstruktion neben der emotionalen und der kognitiven Komponente der Angst (Liebert & Morris, 1967, siehe Kapitel 2.7.3.4.) auch vegetative Symptome, wie plötzliches Herzklopfen, Schwindel und Depersonalisationserleben, berücksichtigt (siehe Kapitel 2.3.4.). Vor der inhaltlichen Itemselektion wurde konsensuell festgelegt, welche Konstrukte von dem Konstrukt der Angst abzugrenzen sind. Hierzu zählen „allgemeine Leistungseinbußen“, „Schlafstörungen“ und „Depression“ (siehe Kapitel 2.5.).

Die Auswahl der angstrelevanten Items geschah anhand eines Delphi-Entscheidungsprozesses (Hasson, Keeney & McKenna, 2000). Jedes Mitglied der Forschungsgruppe (eine Diplom-Psychologin, ein Arzt mit primär wissenschaftlicher Tätigkeit, ein psychologischer Verhaltenstherapeut und ein Facharzt für Innere Medizin mit Zusatzbezeichnung Psychotherapie mit 8 bzw. 10 Jahren klinischer psychotherapeutischer Erfahrung) schätzte unabhängig voneinander ein, welche Items aus den in der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin angewandten bereits etablierten KTT-basierten psychometrischen Verfahren theoretisch für die Angstmessung geeignet sind. Aus einem anfänglichen Itempool von 125 vorselektierten Items (siehe Anhang 9.1.) wurden aufgrund des Iteminhalts 81 Items (mit 2- bis 7-stufigen Likert-skalierten Antwortformaten) von der Forschungsgruppe ausgewählt, welche 13 bewährten klinisch-psychologischen Instrumenten entstammen (ADS[58], ALL[59], BDI[60], BSF[61], GBB[62], GT[63], NI-90[64], PGWI[65], PSQ[66], SF36[67], SKT[68], STAI[69], SWO[70]; siehe Kapitel 5.2.).

[58] ADS: Allgemeine-Depressions-Skala (Hautzinger & Bailer, 1993).
[59] ALL: Fragebogen zum Alltagsleben (Bullinger et al., 1993).
[60] BDI: Beck-Depressions-Inventar (Hautzinger et al., 1994).
[61] BSF: Berliner-Stimmungs-Fragebogen (Hörhold & Klapp, 1993; Rose et al., in Druck).
[62] GBB: Gießener-Beschwerde-Bogen (Brähler & Scheer, 1995).

Tabelle 9: Theoretisch selektierter Itempool (N = 81 Items), welcher zur Testentwicklung des Angst-CATs genutzt wurde.

Itemtext
Ich fühle mich:
Gelöst.
Besorgt.
Beunruhigt.
Kribbelig.
Ausgeglichen.
Unsicher.
Wie fühlen Sie sich jetzt, d. h. in diesem Moment?
Ich bin ruhig.
Ich fühle mich geborgen.
Ich fühle mich angespannt.
Ich bin gelöst.
Ich bin aufgeregt.
Ich bin besorgt, dass etwas schief gehen könnte.
Ich bin beunruhigt.
Ich fühle mich wohl.
Ich fühle mich selbstsicher.
Ich bin nervös.
Ich bin zappelig.
Ich bin verkrampft.
Ich bin entspannt.
Ich bin besorgt.
Ich bin überreizt.
Ich fühle mich durch folgende Beschwerden belästigt:
Herzklopfen, Herzjagen oder Herzstolpern.
Ohnmachtsanfälle.
Schwindelgefühl.
Starkes Schwitzen.
Anfälle.
Übelkeit.
Kloßgefühl im Hals.
Drang zum Wasserlassen.
Schluckbeschwerden.
Gefühl der Benommenheit.
Taubheitsgefühl (Einschlafen, Absterben, Brennen oder Kribbeln in Händen und Füßen).
Hitze, Hitzewallungen.
Durchfälle.
Stiche, Schmerzen oder Ziehen in der Brust.
Zittern.
Leichtes Erröten.
Anfallsweise Atemnot.
Anfallsweise Herzbeschwerden.

[63] GT: Gießen-Test Selbst & Idealselbst (Beckmann et al., 1991).
[64] NI: Narzissmus-Inventar (NI: Deneke & Hilgenstock, 1989; NI-90: Schöneich et al., 2000).
[65] PGWI: Psychological General Wellbeing Index (Ludwig et al., 1990).
[66] PSQ: Perceived Stress Questionnaire (Levenstein et al., 1993).
[67] SF36: Fragebogen zum Gesundheitszustand (Bullinger & Kirchberger, 1998).
[68] SKT: Subjektive-Krankheits-Theorie-Ursachenvorstellung (Faller, 1997).
[69] STAI: State Trait Anxiety Inventory (Laux et al., 1981).
[70] SWO: Fragebogen zu Selbstwirksamkeit,Optimismus und Pessimismus (Scholler et al.,1999).

Tabelle 9 (Fortsetzung): Theoretisch selektierter Itempool (N = 81 Items), welcher zur Testentwicklung des Angst-CATs genutzt wurde.

Itemtext
Die Aussage stimmt...
Ich halte mich für sehr wenig ängstlich.
Ich glaube, ich mache mir verhältnismäßig selten Sorgen um andere Menschen.
Ich habe manchmal plötzlich furchtbare Angst, schwer krank werden zu können.
Es könnte mir schon gefallen, einmal so richtig im Mittelpunkt zu stehen.
Man kann sich furchtbar schämen, wenn man glaubt, versagt zu haben.
Manchmal quält mich das unbestimmte Gefühl, irgendetwas sei mit meinem Körper nicht in Ordnung.
In manchen Zeiten sehe ich alles so schwarz, dass mich eine furchtbare Panik ergreift.
Es gibt Stunden, in denen ich das Gefühl habe, nicht wirklich da zu sein.
Menschenansammlungen schrecken mich eher ab.
Ich beobachte meinen Körper ziemlich genau, um verdächtige Krankheiten möglichst früh zu erkennen.
Ich erlebe mich manchmal wie eine fremde Person.
Die Vorstellung, selbst mal im Rampenlicht zu stehen, ist schon verführerisch.
Es ist mir meistens unheimlich peinlich, wenn ich vor einer Gruppe etwas Dummes gesagt habe.
Mitunter bin ich so von Angst und Unruhe getrieben, dass ich weder ein noch aus weiss.
Ich würde mich auf sehr viel mehr Herausforderungen einlassen, wenn ich nicht Angst hätte, meine Gesundheit würde das nicht durchstehen.
Es macht mich völlig unsicher, wenn sich in einer Gruppe die Aufmerksamkeit aller plötzlich auf mich richtet.
Manchmal erscheint mir mein Körper plötzlich fremd und nicht zu mir dazugehörig.
Es beunruhigt mich, dass heutzutage von so vielen neuen Krankheiten berichtet wird.
Ich erwarte, dass meine Gesundheit nachlässt.
Wie haben Sie sich in dieser Woche einschließlich heute gefühlt?
Ich mache mir so große Sorgen über gesundheitliche Probleme, dass ich an nichts anderes mehr denken kann.
Schwierigkeiten sehe ich gelassen entgegen, weil ich mich immer auf meine Fähigkeiten verlassen kann.
Während der letzten Woche:
Haben mich Dinge beunruhigt, die mir sonst nichts ausmachen.
Hatte ich Mühe, mich zu konzentrieren.
Hatte ich Angst.
Konnten Sie in der letzten Woche:
Es sich bequem machen und sich entspannen?
Wie oft waren Sie in den letzten Wochen sehr nervös?
Wie oft waren Sie in den letzten Wochen ruhig und gelassen?
Haben Sie im vergangenen Monat (i.v.M.) unter Nervosität oder Ihren „Nerven“ gelitten?
Waren Sie im allgemeinen angespannt oder haben Sie irgendwelche Spannungen verspürt?
Haben Sie i.v.M. wegen Ihrer Gesundheit Sorgen oder Befürchtungen gehabt?
Waren Sie i.v.M. ängstlich, besorgt oder aufgeregt?
I.v.M. war ich ausgeglichen und mir meiner selbst sicher.
Haben Sie sich i.v.M. entspannt und gelassen oder angespannt und aufgeregt gefühlt?
Könnten Ihre Beschwerden daher kommen, dass Sie an inneren Ängsten leiden?
Wie häufig trifft diese Feststellung im allgemeinen auf Sie zu?
Sie fürchten, Ihre Ziele nicht erreichen zu können.
Sie fühlen sich ruhig.
Sie fühlen sich angespannt.
Sie fühlen sich sicher und geschützt.
Sie haben viele Sorgen.
Sie haben Angst vor der Zukunft.
Sie sind leichten Herzens.
Sie haben Probleme, sich zu entspannen.

5.3.2. Statistische Itemanalyse und -selektion

Die statistische Itemanalyse und –selektion erfolgte an den drei oben beschriebenen Teilstichproben (siehe Kapitel 5.2.2.). Das methodische Vorgehen lehnt sich an das Vorgehen der US-amerikanisch/dänischen Forschungsgruppe um Ware und Mitarbeiter an, welche die Anwendbarkeit der IRT in Form von CATs im Bereich der Lebensqualitätsforschung verfolgen (Ware et al., 2000, 2003).

5.3.2.1. Unidimensionalität: Faktorenanalysen und Analyse residualer Kovarianzen

Aufgrund des aktuellen Forschungsstands (ungenügende Differenzierbarkeit von Komponenten des Angst-Konstruktes; siehe Kapitel 2.7.3.4.) und der zu diesem Zeitpunkt methodischen Möglichkeiten bzw. praktischen Begrenzungen, sowie aus Gründen der Ökonomie wird die Entwicklung eines unidimensionalen Angst-CATs angestrebt. Daher stellt die Untersuchung der Dimensionalität den ersten Schritt im Prozess der statistischen Itemanalyse und –selektion dar.

Es ist umstritten, welches Verfahren für die Bestimmung der Dimensionalität einer Datenmatrix am geeignetesten erscheint (Hattie, 1984). So hat Hattie bereits 1984 ein Dutzend der derzeit angewandten Verfahren zur Testung der Unidimensionalität überprüft (Hattie, 1984). Diese beruhten auf folgenden Ansätzen: a) der Konsistenz des Antwortmusters der Probanden, b) der Reliabilität des Skalenwertes, c) der Ergebnisse von Faktorenanalysen, d) der Gegenüberstellung linearer und nichtlinearer Faktorenlösungen oder e) anderer Fittinganalysen mit anschließender Beurteilung der residualen Kovarianzen.

Die meisten der eingesetzten Verfahren erschienen Hattie mit großen Mängeln behaftet zu sein. Laut Embretson und Reise (2000) könne man bei der Gesamtsicht der Arbeiten in diesem Bereich (Nandakumar, 1993, 1994; Nandakumar & Stout, 1993; Stout, 1987, 1990) den Schluss ziehen, dass, nachdem die gemeinsame Varianz der Items einem Hauptfaktor zugeordnet würde, der das zu messende Merkmal („latentes Trait“) repräsentiere, eine Analyse der residualen Kovarianzen derzeit die sinnvollste Aussage über die Dimensionalität der Daten erlaube, wobei es offenbar eine nachgeordnete Rolle spiele, mit welcher Methodik der gemeinsame Faktor identifiziert werde. Auch Waller und Mitarbeiter (1996) halten eine Analyse residualer Kovarianzen als Methode zur Dimensionalitätsüberprüfung für sehr reliabel. Und Hambleton,

Swaminathan und Rogers (1991) verweisen insbesondere auf den hohen Stellenwert der Analyse von Residuen im Rahmen der Untersuchung der Unidimensionalität. Sie sehen in dieser Methodik die vielleicht „wertvollste Goodness-of-Fit Data“ überhaupt. Wir haben uns dem Itemselektionsvorgehen von Ware und Mitarbeitern (2000, 2003) angeschlossen, welche vor dem Hintergrund langjähriger Erfahrung mit der Entwicklung IRT-basierter CATs im U.S.-amerikanischen Sprachraum – ähnlich wie oben genannte Autoren es empfehlen - sowohl Faktorenanalysen als auch Analysen residualer Kovarianzen bei der Itemanalyse und –selektion kombinieren. Das methodische Vorgehen zur Untersuchung der Unidimensionalität geschieht demnach in folgender Reihenfolge:

1. eine explorative Faktorenanalyse,
2. eine konfirmatorische Faktorenanalyse
   a) mit einer Analyse residualer Kovarianzen und
   b) der Berechnung von Fit-Indizes.

Das zugrundeliegende Konstrukt wird zunächst mittels explorativer Faktoren-analysen (Programm: SPSS) untersucht.

Da theoretisch zu erwarten ist, dass sich die Datenmatrix durch mehr als einen Faktor abbilden lässt (zur Mehrdimensionalität des Angst-Konstruktes siehe Kapitel 2.7.3.4.), erscheint es sinnvoll, die explorative Faktorenanalyse um eine Untersuchung der Mehrdimensionalität anhand der von Lautenschlager (1989) publizierten „Zufallseigenwerte“, welche aus vielen Monte-Carlo-Studien gewonnen wurden („parallel analysis criterion“; Longman, Cota, Holden & Fecken, 1989; Humphreys & Montanelli, 1975; nach dem Verfahren der Parallelanalyse von Horn, 1965)[71] und dem Everett-Kriterium (1983) zu ergänzen. Mit diesem Vorgehen soll exploriert werden, ob mehrere überzufällige und stabile Faktoren mittels einer Faktorenanalyse extrahiert werden können, welche zu einem Informationsverlust führen könnten, wenn sie nicht in der Itembankkonstruktion berücksichtigt würden.

[71] Es wurden keine eigenen Parallelanalysen über die Daten gerechnet. Jedoch listet Lautenschlager in einem Artikel von 1989 in Tabellen aus vielen Monte-Carlo-Studien generierte „Zufallseigenwerte“ aus Korrelationsmatrizen für $5 \leq p \leq 80$ und $50 \leq n \leq 2000$ auf, die mit Hilfe geeigneter Interpolationstechniken für praktisch alle faktorenanalytischen Anwendungen genutzt werden können, um die Anzahl der bedeutsamen Faktoren zu bestimmen (Bortz, 1999, S. 529).

Da das Ziel der Testkonstruktion die Erstellung einer eindimensionalen Itembank ist, wird - nach inhaltlicher Überprüfung des ersten unrotierten Faktors - dieser als Selektionsgrundlage für die Konstruktion des Angst-CATs genutzt, da er mehr Varianz aufklärt als nachfolgend extrahierte Faktoren. Dies wird durch das der Hauptkomponentenanalyse zugrunde liegende Prinzip der sukzessiven maximalen Varianzaufklärung garantiert. Items, welche auf diesem ersten Faktor hoch laden (als Selektionskriterium dient eine Faktorenladung von $> |+/-.4|$), werden zur weiteren Itemanalyse ausgewählt.

Die so ausgewählten Items werden im Anschluss einer konfirmatorischen Faktorenanalyse mit einer Analyse residualer Kovarianzen unterzogen (Programm Mplus; Muthén & Muthén, 1998). Diese dient der Homogenisierung des Itempools durch den Ausschluss von Items, welche hohe residuale Kovarianzen aufweisen (ausgeschlossen werden Items mit residualen Kovarianzen von r > 0,3 in Anlehnung an Cella / North Western University Chicago: <.40/>.30 bzw. Ware / Tufts & Harvard-University Boston: <.30/>.20).

Anschließend werden über die so selektierten Itemmengen Fit-Indizes zur Beurteilung der konfirmatorischen Ein-Faktor-Lösung mit Hilfe des Computerprogramms AMOS (Arbuckle & Worthke, 1999) berechnet. Das der konfirmatorischen Faktorenanalysen zugrundeliegende Messmodell ist ein multivariates lineares Regressionsmodell, welches die Beziehung zwischen einem Set von abhängigen beobachteten Variablen (hier: die selektierte Itemmenge der jeweiligen Teilstichprobe) und einer latenten Variable (hier: das angenommene „Angst“-Konstrukt) mit Hilfe der Mittelwerte als zu schätzende Parameter beschreibt. Zur Beurteilung der Anpassung eines Ein-Faktor-Modells an die Daten werden im Rahmen konfirmatorischer Faktorenanalysen die folgenden globalen Fit-Indizes berechnet: χ^2-Statistiken, der Root Mean Square Error of Approximation (RMSEA; Steiger & Lind, 1980), der Tucker-Lewis-Index (TLI; Tucker & Lewis, 1973) und der Comparative Fit-Index (CFI; Bentler, 1990). Weil χ^2-Statistiken - wie von vielen Autoren eingeräumt wird (Bentler & Bonett, 1980; Browne & Mels, 1992; Gulliksen & Tukey, 1958; Jöreskog, 1969) - stark stichprobenabhängig sind, ist ihr Nutzen bei der Beurteilung (und Wahl) eines Modells gering. RMSEA, TLI und CFI dagegen sind Fit-Indizes, welche die Stichprobengröße, Freiheitsgrade und eine Reihe von weiteren Parametern bei

ihrer Berechnung berücksichtigen, und daher einen größeren Beurteilungswert als χ^2-Statistiken haben.

5.3.2.2. IRT-Analyse

5.3.2.2.1. Item Response Curves (IRCs)

Die Item Response Theorie (IRT) ermöglicht es, Kategorienfunktionen einzelner Antwortkategorien durch die grafische Betrachtung von Item Response Curves (IRCs) zu untersuchen, Item- und Testinformationskurven zu analysieren sowie Standardmessfehler und Reliabilität einer Skala in Abhängigkeit vom geschätzten Merkmalsausprägungsniveau zu berechnen (siehe Kapitel 3.3.3.). Das Programm TestGraf (Ramsay, 1995) stellt mittels einer nonparametrischen Glättungsfunktion namens „Kernel-Smoothing-Technique“ IRCs grafisch dar und erlaubt die Berechnung oben genannter Statistiken.

Item Response Curves (IRCs) sind grafische Darstellungen der (Antwort-) Kategorienfunktionen von Items (siehe Abbildung 10). Sie veranschaulichen die Antwortwahrscheinlichkeit der einzelnen Antwortkategorien (Ordinate) in Abhängigkeit von der latenten Merkmalsausprägung (Theta) der Angst (Abszisse).

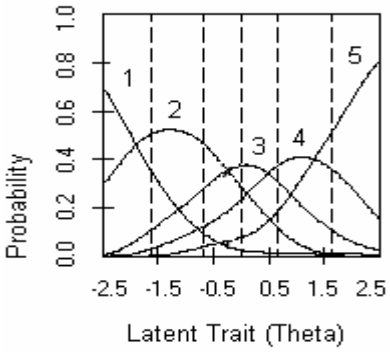


Abbildung 10: Exemplarische Darstellung eines polytomen Items mit modellkonformen Item Response Curves (IRCs).[72]

Das latente Merkmalsausprägungskontinuum der Angst wird in Einheiten einer abweichungsnormierten Standardnormalverteilung[73] dargestellt. In der vorliegenden Untersuchung wurde nicht das Rasch-Modell angewandt, bei dem der Steigungsparameter (a_i) stets auf „1“ fixiert ist, sondern das Generalized

[72] Die Darstelllung der Item Response Curves (IRC) in Abbildung 10 entstammt dem Programm TestGraf. Es modelliert die Daten nonparametrisch. Das in vorliegender Arbeit zur Itemparameterschätzung genutzte Programm Parscale (GPCM-Modellierung) gibt keine grafische Darstellung der IRCs aus.

[73] Dies ist die in den U.S.A. gebräuchliche Variante.

Partial Credit Modell (GPCM; Muraki, 1992) verwendet (siehe Kapitel 3.4.3. und 5.3.2.3.). Dieses erlaubt eine variable Steigung der verschiedenen Kurvenverläufe der einzelnen Itemantwortkategorien.

Die Kategorienfunktionen können nicht nur grafisch dargestellt werden, sondern auch in Form einer mathematischen Gleichung beschrieben werden, welche der darauffolgenden Schätzung der Itemparameter dient (Kapitel 3.3.1., 3.4.3. und 5.4.3.1.). Die zu schätzenden Itemparameter finden sich in der grafischen Darstellung der IRCs wieder. So nennen sich die Schnittpunkte der IRCs „Thresholds“ (Schwellen) und der Mittelwert der Schwellen „Location Parameter“ (Lokationsparameter). Der Lokationsparameter dient der Lokalisation des Items auf dem latenten Traitkontinuum. Die gemittelte Steigung der einzelnen Kurven wird durch den „Slope Parameter“ (Steigungsparameter) ausgedrückt und kann mittels des Programms Parscale (Muraki & Bock, 1999) errechnet werden.

Die grafische Darstellung der Kategorienfunktionen (IRCs) kann zur differenzierten Beurteilung der psychometrischen Qualität der Items genutzt werden. Items mit „guten“ (i. S. von modellkonformen) Kategorienfunktionen zeichnen sich durch IRCs aus, welche pro Antwortkategorie eingipflige, glockenförmige, jedoch nicht unbedingt symmetrische Kurvenverläufe aufweisen, die bis zu einem Kurvenmaximum stetig ansteigen und danach stetig abfallen (Santor & Coyne, 2001). Zudem sollte die *Anordnung* der einzelnen IRCs auf dem geschätzten latenten Kontinuum der Angstausprägung der im Antwortformat vorgegebenen *Abstufung der Ratingstufen* entsprechen. Die IRC der ersten Antwortkategorie verhält sich stets monoton fallend, die der letzten Antwortkategorie stets monoton steigend (siehe Abbildung 10, IRC Nr. 1 und 5).

Als „ungenügend“ werden IRCs beurteilt, wenn sie nicht zwischen unterschiedlichen Ausprägungen der Angst auf dem latenten Kontinuum zu diskriminieren vermögen. Ungenügend sind IRCs also dann, wenn die Kurvenverläufe pro Antwortkategorie mehrgipflig sind und sich die Kurvenverläufe verschiedener Antwortkategorien mehrfach überschneiden (siehe Kapitel 5.4.2.1.).

5.3.2.2.2. Testinformationsfunktion, Standardmessfehler und Reliabilität

Das Programm TestGraf (Ramsay, 1995) ermöglicht ferner die Beurteilung der Item- bzw. Testinformationsfunktion. Eine Iteminformationsfunktion gibt an, wieviel Information ein Item über die Merkmalsausprägungen verschiedener Personen zu liefern vermag, d. h. wie informationsreich ein Item ist.

Die Summe der Iteminformationen der zu einer Skala gehörigen Items ergibt die Testinformation (siehe Kapitel 3.3.3.; Muraki, 1993). Eine Auswahl der Items mit modellkonformen IRCs, welche Indikatoren für eine gute Diskriminationsfähigkeit des Items sind, wirkt sich positiv auf die gesamte Testinformationsfunktion aus, da nur die Items mit einer hohen Iteminformationsfunktion selektiert werden.

Die Informationsfunktion wird im Program TestGraf desweiteren genutzt, um den Standardmessfehler (G.1) zu berechnen und die Reliabilitätsfunktion (G.2) abzuleiten. Gleichung G.1 veranschaulicht den negativen Zusammenhang zwischen Informationsfunktion $I(\theta)$ und Standardmessfehler $s_e(\theta)$. Der Standardmessfehler $s_e(\theta)$ ist in seiner Größe von $I(\theta)$ abhängig.

(G.1): $$s_e(\theta) = 1 / \sqrt{I(\theta)}$$

Aus der Formel G.1 und der in der Klassischen Test-Theorie gebräuchlichen Formel zur Berechnung der Reliabilität (G.2), lässt sich die in der Item Response Theorie (IRT) genutzte Formel zur Reliabilitätsbestimmung (G.3) ableiten.

(G.2): $$\mathrm{Rel}(x) = \frac{s_w^2}{s_w^2 + s_e^2}$$

w = Wahrer Wert; e = (error) Fehler Wert

In der IRT werden keine Aussagen über die in der KTT postulierten „wahren Werte“ (w) getroffen, sondern es werden Schätzungen der Merkmalsausprägung („latent trait“; Theta, θ) vorgenommen (siehe Kapitel 3.3.). Auf die Transformation von Theta auf eine Standardnormalverteilung wurde bereits hingewiesen, woraus sich eine Varianz der wahren Werte von $s_w^2 = 1$ ergibt. Setzt man dies zusammen mit Gleichung G.1, welche die Fehlervarianz bezogen auf θ als $1/I(\theta)$ definiert, in Gleichung G.2 ein, so lässt sich die in Gleichung G.3 dargestellte Reliabilitätsfunktion ableiten (Ramsay, 1995, S. 60).

(G.3): $$\mathrm{Rel}(\theta) = \frac{1}{1 + 1/\mathrm{I}(\theta)}$$

Die Formeln sollen verdeutlichen, dass in der IRT die Informationsfunktion, der Standardmessfehler und die Reliabilität in einer engen Beziehung zueinander stehen.

5.3.2.3. IRT-Modellierung

5.3.2.3.1. Itemparameterschätzung

Welches IRT-Modell das geeigneteste zur Darstellung der Daten ist, hängt im Wesentlichen von der Art der Daten ab (Kapitel 3.4.). So weisen Fragebögen zur Erfassung psychologischer Konstrukte, wie Stimmungen, Beschwerden etc. typischerweise polytome, ordinal geordnete Antwortformate auf. Da hier keine „richtigen“ Antworten geraten werden können, wie dies z. B. bei Leistungstests der Fall ist, kommen prinzipiell sogenannte Ein- und Zwei-Parameter-Modelle in Frage (Kapitel 3.4.1.und 3.4.4.). Diese unterscheiden sich darin, dass bei den Ein-Parameter-Modellen davon ausgegangen wird, dass sich die Items lediglich in ihrem Schwierigkeitsgrad (IRT-Terminologie: „Item Response Thresholds“ bzw. „Location Parameter“) unterscheiden, aber nicht in ihrer Diskriminationsfähigkeit, d. h. der Steilheit der Kurven („Slope Parameter“). Ein solches Modell wäre z. B. das Rating Scale Modell (RSM) von Andrich (1978). Die Anwendung dieses Modells impliziert, dass Items mit unterschiedlichen Antwortformaten in isolierten Gruppen analysiert werden müssen, so dass diese Anwendung für unsere Daten weniger geeignet ist. Als allgemeineres Ein-Parameter-Modell steht das Partial Credit Modell (PCM; Masters, 1982) zur Verfügung. Sowohl das RSM wie auch das PCM können als „Rasch-Modelle für polytome Daten“ charakterisiert werden (Kapitel 3.4.4.). Tatsächlich unterscheiden sich die Items in der von uns untersuchten Stichprobe hinsichtlich ihrer Diskriminationsfähigkeit (Kapitel 5.4.3.1.), so dass es notwendig ist, auch die „Steigungsparameter“ zwischen den Items variieren zu lassen. Von den Zwei-Parameter-Modellen kommen das Graded Response Modell (GRM; Samejima, 1996) und die Modifikation dieses Modells durch Muraki (1992; M-GRM) sowie das Generalized Partial Credit Modell (GPCM; Muraki, 1997) in Frage.[74] Bei den heterogenen Antwortformaten stößt man beim M-GRM auf das gleiche Problem wie beim RSM, dass die Items in isolierten Gruppen analysiert werden

[74] Abkürzungen der IRT-Modelle nach Embretson und Reise (2000).

müssen. Wir haben daher die Itemparameterschätzungen auf der Grundlage des Generalized Partial Credit Modells (GPCM; Muraki, 1997) durchgeführt. Dieses ist in Kapitel 3.4.3. bereits in seinen Grundzügen erörtert worden.

Mit Hilfe des Programms Parscale (Muraki & Bock, 1999) werden anhand der logistischen Item Response Function (IRF; siehe Kapitel 3.4.3. Gleichung G.3.) des GPCMs folgende Itemparameter[75] geschätzt: a_i: „Slope Parameter" (Steigungsparameter), b_{ih}: „Item Threshold Parameter" (Schwellenparameter), b_i: „Location Parameter" (Lokationsparameter) und d_{hi}: „Item Category Parameter" (Antwortkategoriengrenzen). Im Rahmen der Itemparameterschätzung dient als ein Selektionskriterium zur Optimierung der Itembank ein Steigungsparameter von $a_i > 0{,}80$. Dieses Kriterium wurde in Anlehnung an eine Empfehlung von Dr. Bjørner (National Institute of Occupational Health in Kopenhagen) gewählt, um eine möglichst hohe Diskriminationsfähigkeit der Items zu gewährleisten.

5.3.2.3.2. „Differential-Item-Functioning" (DIF)

Voraussetzung für ein „Item-Link-Design" (siehe Kapitel 5.3.2.3.3.) ist das Fehlen von „Differential-Item-Functioning" (DIF; Holland & Wainer, 1990) zwischen den „Anker-Items" verschiedener Teilstichproben. „Anker-Items" sind Items, welche in allen Teilstichproben gleichermaßen vorliegen. Zwischen den Anker-Items verschiedener sich überlappender Teilstichproben darf also keine systematische Antwortverzerrungstendenz (genannt „item bias" oder „DIF") vorliegen. DIF läge z. B. vor, wenn die Itemparameterschätzung der Anker-Items von der Teilstichprobe, in der sie erhoben wurde, abhängig wäre. Eine solche Instabilität in der Itemparameterschätzung würde eine Metrisierung der Itemparameter der Items beider Stichproben anhand der Anker-Items verbieten.

Von den verschiedenen zur Verfügung stehenden Verfahren (Swaminathan & Rogers, 1990; Zumbo, 1999) entschieden wir uns für ein IRT-basiertes Vorgehen. Die Untersuchung wurde mittels des Computerprogramms Parscale (Muraki & Bock, 1999) durchgeführt, mit dem DIF getrennt für Steigungs- und Lokationsparameter berechnet werden kann. Hierzu werden zunächst die genannten Itemparameter für die Anker-Items der zu vergleichenden einzelnen Teilstichproben berechnet, um anschließend mit Hilfe von χ^2-Statistiken die

[75] Zum Verständnis von Itemparametern siehe Kapitel 3.3.1., zur Taxonomie von IRT-Modellen nach der Anzahl der berücksichtigten Itemparameter siehe Kapitel 3.4.1..

Unterschiedlichkeit der Itemparameterschätzungen der Anker-Items zwischen den Teilstichproben auf signifikante Abweichungen von der Nullhypothese überprüfen zu können. Das Fehlen von DIF ist essentiell, da es die Annahme der Invarianz der Itemparameter zwischen den einzelnen Stichproben bekräftigt, und somit die Realisierung eines „Item-Link-Designs“ erlaubt.

5.3.2.3.3. „Item-Link-Design“

Um die Items der drei Teilstichproben, welche den Selektionskriterien genügen, auf einer gemeinsamen Skala abzubilden, so dass sie als eine Itembank des Angst-CATs fungieren können, bedarf es des „Linkings“ („Verkettung“ / „Verbinden“) der Teilstichproben (Embretson & Reise, 2000). Dieses Verbinden erfolgt über ein gemeinsames Set von Items („Anker-Items“), welches in den zu verbindenden Stichproben gleichermaßen vorliegt (siehe Kapitel 5.2.2. und 5.3.2.3.3.).

Die Anker-Items werden genutzt, um eine angemessene lineare „Linking Transformation“ zu ermöglichen, welche die Itemparameter aller selektierten Items der Teilstichproben auf einer gemeinsamen Skala kalibriert. Diese Kalibrierung erfolgt mit dem Programm Parscale (Muraki & Bock, 1999).

Es vergleicht die Itemparameter der Anker-Items der ersten und zweiten (bzw. dritten) Teilstichprobe, indem es die Mittelwertsunterschiede der Itemparameter sowie die Differenzen bezüglich der Standardabweichungen berechnet. Anschließend wird eine Adjustierung der Itemparameter der Anker-Items der zweiten Stichprobe auf die Itemparameter der Anker-Items der ersten Stichprobe vollzogen ($slope_2 = slope_1 x SD_2$; $location_2 = (location_1 – mean_2)/SD_2$; $step_2 = step_1 x SD_2$; step = category threshold; Terminologie nach Parscale, Muraki & Bock, 1999). Dann erfolgt eine Re-Kalibrierung der Itemparameter der verbleibenden sich nicht überlappenden Items zwischen der zweiten (bzw. dritten) und der ersten Teilstichprobe, indem die adjustierten Itemparameter der Anker-Items (Steigungs- und Schwellenparameter) fixiert werden.

5.3.2.3.4. „Item-Fit-Statistiken“

Um die Güte der Anpassung des Generalized Partial Credit Modells (Muraki, 1992) an die Daten zu bestimmen, besteht derzeit kein allgemein akzeptiertes und etabliertes Verfahren (Embretson & Reise, 2000). Während für

Ein-Parameter Modelle einige Fit-Statistiken gebäuchlich sind, ist die Prüfung des Item-Fits bei Zwei-Parameter-Modellen noch in der Entwicklung. Ein besonderes methodisches Problem dieser Item-Fit-Statistiken zur Überprüfung der Modellkonformität zweiparametrischer Modelle liegt in ihrer Abhängigkeit von der untersuchten *Stichprobengröße*, welche von vielen Forschern bemängelt wird (Embretson & Reise, 2000; Hambleton et al., 1991; Van der Linden & Hambleton, 1997 und Muraki, 1997). Simulationsstudien von Hambleton und Mitarbeitern (1991) zeigen beispielsweise, dass die Anzahl zufälliger „Item-Misfits“ mit zunehmender Stichprobengröße steigt. So wurden im Rahmen einer Simulationsstudie mit 50 Items und einer Stichprobengröße von N = 1.200 Personen 10 artifizielle „Item-Misfits“ von den Autoren entdeckt. In einer weiteren empirischen Studie fanden Reise und Waller (1990) im Rahmen einer IRT-basierten Analyse des Multidimensional Personality Questionnaires (MPQ; Tellegen, 1982), dass bei der Analyse von Daten von N = 2.000 Personen, 36 von 300 Items einen signifikanten (artifiziellen) Item-Misfit aufwiesen. Provokativ formulierte McDonald dieses methodische Problem bereits 1989 wie folgt: falls ein IRT-Modell im Rahmen einer Untersuchung *nicht* zurückgewiesen würde, sei dies als ein Zeichen zu werten, dass die Stichprobengröße zu *klein* gewesen sei.

Die mehrfachen empirischen Belege, dass Likelihood-χ^2-Tests sehr sensitiv auf die Stichprobengröße reagieren, veranlassten Embretson und Reise (2000) von der Nutzung dieser Fit-Statistiken als „solid decision-making tools“ (S. 235) im Itemselektionsprozess abzuraten.

Demnach verzichteten in den letzten Jahren zunehmend Forscher, welche 2PL-Modelle (wie das GRM; Samejima, 1969) zur Itemanalyse im Bereich der Persönlichkeitsdiagnostik anwandten, gänzlich auf die Publikation von Fit-Statistiken zur Modellanpassungsgüte (Childs, Dahlstrom, Kemp & Panter, 2000; Gray-Little, Williams & Hancock, 1997; Reise & Henson, 2000).

Da uns aus persönlichen Kontakten zu anderen Forschungsgruppen jedoch bekannt ist, dass die Likelihood-χ^2-Statistiken – aus Mangel an Alternativen – der einzige bislang genutzte Weg zur Beurteilung des Modell-Fits sind, erscheint es uns – obgleich viele Forscher diese nicht (mehr) publizieren – sinnvoll, diese Methodik hier anzuwenden, um die Kommunikation mit anderen Forschungsgruppen über das Fit-Statistik-Problem aufrecht zu erhalten und zu

erleichtern. Dies ist insofern von Belang, als meines Erachtens nur eine Problemfokussierung einen Forschungsanstoß für die Entwicklung besserer Fit-Statistiken zu geben vermag.

Dazu wirde im Folgenden die nach Formel G.5 berechneten Likelihood-χ^2-Statistik (G_i^2) zur Beurteilung des Modell-Fits (für jedes Item) erläutert, welche mit Hilfe des Programms Parscale errechnet wurden (Muraki, 1997, S. 160).

(G.5): $$Gi^2 = 2 \sum_{k=1}^{Ki} \sum_{h=1}^{mi} rkih \ln \frac{rkih}{NkiPih(\bar{\theta}k)}$$

Nachdem für jede Testperson die Angstausprägung (θ) auf der Basis ihres individuellen Antwortmusters mittels des EAP-Algorithmus (Bock & Mislevy, 1982) geschätzt wird, können die θ-Scores jeweils spezifischen Intervallen k auf dem θ-Kontinuum zugeordnet werden. Daraufhin können a) die beobachteten Häufigkeiten der *h*-ten Antwortkategorien eines Items *i* im Intervall K (r_{kih}) und b) die Anzahl der Testpersonen (N_{ki}), welche einem Item *i* im *k*-ten Intervall zugeordnet wurden, berechnet werden. Daraus lassen sich pro Item für jedes K-Intervall m_i Kontigenztabellen erstellen. Es erfolgt eine Reskalierung der θ-Scores in der Form, dass die Varianz der Stichprobenverteilung der latenten Verteilungsannahme, auf der die MML-Schätzung (Marginal Maximum Likelihood; Dempster, Laird & Rubin, 1977) der Itemparameter beruht, gleicht. Für jedes Intervall wird dann die Wahrscheinlichkeit des Mittelwerts $(\bar{\theta}k)$ pro Antwortkategorie und Item auf der Grundlage der reskalierten θ-Scores und der IRF (Item Response Function) des GPCMs $Pih(\bar{\theta}k)$ berechnet. Nach Gleichung G.5 werden sodann Likelihood-χ^2-Tests (G_i^2) errechnet, wobei K_i die Anzahl der Intervalle ist, welche sich aus einer Zusammenfassung benachbarter Intervalle ergibt, die dazu dient, erwartete Werte von $NkiPih(\bar{\theta}k)$ von kleiner als 5 zu vermeiden. Die Zahl der Freiheitsgrade ist das Produkt der Anzahl der Intervalle K_i und m_i –1.

5.4. Ergebnisse

Im Folgenden werden die Ergebnisse der statistischen Itemanalyse und -selektion der drei in Kapitel 5.2.2. beschriebenen untersuchten Teilstichproben zusammengefasst. Die Präsentation der Ergebnisse in diesem Kapitel (5.4.) ist in die einzelnen methodischen Teilschritte untergliedert, welche in Kapitel 5.3. erläutert wurden. Es werden pro Methodenschritt jeweils die Ergebnisse der Untersuchungen an den drei Teilstichproben nacheinander berichtet, da die Itemanalyse und –selektion pro Teilstichprobe separat erfolgte.

Daran schließt sich die Erörterung der Ergebnisse des „Item-Link-Designs“ an, welches die selektierten Items der drei getrennt voneinander analysierten Teilstichproben so miteinander verknüpft, dass sie die Itembank des Angst-CATs konstituieren. Abschließend wird die IRT-Modellierung der gesamten Itembank dargestellt.

5.4.1. Unidimensionalität

Die Itemanalysen vollzogen sich separat an drei verschiedenen Personen- und Itemstichproben (siehe Kapitel 5.2.2.).

Die Dimensionalität wurde zunächst pro Stichprobe mittels explorativer Faktorenanalysen (Hauptkomponentenanalysen) mit dem Programm SPSS untersucht. Es wurden ein- und mehrfaktorielle Faktorenlösungen errechnet. Die Anzahl der extrahierten Faktoren der mehrfaktoriellen Lösungen richten sich nach dem Everett-Kriterium (Everett, 1983) und dem Parallelanalyse-Kriterium („parallel analysis criterion“; Longman, Cota, Holden & Fecken, 1989; Humphreys & Montanelli, 1975; nach dem Verfahren der Parallelanalyse von Horn, 1965). Es wurden keine eigenen Parallelanalysen über die Daten gerechnet. Jedoch listet Lautenschlager in einem Artikel von 1989 in Tabellen aus vielen Monte-Carlo-Studien generierte „Zufallseigenwerte“ aus Korellationsmatrizen für $5 \leq p \leq 80$ und $50 \leq n \leq 2000$ auf, die mit Hilfe geeigneter Interpolationstechniken für praktisch alle faktorenanalytischen Anwendungen genutzt werden können, um die Anzahl der bedeutsamen Faktoren zu bestimmen (Bortz, 1999, S. 529). Die Nutzung dieser „parallel analysis criteria“ wird hier als alternative Methode gegenüber der aufwendigen Berechnung einer Parallelanalyse (Horn, 1965) zur zufallskritischen Bewertung der Faktorenanzahl genutzt.

Zur Konstruktion eines unidimensionalen Angst-CATs wurden die Items ausgewählt, welche auf dem ersten unrotierten Faktor eine hohe Ladung aufwiesen (erster Selektionsschritt).

Anschließend wurden konfirmatorische Faktorenanalysen - wie in Kapitel 5.3.2.1. dargestellt - gerechnet. In diesem Rahmen wurden Analysen residualer Kovarianzen mit dem Programm Mplus (Muthén & Muthén, 1998) zur Homogenisierung des Itempools durchgeführt. Items mit hohen Restkorrelationen wurden aus dem Itempool ausgeschlossen (zweiter Selektionsschritt: $r > 0,3$). Abschließend wurden für die Ein-Faktor-Lösungen der so selektierten Itemmengen verschiedene Fit-Indizes mit dem Programm AMOS (Arbuckle & Worthke, 1999) berechnet.

5.4.1.1. Explorative Faktorenanalysen

5.4.1.1.1. Erste Teilstichprobe

Die explorative Faktorenanalyse der ersten Teilstichprobe zeigt, dass nach dem Parallelanalyse-Kriterium („parallel analysis criterion“, Lautenschlager, 1989; Verfahren der Parallelanalyse nach Horn, 1965) und dem Everett-Kriterium (Everett, 1983) vier Faktoren als zufallskritisch abgesichert gelten können (siehe Tabelle 10).

Da das Ziel die Konstruktion eines unidimensionalen Angst-CATs ist, wurde der erste unrotierte extrahierte Faktor, welcher 40,51% der Varianz aufzuklären vermag, als Selektionsgrundlage ausgewählt. Auf ihm laden 31 Items zwischen 0,43 und 0,77 mit einer durchschnittlichen Faktorenladung von 0,63, wenn wir die absoluten Werte der Faktorenladungen nehmen. Die Anordnung der Items auf dem Faktor lässt ein bipolares Konstruktkontinuum vermuten. Dieses wird durch hoch positiv ladende Items aufgespannt, die erfragen, ob sich eine Person „nervös“, „beunruhigt“, „ängstlich“, „angespannt“ bzw. „unruhig“ fühlt und hoch negativ ladende Items, die erfassen, inwiefern sich eine Person „entspannt“, „gelöst“, „wohl“, „ausgeglichen“ und „ruhig“ fühlt.

Items mit einer Faktorenladung von < 0,4, welche sich auf einem zweiten Faktor gruppierten, wurden ausgeschlossen, da sie offensichtlich die Annahme einer hinreichenden Unidimensionalität verletzten. Die geringe Faktorenladung der ausgeschlossenen Items scheint inhaltlich begründet, da die Mehrzahl dieser Items vegetative Begleiterscheinungen der Angst abbildet, welche offenbar als eigene Dimension betrachtet werden müssen.

Tabelle 10: Die unrotierte Faktorenlösung in der ersten Teilstichprobe (N_{Items} = 37; $N_{Patienten}$ = 1.010).

Abgekürzter Itemtext	Faktorenladungen der vierfaktoriellen unrotierten Lösung			
	1	2	3	4
Bin nervös	,767	-,048	,308	-,267
Bin beunruhigt	,761	-,095	,304	,157
Fühle mich beunruhigt	,716	,081	,217	,342
Hatte Angst	,698	,012	-,072	,263
Fühle mich angespannt	,690	-,076	,185	-,191
War ruhig und gelassen (umgepolt (u.))	,689	-,111	-,198	-,007
Bin verkrampft	,687	-,073	,152	-,187
Bin besorgt, dass etwas schiefgeht	,673	-,124	,323	,172
Fühle mich unsicher	,666	-,104	-,083	,141
Bin besorgt	,661	-,126	,309	,309
Bin aufgeregt	,649	-,003	,361	-,181
Bin überreizt	,630	-,022	,291	-,164
Bin zappelig	,613	,034	,334	-,407
Fühle mich besorgt	,601	,021	,191	,444
Hatte Mühe, mich zu konzentrieren	,577	-,002	-,231	-,019
Fühle mich kribbelig	,571	,167	,265	-,182
Dinge haben mich beunruhigt	,545	,001	,006	,214
Gefühl der Benommenheit	,538	,293	-,227	-,008
Herzklopfen, Herzjagen /-stolpern	,478	,586	-,112	-,040
Sorgen über gesundheitliche Probleme	,466	,105	,086	,361
Stiche, Schmerzen oder Ziehen in der Brust	,426	,606	-,059	-,010
Anfallsweise Herzbeschwerden	,391	,643	-,096	-,015
Schwindelgefühl	,387	,500	-,238	-,007
Engigkeit oder Würgen im Hals	,379	,456	-,181	-,081
Anfallsweise Atemnot	,374	,548	-,044	,018
Übelkeit	,326	,373	-,193	-,038
Erwartung, dass Gesundheit nachlässt (u.)	-,281	,049	-,045	-,290
Schwierigkeiten gelassen entgegen sehen	-,500	,229	,302	-,077
Fühle mich geborgen	-,549	,237	,290	,094
Fühle mich gelöst	-,573	,242	,378	-,029
Fühle mich selbstsicher	-,616	,327	,276	,007
Bin ruhig	-,654	,147	-,063	,307
Fühle mich ausgeglichen	-,657	,237	,314	-,004
Fühle mich wohl	-,686	,179	,286	,026
War in den vergangenen Wochen nervös (u.)	-,688	-,030	-,054	,105
Bin gelöst	-,710	,283	,264	,065
Bin entspannt	-,735	,259	,217	,183

Farbmarkierung: Faktorenladungen: Hellgrau: > 0,4; Mittelgrau: > 0,5; Dunkelgrau: > 0,6.
Eigenwerte: 1. Faktor: 12,81; 2. Faktor: 2,04; 3. Faktor: 1,64; 4. Faktor: 1,36.
Varianzaufklärung (in%): 1. Faktor: 40,51; 2. Faktor: 7,48; 3. Faktor: 5,25; 4. Faktor: 3,74.

5.4.1.1.2. Zweite Teilstichprobe

In explorativen Faktorenanalysen der zweiten Teilstichprobe zeigt sich, dass nach dem Parallelanalyse-Kriterium („parallel analysis criterion";Lautenschlager, 1989) und dem Everett-Kriterium (Everett, 1983) eine fünffaktorielle Lösung möglich ist.

Hier wurden durch den ersten unrotierten Faktor 31,93% der Gesamtvarianz erklärt (Tabelle 11). Auf diesem laden 33 Items zwischen 0,41 und 0,79 mit einer durchschnittlichen Faktorenladung von 0,53 (absolute Werte der Faktorenladungen). Die Items dieser Stichprobe sind auch „bipolar" angeordnet. Hohe positive Faktorenladungen zeigen Items, die erfragen, ob sich eine Person „von Angst und Unruhe getrieben" fühlt, „alles so schwarz sieht, dass sie Panik ergreift", ob sie „unsicher" und „beunruhigt" ist, oder „Angst vor der Zukunft" hat. Zu den hoch negativ ladenden Items zählen Items, die erfassen, inwiefern sich eine Person „ausgeglichen", „sicher", „geschützt", „gelöst", „ruhig" und „entspannt fühlt" sowie „Schwierigkeiten gelassen entgegensieht".

Das Selektionskriterium von $< 0{,}4$ führt in dieser Stichprobe zu einem Ausschluss von insgesamt 10 Items, welche sich auf weiteren Faktoren (2-5) gruppierten.

Die geringen Faktorenladungen der ausgeschlossenen Items scheint inhaltlich begründet, da die Mehrzahl dieser Items vegetative Begleiterscheinungen (Faktor 2) der Angst, körperbezogene spezifische Ängste (Faktor 3) bzw. soziale Ängstlichkeit (Faktor 5; umgepolte Items) abbilden. Diese Komponenten der Angst sind wahrscheinlich eigenständige Aspekte des Angsterlebens. Erstaunlich ist, dass zu den gering auf dem ersten Faktor ladenden Items auch das Item „wenig ängstlich" (aus dem Gießen-Test, GT; Beckmann et al., 1991) zählt. Dies mag daran liegen, dass dieses Item ursprünglich zur Messung einer zeitstabilen Eigenschaft („trait") konzipiert wurde und / oder das Itemantwortverhalten kontextbedingt ist. Dieses Item trägt nämlich im GT zur Erfassung der allgemeinen Skala „Grundstimmung" bei, wird also im Zusammenhang anderer Stimmungsaspekte abgefragt. Ein weiterer Grund mag im Itemantwortformat liegen. Das siebenstufige Antwortformat im GT erweist sich bei Datenanalysen als äußerst unergiebig, da Individuen im Alltag vermutlich nicht zwischen sieben Ausprägungsgraden zu unterscheiden vermögen. Items aus diesem Test wurden dementsprechend ausgeschlossen.

Tabelle 11: Die unrotierte Faktorenlösung in der zweiten Teilstichprobe (N_{Items} = 43; $N_{Patienten}$ = 834).

Abgekürzter Itemtext	Faktorenladungen der fünffaktoriellen unrotierten Lösung				
	1	2	3	4	5
Von Angst und Unruhe getrieben	,792	-,129	,067	-,002	-,011
Alles so schwarz sehen, dass Panik	,773	-,197	,051	,036	,007
Unsicher	,770	-,033	-,058	,100	-,027
Angst vor Zukunft	,758	-,180	-,082	,004	-,011
Beunruhigt	,743	,151	-,056	-,138	-,114
Sie fürchten Ziele nicht zu erreichen	,721	-,203	-,125	,018	,005
Probleme, sich zu entspannen	,701	-,032	-,234	-,128	-,029
Beschwerden wegen innerer Ängste	,688	-,188	-,022	-,044	-,001
Gefühl, nicht wirklich da zu sein	,687	-,207	-,018	,295	-,104
Besorgt	,677	,073	,038	-,127	-,069
Selbsterleben wie fremde Person	,669	-,226	,059	,213	-,147
Viele Sorgen	,663	-,145	-,071	-,002	,017
Angespannt	,642	,062	-,165	-,040	,025
Gefühl der Benommenheit	,640	,372	,003	,089	-,186
Gefühl quält, Körper sei nicht in Ordnung	,598	-,001	,364	-,298	,014
Körper plötzlich fremd und nicht dazugehörig	,557	-,298	,112	,307	-,128
Unsicherheit in Gruppe	,547	-,341	,103	,212	,342
Kribbelig	,544	,189	-,071	,035	-,091
Menschenansammlungen schrecken ab	,543	-,144	,045	,171	,214
Angst, schwer krank zu werden	,506	-,088	,492	-,388	,127
Schwindelgefühl	,488	,482	,121	,041	-,140
Herzklopfen, Herzjagen /-stolpern	,487	,587	,129	,143	,096
Schämen, wenn versagt	,469	-,457	,207	,256	,121
Engigkeit oder Würgen im Hals	,419	,369	,227	,170	-,101
Peinlich, vor Gruppe etw. Dummes zu sagen	,412	-,414	,212	,222	,239
Übelkeit	,411	,330	,081	,189	-,191
Angst, Gesundheit steht das nicht durch	,398	,023	,415	-,356	-,014
Stiche, Schmerzen oder Ziehen in der Brust	,383	,579	,264	,204	,113
Anfallsweise Herzbeschwerden	,358	,626	,277	,178	,176
Anfallsweise Atemnot	,323	,495	,325	,253	,110
Beunruhigung wegen neuer Krankheiten	,308	-,140	,582	-,318	,113
Wenig ängstlich	,255	-,003	-,067	,027	,483
Körper beobachten bzgl. Krankheiten	,190	-,057	,509	-,510	,080
Gefallen, im Mittelpunkt zu stehen	,021	-,335	,450	,127	-,447
Selten Sorgen um andere Menschen	,005	-,026	,010	,137	,512
Im Rampenlicht stehen ist verführerisch	-,004	-,314	,425	,206	-,452
Leichten Herzens	-,563	-,025	,330	,183	-,103
Ruhig	-,619	-,061	,322	,206	,120
Es sich bequem machen / entspannen	-,640	-,061	,253	,156	,037
Gelöst	-,647	-,121	,326	,208	,128
Sicher und geschützt	-,652	,038	,285	,037	,020
Schwierigkeiten gelassen entgegensehen	-,678	,199	,212	-,053	-,105
Ausgeglichen	-,701	-,065	,332	,146	,091

Farbmarkierung: Faktorenladungen: Hellgrau: > 0,4; Mittelgrau: > 0,5; Dunkelgrau: > 0,6.
Eigenwerte: 1. Faktor: 13,73; 2. Faktor: 3,20; 3. Faktor: 2,74; 4. Faktor: 1,69; 5. Faktor: 1,48.
Varianzaufklärung (in%): 1. Faktor: 31,93; 2. Faktor: 7,44; 3. Faktor: 6,37; 4. Faktor: 3,93, 5. Faktor: 3,44.

5.4.1.1.3. Dritte Teilstichprobe

Nutzt man im Rahmen der explorativen Faktorenanalyse der dritten Teilstichprobe das Parallelanalyse-Kriterium („parallel analysis criterion“, Lautenschlager, 1989) und das Everett-Kriterium (Everett, 1983) so zeigt sich, dass eine zweifaktorielle Lösung gegen den Zufall abgesichert ist.

Der erste unrotierte extrahierte Faktor klärt hier 32,98% der Varianz auf (Tabelle 12). Es laden 28 Items zwischen 0,40 und 0,74 mit einer durchschnittlichen Faktorenladung von 0,56 auf ihm (absolute der Faktoren-ladungen).

Auch hier scheinen positiv und negativ ladende Items ein bipolares Konstruktkontinuum aufzuspannen. Zu den positiv ladenden Items gehören Items wie „ich fühle mich beunruhigt“, „angespannt / aufgeregt“, „benommen“ und „unsicher“; negativ ladende Items fragen z. B. nach „Ausgeglichenheit“ und „Selbstsicherheit“ (manche negativ ladenden Items sind in Ihrem Antwortformat umgepolt, siehe Tabelle 12).

Das Ausschlusskriterium der Items liegt wie in den vorangegangenen Itemanalysen bei einer Faktorenladung von < 0,4, was zu einem Ausschluss von zwei vegetativen Items führt. Weitere vegetative Items wurden zunächst in dem Itempool belassen. Es stellt sich aber im Laufe der weiteren Selektionsschritte heraus, dass die meisten dieser Items sukzessive aus dem Itempool ausgeschlossen werden mussten, da sie den weiteren Kriterien der Itemselektion nicht entsprachen.

Tabelle 12: Die unrotierte Faktorenlösung in der dritten Teilstichprobe (N_{Items} = 30; $N_{Patienten}$ = 775).

Abgekürzter Itemtext	Faktorenladungen der zweifaktoriellen unrotierten Lösung	
	1	2
Beunruhigt	,701	-,312
Entspannt und gelassen oder angespannt und aufgeregt fühlen	,664	-,377
Gefühl der Benommenheit	,656	,228
Unsicher	,629	-,299
Besorgt	,592	-,319
Schwindelgefühl	,579	,255
Engigkeit oder Würgen im Hals	,577	,268
Kribbelig	,572	-,145
Herzklopfen, Herzjagen / -stolpern	,571	,395
Zittern	,568	,213
Taubheitsgefühl	,554	,293
Stiche, Schmerzen in der Brust	,554	,388
Anfallsweise Herzbeschwerden	,554	,471
Übelkeit	,536	,043
Aufsteigende Hitze, Hitzewallungen	,522	,383
Starkes Schwitzen	,508	,323
Anfallsweise Atemnot	,469	,450
Ohnmachtsanfälle	,430	,224
Leichtes Erröten	,419	,119
Schluckbeschwerden	,405	,289
Anfälle	,400	,237
Drang zum Wasserlassen	,389	,276
Durchfälle	,267	,066
Gelöst	-,649	,299
Ausgeglichen	-,649	,297
Sorgen wegen Gesundheit (umgepolt (u.))	-,649	,221
Angespannt (u.)	-,664	,414
Ausgeglichen und selbstsicher	-,712	,437
Nervosität (u.)	-,713	,292
Ängstlich, besorgt oder aufgeregt (u.)	-,742	,359

Farbmarkierung: Faktorenladungen: Hellgrau: > 0,4; Mittelgrau: > 0,5; Dunkelgrau: > 0,6.
Eigenwerte: 1. Faktor: 9,89; 2. Faktor: 2,84.
Varianzaufklärung (in%): 1. Faktor: 32,98; 2. Faktor: 9,48.

5.4.1.2. Konfirmatorische Faktorenanalysen

Wie in Kapitel 5.3.2.1. erläutert, werden konfirmatorische Faktorenanalysen eines Ein-Faktor-Modells über die in den Itemmengen der drei Teilstichproben verbliebenen Items gerechnet. In diesem Rahmen wurden zunächst die residualen Kovarianzen mit dem Programm Mpuls (Muthén & Muthén, 1998) errechnet und zur Itemselektion genutzt, sowie anschließend Fit-Indizes mit dem Programm AMOS (Arbuckle & Worthke, 1999) berechnet.

5.4.1.2.1. Analyse residualer Kovarianzen

Die Analysen residualer Kovarianzen dienten der Untersuchung, ob nennenswerte Restkorrelationen zwischen den Items vorliegen, wenn der Faktor, der am meisten Gemeinsames abbildet, statistisch herauspartialisiert wird. Die Herauspartialisierung des ersten Faktors, welcher den größten Teil der gemeinsamen Varianz der Items abbildet, erfolgte, indem von den beobachteten Itemwerten die - mittels des Faktorwertes des ersten Faktors - vorhergesagten Itemwerte abgezogen werden, so dass Item-Residuen resultieren. Dies erfolgte mit dem Programm Mplus (Muthén & Muthén, 1998). Nennenswerte residuale Partialkorrelationen deuten auf das Vorhandensein weiterer Faktoren hin und begründen wegen der damit verbundenen Verletzung der Unidimensionalität den Ausschluss beteiligter Items.

5.4.1.2.1.1. Erste Teilstichprobe

Die Analyse residualer Kovarianzen über die 31 selektierten Items der ersten Teilstichprobe ergab insgesamt wenig Partialkorrelationen. Erwähnenswerte Partialkorrelationen (r = 0,2-0,3) lagen nur im Falle von drei von 451 berechneten Partialkorrelationen vor („zappelig“ / „gelöst“; „besorgt“ / „beunruhigt“; „selbstsicher“/ „gelassen gegenüber Schwierigkeiten“), während eine Partialkorrelation („Herzklopfen“/„Stiche in der Brust“) einen Wert von r = 0,3 überstieg. Während die ersten Partialkorrelationen durch gemeinsame Teilaspekte (wie z. B. motorische Unruhe, kognitive Besorgnis und Gelassenheit) erklärt werden konnten, welche mit dem Angst-Konstrukt in enger Beziehung zu stehen scheinen, stach die letzte Partialkorrelation – auf dem Hintergrund der Ergebnisse der Faktorenanalysen (Ausschluss vegetativer Aspekte der Angst) besonders hervor, so dass letzere als nicht tolerabel angesehen, und das Item „Stiche in der Brust“ aus der Itembank

ausgeschlossen wurde. Die übrigen Partialkorrelationen wurden akzeptiert (siehe Anhang 9.2.1.).

5.4.1.2.1.2. Zweite Teilstichprobe

Die Analyse residualer Kovarianzen über die 33 selektierten Items der zweiten Teilstichprobe führte gegenüber der ersten Teilstichprobe zu mehr Partialkorrelationen. Nennenswerte Partialkorrelationen ($r > 0,2$) fanden sich bei 17 von 538 berechneten Partialkorrelationen. Diese traten zwischen Items, welche vegetative Beschwerden („Herzklopfen“ / „Schwindel“ / „Benommenheit“ / „Übelkeit“) und Items, welche soziale Ängstlichkeit erfragten („Scham, wenn versagt“ / „Unsicherheit in Gruppe“ / „Peinlich, vor Gruppe etwas Dummes zu sagen“), auf. Aus diesem Grund wurden drei „vegetative“ und zwei „sozial ängstliche“ Items sowie ein „köperangstbezogenes“ Item, welche den größten Teil der Partialkorrelationen bedingten, ausgeschlossen. Die übrigen Partialkorrelationen wurden akzeptiert (siehe Anhang 9.2.2).

5.4.1.2.1.3. Dritte Teilstichprobe

Die über die 28 selektierten Items der dritten Teilstichprobe berechnete Analyse residualer Kovarianzen führt zu einer Reihe von erwähnenswerten Partialkorrelationen. 27 Partialkorrelationen von 378 Errechneten überstiegen einen Wert von 0,2, davon vier einen Wert von $r = 0,3$. Eine genaue inhaltliche Betrachtung dieser Ergebnisse zeigte, dass auch hier der wahrscheinliche Grund in der Vielzahl „vegetativer“ Items liegt, so dass die Items „Schwindelgefühl“, „Starkes Schwitzen“, „Schluckbeschwerden“, „Stiche, Schmerzen in der Brust“, „Anfallsweise Atemnot / Herzbeschwerden“, welche die meisten Partialkorrelationen bedingten, auch aus dieser Stichprobe ausgeschlossen wurden. Dies führte zu einer massiven Reduktion der Partialkorrelationen, wie sie in Anhang 9.2.3 dargestellt ist.

5.4.1.2.2. Fit-Indizes

Die Fit-Indizes der konfirmatorischen Faktorenanalysen zur Beurteilung der Datenanpassung an ein Ein-Faktor-Modell wurden separat an den drei Teilstichproben (N_1 = 30 Items; N_2 = 27; N_3 = 23 Items) mit dem Programm AMOS (Arbuckle & Worthke, 1999) berechnet, und sind in Tabelle 13 zusammengefasst.

Tabelle 13: Fit-Indizes der konfirmatorischen Faktorenanalyse der drei Teilstichproben.

Fit-Statistiken	Ein-Faktor-Modell N_1 = 1.010	Ein-Faktor-Modell N_2 = 834	Ein-Faktor-Modell N_3 = 775
Diskrepanz	4243,65	3219,11	1837,11
Freiheitsgrade (df)	405	324	209
p	0,001	0,001	0,001
Parameterzahl	60	54	44
Diskrepanz / df	10,48	9,94	8,79
Root mean square error of approximation (RMSEA)	0,10	0,10	0,10
Tucker-Lewis-Index (TLI)	0,75	0,76	0,76
Comparative fit index (CFI)	0,77	0,78	0,78

Zur Bewertung der Fit-Indizes:
χ^2-Statistiken sind hochgradig sensitiv gegenüber der Stichprobengröße (hier: bis zu N = 1.010 Personen) und daher wenig geeignet zur Modellbeurteilung;
Schermelleh-Engel und Mitarbeiter (2003):
- „guter" Fit: RMSEA: 0-0,05; CFI: 0,97-1,0; p: 0,05-1,0;
- „akzeptabler" Fit: RMSEA: 0,05-0,10; CFI: 0,95-0,97; p: 0,01- 0,05;

Hu und Bentler (1999): „guter Fit": TLI / CFI = 0,90/0,95;
Brown und Cudeck (1993), MacCallum und Mitarbeiter (1996): „guter Fit": RMSEA: < 0,05; „akzeptabel": 0,05-0,08; „mittelmäßig": 0,08-0,1; „schlecht" > 0,1.

Der in Tabelle 13 aufgeführte Root Mean Square Error of Approximation (RMSEA) ist in seiner Höhe akzeptabel. Wenn man die für Strukturgleichungsmodelle üblichen Grenzen (Brown & Cudeck, 1993; MacCallum und Mitarbeiter, 1996) heranzieht, sind die aufgeführten Werte des Tucker-Lewis-Index (TLI) und des Comparative Fit Index (CFI) jedoch zu niedrig. Dies ist ein Befund, der sich nicht nur bei IRT-basierten Reanalysen etablierter Inventare zeigt, sondern auch bei analogen Untersuchungen gut etablierter Fragebögen (STAI State: 20 Items: TLI=0,73, CFI=0,76, RMSEA=0,13; NEO-FFI Neurotizismusskala 12 Items TLI=0,82, CFI=0,86, RMSEA=0,11). Insgesamt erscheint es fraglich, ob die genannten Fit-Indizes im Rahmen einer IRT-Modellierung zur Untersuchung der Unidimensionalität geeignet sind (siehe Kapitel 7.4.1.).

Mit Hilfe linearer Strukturgleichungsmodelle wäre eine angemessenere Spezifikation eines möglichst realitätsgerechten Modells des Angst-Konstruktes denkbar, jedoch in diesem Rahmen nicht realisierbar. Um die in dieser Arbeit angestrebte Konstruktion eines eindimensionalen IRT-basierten CATs zu ermöglichen, wird der geringe Modell-Fit des Ein-Faktor-Modells akzeptiert (siehe Diskussion Kapitel 7.4.1.). Ziel zukünftiger Forschung sollte jedoch die Konstruktion mehrdimensionaler CATs sein, welche aus methodischen und praktischen Begrenzungen an dieser Stelle noch nicht möglich war.

5.4.2. IRT-Analyse

Die IRT-Analyse der Itemeigenschaften umfasst die grafische Analyse der Item Response Curves (IRCs), der Test Informationskurven und die Berechnung von Standardmessfehler und Reliabilität der Itembank und erfolgte mit dem Programm TestGraf (Ramsay, 1995).

5.4.2.1. Item Response Curves (IRCs)

5.4.2.1.1. Erste Teilstichprobe

Die Analyse der Item Response Curves (IRCs) der Items der ersten Teilstichprobe zeigte in der Mehrzahl der Fälle „sehr gute“ (i. S. von modellkonformen) Itemcharakteristiken der ausgewählten Items. Darunter versteht man eingipflige, glockenförmige, jedoch nicht unbedingt symmetrisch verlaufende IRCs, welche pro Antwortkategorie in genau einem Messbereich mit ihrem Maximum alle anderen IRCs des jeweiligen Items übersteigen. Im Falle von Modellkonformität der IRCs verhält sich die IRC der ersten Antwortkategorie stets monoton fallend und die der letzten Antwortkategorie stets monoton steigend (siehe Kapitel 5.3.2.1.). Exemplarisch sei hier in Abbildung 11 (oben) ein Item mit modellkonformen IRCs illustriert.

Die Grafik veranschaulicht die Antwortwahrscheinlichkeit bezüglich der einzelnen Antwortkategorien in Abhängigkeit vom standardnormalverteilten latenten Angstkontinuum (Theta).

Die Schnittpunkte der IRCs nennen sich „Thresholds“ (Schwellenparameter); der Mittelwert der Thresholds wird „Location Parameter“ (Lokationsparameter) genannt. Der Lokationsparameterwert dient der Lokalisation des Items auf dem latenten Angstkontinuum. Die gemittelte Steigung („Slope Parameter“) bedingt die Iteminformation, welche die Diskriminationsfähigkeit eines Items zwischen Testpersonen unterschiedlicher Merkmalsausprägungen ausdrückt.

Die günstigen IRCs des im oberen Teil der Abbildung 11 dargestellten Items sind nicht selbstverständlich, wie z. B. die IRCs des Items „ich fühle mich belästigt durch Herzklopfen“ (Abbildung 11, unten) zeigen. Da eine IRT-Modellierung hierarchisch sortierte Thresholds erfordert, mussten gegebenenfalls Antwortkategorien der Items (z. B. „Herzklopfen“, „Gefühl der Benommenheit“) mit ungenügend diskriminierenden Antwortkategorien – wie in Abbildung 11 dargestellt - so zusammengefasst werden, dass die modifizierten IRCs in genau einem Merkmalsausprägungsintervall ein deutliches Maximum aufwiesen (zum Vorgehen des Zusammenlegens siehe Abbildung 11, unten).

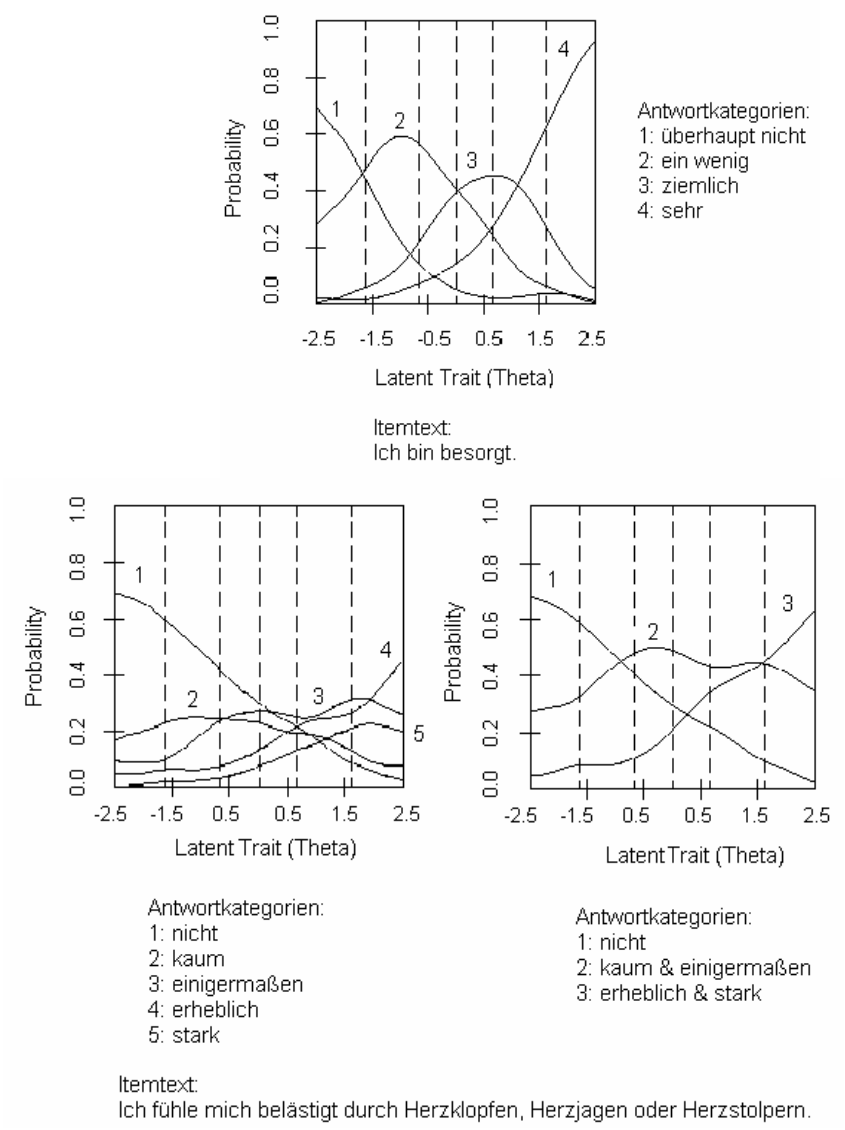


Abbildung 11: IRCs eines Items mit modellkonformer Itemcharakteristik (oben) und eines Items mit nicht modellkonformer Itemcharakteristik[76] (unten links), die ggf. durch das Zusammenlegen von Antwortkategorien verbessert werden kann (unten rechts).

[76] Zur Bewertung der Item Response Curves („gut“ / „schlecht“) i. S. der Modellkonformität siehe Kapitel 5.3.2.2.1.

Das Zusammenlegen der Antwortkategorien hat keine Auswirkungen auf das im späteren CAT-Prozess vorgelegte Antwortformat der Items, sondern hat lediglich Implikationen für die Theta-Schätzung der Personenausprägung. Gelang eine Zusammenlegung benachbarter Antwortkategorien nach grafischer Beurteilung nicht zufriedenstellend, so wurden jeweilige Items (insgesamt drei Items) aus dem Itempool ausgeschlossen. Die IRC-Grafiken der im Itempool nach der gesamten Itemselektion verbliebenen 24 Items befinden sich im Anhangskapitel 9.3.1..

5.4.2.1.2. Zweite Teilstichprobe

Die IRC-Analyse der zweiten Stichprobe ergab insgesamt abgesehen von einem vegetativen Item („Engigkeit im Hals“), welches daher ausgeschlossen wurde, ebenfalls modellkonforme IRCs der ausgewählten Items. Diese überwiegend eingipfligen monoton verlaufenden IRCs der ausgewählten 26 Items sind im Anhang (Kapitel 9.3.2.) abgebildet.

5.4.2.1.3. Dritte Teilstichprobe

Die IRC-Analyse der dritten Teilstichprobe zeigte bei den meisten Items (17 von 23 Items) modellkonforme IRCs (siehe Anhang; Kapitel 9.3.3). Auch in dieser Stichprobe zeigten sich bei einigen Items, welche vegetative Korrelate von Angst erfassen sollen („Ohnmachtsanfälle“, „Anfälle“, „Schluck-beschwerden“, „Erröten“, „Herzklopfen“, „Übelkeit“, „Engigkeit im Hals“, „Benommenheit“ und „Aufsteigende Hitze“), dass die IRCs dieser Items oft in ihrer „Originalversion“ den grafischen Kriterien nicht entsprachen.

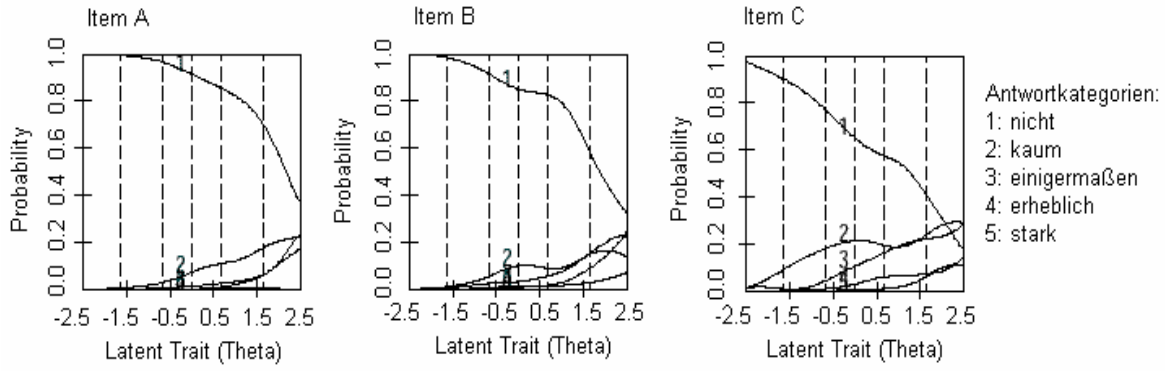


Abbildung 12: Ungenügende IRCs der Items „Ohnmachtsanfälle“ (A), „Anfälle“ (B) und „Leichtes Erröten“ (C).

Die übrigen dieser Items (z. B. „Herzklopfen“, „Gefühl der Benommenheit“) wurden in ihren Antwortkategorien bestmöglichst zusammengefasst. Als ein Beispiel für eine Zusammenfassung der Antwortkategorien sei hier das Item „Kloßgefühl im Hals“ aufgeführt.

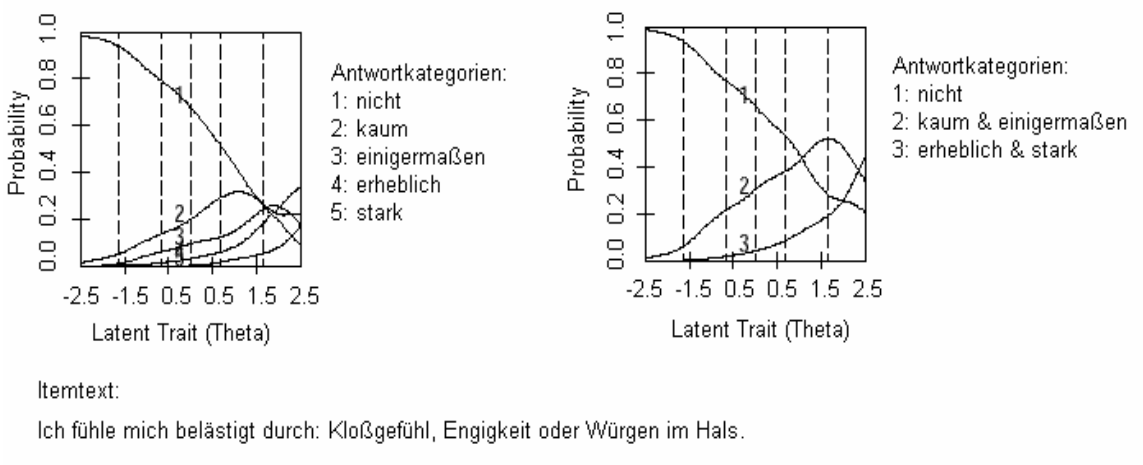


Abbildung 13: Beispiel für eine mögliche Modifikation der IRCs des Items „Kloßgefühl im Hals“.

5.4.2.2. Testinformation und Standardmessfehler

5.4.2.2.1. Erste Teilstichprobe

Der durchschnittliche Iteminformationsgehalt der selektierten und in den Antwortkategorien modifizierten Itemmenge von 24 Items der ersten Teilstichprobe liegt mit Werten zwischen 0,42 und 0,66 sehr hoch. Das daraus resultierende hohe Testinformationsniveau (die Testinformation errechnet sich aus der Summe der Iteminformationen) deutet darauf hin, dass die selektierte Itemstichprobe insgesamt einen hohen Informationsgehalt für das gesamte Merkmalsausprägungsspektrum bietet.[77] Dies ist gerade im Hinblick auf die Entwicklung eines „equal precise test“ (Embretson & Reise, 2000, S. 270), also eines Tests, welcher auf allen Stufen der Merkmalsausprägung gut messen soll, von zentraler Bedeutung. Die Abbildung 14, welche die Testinformationskurve der selektierten Items der ersten Teilstichprobe in Abhängigkeit zum geschätzten Theta-Wert der Angstausprägung in Einheiten

[77] Allerdings muss eingeräumt werden, dass in der Literatur bislang keine etablierten Vergleichsmaßstäbe zur Bewertung vorliegen. Die Bewertung der Höhe der Item- und Testinformation geschieht hier auf der Grundlage des Wissens um die Reliabilität und den Standardmessfehler, welcher in inverser Beziehung zur Testinformation steht.

der abweichungsnormierten Standardnormalverteilung veranschaulicht, zeigt, dass ein insgesamt hoher Informationsgehalt konstatiert werden kann, der jedoch einer gewissen Variation in Abhängigkeit vom Merkmalsausprägungsspektrum unterliegt. Dies ist ein Umstand, der in der empirischen Realität häufig ist, und im Widerspruch zu der Annahme eines merkmalsausprägungs*un*abhängigen Standardmessfehlers steht, welcher in der KTT postuliert wird (siehe Kapitel 3.2.).

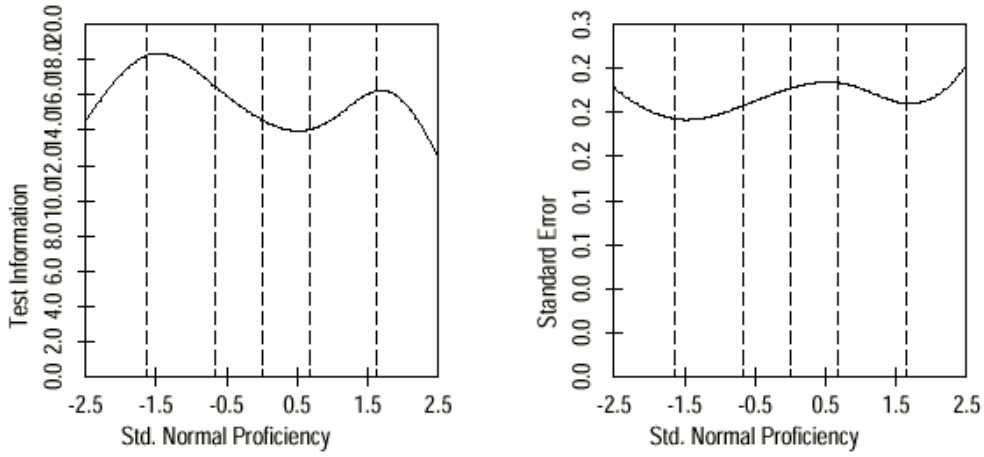


Abbildung 14: Testinformationsniveau (links) und Standardmessfehler (rechts) der selektierten Items der ersten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

Der Möglichkeit, im Rahmen der IRT-Analyse die Merkmalsausprägungsabhängigkeit der Messgenauigkeit einer Skala zu beurteilen, kommt bezüglich der Indikation verschiedener Tests ein hoher Stellenwert zu.

Wie Abbildung 14 verdeutlicht, verhält sich die Testinformationsfunktion zweigipflig. Offensichtlich zeigt sich die leichte Tendenz, dass eine mittlere Angstausprägung bzw. eine mittlere Abwesenheit der Angst etwas besser gemessen werden kann, d. h. die Messung nur mit einem geringen Standardmessfehler behaftet ist.

5.4.2.2.2. Zweite Teilstichprobe

Das Testinformationsniveau und der Standardmessfehler der Itemmenge der 26 ausgewählten Items der zweiten Teilstichprobe (Abbildung 15) liegt geringfügig unter dem der ersten Teilstichprobe, ist aber insgesamt als recht

hoch einzustufen. Der zweigipflige Kurvenverlauf der Testinformation ist hier nicht so deutlich ausgeprägt wie derjenige der ersten Teilstichprobe.

Die Testinformation ist zudem an den extremen Enden des Merkmalsausprägungskontinuums etwas geringer als bei der Skala der ersten Teilstichprobe.

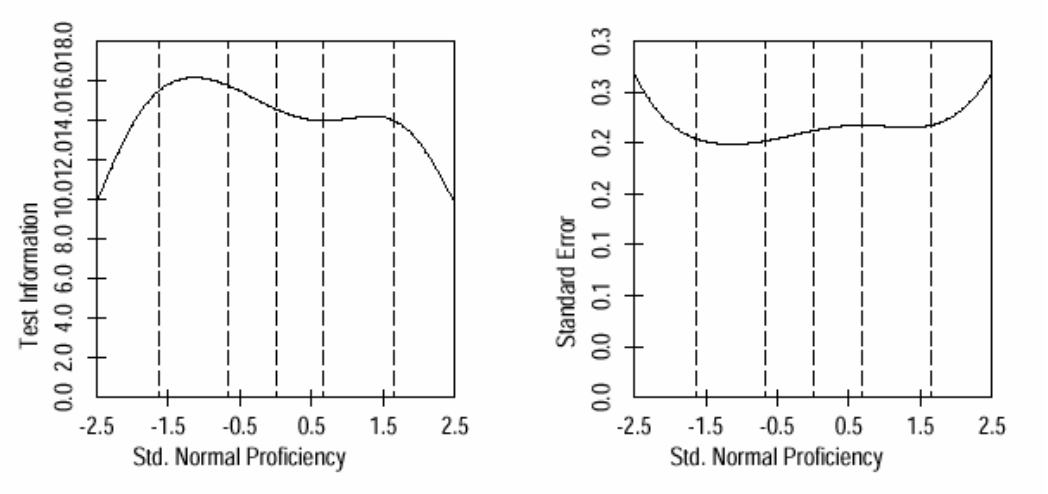


Abbildung 15: Testinformationsniveau (links) und Standardmessfehler (rechts) der selektierten Items der zweiten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung in Einheiten der Standardnormalverteilung).

5.4.2.2.3. Dritte Teilstichprobe

Das Testinformationsniveau in der dritten Teilstichprobe (siehe Abbildung 16) ist verglichen mit den Ergebnissen der ersten beiden Teilstichproben am geringsten, dass heißt die Messung wäre - wenn nur diese Skala zur Angstmessung eingesetzt würde - mit einem größeren Messfehler behaftet.

Während die ersten beiden Teilstichproben Testinformationskurven mit einem tendenziell eher zweigipfligen Kurvenverlauf aufweisen, mutet die Testinformationskurve der dritten Teilstichprobe eher eingipflig mit einem Maximum im mittleren unteren Bereich des zugrundeliegenden Konstruktkontinuums an. Anscheinend beinhaltet dieser Itempool vermehrt Items, welche in diesem Bereich des Merkmalsausprägungskontinuums gut (aber nicht so gut wie die Items der ersten beiden Teilstichproben) differenzieren können.

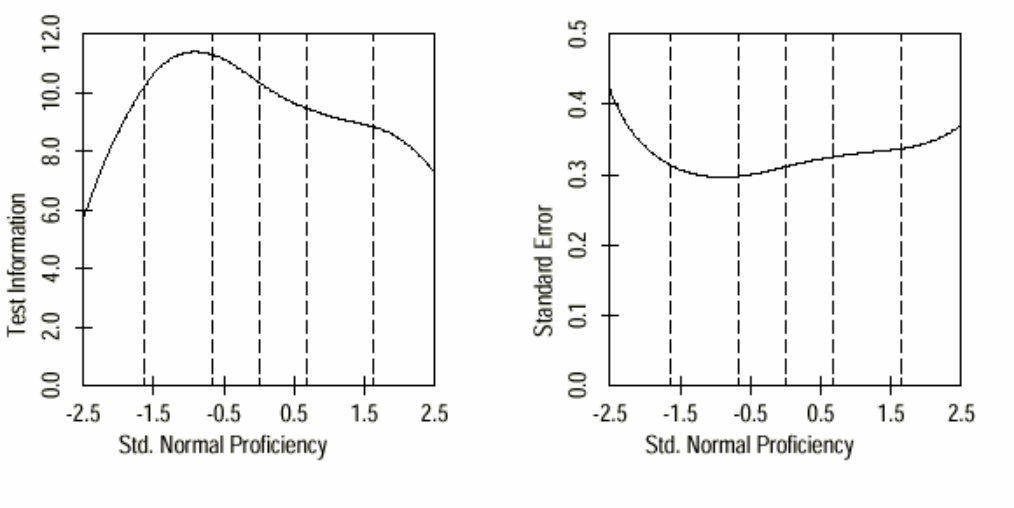
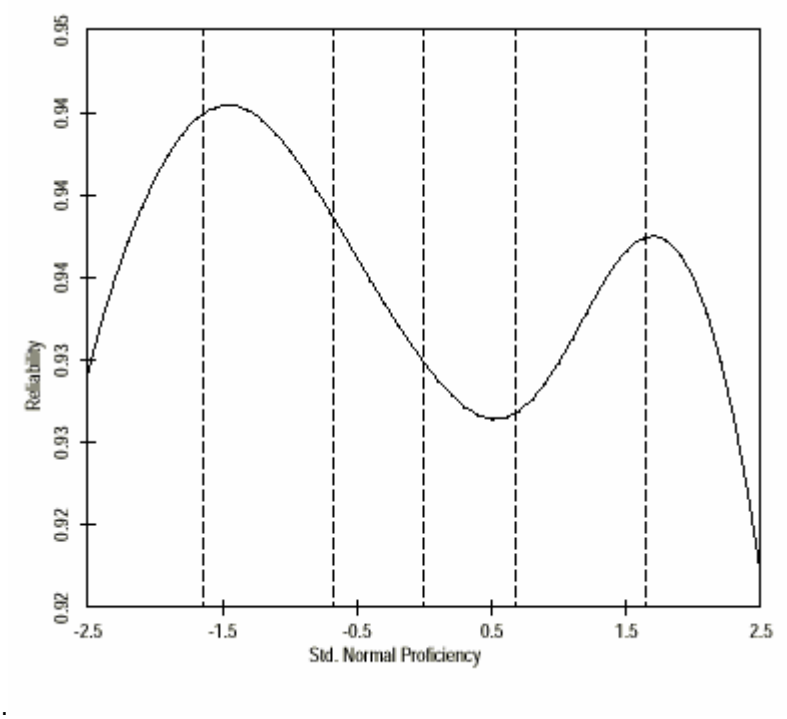


Abbildung 16: Testinformationsniveau (links) und Standardmessfehler (rechts) der selektierten Items der dritten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

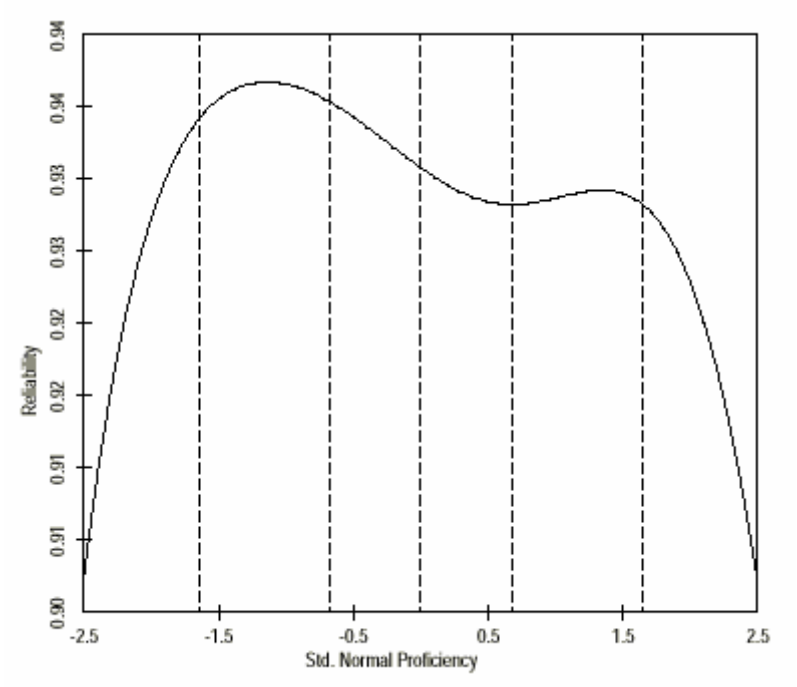
5.4.2.3. Reliabilität

Obgleich die Reliabilität in gegenläufiger Beziehung zum Standardmessfehler steht ($Rel = 1-s_e^2$), werden trotz einer gewissen Redundanz im Folgenden auch die Reliabilitätsfunktionen der drei Teilstichproben grafisch veran-schaulicht. Die enge Beziehung zwischen Testinformationsfunktion, Standardmessfehler und IRT-basierter Reliabilität, wie sie mathematisch in Kapitel 5.3.2.2.2. erläutert wurde, wird bei der vergleichenden Betrachtung der Grafiken des vorherigen und dieses Kapitels deutlich. Die grafische Darstellung der IRT-basierten Reliabilitätsfunktion - wie sie vom Program TestGraf (Ramsay, 1995) ausgegeben wird - bietet, verglichen mit der Reliabilität, welche in der KTT gebräuchlich ist (siehe Kapitel 3.2.), den Vorteil, die Reliabilität in Abhängigkeit vom Merkmalsausprägungskontinuum analysieren und beurteilen zu können.



.

Abbildung 17: Reliabilitäten der selektierten Items aus der ersten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).



.

Abbildung 18: Reliabilitäten der selektierten Items aus der zweiten Stichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

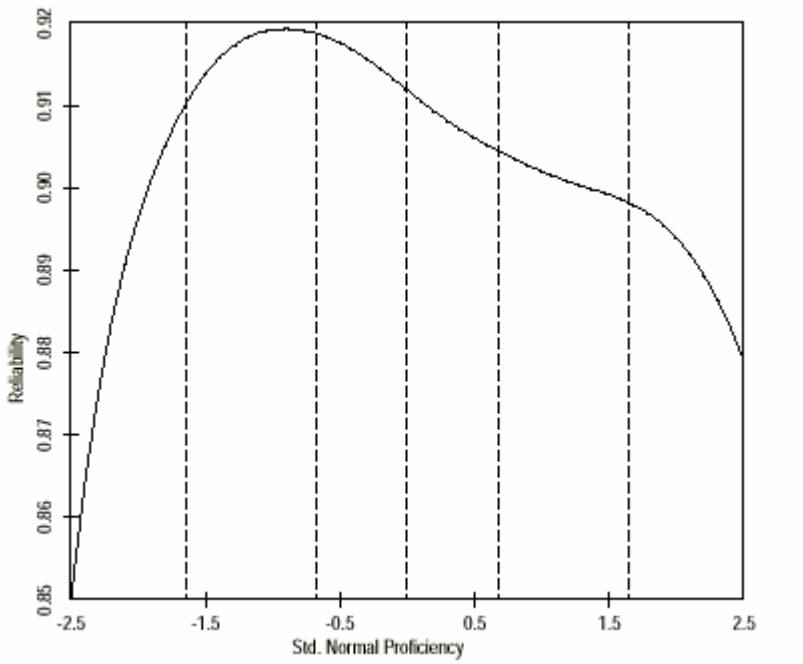


Abbildung 19: Reliabilitäten der selektierten Items aus der dritten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

Die Reliabilitäten der Skalen bestehend aus den jeweils in separaten Analysen selektierten Itemmengen der drei Teilstichproben sind insgesamt mit Werten zwischen Rel = 0,85 (Minimum der dritten Teilstichprobe) und Rel = 0,94 (Maximum der ersten Teilstichprobe) entsprechend der Testinformations-funktion und der Standardmessfehlerwerte aus Kapitel 5.4.2.2. sehr hoch. Während sich auch hier die Kurvenverläufe der Reliabilitätsfunktion der ersten und zweiten Teilstichproben ähneln (tendenziell zweigipfliger Kurvenverlauf), weicht die Reliabilitätsfunktion der dritten Teilstichprobe in Form (eingipflig) und Höhe (geringere Reliabilität) von der der ersten beiden Teilstichproben ab.

5.4.3. IRT-Modellierung

5.4.3.1. Itemparameterschätzung

Im Rahmen der Schätzung der einzelnen Itemparameter auf der Basis des GPCMs wurde als Selektionskriterium ein Steigungsparameterwert von $a_i > 0{,}80$ zur Optimierung der Itembank genutzt (siehe Kapitel 5.3.2.3.). Der Steigungsparameter quantifiziert die gemittelte Steigung aller IRCs eines Items und gilt damit als Indikator für den Iteminformationsgehalt bzw. die Diskriminationsfähigkeit eines Items.

Fünf der Items der ersten Teilstichprobe, ein Item der zweiten und drei Items der dritten Teilstichprobe entsprachen dem oben genannten Selektionskriterium nicht, und wurden daher ausgeschlossen.

Die Steigungsparameterwerte der drei einzelnen Teilstichproben sind in Tabelle 17 zusammengefasst. Die Steigungsparameterwerte (a_i) der verbliebenen 24 Items der ersten Teilstichprobe variieren zwischen 0,80 und 2,60 ($\overline{X}$ = 1,30; SD = 0,38); die der verbliebenen 25 Items der zweiten Teilstichprobe liegen zwischen 0,82 und 1,87 ($\overline{X}$ = 1,30; SD = 0,32) und die der selektierten 13 Items der dritten Teilstichprobe liegen im Bereich von 0,84 bis 2,59 ($\overline{X}$ = 1,40; SD = 0,49).

5.4.3.2. „Differential-Item-Functioning“ (DIF)

Aufgrund des Iteminhalts wurden fünf Items aus dem Berliner-Stimmungs-Fragebogen (BSF)[78] als potentielle „Anker-Items“ untersucht.[79] Differential-Item-Functioning (DIF) wurde IRT-basiert für die fünf Anker-Items getrennt für zwei Parameter - den Steigungs- und den Lokationsparameter - zwischen der ersten und zweiten (bzw. dritten) Teilstichprobe mit dem Computerprogramm Parscale (Muraki & Bock, 1999) berechnet. Das heißt, dass sowohl zur Untersuchung des DIFs zwischen den Itemparameterwerten der Anker-Items der ersten und zweiten Teilstichprobe zehn Einzelvergleichstests (2 Parameter x 5 Anker-Items), als auch zwischen den Itemparametern der Anker-Items der ersten und dritten Teilstichprobe zehn Einzelvergleichstests durchgeführt wurden.

[78] Berliner-Stimmungs-Fragebogen (BSF; Hörhold & Klapp, 1993; Rose et al., in Druck).
[79] Das Item „Gefühl der Benommenheit“ (GBB36) wurde nicht als Anker-Item genutzt, da vorherige Analysen auf Schwierigkeiten vegetativer Items bei der Angst-Messung hindeuteten.

In den somit insgesamt 20 Einzelvergleichstests (χ^2-Statistik) ergaben sich 19 von 20 nicht signifikanten α-Bonferoni[80] korrigierten Ergebnissen (χ^2 zwischen 0,04 – 6,14; $p > 0{,}01$; n.s). Dies erlaubt die Schlussfolgerung, dass - abgesehen von einer Ausnahme - keine bedeutsamen Unterschiede bezüglich der Steigungs- und Schwellenparameterwerte der Anker-Items zwischen den drei Teilstichproben existierten. Bei dem gegenüber anderen Verfahren konservativen Vorgehen zur DIF-Identifizierung (mittels Parscale) entschlossen wir uns, die eine Abweichung zu tolerieren, so dass die Itemparameter dieser Stichproben dementsprechend über ein „Item-Link-Design" auf einer gemeinsamen Skala kalibriert werden konnten.

5.4.3.3. „Item-Link-Design"

Die selektierten Items der drei Teilstichproben, wurden auf einer gemeinsamen Skala abgebildet, indem die Itemparameter der selektierten Items der zweiten und dritten Teilstichproben gemäß dem im Kapitel 5.3.2.3.3. beschriebenen methodischen Vorgehen re-kalibriert wurden.

Tabelle 14: Differenzen zwischen den Itemparameterwerten (Mittelwerte und Standardabweichungen) der getrennt analysierten Teilstichproben, welche in der Re-Kalibrierung des Item-Link-Designs verrechnet wurden.

Abgekürzter Itemtext	Item Parameter	Erste Teilstichprobe (N = 1.010) M ± SD: 0,00 ± 1,00	Zweite Teilstichprobe (N = 834) M ± SD: -0,44 ± 1,37	Dritte Teilstichprobe (N = 779) M ± SD: -0,74 ± 1,12
Fühle mich gelöst	a_i	1,09	1,11	0,97
	b_i	-0,77	-0,76	-0,76
	b_{ih}	0,49 / 0,18 / -0,66	0,77 / 0,40 / -1,18	1,06 / -0,25 / -0,81
Fühle mich besorgt	a_i	1,58	1,69	1,92
	b_i	-1,20	-1,29	-1,19
	b_{ih}	0,27 / -0,27	0,71 / -0,71	0,48 / -0,48
Fühle mich beunruhigt	a_i	1,51	1,87	2,63
	b_i	-0,59	-0,45	-0,48
	b_{ih}	0,97 / -0,35 / -0,62	1,08 / -0,27 / -0,81	0,88 / -0,12 / -0,76
Fühle mich ausgeglichen	a_i	1,52	1,20	0,89
	b_i	-0,79	-0,81	-0,85
	b_{ih}	0,63 / 0,00 / -0,63	0,79 / 0,05 / -0,84	1,18 / -0,45 / -0,72
Fühle mich unsicher	a_i	1,60	1,51	1,48
	b_i	-0,62	-0,65	-0,66
	b_{ih}	0,37 / -0,37	0,54 / -0,54	0,07 / -0,07

Itemparameter: a_i = Steigungsparameter; b_i = Lokationsparameter; b_{ih} = Schwellenparameter.

[80] α-Bonferoni Korrektur nach Bortz (1999, S. 261).

In Tabelle 14 sind die Differenzen zwischen den Mittelwerten und Standardabweichungen der Itemparameterwerte zwischen den Teilstichproben, die in die Re-Kalibrierung mit eingehen, dargestellt.

5.4.3.4. „Item-Fit-Statistiken“

Wie in Kapitel 5.3.2.3.4. erörtert, wurden Likelihood-χ^2-Tests als numerische Item-Fit-Statistiken zur Beurteilung der Modellanpassung der Daten mit dem Programm Parscale (Muraki & Bock, 1999) berechnet. Tabelle 15 veranschaulicht die so berechneten Item-Fit-Statistiken der Itembank (N = 50 Items). Likelihood-χ^2-Tests sind wie in Kapitel 5.3.2.3.4. diskutiert, stark von der Stichprobengröße abhängig, so dass es bei den hier untersuchten Stichprobengrößen von N = 775 bis N = 1.010 nicht erstaunt, dass bei einer Festlegung des Signifikanzniveaus auf $p \leq 0{,}05$ eine Vielzahl von Items (N = 22 Items) als signifikant vom Modell abweichend gewertet werden müssen (siehe Diskussion in Kapitel 7.4.3).

Daraus ergibt sich die Frage nach dem Umgang mit Item-Misfits. Prinzipiell kommen mehrere Möglichkeiten in Frage wie z. B. die Lockerung des Modells (z. B. durch die Wahl eines anderen IRT-Modells) oder der Ausschluss von Items mit Misfit. Diese Konsequenzen erscheinen jedoch nur begründet, wenn den Fit-Statistiken eine zuverlässige und valide Aussagekraft zugestanden wird, die von vielen Autoren angezweifelt wird (Embretson & Reise, 2000; Hambleton et al., 1991; Van der Linden & Hambleton, 1997 und Muraki, 1997).

Aufgrund der Fragwürdigkeit der Fit-Statistiken enthalten sich Van der Linden und Hambleton (1997) bewusst allgemeiner Empfehlungen, da diese abhängig von: a) der Art und Weise des Misfits, b) der Verfügbarkeit von „Ersatz“-Items, c) dem mit dem Konstruieren *neuer* Items verbundenen Aufwand, d) der Verfügbarkeit von Kalibrierungsstichproben und e) dem Testziel seien.

Aus Gründen der Praktikabilität (keine derzeitige Verfügbarkeit weiterer Item- und Personenstichproben) und um die Entwicklung eines IRT-basierten CATs zur Angstmessung (Angst-CAT) zu ermöglichen, entschieden wir uns, der Empfehlung von Embretson und Reise (2000) zu folgen, und diese Fit-Statistik für 2PL-Modelle wie dem hier verwendeten GPCM nicht als „solid decision-making tool“ (S. 235) zu nutzen, d. h. sie nicht als Mittel zum gezielten Itemausschluss heranzuziehen.

Tabelle 15: Item-Fit-Statistiken der die Itembank konstituierenden 50 Items des Angst-CATs.

Abgekürzter Itemtext	df	χ^2 p	
Bin nervös	34	32,16	0,5580
Bin aufgeregt	40	51,47	0,1057
Bin besorgt	38	59,29	0,0151
Bin besorgt, dass etwas schief geht	39	69,58	0,0019
Bin beunruhigt	33	41,55	0,1460
Beschwerden wegen innerer Ängste	37	46,21	0,1425
Bin überreizt	39	43,82	0,2744
Bin verkrampft	37	52,98	0,0429
Bin von Angst und Unruhe getrieben	33	43,06	0,1129
Bin zappelig	40	48,67	0,1634
Fühle mich angespannt	38	54,43	0,0409
Fühle mich besorgt	24	25,85	0,3608
Nervös	53	83,26	0,0050
Fühle mich beunruhigt	35	56,95	0,0109
Fühle mich kribbelig	43	46,92	0,3149
Sich gelassen oder aufgeregt fühlen	44	106,78	0,0000
Fühle mich unsicher	25	27,77	0,3185
Gefühl der Benommenheit	33	43,74	0,1001
Habe Gefühl, nicht wirklich da zu sein	28	29,52	0,3865
Hatte Angst	40	19.24	0,9977
Sie fühlten sich angespannt	32	64,52	0,0006
Sie fühlten sich nervös	29	87,58	0,0000
Sorgen wegen Gesundheit	50	84,17	0,0018
Kloßgefühl im Hals	32	50,68	0,0191
Körper erscheint plötzlich fremd	32	24,79	0,8144
Ängstlich, besorgt oder aufgeregt	35	94,26	0,0000
Menschenansammlungen schrecken mich ab	31	20,63	0,9213
Sehe alles so schwarz, dass mich Panik ergreift	40	53,20	0,0790
Selbsterleben wie fremde Person	27	38,20	0,0747
Sich fürchten, Ziele nicht zu erreichen	33	50,60	0,0257
Sie fühlen sich angespannt	32	42,08	0,1095
Sie haben Angst vor Zukunft	32	47,84	0,0365
Sie haben Probleme, sich zu entspannen	34	32,03	0,5645
Sie haben viele Sorgen	35	41,74	0,2011
Unsicherheit in Gruppe	30	44,54	0,0426
Sie sind leichten Herzens	34	45,70	0,0867
Fühle mich ausgeglichen	33	47,30	0,0510
Bin entspannt	23	30,13	0,1457
Bin gelöst	25	44,84	0,0087
Bin ruhig	38	64,44	0,0047
Es sich bequem machen / entspannen	49	58,33	0,1698
Ausgeglichen und selbstsicher	23	72,25	0,0000
Ruhig und gelassen	40	58,06	0,0322
Fühle mich geborgen	40	77,95	0,0003
Fühle mich gelöst	37	41,66	0,2751
Fühle mich selbstsicher	37	52,26	0,0494
Fühle mich wohl	27	23,81	0,6408
Schwierigkeiten gelassen entgegen sehen	35	40,70	0,2338
Sie fühlen sich ruhig	36	43,16	0,1919
Sie fühlen sich sicher und geschützt	28	49,05	0,0082

5.5. Die Itembank des Angst-CATs: Zusammenfassung

Die Itembank, welche sich nach der Realisierung des „Item-Link-Designs“ ergibt, setzt sich aus den Items der drei Teilstichproben zusammen, welche die einzelnen Kriterien der statistischen Itemanalyse und –selektion in den separat pro Teilstichprobe durchgeführten methodischen Teilschritten erfüllt haben. Insgesamt umfasst die Itembank, welche dem Angst-CAT zugrundegelegt wird, 50 Items, von denen 19 Items der ersten Teilstichprobe, 19 Items der zweiten Teilstichprobe und 7 Items der dritten Teilstichprobe entstammen (siehe Tabelle 16).

Tabelle 16: Überblick über die Herkunft der insgesamt 50 Items der Itembank des Angst-CATs.

Teilstichproben	Anker-Items	+ weitere Items	+ weitere Items	+ weitere Items
1. N = 1.010	5	19	-	-
2. N = 834	5	-	19	-
3. N = 779	5	-	-	7

Anker-Items: Items, welche in allen drei Teilstichproben gleichermaßen vorliegen, um ein Item-Link-Design zu ermöglichen.

Die Items der Itembank sind in Tabelle 17 anhand ihrer Itemparameterwerte (Steigungs-, Lokations- und Schwellenparameterwerte) charakterisiert.

Die Lokationsparameterwerte der Items, welche die Itembank des Angst-CATs konstituieren, liegen zwischen –1,58 und 1,55 ($\overline{X}$ = -0,11; SD = 0,65); die Schwellenparameterwerte (Thresholds) liegen zwischen -2,81 („bin gelöst“) und 3,30 („fühle mich kribbelig“). Die Schwellenparameter der Items streuen also in einem Bereich von ca. 6 Standardabweichungen, so dass angenommen werden kann, dass die Items des Angst-CATs einen großen Teil des Angstkontinuums abzubilden vermögen. Die Verteilung der Schwellenparameterwerte wird in Abbildung 20 veranschaulicht.

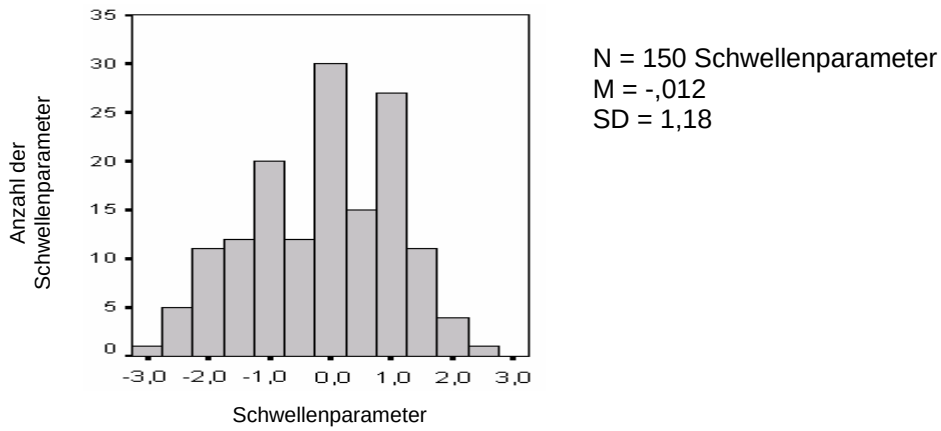


Abbildung 20: Verteilung der Schwellenparameter der Itembank des Angst-CATs.

Tabelle 17: Die Itembank des Angst-CATs (N = 50 Items): Itemparameterschätzung.

Abgekürzter Itemtext	a_i	b_i	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}
Fühle mich besorgt	0,96	-1,58	-1,78	-1,39			
Bin gelöst	1,90	-1,27	-2,81	-0,94	-0,08		
Fühle mich wohl	1,59	-1,17	-2,41	-1,15	0,04		
Bin entspannt	2,13	-1,14	-2,46	-0,92	-0,02		
Fühle mich ausgeglichen	1,20	-0,90	-1,62	-0,91	-0,18		
Fühle mich gelöst	0,86	-0,88	-1,40	-1,10	-0,14		
Sie sind leichten Herzens	0,97	-0,80	-1,86	-1,15	0,62		
Fühle mich geborgen	0,83	-0,77	-2,13	-0,70	0,53		
Fühle mich selbstsicher	1,05	-0,74	-2,23	-0,44	0,46		
Sie fühlen sich angespannt	1,50	-0,71	-2,17	-0,84	-0,20	0,36	
Fühle mich unsicher	1,29	-0,70	-1,13	-0,28			
Fühle mich beunruhigt	1,15	-0,67	-1,84	-0,19	0,02		
Ausgeglichen und selbstsicher	2,60	-0,56	-1,96	-0,12	0,40		
Sie fühlen sich ruhig	1,08	-0,38	-1,23	-0,88	0,96		
Bin ruhig	1,29	-0,37	-2,02	-0,09	0,99		
Unsicherheit in Gruppe	0,88	-0,32	-1,46	0,83			
Sie fühlen sich angespannt	1,26	-0,32	-2,32	0,06	1,31		
Sie haben viele Sorgen	1,24	-0,28	-1,79	0,08	0,86		
Sie fühlen sich nervös	1,84	-0,25	-1,35	0,08	0,52		
Gelassen oder aufgeregt fühlen	1,45	-0,23	-2,50	-1,14	-0,17	0,88	1,76
Sie haben Probleme, sich zu entspannen	1,42	-0,22	-1,38	-0,04	0,76		
Bin besorgt	1,27	-0,20	-1,58	0,12	0,86		
Nervös	0,89	-0,16	-2,46	-0,99	-0,49	1,05	2,10
Schwierigkeiten gelassen entgegensehen	1,47	-0,16	-1,72	0,05	1,19		
Es sich bequem machen / entspannen	0,95	-0,14	-1,78	-0,72	0,69	1,26	
Sie fürchten, Ziele nicht zu erreichen	1,62	-0,14	-1,43	0,18	0,85		
Ängstlich, besorgt oder aufgeregt	2,00	-0,10	-1,40	-0,04	0,14	0,91	
Sie haben Angst vor Zukunft	1,84	-0,10	-1,22	0,27	0,65		
Menschenansammlungen schrecken mich ab	0,82	-0,09	-0,94	0,76			
Fühle mich angespannt	1,44	-0,09	-1,50	0,07	1,15		
Sorgen wegen Gesundheit gehabt	0,96	0,01	-1,98	-1,32	-0,05	1,35	2,02
Hatte Angst	1,01	0,08	-0,40	-0,04	0,68		
Bin beunruhigt	1,97	0,08	-1,10	0,31	1,04		
Beschwerden wegen innerer Ängste	1,28	0,18	-0,94	0,56	0,91		
Bin nervös	2,02	0,19	-0,94	0,36	1,14		
Bin besorgt, dass etwas schiefgeht	1,26	0,26	-0,76	0,34	1,22		
Sie fühlen sich sicher und geschützt	1,46	0,27	-0,33	0,87			
Ruhig und gelassen	1,32	0,30	-0,87	-0,14	1,92		
Gefühl der Benommenheit	0,80	0,31	-0,65	1,27			
Sehe alles so schwarz, dass mich Panik ergreift	1,39	0,31	-0,39	-0,15	0,78	1,01	
Habe Gefühl, nicht wirklich da zu sein	1,63	0,47	-0,37	1,30			
Bin aufgeregt	1,23	0,49	-0,85	0,79	1,51		
Bin verkrampft	1,42	0,58	-0,43	0,84	1,34		
Selbsterleben wie fremde Person	1,60	0,62	-0,04	1,28			
Bin von Angst und Unruhe getrieben	1,69	0,69	0,20	0,79	1,07		
Körper erscheint plötzlich fremd	1,01	0,76	0,10	1,41			
Bin überreizt	1,19	0,93	-0,03	1,10	1,72		
Bin zappelig	1,06	0,94	-0,02	1,09	1,74		
Fühle mich kribbelig	0,83	1,17	-0,38	0,59	3,30		
Kloßgefühl, Engigkeit, Würgen im Hals	0,83	1,55	0,81	2,29			

Itemparameter: a_i = Steigungsparameter; b_i = Lokationsparameter; b_{ih} = Schwellenparameter.

Die Steigungsparameterwerte der Itembank variieren in einem Bereich von $a_i = 0,80$ bis $a_i = 2,60$ ($\overline{X} = 1,34$; SD = 0,40). Diese relativ hohen Steigungsparameterwerte der Items resultieren daher, dass Items mit einem Steigungsparameter $a_i < 0,8$ gezielt aus der Itembank ausgeschlossen wurden,

da ihnen eine geringe Diskriminationsfähigkeit zwischen Personen unterschiedlicher Merkmalsausprägung zugeschrieben wird (Kapitel 5.4.3.1.).

Mit den 50 Items der Itembank soll Zustands-Angst erfasst werden, wobei 70% der Items (N = 35) das Vorliegen der Angst in *positiver* Ausprägung und 30% der Items (N = 15) zur Angst *konträre* Zustände (also das Fehlen der Angst bzw. einen Zustand der „Nicht-Angst“) erfassen (z. B. die Items „selbstsicher“/ „entspannt“/„ruhig und gelassen“/„geborgen“).

Obgleich bei der Instrumentenentwicklung Eindimensionalität angestrebt wurde, und das Ausmaß derselben durch spezifische statistische Itemselektions-kriterien gestärkt wurde, finden sich im Itempool Items, welche verschiedene Aspekte der Angst, erfassen. Diese werden jedoch nicht als statistisch unabhängige Dimensionen behandelt. Zu diesen Aspekten zählen die emotionale und kognitive Komponente der Angst (Liebert & Morris, 1967; siehe Kapitel 2.7.3.4. und 7.3.), sowie alle weiteren Aspekte (abgesehen von dem vegetativen Aspekt der Angst, siehe unten), welche Spielberger (1972) in seiner Definition der Zustands-Angst aufführt (siehe Kapitel 2.3., 2.4.1.1. und 5.3.1.). So besteht die Itembank aus Items, welche speziell den *emotionalen Zustand* der Angst (mit dem Wort „Angst“ im Itemtext) allgemein („von Angst und Unruhe getrieben“/„Hatte Angst“) und im Speziellen („Angst vor Zukunft“/„Furcht, Ziele nicht zu erreichen“) erfragen, und Items, welche explizit die *kognitive Komponente* der Angst („Besorgtheit“) allgemein („besorgt“/„viele Sorgen“) und im Speziellen („besorgt, dass etwas schief geht“/„Sorgen wegen Gesundheit“) erfassen (zue Diskussion der Eindimensionalität siehe Kapitel 2.7.3.4./7.4.1.).

Drei weitere Aspekte, mit denen Spielberger (1972) Zustands-Angst definiert, sind die *Anspannung*, welche in der Itembank durch Items wie „angespannt“ und „Probleme, sich entspannen zu können“ erhoben wird, die *Nervosität* (z. B. „bin nervös“/„fühle mich nervös“) und die *innere Unruhe* (z. B. „aufgeregt“/ „zappelig“/„verkrampft“).

Ausgehend von klinischen Überlegungen (das Körpererleben der Angst steht im klinisch-therapeutischen Alltag oft im Vordergrund) wurden im Rahmen der Itembankkonstruktion auch versucht, Depersonalisationserleben und vegetative Symptome (wie Herklopfen, Schwindel etc.) der Angst in die Itembank mit einzubeziehen. Während Items, welche Aspekte des Depersonalisations-erlebens erfragen („Selbsterleben wie fremde Person“/„Körper erscheint

fremd“), die Kriterien der Itemselektion erfüllten, mussten die meisten Items, welche vegetative Symptome erfragen, aufgrund von Verletzungen der festgelegten statistischen Kriterien ausgeschlossen werden. Zudem mussten auch Items, welche spezifische hypochondrische („Gefühl quält, Körper sei nicht in Ordnung“/„Beunruhigung wegen neuer Krankheiten“) und soziale Ängste („Schämen, wenn versagt“/„Peinlich, vor Gruppe etwas Dummes zu sagen“) erfassen, aus der Itembank ausgeschlossen werden. Die aus der Itembank ausgeschlossenen Items sind in Tabelle 18 zusammengefasst.

Tabelle 18: Überblick über den gesamten Selektionsprozess (31 ausgeschlossene Items).

Abgekürzter Itemtext	Explorative Faktoren-analyse	Analyse residualer Kovarianzen	IRT-Analyse: IRC	IRT-Modellierung: Steigungs-parameter
Schwindelgefühl	X1	X2X3		
Übelkeit	X1	X2		X3
Erwartung, dass Gesundheit nachlässt	X1			
Anfallsweise Herzbeschwerden	X1X2	X3		
Anfallsweise Atemnot	X1X2			
Stiche, Schmerzen oder Ziehen in der Brust	X2	X1X3		
Gefallen, im Mittelpunkt zu stehen	X2			
Im Rampenlicht stehen ist verführerisch	X2			
Selten Sorgen um andere Menschen	X2			
Körper beobachten bzgl. Krankheiten	X2			
Angst, Gesundheit steht das nicht durch	X2			
Beunruhigung wegen neuer Krankheiten	X2			
Wenig ängstlich	X2			
Drang zum Wasserlassen	X3			
Durchfälle	X3			
Schämen, wenn versagt		X2		
Peinlich, vor Gruppe etw.Dummes zu sagen		X2		
Gefühl quält, Körper sei nicht in Ordnung		X2		
Herzklopfen, Herzjagen /- stolpern		X2		X1X3
Schluckbeschwerden		X3		
Starkes Schwitzen		X3		
Anfälle			X3	
Taubheitsgefühl			X3	
Leichtes Erröten			X3	
Ohnmachtsanfälle			X3	
Zittern			X3	
Hatte Mühe, mich zu konzentrieren				X1
Sorgen über gesundheitliche Probleme				X1
Dinge haben mich beunruhigt				X1
Angst, schwer krank zu werden				X2
Aufsteigende Hitze, Hitzewallungen				X3

Selektionsmarkierung:
X: Item wurde in diesem Methodenschritt ausgeschlossen;
1-3: Erste bis dritte Stichprobe, in der jeweiliges Item ausgeschlossen wurde (Items wurden z. T. wegen Stichprobenüberschneidungen mehrfach in verschiedenen Teilstichproben analysiert, um die Stabilität der Ausschlusskriterien zu überprüfen).
Kriterien der Selektion:
1. Explorative F.A. (unrotierte Einfaktorlsg.): Items mit einer Ladung < 0,4 wurden ausgeschlossen;
2. Analyse residualer Kovarianzen: Items mit einer residualen Korrelation > 0,3 wurden ausgeschlossen;
3. IRT-Analyse: Item Response Curves (IRCs): Antwortkategorien, welche nicht genügend zwischen Merkmalsausprägungen zu differenzieren vermochten, wurden ausgeschlossen;
4. IRT-Modellierung: Items mit einem Steigungsparameterwert von $a_i < 0{,}8$ wurden ausgeschlossen.

6. Die Validierung des Computergestützten Adaptiven Tests zur Angstmessung (Angst-CAT)

6.1. Einleitung

Zur Beurteilung der psychometrischen Güte eines Tests ist nach der Testkonstruktion die Validierung des entwickelten Instruments unabdingbar. Die vorliegende empirische Studie widmet sich der Validierung des Angst-CATs, dessen Testentwicklung im vorangegangenen Kapitel 5 beschrieben wurde.

Unter Validierung wird die Überprüfung der Validität eines Tests verstanden (Lienert & Raatz, 1994). Die Validität - nach Bortz und Döring (1995) das wichtigste Testgütekriterien überhaupt – gibt an, „wie gut ein Test in der Lage ist, genau das zu messen, was er zu messen vorgibt“ (S.185). Um die Validität eines Tests zu bestimmen, existieren verschiedene Validitätsansätze, welche unterschiedliche Untersuchungsmethoden erfordern (Cronbach, 1990).

Die Ziele der hier dargestellten Validierungsstudie und die zur Zielerreichung genutzten Validitätsansätze werden in Kapitel 6.2., die untersuchten Hypothesen in Kapitel 6.3 expliziert, gefolgt von einer Beschreibung der untersuchten *Stichprobe* (Kapitel 6.4.) und der an ihr erhobenen *Instrumente* (Kapitel 6.5.). Anschließend werden die zur Untersuchung der verschiedenen Validitätsansätze genutzten statistischen *Methoden* dargestellt (Kapitel 6.6.) und die *Ergebnisse* zusammengefasst und erörtert (Kapitel 6.7.).

6.2. Ziele

Seit Beginn der Validierungsforschung (Anfang des 20. Jahrhunderts durch Spearman, 1904) spielt die *Konstrukt*validität eine dominierende Rolle. Kennzeichnend für diese Art der Validität ist die Erhebung von Konstrukten (z. B. mittels psychometrischer Instrumente) und deren Beziehung zum Testwert des zu validierenden Instruments (hier: das Angst-CAT). Unter der Voraussetzung, dass die ausgewählten und erfassten Konstrukte selbst repräsentativ, reliabel, valide und für die Validierung adäquat sind, können durch die empirische Untersuchung dieser Zusammenhänge Rückschlüsse auf die Gültigkeit des untersuchten Tests gezogen werden.

Das Ziel vorliegender Studie ist die Bestimmung der *Konstruktvalidität* im Sinne einer *Übereinstimmungsvalidität* (*konkurrente* Validität; Lienert & Raatz, 1994, S. 224) von Variablen, von denen aufgrund von theoretischen und empirischen Forschungsbefunden erwartet wird, dass sie in unterschiedlicher Konstruktnähe

zum Angst-CAT positioniert werden können, und welche praktisch zeitgleich – jedoch unabhängig voneinander - mit dem Angst-CAT erhoben werden.

Von Belang ist hierbei nicht nur die Überprüfung, ob mehrere Methoden (psychometrische Tests, Interviews etc.), mit einem ähnlichen Messbereich (Erfassung von Angst), jedoch mit unterschiedlichen Operationalisierungen dieses Messbereichs zu *ähnlichen* Messergebnissen kommen (*konvergente Validität*), sondern auch, ob Hinweise auf *gering* ausgeprägte Zusammenhänge zwischen Tests, welche die Erfassung differierender Konstrukte intendieren, eruierbar sind, so dass Rückschlüsse auf die Fähigkeit des Angst-CATs zur Diskrimination zwischen unterschiedlichen Konstrukten möglich sind (*divergente* bzw. *diskriminante Validität*; Campbell & Fiske, 1959).

6.3. Hypothesen

In der vorliegenden Studie wurden zur Überprüfung der Validität des Angst-CATs neben dem zu validierenden Angst-CAT verschiedene psychometrische Inventare zur Angst- und Depressionserfassung sowie zur Messung verschiedener Persönlichkeitskonstrukte und ein strukturiertes diagnostisches Interview zwischen den Jahren 2002 und 2003 an Patienten der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin angewandt.

Es wird im Sinne einer guten konvergenten Validität erwartet, dass das Angst-CAT mit den erhobenen Angstinventaren in einem engen Zusammenhang steht (hoch korreliert), sowie Patienten mit der Diagnose einer Angststörung im Angst-CAT höhere Werte erzielen als Patienten ohne eine psychische Störung.

Weiterhin wird erwartet, dass sich eine gute divergente Validität des Angst-CATs in Form einer hohen Diskrimination zu anderen Eigenschaftskonstrukten, welche mit den Persönlichkeitsinventaren erfasst werden, und in Form einer Diskrimination zwischen verschiedenen Diagnosegruppen ausdrückt.

Angesichts einer Fülle von theoretischen Forschungsdiskursen (siehe Kapitel 2.5) und einer ausgeprägten empirischen Befundlage (siehe Kapitel 6.5.1.), die darauf hinweist, dass sich die psychometrische Diskrimination zwischen den Konstrukten Angst und Depression bzw. Neurotizismus schwierig gestaltet, wird vermutet, dass eine solche Diskriminationsleistung auch mit dem Angst-CAT nicht hinreichend gelingt.

6.4. Stichprobe

Die Validierungsstichprobe umfasst insgesamt N = 102 Patienten, die in der Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin zur Diagnostik oder Therapie in den Jahren 2002 bis 2003 stationär behandelt wurden. Tabelle 19 fasst die wesentlichsten soziodemografischen und klinischen Charakteristika der Stichprobe zusammen.

Tabelle 19: Soziodemografische und klinische Charakteristika der Validierungsstichprobe.

Charakteristika	Kategorie / Parameter	Angaben
Geschlecht	Weiblich	79,4%
	Männlich	20,6%
Alter	Arithmetischer Mittelwert ($\overline{X}$)	42,28 Jahre
	Standardabweichung (SD)	15,53 Jahre
	Altersspanne	18-77 Jahre
Familienstand	verheiratet	45,1%
	ledig (mit Partner)	14,7%
	ledig (ohne Partner)	25,5%
	geschieden / Verwitwet	14,7%
Diagnosen[81]	Angststörungen (F.40-41)	56,8%
	Depressive Störungen (F.32-34)	58,8%
	Somatoforme Störungen (F.45)	50,0%
	Essstörungen (F.50)	6,9%
	Primär somatische Erkrankungen (nicht F)	9,8%

Leider war es bisher nicht möglich, das Angst-CAT einer gesunden Probandenstichprobe vorzulegen, welche für die Bevölkerung des deutschsprachigen Raumes repräsentativ ist. Jedoch liegen uns von einer Gruppe von N = 35 Medizinstudenten (der Humboldt-Universität zu Berlin) Theta-Werte des Angst-CATs vor, welche im laufenden Sommersemester 2003 erhoben wurden. Diese werden im Folgenden als eine vorläufige Vergleichsstichprobe genutzt.

[81] Die Prozentwerte der Diagnosen summieren sich nicht zu 100%, da Komorbidität zwischen einzelnen Störungen häufig ist.

6.5. Validierungsinstrumente

Zur Validierung wurden im Rahmen der klinisch-psychologischen Routinediagnostik (Testbatterien) das Angst-CAT und fünf psychometrische Verfahren sowie ein strukturiertes diagnostisches Interview angewandt. Diese sollen eine Überprüfung der *Konstrukt*validität des Angst-CATs ermöglichen. Es wird angenommen, dass sie selbst ausreichend valide Repräsentanten der Konstrukte der Angst / Depression und anderer Persönlichkeitsfaktoren darstellen. Folgende psychometrischen Instrumente, welche sich in der klinischen Diagnostik bewährt haben, wurden an oben beschriebener Patientenstichprobe erhoben:

- **das Beck-Angst-Inventar**
  (BAI; Margraf & Ehlers, in Druck),
- **die Hospital Anxiety and Depression Scale**
  (HADS; Hermann, Buss & Snaith, 1995),
- **das Beck-Depressions-Inventar**
  (BDI; Hautzinger, Bailer, Worall & Keller, 1994),
- **das NEO-Fünf-Faktoren-Inventar**
  (NEO-FFI; Borkenau & Ostendorf, 1993) und
- **der Gießen-Test**
  (GT; Beckmann, Brähler & Richter, 1991).

Der Einsatz des Angst-CATs erfolgte an einem stationären Computer; alle weiteren psychometrischen Instrumente wurden computergestützt mittels Handcomputer, sogenannter PDA's (Personal Digital Assistants; Psion), deren Einsatz bereits erprobt ist, erhoben (Rose et al., 1999, 2003; siehe Kapitel 5.2.).

Desweiteren wurde eines der international am weitesten verbreiteten, strukturierten klinischen Interviews (siehe Kapitel 2.7.1. und 4.1.) an oben beschriebener Stichprobe computergestützt angewandt: das M-CIDI (als Papierversion: DIA-X) von Wittchen und Pfister (1996). Dieses unter der Schirmherrschaft der World Health Organization (WHO) und dem National Institute of Mental Health (NIMH) an dem Max-Planck-Institut für Psychiatrie in München entwickelte Instrument dient der strukturierten klinischen Diagnostik der Angst als psychischer Störung nach den Kriterien des ICD-10 (Dilling et al., 2000) und DSM-IV (Saß et al., 1996; siehe Kapitel 6.5.3.).

6.5.1. Klinische Instrumente zur Angst und Depressionsmessung

6.5.1.1. Beck-Angst-Inventar (BAI)

Das Beck-Angst-Inventar (Margraf & Ehlers, 1995) ist ein weit verbreitetes und in vielfältigen klinischen Zusammenhängen eingesetztes Selbstbeurteilungs-Instrument zur Erfassung des Schweregrads klinisch relevanter Angst in Patientengruppen und der Allgemeinbevölkerung (ab 12 Jahren).

Das Instrument, welches 21 Items mit 4-stufigem Antwortformat umfasst, wurde entwickelt, um Angst hinsichtlich der Schwere ihres Auftretens in den letzten 7 Tagen in Anlehnung an die Symptomlisten des DSM-IV (Saß et al., 1996) für Panikanfälle und generalisierte Angst möglicht exakt und ökonomisch zu messen. Die Items repräsentieren weitestgehend somatische Korrelate der Angst (Westhoff, 1993).

Das BAI basiert auf der amerikanischen Originalversion (Beck & Steer, 1993), welche die Erfassung der Ängstlichkeit möglichst unabhängig von depressiver Symptomatik intendiert. Dieser Anspruch wird nur teilweise eingelöst (Korrelationen mit Depressionsmaßen liegen zwischen r = 0,43 (CCL-D[82]) bis r = 0,47 (BDI[83]), N = 281 bzw. N = 287, Margraf & Ehlers, in Druck).

Das BAI (Originalversion) korreliert mit der RCMAS[84] (N = 80 psychiatrische erwachsene Patienten) in einer Höhe von r = 0,58 und mit der Angst-Skala des MMPIs[85] nach statistischer Kontrolle des Zusammenhangs zu den BDI-Scores in einer Höhe von r = 0,30 (N = 125 Jungen) bzw. r = 0,56 (N = 115 Mädchen; Osman et al., 2002). Der Übereinstimmungsvaliditätskoeffizient zu Fremd-ratings der Angst von Klinikern liegt an einer psychiatrischer Stichprobe bei r = 0,40.

Der deutschen Version des BAIs wird eine sehr gute bis gute interne Konsistenz von $\alpha = 0,92$ (N = 291 Patienten mit psychischen und / oder organischen Diagnosen) bis $\alpha = 0,88$ (N = 3.000 Personen aus der Allgemein-bevölkerung) und eine mäßig bis hohe Retest-Reliabilität (N = 1.000; r = 0,68 bei 14 Tagen; r = 0,9 bei 48h) zugeschrieben, so dass eine Sensitivität für Therapieeffekte angenommen werden kann. Sie korreliert mit Angstmaßen zu

[82] CCL-D : Cognition Checklist-Depression (Tönnies, 1995).
[83] BDI : Beck-Depressions-Inventar (Hautzinger, Bailer, Worall & Keller, 1994).
[84] RCMAS: Revised Children's Manifest Anxiety Scale (Reynolds & Richmond, 1978).
[85] MMPI: Minnesota Multiphasic Personality Inventory for Adolescents (Butcher et al., 1992).

r = 0,45 (STAI-State[86]; N = 154), r = 0,48 (STAI-Trait; N = 227), r = 0,50 (CCL[87]; N = 289) und r = 0,73 (SCL-90-R-Angst[88], N = 675; Margraf & Ehlers, in Druck). Populationsnormen für klinische Stichproben und die Allgemeinbevölkerung liegen vor (N = 2.000).

6.5.1.2. Hospital Anxiety and Depression Scale (HADS)

Die Hospital Anxiety and Depression Scale (Hermann, Buss & Snaith, 1995) ist ein kurzer Selbstbeurteilungsfragebogen zur Erfassung von Angst und Depressivität bei Erwachsenen. Er wurde gezielt zum Einsatz bei körperlich Kranken konstruiert (Zigmond & Snaith, 1983) und soll im Kontext somatischer Medizin dazu beitragen, Patienten mit psychischer Morbidität zu identifizieren (Brähler, Holling, Leutner & Petermann, 2002). Die HADS besteht aus 14 Items mit 4-stufigem Antwortformat, aus denen je eine Angst- und Depressivitäts-Subskala (HADS-A /-D) gebildet wird. Angst wird in Anlehnung an die Generalisierte Angststörung (DSM-IV; Saß et al., 1996) und Depressivität wird hinsichtlich „endogenomorpher" Symptome (Freudlosigkeit, Interessenverlust etc.) bezüglich ihres Auftretens in der letzten Woche erfasst.

Die interne Konsistenz der Angst-Subskala liegt bei α = 0,80, die der Depressivitäts-Subskala bei α = 0,81 (N = 6.200 Patienten). Die Retest-Reliabilitäten betragen zwischen r = 0,7 (> 6 Wochen) und r = 0,84 bzw. r = 0,85 (2 Wochen). Korrelationen zu anderen Angst- bzw. Depressionsskalen an N = 1.815 Patienten liegen zwischen r = 0,48 bis r = 0,86 ($\bar{r}$=0,66[89], HADS-A) bzw. r = 0,46 bis r = 0,78 ($\bar{r}$ = 0,59[90], HADS-D; Hinz & Schwarz, 2001). Interkorrelationen zwischen der Angst- und der Depressionsskala des HADS liegen bei r = 0,53. Es existieren Normen für N = 5.579 kardiologische Patienten und vorläufige Normen für N = 278 Gesunde.

---

[86] STAI: State-Trait-Angst-Inventar (Laux et al., 1981).
[87] CCL: Cognition Checklist (Tönnies, 1995).
[88] SCL-90-R: Die Symptom-Checkliste von Derogatis (Franke, 1995).
[89] Gemittelte Korrelation der in der Validierungsstudie des HADS (Hermann et al., 1995) eingesetzten folgenden acht Angstskalen: Angstskala des General Health Questionnaire (GHQ-28); Linear-Analog-Angstskala; Irritability, Depression and Anxiety-Scale (IDA); Zung Angst- und Depressionsskala; Crown-Crisp Experiential Index; Arthritis Impact Measurement Scale (AIMS), de Bonis-Angstskala; State Trait Anxiety Inventory (State).
[90] Gemittelte Korrelation der in der Validierungsstudie des HADS (Hermann et al., 1995) eingesetzten folgenden 6 Depressionsskalen: Depressionsskala des General Health Questionnaire (GHQ-28); Irritability, Depression and Anxiety-Scale (IDA); Zung Angst- und Depressionsskala; Arthritis Impact Measurement Scale (AIMS); Crown-Crisp Experiential Index; Depressionsskala (D-S).

6.5.1.3. Beck-Depressions-Inventar (BDI)

Das Beck-Depressions-Inventar (Hautzinger, Bailer, Worall & Keller, 1994) ist ein seit 30 Jahren international und national weit verbreitetes Selbstbeurteilungsinstrument zur Erfassung des Schweregrades depressiver Symptomatik bei Jugendlichen ab 16 Jahren und bei Erwachsenen. Es entstand vor dem Hintergrund klinischer Beobachtungen depressiver Patienten und erfasst mit 21 Items die häufigsten depressiven Symptome. Seine innere Konsistenz liegt zwischen $\alpha = 0{,}73$ und $\alpha = 0{,}95$, die Retest-Reliabilitäten für eine Woche betragen $r = 0{,}60$ bzw. $r = 0{,}86$.

Korrelationen mit anderen Depressionsinventaren liegen zwischen $r = 0{,}61$ bis $r = 0{,}86$ (HAMA)[91], $r = 0{,}57$ bis $r = 0{,}83$ (SRDS)[92] und $r = 0{,}41$ bis $r = 0{,}70$ (MMPI-D)[93]. Patienten mit *Angst*störungen haben in der Regel ebenfalls erhöhte BDI-Werte, wenngleich Patienten mit Depressionen im BDI meist signifikant höhere Werte als Angstpatienten zeigen (Beck, 1994). Korrelationen mit Angstinventaren liegen auch in spezifischen Artikeln zur Differenzierbarkeit von Angst und Depression nicht vor (Steer, Beck, Riskind & Brown, 1986). Es liegen Normen einer klinischen Stichprobe (N = 477) depressiver Patienten vor.

6.5.2. Persönlichkeitsinventare

6.5.2.1. NEO-Fünf-Faktoren-Inventar (NEO-FFI)

Das NEO-Fünf-Faktoren-Inventar von Borkenau und Ostendorf (1993; Originalversion: NEO-FFI von Costa & McCrae, 1985) ist ein multidimensionaler Persönlichkeitsstrukturtest für Erwachsene, welcher sowohl für Forschungszwecke, als auch für Anwendungen in der Klinischen Psychologie, der Schullaufbahn-, Studien- und Berufsberatung sowie in der Organisations-psychologie genutzt wird.

Er geht auf den sogenannten psycholexikalischen Ansatz zurück (Allport & Odbert, 1936; Cattell, 1943; Angleitner, Ostendorf & John, 1990). Umfangreiche faktorenanalytische Studien zu individuellen Unterschieden in der Persönlichkeit zeigen, dass der Einschätzung von Personen fünf robuste Dimensionen

[91] HAMA: Hamilton-Angst-Skala (Hamilton, 1959).
[92] SRDS: Self-Rating Depression Scale (Zung, 1965).
[93] MMPI-D: Minnesota Multiphasic Personality Inventory-Depression Scale (Hathaway & McKinley, 1983).

(„Big Five“) zugrunde liegen, welche das NEO-FFI mit 60 fünfstufigen Items (pro Skala: 12 Items) mit den folgenden fünf Skalen erfasst:

1. die „*Neurotizismus*“-Skala erfasst emotionale Stabilität versus Labilität, d. h. inwiefern ein Proband z. B. dazu neigt, nervös, ängstlich, traurig, unsicher und verlegen zu sein und sich Sorgen um seine Gesundheit zu machen;
2. mit der „*Extraversions*“-Skala kann das Ausmaß der Geselligkeit, Selbstsicherheit, Aktivität, Gesprächigkeit und der Optimismus einer Person erhoben werden;
3. die Skala „*Offenheit für Erfahrung*“ misst das Interesse an neuen Erfahrungen, Erlebnissen und Eindrücken;
4. die Skala „*Verträglichkeit*“ erfasst die Neigung zu altruistischem Verhalten und das zwischenmenschliche Vertrauen bzw. Harmoniebedürfnis und Nachgiebigkeit und
5. die Skala „*Gewissenhaftigkeit*“ misst das Ausmaß der Impuls- und Selbstkontrolle (Ordentlichkeit, Zuverlässigkeit, Ehrgeiz, Disziplin).

Die internen Konsistenzen der Skalen liegen bei α = 0,78 (N = 2.112), die Retest-Reliabilitäten von zwei Jahren liegen zwischen r = 0,65 (Verträglichkeit) und r = 0,81 (Extraversion). Eine Reihe von Studien zur faktoriellen Validität belegen durch hohe Kongruenzkoeffizienten (r = 0,91 bis r = 0,98) die Stabilität der Faktorenstruktur über unterschiedliche Stichproben. Zur kriteriums-bezogenen Validität werden im Testhandbuch keine Studien erwähnt. Es liegen keine bevölkerungsrepräsentativen Normen, jedoch statistische Kennwerte einer Standardisierungsstichprobe (N = 2.112) vor.

6.5.2.2. Gießen-Test (GT)

Der Gießen-Test von Beckmann, Brähler und Richter (1991) ist ein Selbstbeurteilungsverfahren, welches den Probanden die Gelegenheit gibt, sich selbst in ihrem Realselbst- und Idealselbstbild einzuschätzen. Er dient der klinischen Diagnostik und Therapieverlaufsevaluation und findet unter anderem auch Anwendung in der sozialpsychologischen Forschung. Bei der Erfassung des Realselbst- und Idealselbstbildes werden vor allem die innere Verfassung einer Person und seine psychosozialen Umweltbeziehungen fokussiert.

Der GT besteht aus 40 bipolar formulierten Feststellungen, die auf einer siebenstufigen Skala nach ihrem Zutreffen beantwortet werden sollen, und die zu den folgenden sechs Skalen zusammengefasst werden:

1. die Skala „*Soziale Resonanz*“ dient der Selbsteinschätzung einer Person bezüglich ihrer Wirkung auf die Umwelt. Dazu gehören sowohl äußere Merkmale (Aussehen, Attraktivität) als auch das selbsteingeschätzte eigene Maß an Beliebtheit, Wertschätzung, Achtung und Durchsetzungsfähigkeit;
2. die Skala „*Dominanz*“ bildet Merkmale wie Aggressivität, Eigensinn und Impulsivität versus Gefügigkeit bzw. Unterordnungstendenzen ab;
3. die Skala „Kontrolle“ erfasst das Ausmaß der Selbstkontrolle im Sinne von Ordentlichkeit, Stetigkeit, Eifer und Genauigkeit im Umgang mit Objekten;
4. die Skala „*Grundstimmung*“ dient der Erfassung der allgemeinen Stimmung (u. a. Depressivität, Ängstlichkeit und Ärger);
5. die Skala „*Durchlässigkeit*“ erfasst die zwischenmenschliche Aufgeschlossenheit, Vertrauensseeligkeit und die Fähigkeit, psychosoziale Bedürfnisse im Kontakt mit anderen Menschen zu äußern;
6. die Skala „*Soziale Potenz*“ erhebt das Ausmaß an sozialen Fähigkeiten wie Geselligkeit, Hingabefähigkeit, Konkurrenzfähigkeit etc., welche eine Person sich selbst zuschreibt.

Die mittlere interne Konsistenz der Skalen liegt bei α = 0,86 (N = 235 „neurotische“ Patienten); die Restest-Reliabilitäten für sechs Wochen liegen zwischen r = 0,65 und r = 0,76 (N = 204 „neurotische“ Patienten). Da die Autoren eine konzeptuelle Validität aufgrund der gezielten Auswahl tiefenpsychologisch und sozialpsychologisch relevanter Feststellungen annehmen, liegen Ergebnisse von Kriteriumsvalidierungsstudien an N = 2.182 Probanden vor (zwei eingesetzte Vergleichsinstrumente: a) zum Erziehungsverhalten und b) zu interpersonellen Problemen; Brähler, Schumacher & Brähler, 1999). Eine aktuelle Normierung (1999) an N = 1.008 Ostdeutschen und N = 995 Westdeutschen findet sich bei Brähler und Richter (2000).

6.5.3. Diagnostisches Interview: M-CIDI (DIA-X)

Das Munich Composite International Diagnostic Interview (M-CIDI; Wittchen & Pfister, 1996) ist ein voll standardisiertes computergestütztes Interview-

verfahren zur diagnostisch-klassifikatorischen Erfassung psychischer Störungen, welches sich zum Einsatz in der klinischen Praxis und Forschung (v. a. in epidemiologischen Studien) bei Probanden im Alter von 14 bis 65 Jahren eignet. Als Papierversion nennt es sich *DIA*gnostisches E*X*pertensystem psychischer Störungen (DIA-X).

Es wurde unter der Schirmherrschaft der Weltgesundheitsorganisation (WHO) und dem National Institute of Mental Health (NIMH, U.S.A.) entwickelt und erlaubt die Diagnostik von 64 Störungen nach den Kriterien des ICD-10 (Dilling, et al., 2000) und DSM-IV (Saß et al., 1996). Folgende psychische Störungen werden in 12 Interviewsektionen erfragt: a) Störungen durch Tabak, b) somatoforme und dissoziative Störungen, c) Phobien und andere Angststörungen, d) depressive Störungen und Dysthymie, e) Manie und bipolare affektive Störungen, f) Schizophrenie und andere psychotische Störungen, g) Essstörungen, h) Störungen durch Alkohol, i) Zwangsstörungen, j) Drogenmissbrauch und –abhängigkeit, k) organisch bedingte psychische Störungen, l) posttraumatische Belastungsstörungen.

Das M-CIDI-Programmpaket ermöglicht sowohl eine simultan zum Interview verlaufende computergestützte Datene*ingabe* sowie eine automatische *Auswertung* des Interviews nach den diagnostischen Kriterien des ICD-10 und DSM-IV. Der Diagnosenausdruck umfasst Angaben zu den vorliegenden psychischen Störungen, deren erstes und letztes Auftreten, dem jeweiligen Schweregrad und der Komorbiditätsstruktur.

Die Interrater-Reliabilitäten liegen zwischen κ = 0,81 und κ = 1,0 (*symptom*bezogene Interrater-Reliabilitäten) bzw. κ = 0,82 und κ = 0,98 (*diagnosenbezogene* Interrater-Reliabilitäten). Die Restest-Reliabilitäten von 1-14 Tagen (N = 142 Fälle) liegen zwischen κ = 0,49 (undifferenzierte somatoforme Störung) und κ = 0,83 (Anorexia nervosa); für Angststörungen beträgt sie κ = 0,57 (soziale Phobie) bis κ = 0,92 (Panikattacken).

Die Validität variiert stark zwischen unterschiedlichen Diagnosegruppen. Im Vergleich zu klinischen Konsensus-Diagnosen erfahrener Psychiater ergaben sich Übereinstimmungswerte zu der strukturierten computergestützten Interviewdiagnostik von κ = 0,39 (psychotische Störungen), κ = 0,39 / 0,43 (somatoforme Störungen) bis κ = 0,82 (Panikstörungen).

6.6. Methodisches Vorgehen

Die folgende Beschreibung des methodischen Vorgehens orientiert sich an der Reihenfolge der Darstellung der Ergebnisse.

Der Ergebnisteil, welcher die Validierung des Angst-CATs beinhaltet (Kapitel 6.7.), gliedert sich in einen *ersten* allgemein deskriptiven Ergebnisteil (6.7.1.), einen *zweiten* Teil, welcher der konvergenten Validierung (6.7.2.), und einen *dritten* Teil, welcher der diskriminanten Validierung des Angst-CATs (6.7.3.) dient.

Im *ersten* Teil (Kapitel 6.7.1.) werden die Itemselektion, d. h. die im Angst-CAT dargebotene Anzahl der Items in Abhängigkeit von den geschätzten Theta-Werten mit deskriptiven Statistiken untersucht, mögliche Zeitersparnisse bei Einsatz des Angst-CATs gegenüber herkömmlichen Instrumenten analysiert und Verteilungsparameter der Theta-Werte des Angst-CATs in Abhängigkeit von soziodemografischen Variablen exploriert. Als inferenzstatistische Prüfmethoden werden zur Überprüfung von Mittelwertsunterschieden t-Tests für unabhängige Stichproben (z. B. zur Untersuchung eines Geschlechtseffekts) sowie einfaktorielle Varianzanalysen (zur Untersuchung von Alters- bzw. Familienstandseffekten) durchgeführt.

Der *zweite* Teil (Kapitel 6.7.2.) umfasst die *konvergente* Validierung des Angst-CATs, d. h. es wird die Beziehung zu Instrumenten, deren Messbereiche konstruktnah bzw. -identisch mit dem des Angst-CATs sind, untersucht. Dieser Teil gliedert sich in zwei Unterkapitel. Zunächst wird die konvergente Validität in Bezug auf andere psychometrische Testverfahren und anschließend in Bezug auf das mit dem strukturierten Interview (M-CIDI) erhobene Fremdurteil untersucht.

Zur Überprüfung der konvergenten Validität bezüglich verschiedener Testverfahren wurden Produkt-Moment-Korrelationen (Pearson's Korrelationskoeffizient) mit den erhobenen Summenscores der Angst-Inventare (Angst-CAT, BAI, HADS-A) berechnet. Die konvergente Validität in Bezug auf das Fremdurteil wird untersucht, indem die Mittelwertsunterschiede der Theta-Schätzungen verschiedener Stichproben, welche mit dem M-CIDI klassifikatorisch erfasst wurden (Patienten ohne bzw. mit Angststörungen, Referenzgruppe: Medizinstudenten), inferenzstatistisch mittels einfaktorieller Varianzanalysen überprüft werden.

Der dritte Teil (Kapitel 6.7.3.), welcher der Untersuchung der *diskriminanten* Validität dient, d. h. der Zusammenhangsuntersuchung des Angst-CATs zu Instrumenten, welche die Messung unterschiedlicher Konstrukte intendieren, gliedert sich - wie Kapitel 6.7.2. - in eine Untersuchungsphase zur Überprüfung der Validität in Bezug auf andere Testverfahren und in Bezug auf das diagnostische Fremdurteil (M-CIDI).

Die Überprüfung der diskriminanten Validität in Bezug auf andere Testverfahren geschieht mittels korrelativer Statistiken (Pearson's Korrelationskoeffizient). Hier wird zunächst die Diskriminationsfähigkeit des Angst-CATs zwischen den Konstrukten Angst und Depression untersucht, indem die erhobenen Angst- und Depressionsinventare (BAI, HADS, BDI) in korrelative Beziehung gesetzt werden. Anschließend wird die psychometrische Diskriminationsfähigkeit des Angst-CATs zu anderen Persönlichkeits-konstrukten mittels der Ergebnisse der zwei eingesetzten Persönlichkeits-inventare (NEO-FFI, GT) korrelations- und regressionsstatistisch exploriert.

Die diskriminante Validität in Bezug auf das diagnostische Fremdurteil (M-CIDI) wird bestimmt, indem die Mittelwerte der Theta-Schätzungen von verschiedenen mit dem M-CIDI ermittelten Diagnosegruppen (Patienten mit Angst-, depressiven, Ess- und somatoformen Störungen) verglichen werden. Da bei psychosomatischen Patienten hohe Komorbiditätsraten zwischen den einzelnen Störungsgruppen zu erwarten sind, erfolgt eine Überprüfung der Mittelwertsunterschiede der Theta-Werte zunächst zwischen den in der Realität am häufigsten vorkommenden Diagnosegruppen, bei denen Patienten mehrere Störungen aufweisen (Komorbidität), d.h. aufgrund von Komorbidität kommt es zu Personenüberschneidungen zwischen den Diagnosegruppen. Um den die diskriminante Validität des Angst-CATs möglicherweise beeinträchtigenden Einfluss dieser Komorbidität zu eliminieren, werden anschließend die verschiedenen Diagnosegruppen unter Ausschluss von Komorbidität gebildet (d. h. ohne Überschneidungen zwischen den Diagnosegruppen) und inferenzstatistisch auf Mittelwertsunterschiede in den Theta-Werten überprüft. Als globale Prüfmethode wurde die einfaktorielle Varianzanalyse, als Einzelvergleichsmethode der Scheffé-Test eingesetzt. Alle erläuterten Berechnungen[94] erfolgten unter Einsatz des Programms SPSS 10.0.

[94] Allen erörterten Berechnungen ist gemein, dass sowohl das Intervallskalenniveau als auch

6.7. Ergebnisse

6.7.1. Allgemeine Ergebnisse zum Angst-CAT

6.7.1.1. Die Itemselektion

Die Stoppfunktion des IRT-basierten Itemselektionsalgorithmus (Kapitel 4.3.3.6.) des Angst-CATs wurde im Rahmen vorliegender Validierungsstudie auf eine Reliabilität von Rel (θ) = 0,9 festgelegt. Die Anwendung des Angst-CATs an N = 102 psychosomatischen stationären Patienten (Kapitel 6.4.) zeigte, dass...

1. eine Erfassung der Angstausprägung mit im Durchschnitt 5,3 $\pm$ 1,9 Items ($\overline{X}$ $\pm$ SD) auf diesem Messpräzisionsniveau möglich ist;
2. die durchschnittliche Testdurchführungszeit der Patienten[95] 1 min. und 40 sek. beträgt (SD = 49s)[96];
3. zwischen 4 und 14 Items pro Testdurchlauf zur Angsterfassung genutzt werden (Verteilungsspannweite);
4. die darzubietende Itemanzahl bei einer durchschnittliche Angstausprägung gering ist, jedoch zu den Extremausprägungen hin zunimmt (siehe Abbildung 21).

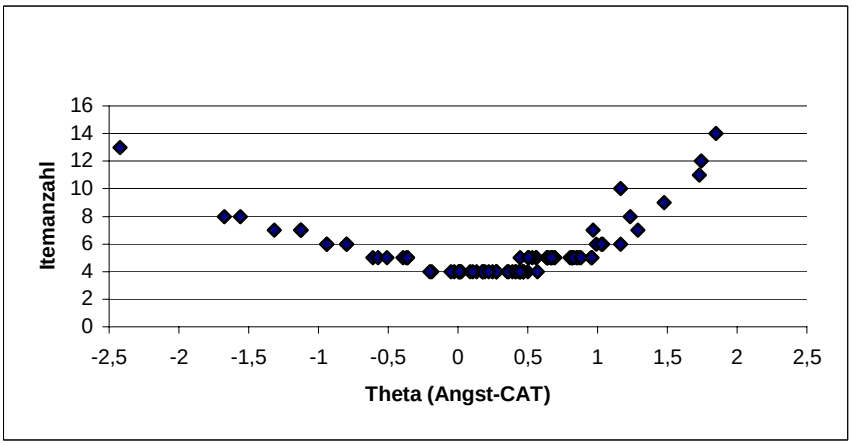


Abbildung 21: Verteilung der im Angst-CAT dargebotenen Anzahl der Items in Abhängigkeit von den durch das Angst-CAT geschätzten Theta-Werten (N = 102 psychosomatische Patienten).

die Normalverteilung vorausgesetzt wird.

[95] Zum Vergleich: Die durchschnittliche Testdurchführungszeit des Angst-CATs beträgt bei Studenten: 1min., 25sek., SD = 46sek..

[96] Zum Vergleich: Die durchschnittliche Testdurchführungszeit des STAIs mit 40 Items wird im Testhandbuch (Laux et al., 1981) auf 6-10 min. geschätzt.

Letzterer Befund resultiert aus der Beschaffenheit der Itembank, welche aus vielen hoch informativen Items besteht, die eine *durchschnittliche* Angstausprägung (bezogen auf das psychosomatische Kollektiv) gut erfassen, jedoch weniger Items aufweist, welche in den *Extrem*bereichen der Angstausprägung eine hohe Iteminformation aufweisen, so dass – wird eine konstante Messgenauigkeit über alle Merkmalsausprägungsbereiche hinweg angestrebt (zur Stoppfunktion siehe Kapitel 4.3.3.6.) – in den Extrembereichen mehr Items dargeboten werden müssen, um diese zu gewährleisten.

Die dargestellten Befunde replizieren Ergebnisse einer Simulations-Vorstudie zur Güte des Angst-CATs (Walter et al., eingereicht), in der computergestützt konventionell (d. h. nicht adaptiv) erhobene psychometrische Daten von N = 2.348 psychosomatischen Patienten mit einem simulierten adaptiven Itemselektionsalgorithmus so reanalysiert wurden, dass für jeden Patienten eine IRT-basierte Theta-Schätzung mit dem Angst-CAT erfolgte. Der in Abbildung 21 veranschaulichte Zusammenhang offenbarte sich bereits in dieser Simulations-Vorstudie. Im Rahmen der Vorstudie konnte weiterhin die im Angst-CAT darzubietende Itemanzahl bei unterschiedlichen Stoppfunktionen simuliert werden. Bei einer Stoppfunktion von Rel(θ) = 0,9 zeigten sich ähnliche Ergebnisse wie in der Validierungsstudie: $6{,}9 \pm 2{,}6$ Items ($\overline{X} \pm$ SD) wurden zur Schätzung der Theta-Werte vom Angst-CAT in Simulationen genutzt. Wurde das Angst-CAT mit einer Stoppfunktion von Rel(θ) = 0,8 simuliert, so benötigte es zur Angsterfassung nur $3{,}1 \pm 0{,}8$ Items ($\overline{X} \pm$ SD).

Erste Hinweise aus dieser Studie widersprechen der naheliegenden Vermutung eines deutlichen Informationsverlusts bei der Darbietung dieser geringen Anzahl von Items. Abbildung 22 veranschaulicht den hohen korrelativen Zusammenhang (r = 0,97)[97] zwischen der simulierten Theta-Schätzung des Angst-CATs auf der Grundlage der gesamten Itembank und der des Angst-CATs (Walter et al., eingereicht).

[97] Bei der Interpretation dieses Ergebnisses muss berücksichtigt werden, dass sich die hohe Ausprägung der Korrelation teilweise aus einer sich überlappenden Itemmenge ergibt.

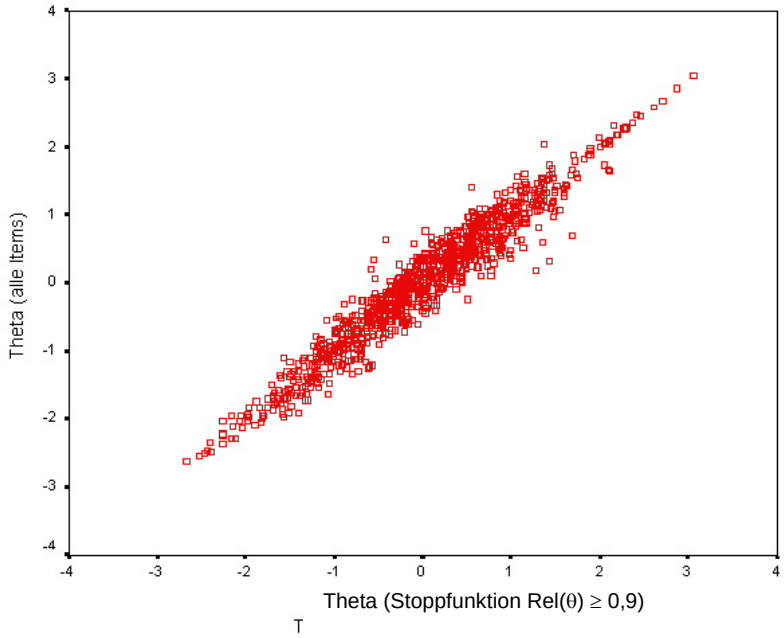


Abbildung 22: Beziehung zwischen der Theta-Schätzung auf der Grundlage aller Items der Itembank und der Theta-Schätzung des Angst-CATs (Stoppfunktion Rel(θ) ≥ 0,9).

6.7.1.2. Statistische Kennwerte in Abhängigkeit von soziodemografischen Variablen

Im Folgenden wurde die Verteilung der Theta-Werte des Angst-CATs von N = 102 psychosomatischen stationären Patienten bezüglich sozio-demografischer Kennwerte wie des Geschlechts, Alters und Familienstands untersucht.

Ein durchgeführter t-Test für unabhängige Stichproben zur zufallskritischen Überprüfung möglicher *geschlechts*bedingter Mittelwertsunterschiede führt zu keinem signifikanten Ergebnis (Tabelle 20).

Tabelle 20: Statistische Kennwerte des Angst-CATs in Abhängigkeit vom Geschlecht.

	Geschlecht	N	$\overline{X}$	SD	SE	Mittlere Differenz	SE Differenz	t-Wert	df	p
Theta Angst-CAT	**weiblich**	81	,326	,735	,082	-0,019	0,177	-0,106	100	0,916
	männlich	21	,345	,674	,147					

Ebenfalls keine signifikanten Ergebnisse resultierten aus einfaktoriellen Varianzanalysen zur Überprüfung der Mittelwertsunterschiede zwischen verschiedenen *Alters*gruppen ($QS_{between}$= 1,31; df = 5; $\overline{Q}$ = 0,26; F = 0,49;

p = 0,78), obgleich Patienten der Altersgruppe der 26-35-Jährigen und der über 75-Jährigen leicht geringere Theta-Werte im Angst-CAT aufweisen als Patienten sonstiger Altersgruppen (siehe Tabelle 21).

Tabelle 21: Statistische Kennwerte des Angst-CATs unterschiedlicher Altersgruppen.

	Alter	N	$\overline{X}$	SD	SE
Theta Angst-CAT	**18-25 Jahre**	20	,466	,650	,145
	26-35 Jahre	16	,149	,819	,205
	36-45 Jahre	23	,388	,808	,169
	46-55 Jahre	20	,358	,819	,183
	56-65 Jahre	15	,319	,480	,124
	> 75 Jahre	8	,132	,611	,216

Auch die Überprüfung der Mittelwertsunterschiede zwischen Gruppen unterschiedlichen *Familienstandes* (einfaktorielle Varianzanalyse) führte zu keinen signifikanten Ergebnissen ($QS_{between}$= 0,33; df = 3; $\overline{Q}$ = 0,11; F = 0,21; p = 0,89; Abbildung 23).

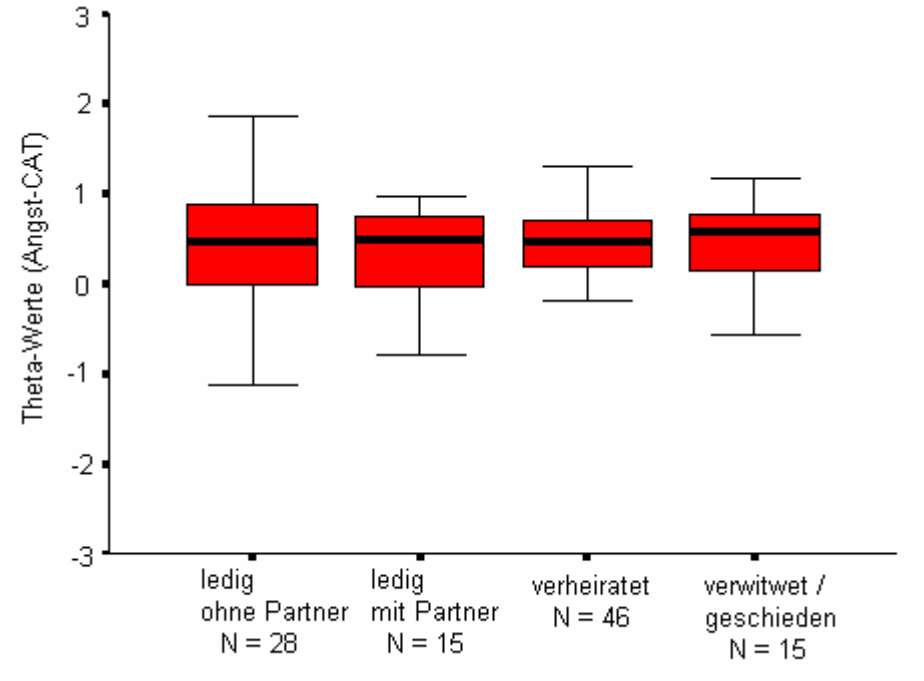


Abbildung 23: Die Theta-Werte-Verteilung des Angst-CATs in Abhängigkeit vom Familienstand.

6.7.2. Konvergente Validierung

In der bereits erwähnten Simulations-Vorstudie zur Güte des Angst-CATs (Walter et al., eingereicht) wurden die Theta-Werte des Angst-CATs psychosomatischer Patienten bereits in Beziehung zur State-Skala des STAI (Laux et al., 1981) gesetzt, um erste Hinweise auf die konvergente Validität des Instruments zu erhalten. Die Simulationsexperimente führten zu einer Korrelation zwischen dem Angst-CAT und der State-Skala des STAIs von r = 0,88. Da die Itembank des Angst-CATs jedoch 15 der 20 Items der State-Skala des STAIs umfasst, kann dieser Befund einer überlappenden Itemmenge geschuldet sein, so dass es der tiefergehenden Untersuchung der Validität - wie sie mit folgender Validierungsstudie realisiert wird - bedurfte.

6.7.2.1. Konvergente Validität in Bezug auf die Angst-Inventare

In der vorliegenden prospektiven Validierungsstudie wurde zunächst das Angst-CAT in korrelationsstatistischen Zusammenhang mit zwei Angstskalen gesetzt (HADS-A und BAI; Kapitel 6.5.1.) . Da aus organisatorischen Gründen nur die Hälfte der Patienten (N = 50) sowohl das Angst-CAT als auch die anderen psychometrischen Instrumente innerhalb von 48 Stunden beantworten konnten (bei der anderen Hälfte der Patienten liegt die Differenz zwischen den Messzeitpunkten bei bis zu 14 Tagen), werden hier Ergebnisse dieser Teilstichprobe (N = 50) und der Gesamtstichprobe (N = 102) berichtet.

Tabelle 22: Korrelationen zwischen dem Angst-CAT und den zwei Angst-Skalen.

	Zeitdifferenz zwischen den Testerhebungen	N	HADS-Angst	BAI
Theta Angst-CAT	< 14 Tage	102	,66*	,51*
	48 h	davon: 50	,76*	,55*

Die Korrelationen (r = 0,51-0,76) deuten - verglichen mit der Interkorrelation der eingesetzten etablierten Instrumente ($r_{HADS\text{-}A / BAI}$ = 0,66; N = 102) oder der bekannten Interkorrelationen dieser Angstinventare zu anderen Angstskalen (Kapitel 6.5.1.1./2.: $r_{BAI / STAI (S/T)}$[98]= 0,45/0,48; $r_{HADS\text{-}A / Angstskalen}$[99] = 0,48-0,86;

[98] In der Validierungsstudie des BAIs (Margraf & Ehlers, in Druck) errechnete Korrelationen.
[99] Gemittelte Korrelation der folgenden in der Validierungsstudie des HADS (Hermann et al., 1995) eingesetzten acht Angstskalen: Angstskala des General Health Questionnaire (GHQ-28); Linear-Analog-Angstskala; Irritability, Depression and Anxiety Scale (IDA); Zung Angst- und

$\bar{r}$ = 0,66) - auf eine *mittelmäßige* bis *gute* konvergente Validität des Angst-CATs hin. Die höheren Korrelationen bei einer zeitnäheren Erhebung ist vor dem Hintergrund der Intention der Messung einer zeitlich variablen *Zustands-Angst* zu erwarten. Insgesamt liegen die Korrelationen des Angst-CATs zur Angstskala des HADS höher als die zum BAI. Für den Unterschied in der Korrelationshöhe sind die Iteminhalte der verschiedenen Angstskalen verantwortlich. Während die Angstskala des HADS inhaltlich eine hohe Itemtextähnlichkeit zu Items des Angst-CATs aufweist (erfragt werden Gefühle der An- und Entspannung, Rastlosigkeit, beunruhigende Gedanken, Zukunftssorgen und Panikzustände), erfragen 13 (von 21) der Items des BAIs somatische Korrelate der Angst (z. B. Taubheits-, Hitze-, Schwindel, Erstickungs- und Schwächegefühle), welche im Rahmen der Konstruktion des Angst-CATs größtenteils aufgrund von Verletzungen der Unidimensionalitätsannahme im Rahmen der statistischen Itemanalyse und -selektion aus der Itembank ausgeschlossen wurden (siehe Kapitel 5.4.1.). Insofern unterscheidet sich die im Angst-CAT realisierte Konzeptualisierung der Angst stärker von derjenigen des BAIs als von derjenigen des HADS.

6.7.2.2. Konvergente Validität in Bezug auf das diagnostische Fremdurteil

Neben psychometrischen Instrumenten wurde ein strukturiertes diagnostisches Interview (M-CIDI; Wittchen & Pfister, 1996; siehe Kapitel 6.5.3.) zur Diagnostik psychischer Störungen an der psychosomatischen Stichprobe eingesetzt. Als Variablen für eine diagnostische Überprüfung der konvergenten Validität werden die im M-CIDI ermittelte Diagnose einer Angststörung (F.40-41.9 nach ICD-10, Dilling et al., 2000; siehe Kapitel 2.6.1.) bzw. das Fehlen der Diagnose einer psychischen Störung (keine F-Kodierung im ICD-10) herangezogen. Abbildung 24 veranschaulicht die Mittelwerte der Theta-Werte der Patientenstichproben ($N_{Angstdiagnose}$ = 58, $N_{keine\ F\text{-}Diagnose}$ = 10), sowie einer nicht diagnostizierten Vergleichsstichprobe von Medizinstudenten (N = 35).[100]

Depressionsskala; Crown-Crisp Experiential Index; Arthritis Impact Measurement Scale (AIMS); de Bonis-Angstskala; State Trait Anxiety Inventroy (State).

[100] Eine Normierung des Angst-CATs an gesunden Probanden ist nicht Gegenstand dieser Arbeit, wird jedoch in naher Zukunft erfolgen.

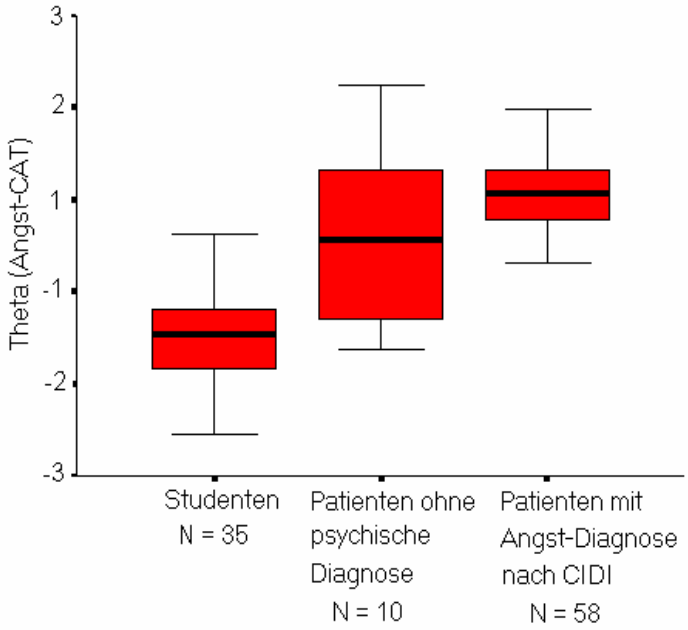


Abbildung 24: Die Theta-Werte-Verteilung des Angst-CATs verschiedener Vergleichsgruppen.

Tabelle 23 berichtet die statistischen Kennwerte des in Abbildung 24 dargestellten Befundes. Eine einfaktorielle Varianzanalyse zur zufallskritischen Absicherung der in der Tabelle dargestellten Mittelwertsunterschiede zeigt, dass sich Patienten mit einer Angststörung von Patienten ohne eine psychische Störung sowie von Studenten in den Theta-Gruppenmittelwerten statistisch bedeutsam unterscheiden (QS = 41,53; df = 2; $\overline{QS}$ = 20,763; F = 35,58, $p \leq 0{,}001$).

Tabelle 23: Statistische Kennwerte verschiedener Vergleichsgruppen.

	Gruppe	N	$\overline{X}$	SD	SE
Theta Angst-CAT	**Patienten mit Angst-Diagnose**	58	,445	,715	,094
	Patienten ohne F-Diagnose	10	,043	,939	,297
	Studenten	35	-,932	,791	,134

6.7.3. Diskriminante Validierung

6.7.3.1. Diskriminante Validität in Bezug auf andere Testverfahren

Zur Exploration der diskriminanten Validität wurden zwei Depressionsinventare (HADS-Depressionsskala; BDI) und zwei Persönlichkeitsinventare (NEO-FFI, GT) an der psychosomatischen Stichprobe (siehe Kapitel 6.4. und 6.5.) angewandt. Diese werden im Folgenden in korrelationsstatistischen Zusammenhang mit dem Angst-CAT gesetzt.

6.7.3.1.1. Angst und Depression

Tabelle 24 veranschaulicht die korrelativen Beziehungen zwischen den eingesetzten Angst- und Depressionsinventaren.

Tabelle 24: Korrelationsgrid: Angst- und Depressionsinventare (N = 102 psychosomatische Patienten).

		Angst			Depression	
		Angst-CAT	**HADS-A**	**BAI**	**HADS-D**	**BDI**
Angst	**Angst-CAT**	1,000	,663*	,514*	,598*	,593*
	HADS–Angst		1,000	,658*	,608*	,619*
	BAI			1,000	,470*	,563*
Depression	**HADS-Depression**				1,000	,711*
	BDI					1,000

Signifikante Korrelationen: *: $p \leq 0{,}05$.

In Einklang mit den theoretischen Ausführungen zur Diskrimination von Angst und Depression (Kapitel 2.5.) und den empirischen Ergebnisse aus anderen Validierungsstudien ($r_{HADS\text{-}A / HADS\text{-}D} = 0{,}53$; $r_{BAI / BDI} = 0{,}47$; $r_{BAI / CCL\text{-}D} = 0{,}43$; siehe Kapitel 6.5.1.) zeigt Tabelle 24, dass eine psychometrische Diskrimination zwischen den Konstrukten „Angst“ und „Depression“ nicht gelingt. Während die Korrelation zwischen dem Angst-CAT und der Angst-Skala des HADS (HADS-A) die Korrelationen zu den Depressions-Skalen (HADS-D, BDI) übersteigt, gilt dies nicht für die Korrelation zwischen dem Angst-CAT und dem BAI. Dieser Befund kann - wie im vorangegangenen Kapitel 6.7.2.1. bereits vermutet - wahrscheinlich durch die unterschiedliche Konzeptualisierung der Konstrukte dieser beiden Skalen (Angst-CAT / BAI) erklärt werden.

Vergleicht man die Korrelationsspannweite (range) des Angst-CATs zu den Depressionsinventaren mit derjenigen der zwei anderen Angst-Skalen

(HADS-A, BAI), so zeigt sich, dass die Korrelationsspannweite des Angst-CATs zu den Depressionsinventaren ($r_{A\text{-}CAT / HADS\text{-}D} = 0{,}59$; $r_{A\text{-}CAT / BDI} = 0{,}60$) im Mittelfeld zwischen derjenigen des BAIs ($r_{BAI / HADS\text{-}D} = 0{,}47$; $r_{BAI / BDI} = 0{,}56$) und derjenigen der HADS-Angstskala zu den Depressionsinventaren ($r_{HADS\text{-}A / HADS\text{-}D} = 0{,}61$; $r_{HADS\text{-}A / BDI} = 0{,}62$) liegt, und damit mit diesen Instrumenten vergleichbar ist.

6.7.3.1.2. Angst und Persönlichkeitskonstrukte

Nach der Erörterung der Korrelationen zwischen den Angst- und Depressionsinventaren, werden nun die korrelativen Beziehungen des Angst-CATs zu den Skalen von zwei Persönlichkeitsinventaren: dem NEO-Fünf-Faktoren-Inventar (NEO-FFI) und dem Gießen-Test (GT) beschrieben (siehe Kapitel 6.5.2.). Diese sind in Tabelle 25 abgebildet. Zur besseren Einordnung des psychometrischen „Standorts“ des Angst-CATs im Gesamt der psychometrischen Instrumente sind die Korrelationen der beiden Angstinventare (BAI, HADS-A) mit in der Tabelle aufgeführt.

Insgesamt sind die korrelativen Beziehungen der Angstinventare (Angst-CAT, BAI, HADS-A) zum NEO-FFI etwas stärker ausgeprägt als die zum GT.

Betrachtet man zunächst die Korrelationen der Angstinventare zum NEO-FFI, so zeigt sich, dass alle drei Angstinventare (Angst-CAT, BAI und HADS-A) insbesondere das Angst-CAT hoch mit der Skala „Neurotizismus“ korrelieren ($r = 0{,}51$ bis $r = 0{,}63$). Dass mit dem Angst-CAT keine bessere Differenzierung zwischen Angst und Neurotizismus gelingt als mit herkömmlichen Instrumenten, ist nicht erstaunlich, da das Angst-CAT (bislang) ausschließlich aus Items etablierter Fragebogen besteht.

Die Berechnung einer einfachen linearen Regression mit dieser Skala führt zu folgender Regressionsgleichung:

Angst-CAT = 0,637 * Neurotizismus-Skala (NEO-FFI)-1,147;

$QS_{Regression} = 20{,}53$; $QS_{Residuen} = 31{,}83$; $R^2 = 0{,}39$; $F = 64{,}48$; $p \leq 0{,}001$.

Diese verdeutlicht, dass die Skala „Neurotizismus“, welche von Costa und McCrae (1985) als stabile Eigenschaft („Trait“) konzipiert wurde, knapp 40% der Varianz der Theta-Werte des Angst-CATs aufzuklären vermag.

Tabelle 25: Korrelationsgrid: Angst- und Persönlichkeitsinventare (N = 102 psychosomatische Patienten).

		A-CAT	BAI	HADS-A	NEO-FFI					GT[101]					
					Neu	Ex	Off	Ver	Gew	SoRe	Do	Ko	Stim	Dur	SoPo
A-CAT		1,000	,514*	,663*	,626*	-,304*	-,086	-,130	-,322*	-,206*	-,053	-,118	,122	,000	-,107
BAI			1,000	,658*	,506*	-,218*	-,174	-,159	-,202*	-,096	-,066	,021	,074	,004	,017
HADS-A				1,000	,591*	-,288*	-,066	-,118	-,226*	-,185*	-,053	-,057	,139	,059	,006
NEO-FFI	**Neu**				1,000	-,547*	-,138	-,263*	-,546*	-,335*	-,075	-,180	,260*	,159	,072
	Ex					1,000	,276*	,165	,408*	,341*	-,174	-,017	-,099	-,424*	-,410*
	Off						1,000	,129	,122	-,030	-,229*	,058	,181	-,017	-,163
	Ver							1,000	,208*	-,020	,279*	,084	,027	-,033	-,042
	Gew								1,000	,371*	,067	,319*	-,184	-,292*	-,251*
GT	**SoRe**									1,000	,152	,120	-,625*	-,662*	-,633*
	Do										1,000	,132	-,351*	-,099	,076
	Ko											1,000	,110	,026	,006
	Stim												1,000	,448*	,393*
	Dur													1,000	,661*
	Sopo														1,000

Farbmarkierung: Korrelationshöhe: hellgrau: r > 0,4; mittelgrau: r > 0,5; dunkelgrau: r > 0,6.;
Signifikante Korrelationen: *: $p \leq 0,05$;
Abkürzungen: NEO-FFI: Neu: Neurotizismus; Ex: Extraversion; Off: Offenheit für Erfahrungen; Ver: Verträglichkeit; Gew: Gewissenhaftigkeit; GT: SoRe: Soziale Resonanz; Do: Dominanz; Ko: Kontrolle; Stim: Grundstimmung; Dur: Durchlässigkeit; SoPo: Soziale Potenz.

Der Messbereich dieser Skalen steht konzeptuell insofern in einem engen Zusammenhang, als bei der Erfassung der emotionalen Stabilität einer Person (Neurotizismus) Items genutzt werden, welche das Erleben negativer Gefühlszustände erfragen. Als negative Gefühlszustände werden Erschütterung, Betroffenheit, Beschämung, Traurigkeit, Sorgen, Unsicherheit, Nervosität und Ängstlichkeit erfragt (Borkenau & Ostendorf, 1993, S. 27), d. h. Begrifflichkeiten verwendet, die zum Teil auch bei der Erfassung der Zustands-Angst eine Rolle spielen. State- und Trait-Angst werden im STAI zwar als zwei Dimensionen konzipiert, die Interkorrelation dieser Skalen liegt jedoch mit r = 0,43 bis r = 0,75 (bei unterschiedlichen Stichproben) recht hoch (Laux et al., 1981). Dieser Befund und schon im Testmanual des STAIs berichtete Ergebnisse (Laux et al., 1981) deuten darauf hin, dass die Zustands- und Eigenschafts-Angst nicht (statistisch) unabhängig voneinander sind (siehe Kapitel 2.4.1. und 2.7.3.2./3.).

[101] Hohe Werte auf den Skalen 1.-4. indizieren eine hohe Ausprägung von 1. Sozialer Resonanz, 2. Dominanz, 3. Kontrolle und 4. (positiver) Grundstimmung, ein hoher Wert auf der Skala 5. indiziert emotionale Verschlossenheit, ein hoher Wert auf der Skala 6. indiziert geringe soziale Kompetenz.

Die geringste Korrelation eines Angstinventars mit der Neurotizismus-Skala findet sich beim BAI. Dies gründet sich wahrscheinlich in der Fokussierung dieser Skala auf der Erfassung somatischer Angstkorrelate (siehe Kapitel 6.5.1.1., 6.7.2.1. und 6.7.3.1.). Die hohe Korrelation zum Angst-CAT und zur HADS deutet auf eine größere Konstruktnähe dieser Skalen hin.
Der in den Korrelationen des Angst-CATs zu den anderen vier Skalen des NEO-FFIs aufgezeigte Zusammenhang lässt sich regressionsstatistisch durch die folgende multiple lineare Regression näher explorieren.

Angst-CAT = 0,686 * NEU + 0,071 * EX – 0,023 * OFF + 0,070 * VER + 0,023 * GEW-1,578;
$QS_{Regression}$ = 20,73; $QS_{Residuen}$ = 31,64; R^2 = 0,40; F = 12,58; $p \leq 0{,}001$.

Obgleich sich signifikante Korrelationen der Angstinventare zu den Skalen Extraversion und Gewissenhaftigkeit zeigen, offenbart oben dargestellte Regressionsgleichung, dass das Hinzufügen der anderen Skalen des NEO-FFI als zusätzliche Prädiktorvariablen zu keiner deutlichen Verbesserung der Varianzaufklärung im Vergleich zur einfachen linearen Regression führt.
Die Korrelationen sind somit wahrscheinlich vor dem Hintergrund relativ hoher Interkorrelationen (im NEO-FFI) zwischen den Skalen „Neurotizismus“ und „Extraversion“ bzw. „Gewissenhaftigkeit“ zu sehen (r = 0,55). Die negative Beziehung des Konstrukts der Ängstlichkeit zu dem der Extraversion ist konstrukttheoretisch durch den Zusammenhang zwischen *sozialer* Ängstlichkeit und der Neigung zu Introversion (Borkenau & Ostendorf, 1993, S. 28) erklärbar. Die negative Beziehung des Konstrukts der Ängstlichkeit zu dem der Gewissenhaftigkeit ist vor dem Hintergrund verständlich, dass „die Unfähigkeit, Impulsen oder Versuchungen zu widerstehen, im NEO-Modell als ein Indikator für Neurotizismus gewertet wird“ (Borkenau & Ostendorf, 1993, S. 28).
Die geringen Korrelationen der Angst-Skalen, eingeschlossen des Angst-CATs, zu den Skalen „Offenheit für Erfahrungen“ und „Verträglichkeit“ sprechen für die diskriminante Validität dieser Skalen.
Die Betrachtung der Korrelationen zum NEO-FFI abschließend, sei angemerkt, dass die hier eruierten Interkorrelationen der fünf Skalen des NEO-FFIs deutlich höher ausfallen als die im Testhandbuch referierten Interkorrelationswerte einer *gesunden* Gesamtstichprobe (r =-0,33 und 0,16; N = 2.112; Borkenau &

Ostendorf, 1993, S. 15). Es bleibt offen, ob dieser Befund – falls generalisierbar – eine Besonderheit psychosomatischer Stichproben sein könnte.

Die Korrelationen der Angstinventare zu *fünf* der sechs Skalen des GT, weisen durch geringe Werte (r = -0,12 bis r = 0,18) auf eine hohe diskriminante Validität zwischen den durch die Skalen erfassten Konstrukten hin. Jedoch bestehen signifikante Korrelationen zwischen der Skala „Soziale Resonanz“ und dem Angst-CAT (r = -0,21) bzw. dem BAI (r = -0,19). Konstrukttheoretisch reflektiert mag dies daraus resultieren, dass diese Skala unter anderem auch eine „narzisstische Gratifikation“, d. h. positive, soziale, selbstwertstärkende Erlebnisse erfasst (Beckmann et al., 1991, S. 39). Vor dem Hintergrund des bereits oben erörterten negativen Zusammenhangs zwischen Extraversion und Ängstlichkeit erscheint eine geringe negative Korrelation plausibel, wenn man zusätzlich annimmt, dass Extrovertierte (Personen mit einer geringeren *sozialen* Ängstlichkeit) mehr soziale Gratifikation erfahren.

Erwähnenswert ist ferner, dass das Angst-CAT und die anderen Angstinventare nur niedrige, nicht signifikante Korrelationen zur Skala allgemeine „Grundstimmung“ des GT aufweisen, so dass es scheint - obgleich eine Diskrimination zwischen den Konstrukten „Angst“ und „Depression“ schwierig ist -, dass sich Angst als ein spezifisches emotionales Konstrukt von einer allgemeinen Grundstimmung psychometrisch diskriminieren lässt.

Betrachtet man die Interkorrelationen der Skalen des GT, so fällt auf, dass vier der sechs Skalen (Soziale Resonanz, Durchlässigkeit, Stimmung und Soziale Potenz) hoch miteinander korrelieren (r > 0,6). Auch die Testautoren (Beckmann et al., 1991) fanden zwischen diesen Skalen Interkorrelationen (bei gesunden und „neurotischen“ Patienten), allerdings in einer deutlich geringeren Ausprägung (r = -0,24 bis r = -0,56). Wahrscheinlich fordert hier die Schwerpunktlegung bei der Testkonstruktion des GT (Beckmann et al., 1991) auf das Erfassen psychosozialer Aspekte durch eine schlechte Diskriminationsleistung zwischen den einzelnen Konstrukten ihren Tribut.

Fokussiert man abschließend die Korrelationen der Skalen beider Persönlichkeitsinventare (NEO-FFI und GT) untereinander, fällt vor allem ein mittelstark ausgeprägter Zusammenhang (r = -0,42 bis r = 0,34) zwischen der Skala „Extraversion“ des NEO-FFIs und den „psychosozialen“ Skalen

(„Soziale Resonanz“, „Durchlässigkeit“[102] und „Soziale Potenz“[103]) des GT ins Auge. Dieser erscheint durch eine relative Konstruktnähe verständlich.

6.7.3.2. Diskriminante Validität in Bezug auf das diagnostische Fremdurteil

Wie bereits zur Überprüfung der konvergenten Validität (siehe Kapitel 6.7.2.2.), werden auch zur Überprüfung der diskriminanten Validität die durch den Einsatz des strukturierten diagnostischen Interviews M-CIDI (siehe Kapitel 6.5.3.) an der psychosomatischen Stichprobe erhobenen Diagnosen psychischer Störungen gemäß ICD-10 genutzt. Tabelle 26 gibt einen ersten groben Überblick über die statistischen Kennwerte der Theta-Werte des Angst-CATs der gesamten psychosomatischen Stichprobe (N = 102), welche sich in Patienten mit bzw. ohne Diagnose einer psychischen Störung nach ICD-10 untergliedern lässt. Desweiteren werden als Vergleichsstichprobe die statistischen Kennwerte von einer Gruppe von Medizinstudenten (N = 35) aufgeführt.

Tabelle 26: Statistische Kennwerte der Theta-Werte des Angst-CATs verschiedener Vergleichsgruppen.

Gruppe	N[104]	$\overline{X}$	SD	SE
Patienten mit der Diagnose einer psychischen Störung (F)	92	,361	,692	,072
Patienten ohne die Diagnose einer psychischen Störung (kein F)	10	,043	,939	,297
Studenten	35	-,932	,791	,134

Eine einfaktorielle Varianzanalyse zeigt, dass das Angst-CAT gut zwischen Patienten mit der Diagnose einer psychischen Störung und ohne eine solche Diagnose (nach ICD-10) bzw. gesunden Personen (Medizinstudenten) zu differenzieren vermag (QS = 42,43; df = 2; $\overline{QS}$ = 21,22; F = 39,08; $p \leq 0{,}001$).

Obgleich das Angst-CAT nicht zur diagnosenspezifischen Differenzierung entwickelt wurde, wurde zusätzlich die Diskriminationsfähigkeit des Angst-CATs bezüglich verschiedener Patientengruppen mit unterschiedlichen Diagnosen einer psychischen Störung untersucht. Hierzu wurden die Patienten der psychosomatischen Gesamtstichprobe (N = 102) mittels der klassifikatorischen

[102] Hohe Werte auf der Skala „Durchlässigkeit“ indizieren emotionale Verschlossenheit.
[103] Hohe Werte auf der Skala „Soziale Potenz“ indizieren eine geringe soziale Kompetenz.
[104] Die Diagnosegruppengrößen summieren sich nicht zur Gesamtstichprobengröße (N = 102), da eine hohe Komorbidität zwischen den Störungen vorliegt.

Diagnostik des M-CIDI verschiedenen diagnosenspezifischen Subgruppen zugeordnet (siehe Tabelle 27).

Tabelle 27: Statistische Kennwerte der Theta-Werte des Angst-CATs verschiedener Diagnosegruppen (mit Komorbidität).

Gruppe	N[105]	$\overline{X}$	SD	SE
Patienten mit einer Angststörung (F.40-41.9)	58	,445	,715	,094
Patienten mit einer depressiven Störung (F.32-34)	60	,491	,638	,082
Patienten mit einer Essstörung (F.50)	7	,407	,661	,250
Patienten mit einer somatoformen Störung (F.45)	51	,369	,644	,090

Tabelle 27 zeigt die statistischen Kennwerte der Theta-Werte des Angst-CATs der verschiedenen diagnostischen Subgruppen, wie sie im psychosomatischen Kollektiv geschätzt wurden. Aufgrund einer hohen Komorbidität zwischen den einzelnen Störungsgruppen (somatoforme Störungen & Angststörungen: 70,6%;[106] Essstörungen & Angststörungen: 42,86%;[107] depressive Störungen & Angststörungen: 83,33%)[108] ist die auf den ersten Blick geringe Diskriminationsfähigkeit des Angst-CATs zwischen den einzelnen Diagnosegruppen nicht erstaunlich.

Um zu überprüfen, wie stark die vorliegende Komorbidität des Patientenkollektivs die Diskriminationsfähigkeit des Angst-CATs beeinträchtigt, wurden Patientengruppen ohne Komorbidität gebildet. Die in Abbildung 25 veranschaulichten Patientengruppen bestehen demnach aus Patienten, welche Störungen aus jeweils nur einer Diagnosegruppe aufweisen, d. h. Patienten mit einer Angststörung *und* einer weiteren Diagnose einer psychischen Störung (= Komorbidität) wurden aus den Patientenkollektiven ausgeschlossen.

Abbildung 25 deutet darauf hin, dass das Angst-CAT bei Patienten ohne Komorbidität besser zwischen verschiedenen Störungen zu differenzieren vermag als bei Vorliegen von Komorbidität. Tabelle 28 gibt Aufschluss über die statistischen Kennwerte der Theta-Werte des Angst-CATs dieser verschiedenen diagnostischen Subgruppen nach Ausschluss von Komorbidität.

[105] Die Diagnosegruppengrößen summieren sich nicht zu der Gesamtstichprobengröße (N = 102), da eine hohe Komorbidität zwischen den Störungen vorliegt.
[106] N = 36 von 51 Patienten mit einer somatoformen Störung haben auch eine Angststörung.
[107] N = 3 von 7 Patienten mit einer Essstörung haben auch eine Angststörung.
[108] N = 50 von 60 Patienten mit einer depressiven Störung haben auch eine Angststörung.

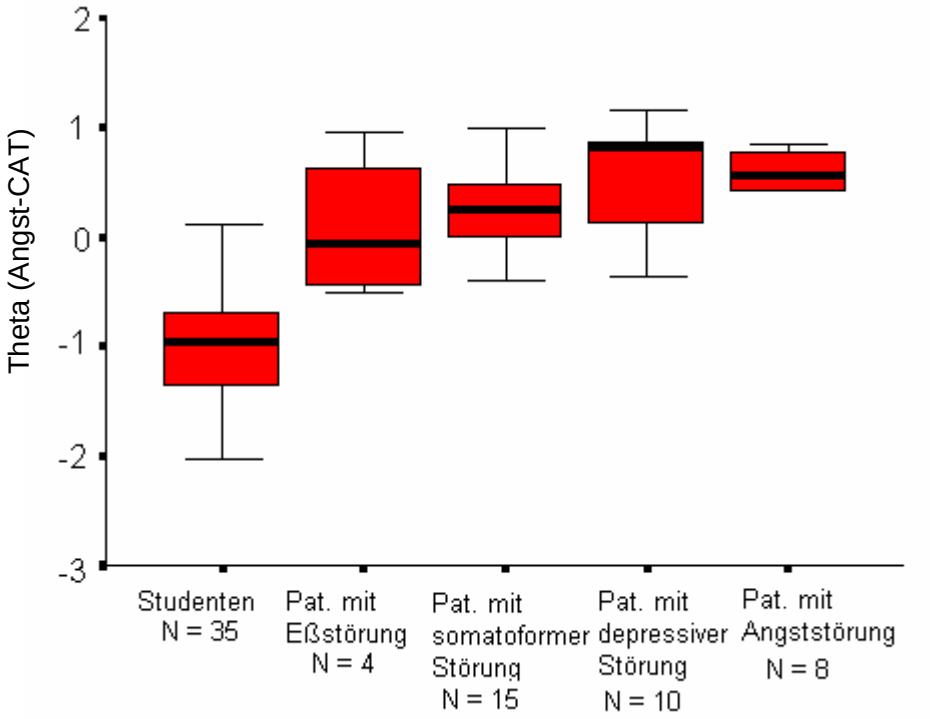


Abbildung 25: Die Theta-Werte-Verteilung des Angst-CATs im Vergleich verschiedener Diagnosegruppen ohne Komorbidität.

Vergleicht man Tabelle 27 (Patientenkollektiv mit Komorbidität) und Tabelle 28 (Patientenkollektiv ohne Komorbidität) so zeigt sich, dass – bei Ausschluss von Komorbidität – die Unterschiede in den Mittelwerten des Angst-CATs verschiedener Patientengruppen deutlicher zu Tage treten. Der zu diskutierende Befund, dass depressive Patienten im Falle von Komorbidität im Durchschnitt etwas höhere Theta-Werte erzielen als ängstliche Patienten (Tabelle 27), wird nach Ausschluss von Komorbidität *nicht* repliziert.

Tabelle 28: Statistische Kennwerte der Theta-Werte des Angst-CATs verschiedener Diagnosegruppen (ohne Komorbidität).

Diagnosegruppe	N [109]	$\overline{X}$	SD	SE	95% Intervall für den Mittelwert: Ober- Grenze	Unter- Grenze
Pat. nur mit einer Essstörung (F.50)	4	,091	,675	,337	-,983	1,164
Pat. nur mit einer somatoformen Störung (F.45)	15	,156	,620	,160	-,187	,500
Pat. nur mit einer depressiven Störung (F.32-34)	10	,552	,514	,171	,157	,947
Pat. nur mit einer Angststörung (F.40-41.9)	8	,553	,762	,270	-,085	1,190

Pat.= Psychosomatische stationäre Patienten.

[109] Die Diagnosegruppengrößen summieren sich nicht zur Gesamtstichprobengröße (N = 102), da eine hohe Komorbidität zwischen den Störungen besteht.

Eine über die Gruppen durchgeführte einfaktorielle Varianzanalyse zur globalen Bewertung der Unterschiede in den Theta-Gruppenmittelwerten des Angst-CATs ergibt, dass sich die Gruppen insgesamt auf einem Signifikanzniveau von $p \leq 0{,}001$ (QS = 30,07; df = 4; $\overline{QS}$ = 7,52; F = 14,50) unterscheiden. Ein anschließend durchgeführter Scheffé-Test zur genaueren Untersuchung der Mittelwertsunterschiede zwischen den einzelnen Gruppen zeigt, dass sich die Patienten mit einer somatoformen bzw. depressiven bzw. Angststörung auf einem Signifikanzniveau von $p \leq 0{,}001$ signifikant von gesunden Personen bzw. der Gruppe der Patienten mit Essstörungen unterscheiden.

Die Unterschiede der Theta-Mittelwerte des Angst-CATs zwischen den drei oben erläuterten Diagnosegruppen (somatoforme, depressive bzw. Angststörung) sind – obgleich sie bei Ausschluss von Komorbidität insgesamt größer sind (siehe Tabelle 27 / 28 im Vergleich) – nicht signifikant. Es bleibt zu diskutieren, ob dies aus den geringen Stichproben*größen* resultiert. Zusammenfassend lässt sich resümieren, dass – obgleich das Angst-CAT nicht dafür konstruiert wurde, verschiedene diagnostische Gruppen voneinander zu trennen – die Ergebnisse in Tabelle 28 und Abbildung 25 – insbesondere die klare Trennung der gesunden Personen von dem psychosomatischen Kollektiv – als Hinweis auf eine gute diskriminante Validität interpretiert werden dürfen.

6.7.4. Zusammenfassung der Validierungsergebnisse

Das Angst-CAT erweist sich in vorliegender Validierungsstudie als ein valides, psychometrisches Verfahren zur Erfassung der Zustands-Angst in einem psychosomatischen Patientenkollektiv.

Eine *mittelmäßige* bis *gute konvergente Validität* des Angst-CATs konnte in Form von mittelmäßig bis hohen Korrelationen zu anderen Angstinventaren (BAI, HADS-A; r = 0,51 bis r = 0,76) belegt werden. Die Höhe der Korrelationen steht im Einklang mit konvergenten Validierungsergebnissen bereits etablierter Angstinventare (r = 0,45 bis r = 0,86). Eine konvergente (diagnosenspezifische) Validität ist insofern gegeben, als Patienten mit der Diagnose einer Angststörung signifikant höhere Theta-Werte im Angst-CAT aufweisen als Patienten ohne die Diagnose einer psychischen Störung bzw. gesunde Personen ($p \leq 0{,}001$).

Die *diskriminante Validität* des Angst-CATs unterscheidet sich im Hinblick auf die untersuchten Konstrukte. Die psychometrische Diskrimination von *Angst*- und *Depression* gestaltet sich – wie theoretisch und empirisch in der Literatur (r = 0,43 bis r = 0,62) bereits vielfach diskutiert – auch im Angst-CAT (BDI, HADS-D; r = 0,59 bis r = 0,60) *schwierig*.

Dagegen kann aufgrund geringer Korrelationen des Angst-CATs zu Skalen von zwei Persönlichkeitsinventaren (NEO-FFI, GT) auf eine *gute* diskriminante Validität bezüglich *anderer Eigenschaftskonstrukte* geschlossen werden. Das Konstrukt der Angst lässt sich von einem allgemeinen Konstrukt der Grundstimmung (r = 0,12) und allen weiteren Skalen der Persönlichkeits-inventare (r = -0,21 bis r = 0,12) gut differenzieren.

Einzige Ausnahme ist die Diskrimination zum Konstrukt „*Neurotizismus*“, welche – angezeigt durch eine Korrelation von r = 0,63 – nicht gelingt, und eine Korrelation des Angst-CATs mit den Skalen „Extraversion“ und „Gewissenhaftigkeit“ (r = 0,3) mitbedingt. Dieser Befund steht in Einklang mit Forschungsbefunden aus der Literatur und ist in die Forschungsdiskussion um eine mögliche Differenzierbarkeit zwischen einer Eigenschafts- und einer Zustands-Angst einzuordnen (zur Diskussion siehe Kapitel 7.5.1.).

Obgleich das Angst-CAT nicht zur diagnosenspezifischen Diskrimination entwickelt wurde, legen die berichteten Ergebnisse nahe, dass eine diskriminante Validität bezüglich verschiedener Diagnosegruppen bedingt gegeben ist. Eine Differenzierung zwischen verschiedenen Diagnosegruppen ist tendenziell möglich, jedoch nur bei Patienten, welche keine Komorbidität aufweisen. Insofern sollte das Angst-CAT stets im Zusammenhang weiterer klinischer Diagnostik interpretiert werden.

7. Diskussion

7.1. Einleitung

Die vorliegende Forschungsarbeit zur Entwicklung und Validierung eines auf der Grundlage der Item Response Theorie (IRT) konstruierten Computergestützten Adaptiven Tests zur Angstmessung (Angst-CAT) stellt im deutschen Sprachraum eine *klinisch-psychologische Pionierarbeit* dar. Während im internationalen Sprachraum meines Wissens bislang nur zwei IRT-basierte CAT-Versionen etablierter Instruments (NEO-PIR; Reise & Henson, 2000; Simms & Clark, in Vorbereitung) im Bereich der Persönlichkeits-diagnostik existieren, werden IRT-basierte CATs im klinischen Bereich derzeit vor allem von zwei Forschergruppen, von denen sich eine mit der Messung von Lebensqualität befassen (Ware et al., 2000, 2003) und eine die mehrdimensionale Erfassung pädiatrischer Symptome fokussiert (Gardner et al., 2002), entwickelt und erprobt. Weitere IRT-basierte Anwendungen konzentrieren sich in der Persönlichkeitsdiagnostik vor allem auf die IRT-basierte (Re-) Analyse und Evaluation bereits etablierter Instrumente (siehe Kapitel 3.5.2.).

Im Vergleich zu der weiten Verbreitung von IRT- und / oder CAT-Anwendungen im Bereich der *Leistungs*diagnostik, welche sowohl im deutschsprachigen (Hornke, 1993, 1994, 1996; 1999; Hornke et al., 2000; Kubinger & Wurst, 1986; 1993; 2000; Rost, 1999; Rost & Carstensen, 2002) als auch im internationalen Sprachraum stark vorangeschritten ist (z. B. Graduate Record Examination, GRE des Educational Testing Service oder Computerized Placement Test des College Boards, siehe Kapitel 3.5.1.), findet sich im Bereich der *Persönlichkeits*diagnostik ein deutliches Forschungsdefizit bezüglich der Entwicklung IRT-basierter CATs.

Da die Persönlichkeitspsychologie auf eine lange Tradition in der Testentwicklung *umfangreicher* Inventare zurückblickt und zur Entwicklung von IRT-basierten CATs *große* Itemmengen und Personenstichproben nötig sind, liegt angesichts umfangreicher bereits erhobener Persönlichkeitsdatenmengen, jedoch gerade in diesem Bereich ein besonderes Potential (Embretson & Hershberger, 1997).

Dieses Potential und das zunehmende Wissen um die vielfältigen Vorteile der IRT, die einige im Rahmen der Klassischen Test-Theorie (KTT) aufgeworfenen

messtheoretischen Probleme zu lösen verspricht, sowie erweiterte Möglichkeiten der statistischen Analyse von Antwortkategorien, Items und Skalen bietet (z. B. IRC-Analyse, Untersuchung von Itemparametern, Item- und Testinformationen, Differential-Item-Functioning (DIF), Personen- und Modell-Fit, Entwicklung von instrumentenübergreifenden Metriken durch Equating- oder Linking-Prozeduren; siehe Kapitel 3.3.3.), evozierte innerhalb der letzten Jahrzehnte eine stetige Zunahme der Nutzung der IRT bei der Erforschung von Persönlichkeitsinventaren (Orlando & Marshall, 2002; Cooke et al., 2001; Ferrando, 2001; Chernyshenko et al., 2001; Childs et al., 2000; Santor & Coyne, 2000; Orlando et al., 2000; Reise & Henson, 2000; Rouse et al., 1999).

Obgleich diese rege Forschungsaktivität von dem Potential der IRT bezüglich der methodischen Weiterentwicklung von Persönlichkeitsinstrumenten zeugt, konnte sich die Anwendung dieser Methoden in der *klinischen Praxis* der Testentwicklung bisher nicht durchsetzen. Mögliche Gründe können in der methodischen Unsicherheit angesichts der mathematischen Komplexität der IRT-Modelle und in einem Zweifel bezüglich des allgemeinen Nutzens dieser Methodik im Bereich der Persönlichkeitsforschung liegen (siehe Kapitel 3.5.2.).

Da zu der geringen Nutzung von IRT-Methoden in der klinischen Testpraxis eine relativ geringe Verbreitung von *Computerdiagnostik* im europäischen Raum (Jäger & Krieger, 1994; Hänsgen & Bernascomi, 2000; siehe Kapitel 4.1.) – und somit auch von computergestützten Angstinventaren (siehe Kapitel 2.4.) hinzukommt -, stehen der Erforschung und Verbreitung IRT-basierter CATs (Meijer & Nering, 1999) – und somit auch des Angst-CATs – gleich mehrere Hürden entgegen. Während die zunehmende Verbreitung und Kostenreduktion von Hard- und Software den Trend zur Computerisierung begünstigt, gilt es einer allgemeinen technokratische Skepsis durch offene Kommunikation der Vor- und Nachteile von Computerdiagnostik (siehe Kapitel 4.2.2./3.) zu begegnen.

Zweifel bezüglich des Nutzens *IRT-basierter CATs* im Allgemeinen mögen sich zerstreuen, wenn man den zunehmenden Trend zur erfolgreichen Nutzung von CATs zur Leistungsdiagnostik in größeren Institutionen (BRD: Hornke, 1999; USA: ETS, 1996; siehe Kapitel 4.6.1.) und die ersten fruchtbaren Arbeiten zu IRT-basierten CAT-Entwicklungen in der klinischen Diagnostik reflektiert

(Reise & Henson, 2000; Simms & Clark, in Vorbereitung; Ware et al., 2000, 2003; Gardner et al., 2002).

Fokus vorliegender Arbeit war angesichts des großen Forschungsdefizits das Aufzeigen und Erproben eines möglichen methodischen Wegs der Entwicklung und Validierung eines IRT-basierten CATs im klinisch-psychologischen Bereich. Aufgrund einer hohen Prävalenz von *Angststörungen*, insbesondere im psychosomatischen Bereich (24-29%; Fliege et al., 2002; siehe Kapitel 2.6.2.), in dessen Rahmen diese Forschungsarbeit geschrieben wurde, verfolgt die Studie das Ziel, mit der Entwicklung eines *Angst-CATs* zu erproben, ob die praktischen, ökonomischen und testtheoretischen Vorteile, welche die IRT verspricht (siehe Kapitel 3.3.3.), tatsächlich eingelöst werden können. Von besonderem Interesse ist hier die Frage, ob mit einem IRT-basierten CAT ein kurzes Screening-Instrument konstruiert werden kann, welches die Messung von Zustands-Angst auf einem konstant hohen Messpräzisionsniveau mit einer adaptiv verringerten Anzahl von dargebotenen Items erlaubt (siehe Kapitel 4.3.3. / 4.4.). Hiermit verbindet sich die Hoffnung, die Psychodiagnostik sowohl für den Diagnostiker (durch Zeit- und Kosteneinsparungen) als auch für den Patienten (durch eine Reduktion der zeitlichen und emotionalen Beanspruchung) weniger belastend gestalten zu können.

In diesem Zusammenhang stellt sich die Frage, worin der *spezifische Vorteil* (Zugewinn) einer Itemreduktion mittels eines *CATs* liegt, da für die meisten herkömmlichen psychometrischen Instrumente KTT-basierte *Kurzversionen* bereits existieren. Der Vorteil einer IRT-basierten Itemreduktion besteht einerseits darin, dass Patienten während eines CAT-Prozesses nur diejenigen Items dargeboten bekommen, welche ihrem Merkmalsausprägungsniveau optimal entsprechen, d. h. bei Leistungstests wird z. B. eine Unter- oder Überforderung der Testperson vermieden, andererseits ermöglicht ein CAT die Gleichhaltung einer hohen Messpräzision, welche bei Kurzinstrumenten in dieser Form nicht möglich ist. Denn während Screening-Verfahren mit wenigen globalen Items ein weites Merkmalsausprägungsspektrum erfassen müssen und damit häufig psychometrische „Decken- und Bodeneffekte“ resultieren, können diese in einem CAT dadurch vermieden werden, dass nach wenigen globalen Start-Items, welche das gesamte Merkmalsausprägungsspektrum abdecken, hoch diskriminative Items zur Messung der individuellen

Merkmalsausprägung durch einen spezifischen Itemselektionsalgorithmus (siehe Kapitel 4.3.3.3.) adaptiv ausgewählt werden.

7.2. Aufbau des Diskussionsteils

Im Folgenden wird die Entwicklung und Validierung des Angst-CATs diskutiert. Zunächst erfolgt eine konzeptuelle Diskussion um den *Geltungs-* und *Gültigkeitsbereich* sowie den intendierten und realisierten Messbereich des Angst-CATs (Kapitel 7.3.). Dieser folgt eine kritische Auseinandersetzung über die im Rahmen der Testkonstruktion eingesetzten *Methoden* und *Ergebnisse der Itemanalyse und –selektion* (Kapitel 7.4.). Daran schließt sich eine Diskussion der *Ergebnisse der Validierungsstudie* an (Kapitel 7.5.), in deren Rahmen auch *zentrale* Aspekte der realisierten *computergestützten adaptiven* Diagnostik reflektiert werden. Abschließend wird ein *Resumée* gezogen und ein *Ausblick* versucht (Kapitel 7.6.).

7.3. Zum Geltungs- und Gültigkeitsbereich des Angst-CATs

Zunächst steht der Geltungs- und Gültigkeitsbereich des Angst-CATs zur Diskussion. Im Sinne eines eindimensionalen Breitbandverfahrens soll es sowohl für den Einsatz an *psychosomatischen*, als auch an *psychiatrischen* Patienten, an Patienten mit rein *somatischen* Erkrankungen und an *gesunden* Probanden geeignet sein. Kritisch einzuräumen ist hier, dass die Nutzung von Itemparametern, welche an einer psychosomatischen Stichprobe vorkalibriert wurden, zur Schätzung der Personenparameter von Personen anderer Stichproben nur dann problemlos ist, wenn eine IRT-Modellierung gelingt, und somit die Itemparameterinvarianz angenommen werden kann (siehe Kapitel 3.3.1./2.). Um die Itemparameterinvarianz der Itembank des Angst-CATs zu überprüfen, sind langfristig weitere emprische Studien an anderen Personenstichproben geplant.

Das Konstrukt der Angst wurde zu Beginn der Testentwicklung in Anlehnung an die Definition der *Zustands-Angst* von Spielberger (1972) definiert, der ähnlich wie Liebert und Morris (1967) sowohl *emotionale* (z. B. innere Unruhe) als auch *kognitive* Aspekte (z. B. Besorgtheit) der Angst beschreibt, sowie zusätzlich *vegetative* Symptome (z. B. Überregbarkeit) als kennzeichnend für die Zustands-Angst ansieht (siehe Kapitel 2.4.1.1.). Diese Aspekte entsprechen weitgehend den Kriterien, die in der ICD-10 (Dilling et al., 2000) für die Generalisierte Angststörung (F41.1; siehe Kapitel 2.6.1.) aufgeführt werden.

Die *Itembankentwicklung* erfolgte in mehreren Schritten der Itemanalyse und -selektion (siehe Kapitel 5.3.) an drei psychosomatischen Patientenstichproben (N_1 = 1.010; N_2 = 834; N_3 = 775). Sie verfolgte das Ziel, die Items zu identifizieren, welche aus psychometrischer Sicht als die „besten“ erscheinen, da sie unter anderem einen hinreichend großen Teil gemeinsamer Varianz des Angst-Konstruktes erfassen. Hierzu wurden aus einem inhaltlich vorselektierten Itempool von 81 Items sukzessiv diejenigen Items ausgeschlossen, die den gesetzten psychometrischen Qualitätskriterien nicht entsprachen, so dass sich schließlich die endgültige Itembank des Angst-CATs aus 50 Items konstituierte.
Die bereits im Rahmen der Vorselektion in einem Delphi-Entscheidungsprozess *ausgeschlossenen* Items erfragen allgemeine Leistungseinbußen, Schlafstörungen und Depression, welche konsensuell als vom Konstrukt der Angst abzugrenzende Konstrukte festgelegt wurden (siehe Kapitel 5.3.1.).
Die anschließende statistische *Itemselektion* resultierte in einem Ausschluss von 30 Items, von denen die meisten *somatische* Korrelate der Angst, manche auch *gesundheitsspezifische* Sorgen oder spezifische *soziale* Ängste erfassen. Der Ausschluss spezifischer Ängste und Sorgen ist im Sinne des Bemühens um eine möglichst *situationsübergreifende* Messung der Zustands-Angst erwünscht (siehe Kapitel 2.3.2., 2.6.1., 2.7.3.3. und 5.3.1.).
Der Befund, dass der überwiegende Teil der ausgeschlossenen Items somatische Korrelate der Angst erfragt (z. B. Herzjagen, Zittern, Schwitzen etc.), kann vor dem Hintergrund von Forschungsmodellen, welche die faktorenanalytische Differenzierung der Konstrukte der Angst und Depression fokussieren, diskutiert werden (siehe Kapitel 2.5.).
Während die Itemselektion des Angst-CATs zu einer Konzeptualisierung der Angst *weitgehend ohne vegetative* Aspekte führte, konzipierten Forscher in den 80ern bis Mitte der 90er Jahre den Angst-Faktor noch als einen, der sich *vor allem* durch Symptome somatischer Anspannung und vegetativer Übererregbarkeit auszeichnet (neben einem globalen Faktor der negativen Affektivität, der die hohe gemeinsame Varianz zwischen Angst und Depression erklären sollte; Clark & Watson, 1991; Watson & Clark, 1984; Watson et al., 1995). Erst vor einigen Jahren wurde diese Vorstellung im Einklang mit der hier erfolgten Itemselektion revidiert bzw. weiterentwickelt.

Barlow und Mitarbeiter (1996) konzipierten in einem „Drei-Faktoren-Modell“ das Konstrukt der Angst in Form einer negativen Affektivität und grenzen diese als eigenständige Grundemotion von einem *autonomen Erregungszustand*, den sie für einen *spezifischen Indikator* von *Panik* bzw. Furcht halten (und von der Depression, welche vor allem durch Anhedonie gekennzeichnet sei), entschieden ab.

Die Konzeption eines für Panikzustände spezifischen *separaten* vegetativen Indikators, der nicht im Sinne eines globalen, breiten Angstfaktors zusammen mit allen anderen Angstsymptomen zu interpretieren sei, setzte sich gestützt durch empirische Belege aus umfangreichen Strukturgleichungsanalysen (Brown et al., 1997; Chorpita et al., 1998) in einem integrativen hierarchischen Modell der Angst (und Depression) im Forschungskontext durch (Mineka et al., 1998). Auch im klinischen Kontext werden intensiv ausgeprägte vegetative Angstsymptome, welche attackenweise auftreten, als für Panikstörungen (F.41.0) charakteristisch erachtet (ICD-10; Dilling et al., 2000; DSM-IV; Saß et al., 1996; siehe Kapitel 2.5. und 2.6.1.).

Insofern entspricht die Operationalisierung der Angst – wie sie bei der Itembankentwicklung des Angst-CATs erfolgte – dem derzeitigen Stand der Forschung und klinischen Diagnostik. Von der ursprünglichen Definition der Zustands-Angst nach Spielberger (1972), die neben dem *emotionalen* auch den *kognitiven* und *vegetativen* Aspekt der Angst betonte, wird also durch den Ausschluss vegetativer Items aufgrund von Unidimensionalitätsverletzungen abgewichen.

Das entwickelte Angst-CAT intendiert somit die Erfassung einer *situations*- und *objektübergreifenden*, *generalisierten Zustands-Angst* und *nicht* die Erhebung eines *akuten Panik*zustandes mit ausgeprägter *vegetativer* Symptomatik.

Die endgültige *Itembank* besteht zu 70% aus Items (N = 35; z. B. „ängstlich“ oder „besorgt“), welche das Vorliegen von Zustands-Angst in positiver Ausprägung und 30% aus Items (N = 15), welche zur Angst konträre Zustände (i. S. eines Zustands der „Nicht-Angst“; z. B. „selbstsicher“ oder „entspannt“) erfassen.

Bei einer Sichtung der *Itemtexte* der die Itembank (N = 50) konstituierenden Items fällt auf, dass die Itemselektion dazu führte, dass Items in der Itembank verblieben, welche sowohl emotionale (i. S. einer inneren Unruhe) als auch

kognitive (i. S. einer Besorgtheit) Aspekte der Angst sowie ein für Angstphänomene im klinischen Bereich typisches Entfremdungserleben (i. S. einer Depersonalisation) erfassen (siehe Kapitel 5.4.4.). Dieser Befund steht im Einklang mit den auf der Basis von empirischen Studien von Liebert und Morris (1967; Morris et al., 1970, 1981, 1983) geäußerten Schlussfolgerungen von Forschern (Benson et al., 1992; Krohne, 1996), dass eine Differenzierung der *emotionalen* und *kognitiven* Komponente der Angst, wie sie ursprünglich von Liebert und Morris (1967) angedacht war, empirisch nicht gelingt (siehe Kapitel 2.7.3.4.).

Wie verschiedene Studien zeigen, stehen die Konstrukte der Zustands-Angst (State) und der Eigenschafts-Angst (Trait), die im *State-Trait-Modell* der Angst (Spielberger, 1972) differenziert werden, in einem engen Zusammenhang ($r_{State/Trait\text{-}Angst}$ = 0,56 – 0,75; Laux et al., 1981; siehe Kapitel 2.7.3.4.). Da vorliegende Arbeit sich auf die Entwicklung eines kurzen Screening-Instruments zur Erfassung der *Zustands-Angst* konzentriert (zu bereits etablierten State-Angst-Verfahren siehe Kapitel 2.7.3.3.), schließen wir uns dem Vorschlag von Uhlenhuth (1985) an, gegebenenfalls die *Trait-Angst* aus der Mittelung wiederholter State-Angst-Messungen abzuleiten, und streben *keine separate* Erfassung der Trait-Angst durch eine eigene Skala an.

Nach der konzeptuellen *inhaltlichen* Diskussion stellt sich schließlich die Frage, ob die Erfassung der Zustands-Angst mit den Items des Angst-CATs *formal* angemessen realisiert wird. Betrachtet man die *Iteminstruktionen* so fällt ein Aspekt auf, der demgegenüber kritisch angeführt werden kann. Während sich klassischerweise Instrumente, welche Zustands- und Eigenschafts-Angst erfassen unter anderem durch unterschiedliche Selbsteinschätzungszeiträume unterscheiden, wird bei der Sichtung der Iteminstruktionen des Angst-CATs offensichtlich, dass die Selbsteinschätzungszeiträume der Items zwischen „Wie fühlen Sie sich jetzt, d. h. in diesem Moment...“ über „während der letzten Woche...“ bis „in den vergangenen Wochen bzw. im vergangenen Monat...“ variieren. Diese Unterschiede im Erfragungszeitraum resultieren aus dem Umstand, dass die psychometrischen Instrumente, aus denen die Items rekrutiert wurden, verschiedene Zeitkriterien definieren. Die Entscheidung, nur Items der Instrumente zu nutzen, die sich auf einen kurzen Erfragungszeitraum beziehen, hätte zu einer Reduktion der Größte der Item- und Personen-

stichprobe geführt, welche die Stabilität der Parameterschätzung hätte gefährden können. Nach der erfolgreichen Erprobung des Angst-CATs ist nun eine Revision der Iteminstruktion geplant, welche den Erfragungszeitraum für alle Items auf zwei Wochen eingrenzt. Zusätzlich wird eine erneute Itemparameterkalibrierung des Angst-CATs nötig, da der Effekt einer Revision von Iteminstruktionen auf die Stabilität der Itemparameterschätzung bislang nicht ausreichend kalkulierbar ist (Knowles & Condon, 2000; Stocking, 1997).

Nach diesem *konzeptionellen*, den Messbereich fokussierenden Diskussionsteil folgt nun eine Diskussion um die im Rahmen der Testkonstruktion des Angst-CATs verwendeten *Methoden* und *Ergebnisse*.

7.4. Diskussion der Methoden und Ergebnisse

Das in der Einleitung erörterte Forschungsdefizit bringt es mit sich, dass bezüglich der praktischen Umsetzung der Testentwicklung eines IRT-basierten CATs noch viele Fragen offen sind. Es besteht derzeit kein allgemeiner Forschungskonsens über eine grundlegende *methodische Strategie der CAT-Entwicklung*, so dass in Anlehnung an Lehrbücher (Embretson & Reise, 2000; Embretson & Hershberger, 1997; Hambleton et al., 1991; Hambleton & Slater, 1997), Übersichtsartikel (Hattie, 1984; Nandakumar, 1994 etc.) und an eine Testentwicklungsstrategie einer US-amerikanischen Forschungsgruppe (Ware et al., 2000, 2003) bei der hier vorliegenden CAT-Entwicklung ein methodischer Weg beschritten wurde, in dessen Rahmen unterschiedliche Methoden zur sukzessiven Itemselektion angewandt werden, die jeweils Teil einer lebhaften und langanhaltenden Diskussion sind. Im Folgenden werden die Methoden und Ergebnisse in der chronologischen Reihenfolge ihrer Anwendung diskutiert.

7.4.1. Unidimensionalität

In der Literatur herrscht ein breiter Konsens, dass die Messung von Konstrukten *Unidimensionalität* erforderlich macht (McNemar, 1946; Bond & Fox, 2001). Obgleich eine angesichts verschiedener Facetten der Angst erscheinende *multidimensionale Differenzierung* der Angst sinnvoll wäre, gelingt sie wie bereits diskutiert (Kapitel 2.7.3.4) empirisch nicht im Sinne einer statistischen Unabhängigkeit von Angst*komponenten* (emotionale vs. kognitive Aspekte der Angst) bzw. Angst*konstrukten* (State-/Trait-Angst). Um unterschiedliche (voneinander abhängige) Facetten der Angst differenzierter und erschöpfender zu erforschen, wäre die Anwendung von Strukturgleichungsmodellen (Kaplan,

2000), wie sie in zahlreichen Studien bereits erfolgt, sinnvoll. Diese hätte jedoch den Rahmen vorliegender Arbeit überschritten, und wäre nicht zielführend im Sinne der Konstruktion eines unidimensionalen Angst-CATs gewesen.

Allerdings könnte zukünftig in Ableitung von Erkenntnissen aus der Strukturgleichungsforschung ein Forschungsziel in der *multidimensionalen* IRT-Modellierung (Reckase, 1997; Rost & Carstensen, 2001; Segall, 1996) und CAT-Erfassung mehrerer, voneinander abhängiger Angstkomponenten liegen. Sie kann jedoch aufgrund zunächst begrenzter technischer und fachlicher Möglichkeiten erst als „nächster Schritt" nach der hier vorliegenden erfolgreichen Erprobung der Entwicklung eines eindimensionalen CATs erfolgen.

Der erste Schritt der Testkonstruktion des Angst-CATs galt somit der Überprüfung der *Unidimensionalität*. Zur Bestimmung der Dimensionalität einer Datenmatrix wird häufig die explorative Faktorenanalyse genutzt, welche auf der Basis einer Inter-Item-Korrelationsmatrix die linearen Beziehungen zwischen Variablen und Items untersucht. Alternativ dazu schlagen manche Forscher, welche betonen, dass zweiparametrische Modelle zwar eine lineare Regression der latenten Itemantworten auf dem zu messenden latenten Kontinuum („latent trait") voraussetzen, aber die Regression der beobachtbaren Itemantworten auf dem latenten Kontinuum (d. h. die IRCs) nonlinear sei, zugunsten eines größeren Informationsgewinns sogenannte „nonlinear factor analysis of the normal ogive model" (Ferrando, 2001), Faktorenanalysen auf der Basis polychorischer Korrelationsmatrizen (Jöreskoog & Sörbom, 2002) oder „full information factor analysis" (Embretson & Reise, 2000; Software: TESTFACT; Wilson, Wood & Gibbons, 1991) vor, da lineare Faktorenanalysen vor allem bei der Anwendung auf dichotome Items zu abgeschwächten Faktorenladungen und Scheinbelegen von Multidimensionalität führen könnten (Waller et al., 1996; Ferrando, 2001).

Da jedoch die *lineare Faktorenanalyse* als historischer Standard der Itemanalyse in der Persönlichkeitsforschung gilt, die aktuell in der Forschung verbreiteteste und am häufigsten empfohlene Methodik zur Untersuchung der Unidimensionalität ist (Hambleton & Swaminathan, 1985; Lumsden, 1976) und zum Zeitpunkt der Testentwicklung abteilungsinterne Erfahrungen mit der Software zur Durchfühung nonlinearer Faktorenanalysen (z. B. NOHARM,

Fraser & McDonald, 1988) fehlten, wurde sie als erster Schritt bei der CAT-Entwicklung genutzt. Zur *Bestimmung der Dimensionalität* existieren eine Vielzahl von Kriterien wie das Kaiser-Guttman-Kriterium (Guttman, 1954), der Scree-Test (Cattell, 1966), das Parallelanalyse-Kriterium („parallel analysis criterion“ nach Lautenschlager, 1989; Verfahren der Parallelanalyse nach Horn, 1965) sowie modizierte Verfahren der Parallelanalyse (Drasgow & Lissak, 1983; Humphrey & Montanelli, 1975), das Everett-Kriterium (Everett, 1983) oder die „Lisrel-Enscheidungstabelle“ (Jöreskoog, Sörbom, du Toit & du Toit, 2000). Dabei sind sich die Forscher seit Jahrzehnten uneinig, welche Methode als die Beste zur Einschätzung der Dimensionalität einer Datenmatrix gilt.

In IRT-Anwendungsstudien im Bereich der Persönlichkeitsdiagnostik wird sowohl das Kriterium eines Eigenwerts > 1 genutzt (Reise & Waller, 1990; Reise & Henson, 2000; Gray-Little et al., 1997; Waller, 1997), welches laut Cliff (1988) theoretisch nicht gerechtfertigt sei, und in Simulationsstudien die Faktorenanzahl um 30-50% überschätze (Zwick & Velicer, 1986), als auch residuale Korrelationen (Reise & Henson, 2000) und sogar Steigungsparameter (Childs et al., 2000) als Belege für Unidimensionalität herangezogen.

Hattie (1984), welcher die gesamte Literatur zu den angewandten Methoden der Überprüfung der Unidimensionalität sichtete, und über ein Dutzend verschiedener Verfahren überprüfte, erschienen die meisten Verfahren zur Bestimmung der Dimensionalität mit großen Mängeln behaftet zu sein (siehe Kapitel 5.3.2.1.). Embretson und Reise (2000) kommen nach einer Gesamtsicht der Arbeiten in diesem Bereich (u. a. Stout, 1987, 1990; Nandakumar & Stout, 1993) zu dem Schluss, dass man die bestmögliche Information hinsichtlich der Dimensionalität der Daten erhält, wenn die gemeinsame Varianz einem *dominanten Faktor* zugeordnet wird, um danach die verbliebenen Residual-kovariationen zu analysieren. Dabei erscheint es ihnen nachrangig, mit welcher Methodik der gemeinsame Faktor identifiziert werde.

In vorliegender Arbeit wurden zur Untersuchung der Dimensionalität zunächst ein- und mehrfaktorielle explorative Faktorenanalysen an den drei der Testentwicklung zugrundeliegenden Itemteilstichproben durchgeführt. Die Exploration der Dimensionalität erfolgte anhand des Everett-Kriteriums (Everett, 1983) und des Parallelanalyse-Kriteriums („parallel analysis criterion“; genutzte Referenzwerte aus simulierten Monte-Carlo-Studien nach Lautenschlager,

1989; Verfahren der Parallelanalyse nach Horn 1965). Sie führte in den untersuchten Teilstichproben zur Extraktion von zwei bis fünf überzufälligen Faktoren, welche Multidimensionalität vermuten lassen. Die Betrachtung der Varianzaufklärung dieser Faktoren sowie der Eigenwerte zeigte, dass jeweils der erste Faktor den größten Teil der Gesamtvarianz aufklärt (N_1: 40,5%; N_2: 31,9%; N_3: 32,9%) und die höchsten Eigenwerte aufwies. Alle weiteren Faktoren trugen deutlich weniger zur Aufklärung der Gesamtvarianz bei. Diese Werte stehen im Einklang mit einer mündlichen Empfehlung von Chang und Reeve (2003), die einen Faktor als hinreichend dominant und unidimensional im Hinblick auf eine unidimensionale IRT-Modellierung ansehen, wenn der erste Faktor mehr als 20% der Gesamtvarianz aufklärt, und sich sein Eigenwert in einer Relation von 3:1 zum Eigenwert des zweiten Faktors verhalte. Neben dieser groben Empfehlung entwickelten Forscher in jüngster Zeit auch Konzepte und Methoden zur Überprüfung einer für IRT-Anwendungen *hinreichenden* Unidimensionalität im Sinne einer „essential dimensionality“ (Stout, 1987, 1990), auf die später noch eingegangen wird.

Wie im konzeptuellen Teil bereits zusammengefasst, gruppierten sich zumeist Items, welche vegetative und somatische Angstkorrelate erfragen, auf den zusätzlichen Faktoren der Faktorenlösungen, so dass diese Items, welche auf dem ersten Faktor gering luden, offensichtlich die Annahme der Unidimensionalität verletzten, und somit aus der Itemmenge ausgeschlossen wurden. Als Selektionskriterium wurde eine *Faktorenladung* > 0,4 festgelegt. Dieses entspricht den in der Persönlichkeitsforschung üblichen Cut-Off-Werten (Finch & West, 1997, S. 448: r > 0,4; Waller et al., 1996: r > 0,3).

In Anlehnung an Embretson und Reise (2000) sowie Hambleton und Mitarbeiter (1991), welche in der Analyse residualer Korrelationen die vielleicht „wertvollste Goodnes-of-Fit-Data“ überhaupt sehen (siehe Kapitel 5.3.2.1.), schloss sich an die explorative Faktorenanalyse eine konfirmatorische Faktorenanalyse an, in deren Rahmen die Analyse residualer Korrelationen erfolgte. Hohe *residuale Korrelationen* zwischen Items (r > 0,3), welche laut Thissen und Mitarbeitern (1983) auf einen Mangel lokaler Unabhängigkeit hindeuten können, führten zum Ausschluss zusätzlicher (v.a. vegetativer) Items.

Die Analyse residualer Korrelationen wird unter anderem auch im Rahmen der Entwicklung des NEO-PI-R-CATs von Reise und Henson (2000) geschildert,

allerdings ohne dass die Autoren das genaue diesbezügliche *Selektionskriterium* explizieren. Es ist allgemein anzumerken, dass es im Sinne einer besseren Verständigung zwischen Forschergruppen wünschenswert wäre, wenn in zukünftigen IRT-Studien Bewertungsmaßstäbe zur Itemselektion kommuniziert würden. Die hier in den einzelnen Testentwicklungsschritten genutzten Selektionskriterien entstammen entweder Hinweisen aus der Literatur oder mündlichen, erfahrungsbasierten Empfehlungen von Experten, die damit sicher immer zu einem Teil willkürlich sind.

Wenig kommuniziert bzw. angewandt werden im Bereich der IRT-basierten Re-Analyse von Persönlichkeitsskalen auch *Fit-Indizes* unidimensionaler Modelle, welche im Rahmen *konfirmatorischer Faktorenanalysen* gerechnet werden können. Nur sechs mir bekannte Arbeiten publizieren faktorenanalytische Fit-Indizes im Vorfeld ihrer IRT-Modellierungen in der Persönlichkeitsdiagnostik (siehe Tabelle 29).

Tabelle 29: Überblick über publizierte Fit-Indizes unidimensionaler faktorenanalytischer Modelle.

Autoren	Jahr	Inventar	Item-anzahl pro Skala	Fit-Indizes RMSEA	CFI	p
Cooke et al.	2001	HPCL	13	0,07	0,92	0,001
Marshall et al.	2002	PDEQ	15	0,07	0,91	0,01
Orlando & Marshall	2002	PTSD Checklist	17	0,09	0,81	-
Chernyshenko et al.	2001	Goldberg's Big Five	10 (50)*	0,06-0,10	0,90-0,96	-
		16 PF	10-15 (185)*	0,05-0,08	0,75-0,95	-
Becker	2003	Angst-CAT	22-37 (50)*	0,10	0,77-0,78	0,001

Inventare: HPCL= Hare Psychopathy Checklist; PDEQ = Peritraumatic Dissociative Experience Scale; 16PF = 16-Persönlichkeits-Faktoren-Inventar; PTSD Checklist = Post-Traumatic-Stress-Disorder-Checklist, NEO-PIR = Neuroticism-Extraversion-Openess-Psychoticism-Inventory-Revised.
' * ' = die in Klammern aufgeführte Zahl gibt Aufschluss über die Anzahl der Items des gesamten Instruments.
Farbmarkierung: hellgrau: Angst-CAT; dunkelgrau: Fit-Indizes**: nicht** „guter" bzw. **nicht** „akzeptabler" Fit nach folgenden Autoren: Schermelleh-Engel et al. (2003): „guter" Fit: RMSEA: 0 – 0,05; CFI: 0,97-1,0; p: 0,05 – 1,0; „akzeptabler" Fit: RMSEA: 0,05 – 0,10; CFI: 0,95-0,97; p: 0,01- 0,05; Brown & Cudeck (1993); MacCallum et al. (1996): „guter" Fit: RMSEA < 0,05: „akzeptabler" Fit; RMSEA: 0,05-0,08; "mittelmäßiger" Fit: RMSEA: 0,08-0,1; "schlechter Fit": RMSEA > 0,1. χ^2-Statistiken sind hochgradig sensitiv gegenüber der Stichprobengröße (hier: bis zu N = 1.010 Personen) und daher wenig geeignet zur Modellbeurteilung.

Den Bewertungsrichtlinien von mehreren Autoren (Brown & Cudeck, 1993; MacCallum et al., 1996; Schermelleh-Engel et al., 2003; siehe

Kapitel 5.4.1.2.2.) folgend, können die meisten der Fit-Indizes, welche bei eindimensionalen faktorenanalytischen Modellierungen verschiedener klinischer und Persönlichkeitsskalen im Vorfeld einer IRT-Modellierung berechnet wurden, als nicht „akzeptabel“ (siehe graue Farbmarkierung in Tabelle 29) bewertet werden. Dies ist ein Befund, der sich nicht nur bei IRT-basierten Reanalysen etablierter Inventare zeigt, sondern auch bei analogen Untersuchungen gut etablierter Fragebögen (STAI State: 20 Items: TLI=0,73, CFI=0,76, RMSEA=0,13; NEO-FFI Neurotizismusskala 12 Items TLI=0,82, CFI=0,86, RMSEA=0,11).

Es fällt auf, dass die Fit-Indizes *schlechter* ausfallen, je *mehr* Items zur eindimensionalen Modellierung genutzt werden. Da zur Analyse der Itembank des Angst-CATs selektierte Itemmengen zwischen 22 und 37 Items (in drei verschiedenen Teilstichproben) genutzt wurden, welche jeweils umfangreicher als die Itemanzahl der anderen in Tabelle 29 aufgeführten Skalen sind, erstaunt das Ergebnis, dass die Fit-Indizes vorliegender Arbeit nur als knapp „akzeptabel“ gewertet werden können, nicht.

Angesichts der insgesamt über alle analysierten Skalen hinweg tendenziell eher als knapp *akzeptabel* bis *schlecht zu bewertenden Fit-Indizes* und eines allgemeinen Zweifels, ob sich die konfirmatorische Faktorenanalyse mit den Fit-Indizes als Methode und Statistik zur Bestimmung einer für erfolgreiche IRT-Modellierungen *hinreichenden Unidimensionalität* überhaupt eignet (Chernyshenko et al., 2001), ist erklärbar, warum das Gros der IRT-Forschungsarbeiten Fit-Indizes konfirmatorischer Faktorenanalysen nicht publiziert (Childs et al., 2000; Cooke et al., 1999; Gray-Litte et al., 1997; Orlando et al., 2000; Reise & Waller, 1990; Santor & Coyne, 2000). Seit einiger Zeit scheint sich in methodisch versierten Forscherkreisen (Stout, 1987, 1990; Nandakamour, 1993, 1994; Nandakamour & Stout, 1993) zunehmend die Meinung durchzusetzen, dass für eine erfolgreiche unidimensionale IRT-Modellierung keine *perfekte* Unidimensionalität, sondern lediglich eine „*approximative*“ (McDonald, 1994) oder „*essentielle*“ *Unidimensionalität* erforderlich sei (Ferrando, 2001). Das bedeutet, dass für eine IRT-Modellierung die Anforderungen an die Unidimensionalität nicht so streng sein müssen wie es in der Strukturgleichungsforschung üblich ist, sondern dass eine IRT-Modellierung bereits dann erlaubt sei, wenn eine „major dimension“ im Sinne

eines dominanten Faktors existiere (unabhängig von der Existenz von mehreren „minor dimensions“; Ferrando, 2001), der den größten Teil der gemeinsamen Varianz aufkläre (Reise & Waller, 1990; Embretson & Reise, 2000). Nach Stout (1990) ist es psychometrisch begründet und angemessen, die *strenge* Forderung nach lokaler Unabhängigkeit der Daten durch die Forderung nach „essentieller“ Unidimensionalität abzuschwächen. Nandakumar (1993; Nandakumar & Stout, 1993) entwickelte zur Überprüfung dieser essentiellen Unidimensionalität, welche von Stout (1990) mathematisch definiert ist, auch einen Test (DIMTEST; Stout, Douglas, Junker & Roussos, 1993), der jedoch zum Zeitpunkt der vorliegenden Testentwicklung nicht verfügbar war. In zukünftigen Studien gilt es, die Diskussion um die angemessene Methode zur Bewertung der für IRT-Modellierungen hinreichenden Unidimensionalität aufrechtzuerhalten und oben genannten neuen Test anzuwenden.

In der vorliegenden Studie wird aufgrund der Ergebnisse der explorativen Faktorenanalysen und der residualen Korrelationsanalysen angenommen, dass eine für eine erfolgreiche IRT-Modellierung nötige „hinreichende“ Unidimensionalität der Items zur Messung von Angst vorliegt, welche durch die realisierten Itemselektionskriterien (Faktorenladungen > 0,4; Residuale Korrelationen < 0,3) weiter gestärkt wurde.

7.4.2. IRT-Analyse

Nach der Diskussion um die zur Unidimensionalitätsuntersuchung angewandten Methoden und Ergebnisse (siehe Kapitel 5.3.2.1. und 5.4.1.) folgt nun eine kritische Reflektion der in vorliegender Arbeit durchgeführten *IRT-Analyse* (siehe Kapitel 5.3.2.2. und 5.4.2.). Diese umfasst die grafische Inspektion der Item Response Curves (IRCs) und die Untersuchung der Testinformationsfunktion sowie des Standardmessfehlers und der Reliabilität.

Insbesondere die *Untersuchung der IRCs* stellt gegenüber den in der KTT eingesetzten Analysemethoden eine fortgeschrittene Methodik zur psychometrischen Beurteilung einzelner Items und Antwortkategorien dar (zu den Vorteilen der IRT siehe Kapitel 3.3.3.). Sie wird von vielen Forschern zur Beurteilung der Modellkonformität und Diskriminationsfähigkeit von dichotomen und polytomen Items genutzt (Cooke et al., 1997, 1999, 2001; Gray-Little et al., 1997; Orlando & Marshall, 2002; Reise & Waller, 1990; Reise & Henson, 2000; Santor et al., 1994, 1995, 2000; Orlando et al., 2000). Über die allgemeinen

grafischen Kriterien,[110] welche die IRCs optimalerweise erfüllen sollten, besteht in der Literatur allgemeiner Konsens. Jedoch existieren keine *eindeutigen* grafischen Selektionskriterien, welche IRCs als „schlecht" bewertet können, und damit einen Itemausschluss notwendig machen. Da die meisten Autoren in Publikationen in Fachzeitschriften nur zu illustrativen Zwecken eine Auswahl modellkonformer IRCs weniger Items präsentieren, kann aufgrund dieses publikatorischen Mangels ein formaler grafischer Vergleich zwischen IRCs von verschiedenen Tests an dieser Stelle nur sehr begrenzt erfolgen. Es liegen nämlich nur die IRCs *aller* Items einer Skala in einer Publikation über die „Hamilton Rating Scale for Depression" (HRSD; Santor & Coyne, 2000) vor, die mit den IRCs der Items der gesamten Itembank des Angst-CATs verglichen werden können (siehe Anhang 9.3.). In der Studie von Santor und Coyne, in der die IRCs der 21 Items des HRSD grafisch untersuchen wurden, fanden die Autoren bei einer Reihe von Items Schwierigkeiten im Kurvenverlauf der IRCs, welche die Autoren zu der Schlussfolgerung bewogen, dass diese Items zur eindimensionalen Erfassung der Depression nicht geeignet seien. Der formale grafische Vergleich der IRCs der Items des Angst-CATs (N = 50) und der Items des HRSD (N = 21) fällt dementsprechend zugunsten einer höheren Modellkonformität der IRCs der Items des Angst-CATs aus. Eine Beurteilung der IRCs der Items des HRSD mit den bei der Entwicklung des Angst-CATs realisierten grafischen Selektionskriterien hätte bei der HRSD zu der Empfehlung eines Ausschlusses von 12 (von 21) Items geführt.

Nach der Analyse der IRCs schließt sich in der Testentwicklung des Angst-CATs die *Untersuchung der Item- und Testinformationsfunktion* (siehe Kapitel 3.3.3. und 5.3.2.2.2.) an. Diese bietet den Vorteil, die Messpräzision einer Skala in Abhängigkeit vom Merkmalsausprägungskontinuum zu beurteilen, und kann damit einen wichtigen Beitrag zum Vergleich verschiedener Testverfahren bezüglich ihrer Indikation leisten. Obgleich eine Reihe von Autoren Item- und Testinformationskurven zur IRT-basierten Re-Analyse bereits etablierter psychometrischer Instrumente nutzen, fehlt bislang ein Beurteilungsmaßstab zur Einschätzung der Höhe dieser Statistik.

[110] Grafische Kennzeichen eines guten IRT-Modell-Fits von polytomen Items: glockenförmiger Kurvenverlauf der einzelnen Antwortkategorienkurven, Kurvenmaximum überschneidet alle anderen Kurvenverläufe in genau einem Merkmalsausprägungsbereich, aufsteigend angeordnete Schwellenparameter, monoton absteigende erste Antwortkategorienkurve und monoton ansteigende letzte Antwortkategorienkurve (siehe Kapitel 5.3.2.2.1. und 5.4.2.1.).

Insbesondere verwirrt, dass die Testinformationen meist ohne Angabe der Anzahl der Items eines Tests publiziert werden. Dies erschwert den Vergleich von Testinformationen unterschiedlicher Instrumente, da die Testinformation in ihrer Höhe direkt von der Itemanzahl abhängig ist (Addition der Iteminformation aller Items = Testinformation). Um die Höhe der Testinformationen der drei in vorliegender Arbeit analysierten Itemstichproben N_1 – N_3 (siehe Kapitel 5.4.2.2.) des Angst-CATs bewerten zu können, wurde aus den gesichteten IRT-Publikationen die Spannweite der jeweils präsentierten Testinformationen (range (TI)) herausgesucht und – falls angegeben – durch die Anzahl der analysierten Items dividiert. So konnte die durchschnittliche Spannweite der Iteminformation ($\overline{I\!I}$) pro Skala errechnet werden und ein Vergleich der Iteminformationen zwischen den Skalen erfolgen.

Meines Wissens liegen derzeit sechs IRT-Publikationen in der Persönlichkeitsdiagnostik mit Angaben zur Testinformation vor. Tabelle 30 verdeutlicht, dass die durchschnittliche Spannweite der Iteminformationen des Angst-CATs mit der anderer untersuchter Instrumente vergleichbar ist.

Tabelle 30: Überblick über verschiedene Test- und Iteminformationsniveaus verschiedener Skalen.

Autoren	Jahr	Inventar	Item-anzahl pro Skala	TI range	AM II range
Reise & Henson	2000	NEO-PI-Neuroticism Scale	8	1 – 4	0,1 – 0,5
Gray-Little et al.	1997	Rosenberg Self-Esteem Scale	10	1 – 11	1,1
Marshall et al.	2002	Peritraumatic Dissociative Questionnaire	8 [111]	-	0,1-0,8*
Ferrando	1994	EPI Impulsivity Scale	6 [112]	0-13	0,0-2,2
Cooke et al.	2001	Hare Psychopathy Checklist	20	5 – 15	0,3-0,8
Santor & Ramsay	1998	BDI CES-D	21 20	5 – 9 2 – 15	0,2-0,4 0,1-0,8
Childs et al.	2000	MMPI-2 Depression Scale	10	1-10	0,1-1,0
Becker	2003	Angst-CAT: N_1 N_2 N_3	24 26 17	14-18 10-16 6-12	0,6-0,8 0,4-0,6 0,4-0,7

Inventare: NEO-PI: Neuroticism Extraversion Openess Psychoticism Inventory; EPI: Eysenck Personality Inventory; BDI: Beck Depression Inventory; CES-D: Center of Epidemiological Studies-Depression Scale; MMPI: Minnesota Multiphasic Personality Inventory;
TI range: Spannweite der Testinformationsfunktion;
AM II range: Spannweite der durchschnittlichen Iteminformationsfunktion, d. h. TI range / Itemanzahl pro Skala;
`*´: reine Spannweite der Iteminformation (direkt von Marshall et al., 2002, angegeben, keine arithmetische Mittelwertsbildung).

[111] Diese Items wurden aus der EPI Impulsivity Scale von insgesamt 11 Items selektiert.
[112] Diese Items wurden aus dem PDEQ von insgesamt 10 Items selektiert.

Ein Vergleich der Testinformationskurvenverläufe der einzelnen Publikationen ergibt, dass Testinformationskurven etablierter Instrumente sowohl eingipflig (CES-D, PDEQ) als auch mehrgipflig (NEO-PI-Neuroticism; BDI) sein können, d. h. die Diskriminationsfähigkeit einer Skala in Abhängigkeit zum Merkmalsausprägungskontinuum in der Regel variiert. Diese Beobachtung fand sich auch in vorliegender Studie (siehe Kapitel 5.4.2.2.). Analog dazu verhält sich die Variation des Standardmessfehlers ($SE_{(N1-N3)}$ = 0,2 bis 0,4)[113] und der Reliabilitäten ($Rel_{(N1-N3)}$ = 0,85 bis 0,94) der untersuchten Itemstichproben des Angst-CATs (N_1-N_3) ebenfalls in Abhängigkeit zum latenten Merkmalsausprägungskontinuum (siehe Kapitel 5.4.2.3.).

An die IRT-Analyse, welche die Berechnung verschiedener Statistiken umfasste (Item- und Testinformation, Standardmessfehler und Reliabilität in Abhängigkeit des „latent traits"), schloss sich die *IRT-Modellierung* als letzter Untersuchungsschritt in der Entwicklung des Angst-CATs an (siehe Kapitel 5.3.2.3. und 5.4.3.). Diese wird im Folgenden diskutiert.

7.4.3. IRT-Modellierung

Im Hinblick auf die IRT-Modellierung stehen die Modellwahl, die Fit-Statistiken, das Differential-Item-Functioning (DIF) sowie das Item-Link-Design („Linking"), und die Stabilität der Itemparameterschätzung zur Diskussion.

Vorliegende Arbeit wählte das *Generalized Partial Credit Modell* (GPCM, Muraki, 1997; siehe Kapitel 3.4.3.) aus den möglichen IRT-Modellen aus (siehe Kapitel 3.4.1./4.), da es eine unidimensionale zweiparametrische IRT-Modellierung polytomer Daten mit einer simultanen Analyse unterschiedlicher Antwortformate erlaubt sowie die Variation der Diskriminationsfähigkeit unterschiedlicher Antwortkategorien und unterschiedlicher Items bei der Modellierung berücksichtigt. Es gilt als wenig restriktiv. Nachteilig ist am GPCM, dass sich der Schätzalgorithmus mathematisch aufwendiger als bei klassischen Rasch-Modellierungen gestaltet, und es für eine stabile Parameterschätzung – wie alle komplexeren IRT-Modelle – große Personenstichproben voraussetzt (siehe Kapitel 3.4.5.).

Das GPCM wurde zur Modellierung von Persönlichkeitsskalen bislang wenig genutzt. 10 von 26 IRT-Anwendungsstudien im Bereich der Persönlichkeitsdiagnostik wenden das ältere, bereits „etablierte" Graded Response Model

[113] SE = Standard Error of Measurement; Standardmessfehler.

(GRM) von Samejima (1969) und sieben Studien das 2PL-Modell von Birnbaum (1968) an (siehe Tabelle 5 in Kapitel 3.5.2.), obgleich beide Modelle (GRM und 2PLM) restriktiver als das „neuere“ GPCM sind. Während nämlich das 2PL-Modell von Birnbaum (1968) keine variierende Antwortkategorien-schwellenparameter berücksichtigt, erlaubt das GRM keine Variation der Steigungsparameter unterschiedlicher Antwortkategorien und kann Items nur in isolierten Gruppen von Items mit gleichen Antwortformaten modellieren.

Obgleich erste Hinweise auf vergleichbare Ergebnisse zwischen dem PCM (Masters, 1982), auf dessen Basis das GPCM (Muraki, 1997) entwickelt wurde, und dem von Thissen und Steinberg (1986) erweiterten GRM (Samejima, 1969) vorliegen (Maydeu-Olivares, Drasgow & Mead, 1994; Childs & Chen, 1999), sollte eine mögliche Übereinstimmung dieser Modelle durch entsprechende Studien weiter erforscht werden. Dies ist gerade vor dem Hintergrund eines eklatanten Forschungsdefizits an *IRT-Modellvergleichsstudien* (besonders im Bereich der Persönlichkeitsdiagnostik) relevant. Solche Studien, welche simultan verschiedene polytome IRT-Modelle erproben, werden von mehreren Autoren gefordert, da man sich von ihnen ein besseres Verständnis der Struktur von Tests (de Koning et al., 2002), sowie eine Reduktion bislang bestehender Unsicherheiten bei der Wahl des „richtigen“ Modells (Embretson & Reise, 2000) und eine Verbesserung in der Beurteilung (und ggf. eine Weiterentwicklung) von Modell-Fit-Statistiken verspricht (Hambleton et al., 1991).

Dies leitet zu einem weiteren Problemfeld bei der Anwendung polytomer IRT-Modelle im Bereich der Persönlichkeitsforschung über. Während statistische *Modellgeltungstests* für Rasch-Modelle weitgehend erforscht und etabliert sind (Andersen, 1973; Glas, 1988; Keldermann, 1984; Molenaar, 1974), gilt dies nicht für zwei- bzw. dreiparametrische Modelle (wie das GPCM). Diese gelten als wenig entwickelt und defizitär (Van der Linden & Hambleton, 1997, siehe Kapitel 3.4.5.).

Dies führt in zahlreichen Publikationen zu einem Verzicht der Darstellung von IRT-spezifischen Item-Fit-Statistiken bei der IRT-Analyse von Persönlichkeits-inventaren (Childs et al., 2000; Cooke & Michie, 1997; Cooke et al., 1999; Ellis et al., 1989; Gray-Little et al., 1997; Marshall et al., 2002; Orlando & Marshall, 2002; Reise & Henson, 2000; Rouse et al., 1999; Santor et al., 1995; Santor & Ramsay, 1998; Santor & Coyne, 2000; Schmit & Ryan, 1997). Während zur

Überprüfung des GRMs (Samejima, 1969) meines Wissens keinerlei Fit-Methoden und -Ergebnisse publiziert sind, wird die erfolgreiche Anwendung des 2-PL-Modells von Birnbaum (1968) durch mehrere Publikationen mit guten numerischen Fit-Ergebnissen (Software: BILOG 3; Mislevy & Bock, 1990) belegt (Ferrando, 1994; Ferrando, 2001; Finch & West, 1997; Reise, 1999; Reise & Waller, 1990; Waller et al., 1996).

Werden Item-Fit-Methoden verwendet, so dominieren im Allgemeinen die numerischen Fit-Statistiken über die grafischen Untersuchungen zur Modellanpassung.

Die in vorliegender Studie präsentierten *Likelihood-χ^2-Fit-Statistiken* der Items der Itembank des Angst-CATs (siehe Kapitel 5.4.3.4.) ergaben eine Vielzahl von Items (N = 22), welche als signifikant vom GPCM abweichend gewertet werden müssten ($p \leq 0{,}05$). Dies ist angesichts des großen Stichprobenumfangs der hier analysierten Teilstichproben ($N_1 = 1.010$; $N_2 = 834$; $N_3 = 775$) und der vielfach kritisierten methodischen Schwäche dieser Item-Fit-Statistik, welche in ihrer starken Abhängigkeit von der Stichprobengröße liegt (Embretson & Reise, 2000; Hambleton et al., 1991; Van der Linden & Hambleton, 1997; McDonald, 1989; Muraki, 1997; Rost et al., 1999; siehe Kapitel 5.4.3.), nicht weiter erstaunlich. Hambleton und Mitarbeiter (1991) fanden in mehreren Simulationsstudien zur Überprüfung ähnlicher Modellgeltungstests bei einer systematischen Vergrößerung der Personen- und Itemstichprobe eine zunehmende Anzahl von Item-Misfits, welche sie als statistische „Artefakte“ bewerteten (siehe Kapitel 5.3.2.3.4.).

Auch Rost und Mitarbeiter (1999) machen auf die Stichprobenabhängigkeit von Likelihood-χ^2-Fit-Statistiken – allerdings zur Überprüfung des Rasch-Modells – aufmerksam und fanden, dass die fünf Skalen des NEO-FFIs den Kriterien für die Geltung des Rasch-Modells nicht genügten. Es bleibt zu spekulieren, ob der von ihnen gefundene Item-Misfit aus einer mangelhaften Fit-Methodik oder der Inadäquatheit des Modells resultiert, denn ein Jahr später gelang Reise und Henson (2000) die Modellierung und Entwicklung einer CAT-Version des NEO-PIR anhand des GRM (siehe Kapitel 3.5.2.). Dies könnte auch so interpretiert werden, dass das GRM besser zur Modellierung von Persönlichkeitsskalen wie dem NEO-PI-R geeignet ist als das Rasch-Modell.

Aufgrund der Unsicherheiten, welche sich aus der Stichprobenabhängigkeit von Likelihood-χ^2-Fit-Statistiken ergeben, wurde in vorliegender Arbeit der Empfehlung von Embretson und Reise (2000) gefolgt, die Likelihood-χ^2-Fit-Statistik nicht als „solid-decision-making tool“ (S. 235) zur Itemselektion zu nutzen. Dies ist insofern sinnvoll, als Chernyshenko und Mitarbeiter (2000) darauf hinweisen, dass – im Falle Forscher ließen sich in der Itemselektion von signifikanten χ^2-Ergebnissen leiten – damit eine Variablenkonfundierung erfolge, da nicht beurteilt werden könne, ob der mangelhafte Item-Fit bei der IRT-Modellierung auf eine *schlechte Qualität* der *Items*, des *Modells* oder der angewandten *Fit-Statistik* hinweise (siehe oben erläuterte NEO-PI-Modellierung von Rost et al., 1999, bzw. Reise & Henson, 2000).

Gründe er sich auf einer schlechten Qualität der Items, so können nach Chernyshenko und Mitarbeitern (2000) mehrere Ursachen verantwortlich sein. So könnten spezifische formale (z. B. negative Itemformulierungen) oder inhaltliche Eigenschaften von Items (Itemtextinhalt), Verletzungen von grundlegenden IRT-Voraussetzungen wie der Unidimensionalität oder der lokalen stochastischen Unabhängigkeit und grundlegende Unterschiede bei der Beantwortung von Persönlichkeitsitems im Vergleich zur Beantwortung von Leistungsitems eine Rolle spielen.

Während eine genaue Inspektion formaler und inhaltlicher Eigenschaften der Items, denen ein signifikanter Misfit in vorliegender Arbeit zugeschrieben wurde, keine Auffälligkeit offenbarte, die den Misfit hätte erklären können, und die Erfüllung der Unidimensionalität bereits weiter oben diskutiert wurde, sowie die lokale stochastische Unabhängigkeit in der Regel nicht direkt überprüfbar ist, bleibt weiter zu erforschen, ob die von Chernyshenko und Mitarbeitern (2000) vermutete Andersartigkeit von Persönlichkeitsitems verglichen mit Leistungsitems eine IRT-Modellierung erschwert.

Zur Beurteilung, ob spezifische IRT-Modelle zur Modellierung bestimmter Daten (z. B. Persönlichkeitsdaten) nicht adäquat sind, fordert Rost (1999) die Entwicklung von „Overall-Fit-Statistiken“ (S. 152) zum Vergleich der Modellgültigkeit mehrerer konkurrierender IRT-Modelle. Weiterhin regt er an, neben der statistischen Signifikanz von Modellabweichungen auch Modellabweichungen nach ihrer psychologischen Bedeutsamkeit zu beurteilen.

Die numerischen Item-Fit-Statistiken wurden trotz reflektierter Mängel in vorliegender Arbeit präsentiert (siehe Kapitel 5.4.3.4.), um die Kommunikation mit anderen Forschungsgruppen über dieses Problem zu erleichtern. Es bleibt zu hoffen, dass sich in den nächsten Jahren für zweiparametrische IRT-Modelle gegenüber der Stichprobengröße *robuste* und bezüglich spezifischer *Formen* des Misfits *aufschlussreichere* Verfahren zur Beurteilung des *spezifischen Item-* und des *globalen Modell-Fits* etablieren (Chernyshenko et al., 2001).

Um dem vorausgegangen erörterten Fit-Statistik-Problem zu begegnen, plant die Forschungsgruppe, in dessen Rahmen die vorliegende Arbeit entstand, zum einen die Erprobung weiterer numerischer sowie grafischer Methoden zur Untersuchung des Modell-Fits. Weiterhin ist geplant, den Empfehlungen von Van der Linden und Hambleton (1991; siehe Kapitel 5.4.3.4.) zu folgen, und den Item-Fit sowie die Modellvorhersage und Itemparameterinvarianz an anderen realen und simulierten Personenstichproben zu überprüfen, sowie schließlich zur Optimierung der Itembank des Angst-CATs auch neue modellkonforme Items zu konstruieren. Langfristig wäre auch die Erprobung des GRM (Samejima, 1969) und des 2PLM (Birnbaum, 1968) an den vorliegenden Daten interessant, um einen Vergleich unterschiedlicher zweiparametrischer Modelle und ihrer Modellgültigkeit zu ermöglichen.

Die Diskussion um den Item- bzw. Modell-Fit ist essentiell, da eine Anwendung von IRT-Modellen ohne den Beleg der Modellgültigkeit „suspekt“ bleibt (Chernyshenko et al., 2000, S. 524). Schon Lord (1980) betonte, dass der Gebrauch jedes Modells empirisch zu begründen sei, und ein Vorteil der IRT liegt ja gerade – verglichen mit der KTT – in der potentiellen Falsifizierbarkeit von spezifizierten Modellen (siehe Kapitel 3.3.1.), die durch die Diskussion um angemessene Fit-Statistiken letztendlich nicht untergraben werden darf.

Schließlich ist die Diskussion um den empirischen Nachweis der Modellgültigkeit so brisant, da dieser impliziert, dass zentrale Charakteristika der IRT wie die Annahme der Modellierung der Itemantworten mittels der Item Response Function (IRF) und die Itemparameterinvarianz gelten (siehe Kapitel 3.3.1.). Insbesondere die Erfüllung der Annahme der *Itemparameter-invarianz* ist für die Funktionsfähigkeit von CATs (wie hier des Angst-CATs) notwendig, da die Itemselektion und Personenparameter-schätzung späterer Personenstichproben auf der Basis von Itemparametern erfolgt, welche an einer

Vorkalibrierungsstichprobe geschätzt wurden. Hier sei kritisch einzuräumen, dass zu den methodischen Unwägbarkeiten der IRT derzeit auch zählt, dass bezüglich der Itemparameterinvarianz widersprüchliche Studienergebnisse vorliegen. So fanden eine Reihe von Forschern (Dorans & Kingston, 1985; Forsyth, Saisangjan & Gillmer, 1981; Rentz & Barshaw, 1977), dass das Rasch-Modell relativ robust gegenüber Verletzungen seiner Voraussetzungen reagiert, während andere Forscher (Cook, Eignor & Taft, 1984; Loyd & Hoover, 1980; Slinde & Linn, 1978) dies nicht bestätigen konnten. Abgesehen von einigen wenigen neueren Forschungsarbeiten (z. B. Knowles & Condon, 2000; Sinar & Zickar, 2002) herrscht hier noch ein großes Forschungsdefizit vor allem bei der systematischen Erforschung der Itemparameterstabilität von mehr-parametrischen IRT-Modellen (wie dem GPCM). Als allgemeine Einflussfaktoren, welche die Robustheit der Itemparameterschätzung bedingen, gelten neben der Erfüllung spezifischer IRT-Voraussetzungen (wie der Unidimensionalität bzw. der lokalen stochastischen Unabhängigkeit), die *Größe* der Personenstichprobe zur IRT-Kalibrierung (Ferrando, 2001). Die Größen der in vorliegender Studie analysierten Personenstichproben ($N_1 = 1.010$; $N_2 = 834$; $N_3 = 775$) sind angesichts der von von zwei Forscherkreisen ausgesprochenen Empfehlungen bei der Anwendung des GPCMs als hinreichend zu bewerten (Muraki & Bock, 1999: n = 500-1.000; Cella & Chang, 2000: n > 1.000; siehe Kapitel 3.3.4.).

Eine empirische Überprüfung der Itemparameterinvarianz ist nach Suen (1990) sehr zu empfehlen und kann nach Knowles und Condon (2000) auf drei prinzipiellen Wegen erfolgen: der Untersuchung von *Differential-Item-Functioning* (DIF) a) mittels KTT-basierter Methoden, b) mittels IRT-basierter Methoden (siehe Kapitel 5.3.2.3.2.) und c) mittels Strukturgleichungsmodellen.

In vorliegender Studie erfolgte sie IRT-basiert mit dem Ziel, unerwünschten DIF bei *Anker-Items*, welche zum Item-Link-Design genutzt wurden, zu explorieren. Wie im Ergebnisteil (Kapitel 5.4.3.2.) dargestellt, eigneten sich die ausgewählten Anker-Items zum „Linking“, da (abgesehen von einem) bei 20 Einzelvergleichstests keine Hinweise auf signifikante Unterschiede in der Itemparameterschätzung der Items eruiert werden konnten. Hier sei kritisch anzumerken, dass – obgleich die Anker-Items des Angst-CATs, wie von Hambleton und Mitarbeitern (1991) gefordert, dem intendierten Inhaltsbereich

der Itembank des Angst-CATs inhaltlich gut entsprechen – sie in ihrer Anzahl (6 von insgesamt 50 Items der Itembank) *unter* den Empfehlungen (20-25% der Gesamtitemzahl eines Tests) genannter Autoren bleiben. Embretson und Reise (2000) geben in dieser Hinsicht zu bedenken, dass ein kleines Set von Anker-Items beim Linking ein „source of problems“ (S. 256) sein könnte und weisen auf ein Forschungsdefizit hinsichtlich der für ein gutes Linking erforderlichen Anzahl von Anker-Items hin (S. 260).

Um die potentielle Gefährdung der Robustheit und Güte der Itemparameter-schätzung durch ein *Item-Link-Design* (Kaskowitz & DeAyala, 2001; siehe Kapitel 5.3.2.3.3.), in dessen Rahmen eine mathematische Neu-Adjustierung der Itemparameter verschiedener Itemstichproben auf einer gemeinsamen Metrik erfolgt, auszuschließen, wird die Entwicklung zukünftiger CATs von der Forschergruppe, in dessen Rahmen vorliegende Arbeit geschrieben wurde, nur noch auf der Basis *einer* großen Item- und Personenstichprobe stattfinden (und nicht wie in vorliegender Studie auf der Basis von *drei* Teilstichproben, welche es über ein Item-Link-Design zu verbinden gilt).

Nichts desto trotz könnte es an dieser Stelle auch sinnvoll sein, das Potential, welches die IRT mit der Möglichkeit des „Linkings“ überhaupt erst Forschern eröffnet (siehe Kapitel 3.3.3.), weiter zu explorieren und einen Beitrag hinsichtlich der Methodenentwicklung des Linkings zu leisten, welcher in der Erprobung anderer Anker-Items und Anker-Itemsetgrößen sowie verschiedener Linking-Methoden liegen könnte („mean and sigma“ oder „characterstic curve methods“, Embretson & Reise, 2000).

Vorliegende Studie beschränkte sich auf die Überprüfung der Itemparameterinvarianz bezüglich eines Sets von Anker-Items. In zukünftigen Studien wird die Itemparameterinvarianz der gesamten Itembank an verschiedenen Personenstichproben - auch hinsichtlich spezifischer soziodemografischer Stichprobencharakteristika (Alter, Geschlecht, etc.) - weiter untersucht werden müssen.

Die in vorliegender Studie angewandte Linking-Prozedur (siehe Kapitel 5.3.2.3.3. und 5.4.3.3.) führte zur Itemparameterschätzung aller selektierter Items, welche nachfolgend als die Itembank konstituierend angesehen werden. Die Güte der Itembank des Angst-CATs, deren

Inhaltsbereich bereits zu Beginn dieses Kapitels konzeptuell diskutiert wurde, wird im Folgenden aus methodischer Sicht bewertet.

7.4.4. Evaluation der Itembank des Angst-CATs

Wie in Kapitel 4.3.3.1. dargestellt, existieren mehrere *psychometrische Anforderungen an eine „gute" Itembank*. Bezüglich der erwünschten *Größe einer Itembank* liegen nur Erfahrungswerte aus der Leistungsdiagnostik vor. Dort variieren die Empfehlungen zwischen 70 und 200 Items (Weiss, 1985; Hornke, 1993), während in der Persönlichkeitsdiagnostik von mehreren Autoren vermutet wird, dass hier die Itembank durchaus aus weniger Items bestehen kann, da die Items größtenteils ein polytomes Antwortformat aufweisen (Dodd et al., 1995; Embretson & Reise, 2000; Master & Evans, 1986). In der vorliegenden Arbeit wird angenommen, dass die Itembankgröße (N = 50 Items) des Angst-CATs ausreicht. Im Sinne einer Itembankoptimierung ist langfristig von der Forschungsgruppe geplant, die Itembank des Angst-CATs durch die Konstruktion und Kaliberierung neuer Items zu erweitern, und damit Auswirkungen der systematischen Vergrößerung der Itembank zu explorieren.

Neben der Größe der Itembank ist die *Diskriminationsfähigkeit* und *Breite des Messbereichs* entscheidend bei der psychometrischen Evaluation einer Itembank.

Eine hohe Diskriminationsfähigkeit des Angst-CATs wurde durch einen gezielten Ausschluss von Items mit einem Steigungsparameter von $a_i < 0,8$ hergestellt (siehe Kapitel 5.3.2.3.1. und 5.4.3.1.). Dieses Selektionskriterium ist dem von Waller und Mitarbeitern (1996) genutzten Kriterium von $a_i < 1,0$ ähnlich. Waller und Mitarbeiter (1996) weisen ferner darauf hin, dass die Steigungsparameterwerte (a_i) typischer Persönlichkeitsitems zwischen $a_i = 0,5$ bis 1,5 lägen, und grob Faktorenladungen von 0,4 bis 0,8 entsprächen. Die Steigungsparameterwerte der Itembank des Angst-CATs liegen in einem Bereich von $a_i = 0,80$ bis $a_i = 2,60$ ($\overline{X} = 1,34$; SD = 0,40). Mit einem durchschnittlichen Steigungsparameterwert von $\overline{a_i} = 1,34$ (siehe Kapitel 5.4.4.) und Faktorenladungen von 0,4 – 0,8 (siehe Kapitel 5.4.1.1.) steht das Angst-CAT im Einklang mit diesen Beobachtungen, obgleich einschränkend betont werden muss, dass Waller und Mitarbeiter das zweiparametrische Birnbaum Modell zur IRT-Modellierung anwandten und es zu diskutieren ist, ob

Steigungsparameterwerte über verschiedene Modellierungen hinweg miteinander verglichen werden können.

Da in klinischen IRT-Anwendungsstudien (Kapitel 3.5.2.) unterschiedliche IRT-Modelle genutzt werden (Rasch-Modell, Birnbaum-Modell, GRM, PCM, GPCM), fällt ein Vergleich und damit eine Bewertung der Schwellen- und Lokationsparameterwerte zwischen verschiedenen Studien ebenfalls schwer. Die Lokationsparameterwerte der Items des Angst-CATs liegen zwischen –1,58 und 1,55 ($\overline{X}$ = -0,11; SD = 0,65); die Schwellenparameter (Thresholds) variieren zwischen –2,81 („bin gelöst“) und 3,30 („fühle mich kribbelig“). Da die Schwellenparameter der Items folglich in einem Bereich von ca. 6 Standardabweichungen streuen, kann angenommen werden, dass die die Itembank des Angst-CATs konstituierenden Items einen großen Teil des Angstkontinuums abzubilden vermögen.

Zusammenfassend lässt sich resümieren, dass die hohen Steigungsparameterwerte und die Spannweite der Schwellenparameterwerte der Items des Angst-CATs erwarten lassen, dass das Angst-CAT eine hoch diskriminative Erfassung eines weiten Merkmalsausprägungsbereichs der Angst ermöglicht.

7.5. Zur Validierung des Angst-CATs

7.5.1. Zur allgemeinen Funktionsweise des Angst-CATs

Um die psychometrische Güte des Angst-CATs zu überprüfen, befasst sich der zweite empirische Teil der vorliegenden Arbeit mit der *Validierung* des entwickelten Instruments (siehe Kapitel 6.).

Die Validierungsstudie an N = 102 psychosomatischen, stationär behandelten Patienten ergab, dass mit dem Angst-CAT, dessen Stoppfunktion „a priori“ auf eine Reliabilität von Rel(θ) = 0,9 festgesetzt wurde, eine Erfassung der Angstausprägung mit im Durchschnitt 5,3 $\pm$ 1,9 Items ($\overline{X}$ $\pm$ SD) möglich ist.

Dieser Befund zeigt, dass der theoretisch von CATs erwartete Vorteil einer größeren Testökonomie durch *maßgebliche Itemeinsparungen* (Wainer, 1990; Meijer & Nering, 1999; Kapitel 4.4.) eingelöst werden kann. In der Literatur zu IRT-basierten CATs werden Itemeinsparungen von 25% bis 66% berichtet (Gardner et al., 2002; Handel et al., 1999[114]; Hornke, 1999[115]; Koch et al.,

[114] In der Studie von Handel und Mitarbeitern (1999) wurde eine CAT-Version des MMPI, welche auf der Basis der „Countdown Method“ (siehe Kapitel 4.3.2.) entwickelt wurde, evaluiert. Alle anderen in diesem Kapitel erwähnten CATs sind IRT-basiert.

1990[116]; Reise & Henson, 2000; Singh, 1993; Waller, 1997; Waller & Reise, 1989; Weiss, 1985).

Diese offenbaren sich in CATs, welche durchschnittlich zwischen 3 und 8 Items (Gardner et al., 2002; Hornke, 1999; Reise & Henson, 2000; Simms & Clark, in Vorbereitung; Waller & Reise, 1989) darbieten. Die in vorliegender Studie erreichte Itemreduktion auf $5{,}3 \pm 1{,}9$ Items ($\overline{X} \pm \mathrm{SD}$) steht im Einklang mit diesen Ergebnissen zu IRT-basierten CATs in der Leistungs- (Hornke, 1999) und klinisch-psychologischen Diagnostik (Gardner et al., 2002; Reise & Henson, 2000; Simms & Clark, in Vorbereitung; Waller & Reise, 1989).

Die Itemersparnis kann natürlich auch zu *Zeit*- und *Kosteneinsparungen* führen, die von einigen Forschern (Butcher, 1987; Gregory, 1996; Hornke, 1993, 1996; Rose et al., 1999, 2003; Weiss & Vale, 1987) auf 15 – 80% geschätzt werden. Die Kosteneinsparungen wurden in vorliegender Studie nicht berechnet. Angesichts der im weiteren dargestellten hohen Item- und Zeiteinsparungen ist jedoch die Vermutung gerechtfertigt, dass nach einer einmaligen Anschaffungsgebühr (Soft- und Hardware IRT-basierter CATs), durch den Einsatz IRT-basierter CATs langfristig eine hohe Kostenreduktion erreicht werden kann, da sowohl laufende Materialkosten gesenkt, als auch Personal durch die Entlastung von diagnostischer Routinetätigkeit für anderweitige anspruchsvollere Tätigkeiten verfügbar wird.

Vergleicht man die Testbearbeitungszeit des Angst-CATs mit derjenigen eines etablierten Instruments wie beispielsweise des STAI, dessen durchschnittliche Bearbeitungsdauer zwischen 6 und 10 Minuten liegt (Laux et al., 1981), so ergibt sich eine durchschnittliche Zeitersparnis von 72 bis 86%, da psychosomatische Patienten durchschnittlich lediglich eine Minute und 40 Sekunden und gesunde Personen (N = 35 Studenten) eine Minute und 25 Sekunden zur Bearbeitung des Angst-CATs benötigen. Summieren sich solche Zeitersparnisse bei mehreren Instrumenten einer Testbatterie, so kann dies sowohl eine erhebliche *zeitliche* als auch *emotionale Entlastung* (i. S. einer Vermeidung von Langeweile, Überforderung oder Frustration etc.) für den Patienten und den Diagnostiker (z. B. auch durch ein direktes Ergebnis-

[115] Die Studie von Hornke (1999) untersuchte eine CAT-Version des Adaptiven Matrizentests (Leistungsdiagnostik).

[116] Die Studie von Koch und Mitarbeitern (1990) untersuchte eine CAT-Version zur Einstellungsmessung.

Feedback) bedeuten (siehe Kapitel 4.2.1. und 4.4.). Eine Erhöhung der Bearbeitungszeit, welche von Kubinger (1996) bei CATs aufgrund eines Wechsels der Iteminstruktionen und Antwortformate vermutet wird, die jedoch im Verlauf des CAT-Prozesses abnähme, konnte hier nicht festgestellt werden.

Die im Zusammenhang mit der Kürze von CATs aufzuwerfende Frage nach einem Informationsverlust, wird von den meisten Forschern auf diesem Gebiet mit Korrelationsstudien beantwortet, die darauf hinweisen, dass eine CAT-Version keinen wesentlichen Informationsverlust gegenüber einer „Vollversion" impliziere (z. B. Gardner et al., 2002; Hornke, 1993, 1996). Auch eine Simulations-Vorstudie (Walter et al., eingereicht) zur Erforschung eines möglichen Informationsverlusts beim Einsatz des Angst-CATs weist auf *keinen wesentlichen Informationsverlust* hin ($r_{\text{Angst-CAT / STAI-S}} = 0{,}97$). Da jedoch die Instrumente, welche in der Simulations-Vorstudie in einen korrelations-statistischen Zusammenhang gesetzt wurden, sich in ihrer Itemmenge überschneiden (das Angst-CAT enthält 15 Items der State-Angst-Skala des STAIs), sind weitere Belege gegen einen Informationsverlust in zukünftigen Studien zu erbringen, in denen sowohl die gesamte Itembank (N = 50 Items), des Angst-CATs als auch das Angst-CAT als adaptive Version erhoben werden und korrelationsstatistisch verglichen werden sollte.

Obgleich den meisten Patienten bei einer Bearbeitung des Angst-CATs nur wenige Items dargeboten werden, replizierte sich ein Befund, der sich bereits in einer Simulations-Vorstudie (Walter et al., eingereicht) zeigte. Es ist ein u-förmiger Zusammenhang zwischen der Merkmalsausprägung und der dargebotenen Itemzahl (siehe Kapitel 6.7.1.1.). Zur Schätzung der Angstausprägung in den *Extrembereichen* müssen aufgrund eines gewissen Mangels hoch diskriminativer Items in diesen Bereichen, den Testpersonen *mehr Items dargeboten* werden, wenn das „a priori" festgesetzte *Messgenauigkeitsniveau* ($\text{Rel}(\theta) = 0{,}9$) eingehalten werden soll. Inwiefern die angestrebte Messgenauigkeit in diesen Bereichen tatsächlich erreicht wird, bleibt zu erforschen. Der Befund steht im Einklang mit der bereits diskutierten Abhängigkeit des Standardmessfehlers und der Reliabilität vom Angstaus-prägungsniveau, wie er im Rahmen der Testentwicklung (zur IRT-Analyse siehe Kapitel 3.3.1., 5.3.2.2.2. und 5.4.2.2.2.) grafisch belegt wurde.

Mit der Möglichkeit der Offenlegung dieser Abhängigkeit der Messgenauigkeit von der Merkmalsausprägung und der *Kontrolle der Messgenauigkeit* durch die implementierte Stoppfunktion löst das hier entwickelte Angst-CAT einen der wesentlichsten Vorteile, welche sich mit IRT-basierter computergestützter adaptiver Messung verbindet, ein (siehe Kapitel 4.4.).

Im Hinblick auf die weitere Exploration der Messgenauigkeit in den Extrembereichen erscheint es sinnvoll, eine Studie zur Überprüfung der Reliabilität – auch im Sinne einer Veränderungsmessung, da das Angst-CAT ja intendiert, variable Angstzustände zu erfassen – zu planen, um unter anderem Messgenauigkeitseinbußen in den Extrembereichen eruieren, und durch eine gezielte Konstruktion und Kalibrierung neuer Items, welche in diesen Bereichen hoch diskriminativ sind, die Itembank des Angst-CATs optimieren zu können.

Bevor die weiteren Ergebnisse der Validierungsstudie des Angst-CATs diskutiert werden, stehen noch Aspekte, die für einen *CAT* – wie es hier entwickelte wurde – *spezifisch* sind, zur Diskussion.

7.5.2. CAT-spezifische Aspekte

Besonders zentral ist die computergestützte adaptive *Itemselektion*, welche die Anpassung der Items an das Fähigkeitsniveau der Testperson – mittels des Zugriffs auf eine in der Testentwicklungsphase kalibrierte Iteminformations-tabelle – gewährleistet. Die Itemselektion (siehe Kapitel 4.3.3.3.) erfolgte hier mittels des *Maximum-Information-Verfahrens* (MI) auf der Basis der *Fisher-Information*, da dies die zur Zeit der Testkonstruktion am häufigsten angewandte Methode der Itemselektion darstellte. Es liegen jetzt jedoch Hinweise dafür vor, dass das MI-Verfahren auf der Basis anderer Statistiken (z. B. Fisher-Intervall-Information oder Kullbach-Leibler Information, Cheng & Liou, 2000; Chen, Ankenmann & Chang, 2000) zumindest bei kürzeren Tests (< 10 Items) vorteilhafter sein könnte. Neben den MI-Verfahren existieren weitere Verfahren wie das Bayes'sche Sequentialverfahren (BE; Owen, 1969) zur Itemselektion, welches im Falle des Nutzens der „a posteriori"-Verteilung bei kurzen CATs (5-20 Items) mit einem geringeren durchschnittlichen Standardmessfehler als das MI-Verfahren behaftet sein soll (Meijer & Nering, 1999). Dieser Unterschied nivelliere sich jedoch mit zunehmender Testlänge. Vor diesem Hintergrund und angesichts der relativen Kürze des Angst-CATs wäre eine Erprobung der Itemselektion mit dem MI-Verfahren auf

der Basis anderer Statistiken oder des BE-Verfahrens in zukünftigen Studien sinnvoll.

Allgemein führt eine solche adaptive Itemselektionsstrategie zu einer interindividuell variablen Darbietung der Items und somit zu Unterschieden im: a) Itemset, b) der Itemreihenfolge und c) den Antwortformaten (siehe Kapitel 4.3.3.). Wird – wie hier angenommen – die IRT-Modellierung (mit dem GPCM) trotz kritisch diskutierter Fit-Statistiken für gültig erklärt, so dürfen (a) *Unterschiede im Itemset* wegen der Erfüllung der Stichproben-invarianzannahme (siehe Kapitel 3.3.1.) keine verzerrende Auswirkung auf die Item- und Personenparameterschätzung haben.

Inwiefern (b) die von Papier-und-Bleistift-Verfahren grundsätzlich verschiedene Itemdarbietung durch mögliche Itemreihenfolge/-p*ositions- bzw. ~kontexteffekte* die Validität der Item- und Personenparameterschätzung gefährdet, wird derzeit lebhaft diskutiert (Dahlstrom, Brooks & Peterson, 1990; Embretson & Reise, 2000; Knowles, 1988; Knowles et al., 1992; Knowles & Condon, 1999, 2000; Reise & Henson, 2000; Reise & Waller, 1990; Steinberg, 1994; Tourangeau & Rasinski, 1988). Knowles und Mitarbeiter (1988, 1992, 2000) fanden beispielsweise, dass ein Item bei einer frühen Darbietung im CAT-Prozess höher mit der endgültigen Personenparameterschätzung korreliere als bei einer späteren Darbietung. Dies erklären sie sich im Sinne einer „self-generated validity“ (Feldman & Lynch, 1988; Feldman, 1992), d. h. einer selbsterfüllenden Antworttendenz von Personen. Weiterhin zeigten sie z. B. bei der Untersuchung eines Instruments zur Erfassung von Angst (!) einen Itemschwierigkeits-Shift der Items in Abhängigkeit von der Itemposition („Windle Effect“; Windle, 1954). So reduzierte sich die Angst im Laufe des CAT-Prozesses, jedoch nur im Sinne einer abnehmenden spezifischen *Test*angst. Vor dem Hintergrund dieser Ergebnisse ist eine zukünftige Untersuchung von Kontexteffekten auf die Item- und Personenparameterschätung bei der Darbietung des Angst-CATs essentiell, da das Vorliegen von Kontexteffekten die in der IRT-Modellierung geforderte Annahme der lokalen stochastischen Unabhängigkeit verletze (siehe Kapitel 3.3.2.) und somit eine Gefahr für die Validität des CATs berge.

Welche Auswirkungen (c) der *Wechsel im Antwortformat*, der bei einem IRT-basierten CAT allgemein möglich ist, und auch im Angst-CAT vorliegt, auf die Item- und Personenparameterschätzung hat, ist ebenfalls diskussionswürdig.

Es kann sowohl vermutet werden, dass er die Datenqualität beeinträchtigt, da er eine höhere Konzentrationsleistung erfordert, und damit schneller zu Ermüdung führt, andererseits kann auch vermutet werden, dass er die Datenqualität verbessert, da er mechanischem Antwortverhalten und der Gefahr vorschnellen Antwortens – wie es von Hornke (1993) und Kubinger (1999) bei CAT-Versionen beobachtet wurde – entgegenwirkt (siehe Kapitel 4.2.2.).

Ebenfalls weiter zu erforschen ist der mögliche Einfluss der *Start*- und *Stoppfunktion* (Dodd et al., 1993; Thissen & Mislevy, 1990; Tonidandel, Quinones & Adams, 2002; siehe Kapitel 4.3.3.2./6.) und möglicher *Itemdarbietungskontrollen* (z. B. die unterschiedliche visuelle Gestaltung der Itemdarbietung, Möglichkeiten des Vor- bzw. Zurückblätterns, des Korrigierens[117] oder Auslassens von Items; siehe Kapitel 4.3.3.5.). Bei der Bearbeitung des Angst-CATs ist weder ein Vor- noch Zurückblättern noch eine Korrektur der Itemantwort durch die Testperson möglich, da ein dadurch möglicher Verwirrungseffekt (durch unterschiedliche Itemdarbietungen je nach Beantwortung) der Testpersonen vermieden werden sollte. Weiterhin war zur Vermeidung von „missing data“ das Auslassen der Bearbeitung von Items nicht möglich.

Neben der diskutierten Güte der gewählten *Itemselektionsstrategie* und deren potentiellen Folgen hängt die Qualität eines CATs auch maßgeblich von der Güte der *Personenparameterschätzung* ab (Theta-Schätzung; siehe Kapitel 4.3.3.4.).

Zur Personenparameterschätzung liegen eine Reihe von unterschiedlichen Schätzverfahren[118] vor, die jeweils spezifische Vor- und Nachteile haben. Die Theta-Schätzung im Angst-CAT erfolgt mittels des *Bayes'schen-Expected-A-Posteriori-Schätzverfahrens (EAP)*, da diesem einerseits eine bessere Testeffizienz als dem Weighted-Maximum-Likelihood- (WLE) oder dem

[117] Lunz, Bergstrom und Wright (1992) untersuchten den Einfluss des Zurückblätterns von Items innerhalb eines CATs (in der Leistungsdiagnostik) auf die Schätzung der Merkmalsausprägung und Testeffizienz und fanden, dass die Theta-Schätzungen von CATs mit vs. ohne Zurückblättern zu r = 0,98 korrelierten und das Zurückblättern zu einer Verbesserung der Testleistung von 1% führte.

[118] Zu den vier etablierten Personenparameterschätzverfahren: Maximum Likelihood Estimation (MLE), Weighted Maximum Likelihood Estimation (WLE), Expected A Posteriori Estimation (EAP), Maximum A Posteriori Estimation (MAP) siehe Kapitel 4.3.3.4..

Maximum-Likelihood-Schätzverfahren (MLE)[119] zugeschrieben wird, andererseits es unter vielen Bedingungen messgenauere Schätzungen erlaubt als das Maximum-A-Posteriori-Schätzverfahren (MAP) (Wang & Wang, 2001) und in der CAT-Anwendungsforschung bereits gut etabliert ist.

Kritisch beim EAP-Schätzverfahren ist allerdings einzuräumen, dass einerseits ein potentiell verzerrender Einfluss von der zur Schätzung genutzten „a priori" Verteilungsannahme ausgehen kann, der jedoch mit zunehmender Testlänge abnimmt (Cheng & Liou, 2000; Meijer & Nering, 1999), und andererseits dieses Schätzverfahren eine leichte Theta-Schätztendenz zur Mitte aufweist. Um diesen Verzerrungstendenzen zu begegnen, wurden mehrere neue Schätzverfahren entwickelt: das WLE-Schätzverfahren (Warm, 1989), welches zwar eine geringere Verzerrungstendenz, jedoch einen größeren Standardmessfehler als das EAP-Verfahren aufzuweisen scheint, das MAP- (Wang & Wang, 2001) und das EU-MAP Schätzverfahren („Essentially Unbiased Maximum Expected A Posteriori"; Wang, Hanson & Che-Ming, 1999). Das erste Verfahren (WLE) ist in Simulationsexperimenten von der Forschungsgruppe, in dessen Rahmen das Angst-CAT entwickelt wurde, bereits mit gutem Erfolg angewandt worden. Eine Simulation der anderen Schätzverfahren (MAP, EU-MAP) bzw. von Kombinationen dieser Schätzverfahren (Embretson & Reise, 2000) steht noch aus. Allgemein gelten alle Ansätze – im Falle der Gültigkeit der IRT-Modellierung – als konsistent und effektiv in ihrer Anwendung (Chen, 1997; Meijer & Nering, 1999; Nicewander & Thomasson, 1999) und erlauben eine hohe Messgenauigkeit bei der Theta-Schätzung. Da die verschiedenen Schätzverfahren in ihrer Theta-Schätzung mit zunehmender dargebotener Itemzahl konvergieren, scheint laut Wang und Wang (2001) weniger der spezifische Schätzalgorithmus sondern vielmehr die Stoppfunktion, welche die Testlänge determiniert, entscheidend zu sein.

Ein weiterer für CATs spezifischer Diskussionspunkt ist die Überprüfung der *Äquivalenz* zwischen *Papier-und-Bleistift-Verfahren*, *computergestützten Tests* und *CATs*. Diese wird von vielen Forschern gefordert (Schwenkmezger & Hank, 1993), da vermutet wird, dass sich sowohl Item- als auch Personenparameterschätzungen je nach Erhebungsmodus unterscheiden (Hetter, Segall & Bloxom, 1994). Eine Äquivalenzprüfung ist bei der Entwicklung eines CATs

[119] Dem MLE wird eine Schätztendenz zu den Extremen zugeschrieben (Lord, 1983).

von Belang, da die Kalibrierung der dem CAT zugrundegelegten Itemparameter meist auf *konventionell* erhobenen Testdaten beruht (Papier-und-Bleistift-Verfahren). Hier befindet sich die vorliegende Studie in der günstigen Lage, dass die Itemparameterschätzung auf der Basis von bereits *computergestützt* erhobenen Daten erfolgen konnte, da in der psychosomatischen Klinik, in der das Angst-CAT entwickelt wurde, die psychometrische Diagnostik bereits seit 1990 computergestützt erfolgt, d. h. jede Frage auf dem Bildschirm eines Handcomputers gesondert dargestellt wird (Rose et al., 1999; siehe Kapitel 5.2.1.). Diese „Item-by-Item"-Präsentation ist mit derjenigen im späteren CAT-Prozess identisch. Embretson und Reise (2000) machen übrigens bei dieser Art der Präsentation darauf aufmerksam, dass dies die Gefahr des „Verrutschens" in der Antwortkategorie oder Itemtextzeile, welche bei Papier-und-Bleistift-Verfahren gegeben ist, reduziere.

Da das Angst-CAT nicht in *Papier-und-Bleistift-Form* vorliegt, stellt sich hier auch nicht die viel diskutierte Frage nach einer klassischen Äquivalenz-überprüfung (Embretson & Reise, 2000, S. 265). Die Äquivalenzprüfung ist vor allem bei der Entwicklung von CAT-Versionen bereits etablierter Papier-und-Bleistift-Verfahren wie z. B. der IRT-basierten CAT-Version des NEO-Pis (Reise & Henson, 2000) oder der „Countdown-Strategie-basierten" CAT-Version des MMPIs (Handel et al., 1999; siehe Kapitel 4.3.2.) wichtig. Diese Autoren fanden, dass die Item- und Skalenmittelwerte von State-Inventaren (z. B. STAI und STÄI)[120] bei einer computergestützten Datengewinnung höher ausfielen als bei der Papier-und-Bleistift-Vorgabe; die Trennschärfen, Reliabilitäten, Verteilungs-formen und Skaleninterkorrelationen jedoch keine Unterschiede zwischen den unterschiedlichen Erhebungsmodi aufwiesen. Da das Angst-CAT die Erfassung der State-Angst intendiert, ist dieses Ergebnis unter Umständen beim Vergleich der Theta-Werte des Angst-CATs mit den Angstsummenscores etablierter Instrumente zu beachten. Bei der Entwicklung des Angst-CATs erscheint vor allem eine Äquivalenzprüfung zwischen dem CAT und der gesamten Itembank, wie sie in Simulationsstudien bereits mit guten Ergebnissen erfolgte, deren Replikation an realen Daten jedoch noch aussteht, zentral.

[120] STÄI = State-Trait-Ärgerausdrucks-Inventar (Schwenkmezger, Hodapp & Spielberger, 1992).

7.5.3. Konvergente und diskriminante Validität

Der Vergleich IRT-basierter Angst-CAT-Scores (Theta-Werte) mit Summenscores konventioneller Angstinventare fand im Rahmen der *Validierungsstudie* statt, welche sich an die Entwicklung des Angst-CATs (siehe Kapitel 5) anschloss und deren Ergebnisse (siehe Kapitel 6) nun näher diskutiert werden.

Die Untersuchung der Abhängigkeit der Theta-Werte von *soziodemografischen* Variablen ergab, dass weder das Geschlecht, noch das Alter oder der Familienstatus signifikant zur Varianzaufklärung beitragen (siehe Kapitel 6.6.1.2.). Allerdings weisen die Altersgruppe der 26-35-Jährigen und die der über 75-Jährigen durchschnittlich leicht geringere Theta-Werte als sonstige Altersgruppen im Angst-CAT auf.

Die Untersuchung der *konvergenten Validität* ergab mittelmäßig bis hohe *Korrelationen zu anderen Angstinventaren* (BAI, HADS-A; r = 0,51* bis r = 0,76*, siehe Kapitel 6.7.2.). Die Korrelations*höhe* ist nach Lienert und Raatz (1994), welche erörtern, dass man in der Praxis mit signifikanten Validitätskoeffizienten von r > 0,6 „sehr zufrieden“ sein könne und – sich die an die Höhe des Validitätskoeffizienten gestellten Anforderungen bei der Nutzung von weiteren klinischen Informationen zur diagnostischen Beurteilung in der Praxis auf r > 0,5 reduzierten – als *gut* einzuschätzen.

Diese gute konvergente Validität ist vor allem vor dem Hintergrund der relativen Kürze des Angst-CATs hervorzuheben, da in der KTT eine Testverkürzung häufig auch mit Reliabilitäts- und Validitätseinbußen einhergeht. Hier gilt, dass sich die Validität eines Tests umgekehrt proportional zu seiner Ökonomie verhält (Lienert & Raatz, 1994), d. h. je länger ein Test ist, desto höheren Ansprüchen an die Höhe des Validitätskoeffizienten sollte er genügen, oder umgekehrt ein CAT muss nicht extrem hohe Validitätskoeffizient aufweisen, um als valide zu gelten, da er relativ kurz ist.

Die Höhe der in der vorliegenden Studie ermittelten konvergenten Validitäts-koeffizienten steht im Einklang mit der Höhe von Validitätskoeffizienten etablierter Angstinventare (BAI / STAI / HADS; r = 0,45 bis r = 0,86) in anderen Validierungsstudien (Margraf & Ehlers, in Druck; Hinz & Schwarz, 2001; siehe Kapitel 6.5.1) und ist damit als *sehr gut* zu beurteilen.

Interessant ist, dass die Theta-Werte des Angst-CATs höher mit der Angstskala des HADS als mit dem BAI korrelieren. Dies erklärt sich durch den unterschiedlichen Messbereich dieser Instrumente. Während das BAI eher für akute Panikzustände charakteristische vegetative Angstsymtpome erfasst, intendiert das Angst-CAT die Messung einer aktuellen objektübergreifenden, generalisierten Zustands-Angst weitgehend *ohne vegetative* Begleitsymptome.

Weitere Belege für die konvergente Validität des Angst-CATs ergaben sich bei der Analyse der mit dem Angst-CAT ermittelten durchschnittlichen Theta-Werte verschiedener *diagnosenspezifischer Gruppen*. Patienten mit einer diagnostizierten Angststörung (F.40/41) wiesen im Vergleich zu Patienten ohne psychische Störung bzw. gesunden Studenten durchschnittlich signifikant höhere Theta-Werte auf (QS = 41,35; df = 2; $\overline{QS}$ = 20,76; F = 35,58; $p \leq 0{,}001$), d. h. es liegt eine relative diagnosenspezifische Konvergenz zwischen dem Angst-CAT und einer mit einem strukturierten computergestützten klinischen Interview (M-CIDI; siehe Kapitel 2.7.1. und 6.5.3.) erhobenen klinischen Diagnose einer Angststörung vor.

Um die *diskriminante Validität* zu untersuchen, wurden das Konstrukt der Depression und verschiedene Persönlichkeitskonstrukte (Neurotizismus, Extraversion etc.) psychometrisch erfasst (siehe Kapitel 6.7.3.).

Die *Diskrimination* zwischen den Konstrukten *Angst* und *Depression* gestaltet sich – wie theoretisch erwartet – mit dem Angst-CAT ähnlich wie bei anderen Angstinventaren schwierig (STAI; siehe Kapitel 2.5.; BAI; HADS; siehe Kapitel 6.5.). Der enge Zusammenhang zwischen den Konstrukten der Angst und der Depression wird sowohl konzeptionell von einer Reihe von Forschern modelliert (Clark & Watson, 1991; Krueger & Finger, 2001; Mineka et al., 1998; Watson et al., 1984, 1995) als auch im Sinne einer diagnostischen Komorbidität (Neumer 2000, S. 53: 14,6-45,9%; DSM-IV, Saß et al., 1996: 50-65%) bzw. Überlappung von Symptomen (Garber et al., 1980) vielfach diskutiert (siehe Kapitel 2.5.), so dass es nicht erstaunt, dass eine gute psychometrische Differenzierung zwischen Angst und Depression mit dem Angst-CAT nicht gelingt. Einige Autoren erklären dies damit, dass diesen Konstrukten ein gemeinsamer globaler Faktor, der je nach Forschergruppe „negative Affektivität“ (Watson & Clark, 1984), „negative Emotionalität“ (Tellegen & Waller, 2001), „internalizing factor“ (Krüger & Finger, 2001) oder „general

neurotic syndrome“ (Andrews, Stewart, Morris-Yates, Holt & Henderson, 1990; Andrews, 1996) genannt wird, zugrunde liege. Letzterer Faktorenname deutet bereits auf den nächsten Befund hin: erwartungsgemäß gelingt dem Angst-CAT in Einklang mit anderen etablierten Angstinventaren eine Diskrimination zum Eigenschaftskonstrukt „*Neurotizismus*“ ebenfalls nicht.

An dieser Stelle sei auf den engen Zusammenhang zwischen dem Cattell'schen Konstrukt der „Ängstlichkeit“ (Cattell & Scheier, 1960) und dem Eysenck'schen Faktor „Neurotizismus“ (Eysenck, 1947) und der Uneinigkeit in Forscherkreisen, ob Ängstlichkeit und Neurotizismus ähnliche oder sogar identische Persönlichkeitskonstrukt nur auf unterschiedlichen Abstraktionsniveaus darstellen, hingewiesen. Somit wird die konzeptuelle Trennung zwischen einer Eigenschafts- (Trait-) und einer Zustands-Angst (State) – wie sie im State-Trait-Modell der Angst formuliert wird (Spielberger, 1972; siehe Kapitel 2.4.1.1.) – mit vorliegenden Befund der mangelnden Diskrimination zwischen einer State- (hier: Angst-CAT) und einer Trait-Angst (hier: Neurotizismus-Skala des NEO-FFIs) – erneut in Frage gestellt. Dies steht im Einklang mit anderen Studien, die eine mangelnde Differenzierung zwischen einer State- und Trait-Angst belegen (Endler et al., 1976; Hermann et al., 1991; Spielberger, 1972; Steyer et al., 1999; siehe Kapitel 2.4.1.).

Manche Autoren (Eysenck & Eysenck, 1985; Gray, 1981) konzipieren Ängstlichkeit auch als eine Kombination aus Neurotizismus und *niedriger Extraversion*. Diese Überlegung ist konform mit dem Befund der vorliegenden Validierungsstudie, dass nicht nur die psychometrische Diskrimination zum Konstrukt Neurotizismus schwierig ist, sondern dadurch begründet auch die Diskrimination zu sozialen Skalen (NEO-FFI: Extraversion, GT: Soziale Resonanz, Soziale Potenz) reduziert wird (siehe Kapitel 6.6.3.1.2.).

Abgesehen von dem erwartungsgemäß geringen Diskriminationsvermögen des Angst-CATs bezüglich der Konstrukte Depression und Neurotizismus, offenbarte sich insgesamt eine *gute diskriminante Validität* des Angst-CATs bezüglich einer Vielzahl *anderer Eigenschaftskonstrukte* (NEO-FFI: Offenheit, Verträglichkeit; GT: Dominanz, Zwanghaftigkeit, allgemeine Grundstimmung etc.).

Weitere Belege für die *diskriminante Validität* des Angst-CATs ergaben sich bei der Analyse der mit dem Angst-CAT ermittelten durchschnittlichen Theta-Werte

verschiedener *Diagnosegruppen*. Patienten mit einer diagnostizierten Angststörung (F.40/41) bzw. depressiven Störung (F.32-34) wiesen im Vergleich zu Patienten mit somatoformen Störungen (F.45) oder Essstörungen (F.50) signifikant höhere Theta-Werte auf (QS = 30,07; df = 4; $\overline{QS}$ = 7,52; F = 14,50; $p \leq 0{,}001$). Obgleich das Angst-CAT nicht zur diagnosenspezifischen Diskrimination (zur Angst als Störung siehe Kapitel 2.6.) entwickelt wurde, ist eine Differenzierung zwischen verschiedenen Diagnosegruppen tendenziell möglich – jedoch nur bei Patienten, welche keine Komorbidität (mit Angststörungen) aufweisen. Das Angst-CAT sollte folglich stets im Zusammenhang weiterer klinischer Diagnostik *interpretiert* werden.

Dies wirft einen weiteren Diskussionspunkt auf: die Interpretation und Kommunikation der Theta-Werte des Angst-CATs. Embretson und Reise (2000) sehen in der Möglichkeit einer *iteminhaltsbezogenen Interpretation* der Theta-Werte (siehe Kapitel 3.3.3.) eine informationsreiche Ergänzung zur normbezogenen Interpretation von Testwerten wie sie in der KTT üblich ist. Wie sich eine solche inhaltsbezogene Interpretation der Theta-Werte (hier: des Angst-CATs) pragmatisch umsetzen lässt, ist bislang jedoch noch wenig erforscht.

7.6. Zusammenfassung und Ausblick

Da die vorliegende Arbeit über die Entwicklung und Validierung eines auf der Grundlage der Item Response Theorie (IRT) realisierten computergestützten adaptiven Tests zur Angstmessung (Angst-CAT) im deutschen Sprachraum als eine *klinisch-psychologische Pionierarbeit* angesehen werden kann (siehe Kapitel 3.5.2.), wurden im Diskussionsteil eine Reihe von Fragen aufgeworfen, welche angesichts des jungen Forschungsstandes offen bleiben müssen.

Es lässt sich resümieren, dass das Angst-CAT als ein IRT-basierter computergestützter adaptiver Test eine methodische Fortentwicklung der rein computergestützten Versionen etablierter Angstinventare (siehe Kapitel 4.2.4.) darstellt. Sowohl die Itembankentwicklung als auch die Itemselektion und Personenparameterschätzung des Angst-CATs erfolgte *IRT-basiert*, so dass sich eine Reihe von theoretisch erwarteten Vorteilen einlösen ließen.

So erwies sich das Angst-CAT sowohl in Simulationsexperimenten einer Vorstudie als auch in der hier dargestellten Validierungsstudie als ein *kurzes, messpräzises Screening-Instrument* zur Messung einer objekt- und situations-

übergreifenden *aktuellen Zustands-Angst*. Es ermöglicht eine mobile, ökonomische und messgenaue Erfassung der Angstausprägung, indem es Testpersonen nur die Items darbietet, die ihrem individuellen Angstausprägungsniveau optimal entsprechen. Die durch einen adaptiven Itemselektionsalgorithmus realisierte *Reduktion der dargebotenen Itemzahl* vermag die *psychodiagnostische Belastung* der Testpersonen und Diagnostiker zu *reduzieren*, sowie zu erheblichen *Zeit-* und *Kosteneinsparungen* beizutragen. Inwieweit diese Vorteile zu einer positiven Rezeption und gegebenenfalls Verbreitung des Angst-CATs oder der weiteren Erforschung und Entwicklung IRT-basierter CATs in der klinisch-psychologischen Diagnostik führen, hängt maßgeblich von der Einstellung der Anwender zur *IRT* und zur *Computerdiagnostik* ab und bleibt abzuwarten (Gitzinger, 1990). Hier gilt es - falls sich die auf der Forschungsebene bereits etablierte Erkenntnis von den Potentialen IRT-basierter Methoden und CAT-Verfahren auch in der Praxis durchsetzen möchte – Unsicherheiten ob des Nutzens der IRT in diesem Bereich (verglichen mit dem Bereich der Leistungsdiagnostik; siehe Kapitel 3.5.) durch eine vermehrte Forschungstätigkeit, und technokratischer Skepsis gegenüber Computerdiagnostik (siehe Kapitel 4.2.2./3.) durch offene, transparente Kommunikation zwischen Forschern und Anwendern zu begegnen. Dieses ist, gerade weil sich die IRT-Modellierung und CAT-Entwicklung von Persönlichkeitsskalen – wie es Chernyshenko und Mitarbeiter (2001) in einem Überblicksartikel zusammenfassen und wie es auch vorliegende Studie belegt – komplizierter gestaltet als vermutet, von zentraler Bedeutung.

Die IRT ist kein Wundermittel, welches alle testtheoretischen Probleme, die im Rahmen der KTT aufgeworfen werden, zu lösen vermag. Langfristig liegt wohl – wie viele Autoren in jüngster Zeit betonen (Embretson & Hershberger, 1997; Embretson & Reise, 2000; Rost, 1999; Verstralen et al., 2001) – im *kombinierten* Gebrauch bewährter KTT-basierter und neuer, innovativer IRT- und CAT-Methoden die Chance, die klinisch-psychologische Diagnostik zu verbessern und in ihren Möglichkeiten zu erweitern.

8. Literatur

Abramson, L. Y., Seligman, M. E. P. & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. Journal of Abnormal Psychology, 87, 49-74.

Allport, G. W. & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. Psychological Monographs, 47, 211.

Amelang, M. & Bartussek, D. (2001). Differentielle Psychologie und Persönlichkeitsforschung (5. Aufl.). Berlin: Kohlhammer-Verlag.

Amelang, M. & Zielinski, W. (1996). Psychologische Diagnostik und Intervention (2. Aufl.). Berlin: Springer-Verlag.

American College Testing (ACT, 1993). Collegiate assessment of academic proficiency writing skill tests. Iowa City, I.A.: Authors.

American Psychological Association (APA; 1986). Guidelines for computer-based tests and interpretations. Washington D.C.: Authors.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andrews, G. (1996a). Comorbidity and the general neurotic syndrome. British Journal of Psychiatry, 168, 76-84.

Andrews, G. (1996b). Comorbidity in neurotic disorders: The similarities are more important than the differences. In R.M. Rapee (Ed.), Current controversies in the anxiety disorders (pp. 3-20). New York: Plenum.

Andrews, G., Stewart, G., Morris-Yates, A., Holt, P. & Henderson, S. (1990). Evidence for a general neurotic syndrome. British Journal of Psychiatry, 157, 6-12.

Andrich, D. (1978). Application of a psychometric model to ordered categories which are scored with successive integers. Applied Psychological Measurement, 2, 581-594.

Angleitner, A., Ostendorf, F. & John, O. P. (1990). Towards a taxonomy of personality descriptors in German: A psycho-lexical study. European Journal of Personality, 4, 89-118.

Arbeitsgemeinschaft für Methodik und Dokumentation in der Psychiatrie (AMDP; 1997). Das AMDP-System. Manual zur Dokumentation psychiatrischer Befunde (6. Aufl.). Göttingen: Hogrefe.

Arbuckle & Worthke (1999). Amos. User's Guide (Version 4.0). Chicago: Small Waters Cooperation.

Barlow, D. H., Chorpita, B. F. & Turovsky, J. (1996). Fear, panic, anxiety and disorders of emotion. Nebrasca Symposium of Motivation, 43, 251-328.

Battegay, R. (1970). Angst und Sein. Stuttgart: Hippokrates Verlag.

Beck, A. T. (1994). Beck-Depression-Inventory: BDI. Toronto: Huber.

Beck, A. T. & Steer, R.A. (1993). Beck Anxiety Inventory: BAI. San Antonio: The Psychological Cooperation.

Becker, C. (1997). Interaktions-Angst-Fragebogen: IAF (3. Aufl.). Göttingen: Beltz-Verlag.

Becker, J., Walter, O. B., Fliege, H., Bjorner, J., Ravens-Sieberer, U., Walter, M., Klapp, B. F. & Rose, M. (2002). Using the item response theory to develop a computer adaptive test for anxiety. Quality of Life Research, 11, 670.

Becker, J., Walter, O. B., Fliege, H., Klapp, B. F. & Rose, M. (submitted). Using item response theory to develop a Computerized Adaptive Test (CAT): Anxiety-CAT. Psychological Assessment.

Beckmann, D., Brähler, E. & Richter, H. E. (1991). Der Gießen-Test: GT. Ein Test für Individual- und Guppendiagnostik. Bern: Huber.

Beckmann, J. F. & Guthke, J. (1999). Psychodiagnostik des schlussfolgernden Denkens. Handbuch zur Adaptiven Computergestützten Intelligenz-Lerntestbatterie für Schlussfolgendes Denken: ACIL. Göttingen: Hogrefe.

Ben-Porath, Y. S., Slutske, W. S. & Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the MMPI. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1, 18-22.

Benesch, H. (1995). Enzyklopädisches Wörterbuch – Klinische Psychologie und Psychotherapie. Weinheim: Beltz-Verlag.

Benson, J., Moulin-Julian, M., Schwarzer, C., Seipp, B. & El-Zahhar, N. (1992). Cross-validation of a revised test anxiety scale using multi-national samples. In K.A. Hagtvet (Ed.), Advances in test anxiety research (pp. 62-83). Lisse, Niederlande: Swets & Zeitlinger.

Bentler, P.M. (1990). Comparative fit indexes in structural equation models. Psychological Bulletin, 107, 238-246.

Bentler, P.M. & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Billings, A. G. & Moos, R. H. (1984). Coping, stress and social resources among adults with unipolar depression. Journal of Personality and Social Psychology, 46, 877-891.

Binet, A. (1909). Les idées modernes sur les enfants. Paris: Ernest Flammarion.

Birbaumer, N. & Schmidt, R. F. (1996). Biologische Psychologie (3. Aufl.). Berlin: Springer-Verlag.

Birbaumer, N., Tunner, W., Hölzl, R. & Mittelstädt, L. (1973). Ein Gerät zur kontinuierlichen Messung subjektiver Veränderung. Zeitschrift für experimentelle und angewandte Psychologie, 20, 173-181.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), Statistical theories of mental test scores. Reading. MA: Addison-Wesley.

Bjorner, J. B., Kosinski, M. & Ware, J. E. (2003). The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. Quality of Life Research, 12, 887-902.

Bloom, B. L. (1992). Computer-assisted psychological intervention: A review and commentary. Clinical Psychology Review, 12, 160-197.

Bock, R. D., Gibbons, R. & Muraki, E. J. (1988). Full information item factor analysis. Applied Psychological Measurement, 12, 261-180.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. Applied Psychological Measurement, 12, 261-280.

Bock, R. D. & Mislevy, R. J. (1988). Comprehensive educational assessment for the states: The duplex design. Educational Evaluation and Policy Analysis, 10, 89-105.

Börner, R.J., Gülsdorff, Margraf, J., Osterheider, M., Philipp, M. & Wittchen, H.-U. (1997). Die Panikstörung – Diagnose und Behandlung. Stuttgart: Schattauer-Verlag.

Bond, T. G. & Fox, C. M. (2001). Applying the Rasch model. Mahwah, N.J.: Lawrence Erlbaum.

Borkenau, P. & Ostendorf, F. (1993). NEO-Fünf-Faktoren Inventar: NEO-FFI. Göttingen: Hogrefe.

Bortz, J. & Döring, N. (1995). Forschungsmethoden und Evaluation (2. Aufl.). Berlin: Springer-Verlag.

Bortz, J. (1999). Statistik für Sozialwissenschaftler (5. Aufl.). Berlin: Springer-Verlag.

Bouman, T. K. & Kok, A. R. (1987). Homogeneity of Beck's depression inventory (BDI): Applying Rasch analysis in conceptual exploration. Acta Psychiatrica Scandinavica, 76, 573.

Bös, K. & Mechling, H. (1985). Bilder-Angst-Test für Bewegungssituationen: BAT. Göttingen: Hogrefe.

Brähler, E., Holling, H., Leutner, D. & Petermann, F. H. (2002). Brickenkamp. Handbuch psychologischer und pädagogischer Tests. Göttingen: Hogrefe.

Brähler, E. & Richter, H. E. (2000). Das psychologische Selbstbild der Deutschen im Gießen-Test zur Jahrhundertwende. In O. Decker & E. Brähler (Hrsg.), Deutsche – 10 Jahre nach der Wende (S. 47-51). Gießen: Psychosozial-Verlag.

Brähler, E. & Scheer, J. W. (1995). Gießener Beschwerdebogen: GBB. (2. Aufl.). Bern: Huber.

Brähler, E., Schumacher, J. & Brähler, C. (1999). Erste gesamtdeutsche Normierung und spezifische Validitätsaspekte des Gießen-Tests. Zeitschrift für Differentielle und Diagnostische Psychologie, 20, 231-243.

Breuer, J. & Freud, S. (1960). Studies on hysteria. Oxford: Beacon. (Original erschienen 1895).

Brown, M.W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In: K.A. Bollen & J.S. Long (Eds.), Testing structural equation models (pp.136-162). Newbury Park, CL: Sage.

Brown, M.W. & Mels, G. (1992). RAMONA user's guide. The Ohio State University: Department of Psychology.

Brown, J. M. & Weiss, D. J. (1977). An adaptive testing strategy for achievement test batteries (Research Report No. 77-6). Minneapolis: University of Minnesota, Psychometric Methods Program.

Brown, T. A., Chorpita, B. F. & Barlow, D. H. (1997). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative effect, positive affect and autonomic arousal. Journal of Abnormal Psychology, 107, 2, 179-192.

Bullinger, M. & Kirchberger, I. (1998). SF-36 Fragebogen zum Gesundheitszustand MOS Short-Form-36 Health Survey. Göttingen: Hogrefe. (Original erschienen 1993: SF-36; Ware, J.E., Snow, K.K., Kosinski, M. & Gandek, B.).

Bullinger, M., Kirchberger, I. & Steinbüchel, N. V. (1993). Der Fragebogen Alltagsleben – Ein Verfahren zur Erfassung der gesundheitsbezogenen Lebensqualität. Zeitschrift für Medizinische Psychologie, 2, 121-131.

Butcher, J. N. (1987). Computerized psychological assessment. New York: Basic Books.

Butcher, J. N., Keller, L. S. & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. Journal of Consulting and Clinical Psychology, 53, 803-815.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S. & Kämmer, B. (1992). Manual for administration, scoring, and interpretation of the Minnesota Multiphasic Personaltiy Inventory for Adolescents: MMPI-A. Minneapolis: University of Minnesota Press.

Byrne, D. (1961). The repression-sensitization scale: rationale, reliability and validity. Journal of Personality, 29, 334-349.

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

Cannon, W. B. (1975). Wut, Hunger, Angst & Schmerz. Eine Physiologie der Emotionen. Berlin: Urban & Schwarzenberg.

Carstensen, C. H. (2000). Mehrdimensionale Testmodelle mit Anwendungen aus der pädagogisch-psychologischen Diagnostik. Dissertation, Kiel: Universität Kiel.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. Journal of Abnormal and Social Psychology, 38, 426-506.

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate Behavioral Research, 1, 245-276.

Cattell, R. B. (1974). How good is the modern questionnaire? General principles for evaluation. Journal of Personality Assessment, 38, 115-129.

Cattell, R. B. & Scheier, I. H. (1960). Handbook for the Objective Analytic (O-A) anxiety battery. Champaign, IL.: Institute for Personality and Ability Testing.

Cattell, R. B. & Scheier, I. H. (1963). Handbook for the IPAT anxiety scale questionnaire. Champaign, IL.: Institute for Personality and Ability Testing.

Cella, D. & Chang, C.-H. (2000). A discussion of item response theory and its applications in health status assessment. Medical Care, 38, (2), 66-72.

Cella, D. & Nowinski, C.J. (2002). Measuring quality of life in chronic illness: the functional assessment of chronic illness therapy measurement system. Archive of Physical and Medical Rehabilitation, 83, 12, (2), 10-17.

Chang, C.-H. & Reeve, B. B. (2003). Item response theory (IRT) modeling and its applications to health outcomes measurement. Workshop at the conference of the international society for quality of Life Research, Orlando, FL.

Chen, S.-K. (1997). A comparison of maximum likelihood estimation and expected a posteriori estimation in computerized adaptive testing using the generalized partial credit model. Dissertation, Austin, TX.: University of Texas.

Chen, S.-K., Ankenmann, R. D. & Chang, H.-H. (2003). A comparison of item selection rules at the early stages of computerized adaptive testing. Applied Psychological Measurement, 24, 241-255.

Cheng, P. E. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. Applied Psychological Measurement, 24, 257-265.

Chernyshenko, O. S., Stark, S., Chan, K.-Y., Drasgow, F. & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. Multivariate Behavioral Research, 36, 523-562.

Childs, R. A. & Chen, W.-H. (1999). Obtaining comparable item parameter estimates in Multilog and Parscale for two polytomous IRT models. Applied Psychological Measurement, 23, 371-379.

Childs, R. A., Dahlstrom, W. G., Kemp, S. M. & Panter, A. T. (2000). Item response theory in personality assessment: A demonstration using the MMPI-2 depression scale. Psychological Assessment, 7, 37-54.

Chorpita, B. F., Albano, A. M. & Barlow, D. H. (1998). The structure of negative emotions in a clinical sample of children and adolescents. Journal of Abnormal Psychology, 107, 74-85.

Clark, L. A. (1993). Schedule for nonadaptive and adaptive personality (SNAP). Manual for administration, scoring and interpretation. Minneapolis: University of Minnesota Press.

Clark, L. A. & Watson, D. (1991). Tripartite model of anxiety and depression: Evidence and taxonomic implications. Journal of Abnormal Psychology, 103, 3-16.

Cliff, N. (1988). The eigenvalue greater than one rule and the reliability of components. Psychological Bulletin, 103, 276-279.

Colby, K. M., Watt, J. B. & Gilbert, J. P. (1966). A computer method of psychotherapy: Preliminary communication. Journal of Nervous and Mental Disease, 142, 148-152.

College Board (1993). Coordinator's notebook for the computerized placement test. Princeton, N.Y.: Educational Testing Service.

Collegium Internationale Psychiatriae Scalarum (CIPS; 1996). Internationale Skalen für Psychiatrie (4. Aufl.). Göttingen: Beltz-Verlag.

Cook, L. L., Eignor, D. R. & Taft, H. L. (1984). A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates. Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Cooke, D. J., Kosson, D. S. & Michie, C. (2001). Psychopathy and ethnicity: Structural, item and test generalizability of the Psychopathy Checklist-Revised (PCL-R) in Caucasian and African American Participants. Psychological Assessment, 13, 531-542.

Cooke, D. J. & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist – Revised. Psychological Assessment, 9, 3-14.

Cooke, D. J., Michie, C., Hart, S. D. & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist – Revised (PCL): An item response theory analysis. Psychological Assessment, 11, 3-13.

Costa, P. T. & McCrae, R. R. (1985). The NEO Personality Inventory: NEO-PI. Odessa: Psychological Assessment Resources.

Cronbach, L. J. (1990). Essentials of psychological testing (5th ed.). New York, N.Y.: Harper Collins.

Curran, L. T. & Wise, L. L. (1994). Evaluation and implementation of CAT-ASVAB. Paper presented at the annual meeting of the American Psychological Association (APA), Los Angeles.

Dahlstrom, W. G., Brooks, J. D. & Peterson, C. D. (1990). The Beck Depression Inventory: Item order and the impact of response sets. Journal of Personality Assessment, 55, 224-233.

Darwin, C. (1965). The expression of the emotions in man and animals. Chicago: University of Chicago Press.

De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. Educational and Psychological Measurement, 49, 789-805.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. Applied Psychological Measurement, 16, 327-343.

De Beer, M. (2001). The construction and evaluation of a dynamic computerized adaptive test for the measurement of learning potential. Dissertation, Johannesburg: University of South Africa.

De Koning, E., Sijtsma, K. & Hamers, J. H. M. (2002). Comparison of four IRT models when analyzing two tests for inductive reasoning. Applied Psychological Measurement, 26, 302-320.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B, 1-38.

Deneke, F. W. & Hilgenstock, B. (1989). Narzissmus-Inventar: NI. Bern: Huber.

Dilling, H., Mombour, W. & Schmidt, M. H. (2000). Internationale Klassifikation psychischer Störungen. ICD-10 Kapitel V (F). Klinisch-diagnostische Leitlinien (3. Aufl.). Bern: Huber.

Dodd, B. D. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. Applied Psychological Measurement, 14, 355-366.

Dodd, B. D., De Ayala, R. J. & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. Applied Psychological Measurement, 19, 5-22.

Dodd, B. D., Koch, W. R. & De Ayala, R. J. (1988). Computerized adaptive attitude measurement: A comparison of the graded response and rating scale models. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Dodd, B. D., Koch, W. R. & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. Applied Psychological Measurement, 13, 129-143.

Dodd, B. D., Koch, W. R. & De Ayala, R. J. (1993). Computerized adaptive testing using the partial credit model: Effects of item pool characteristics and different stopping rules. Educational and Psychological Measurement, 53, 61-77.

Dorans, N. J. & Kingston, N. M. (1985). The effect of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE Verbal scale. Journal of Educational Measurement, 22, 249-262.

Drasgow, F. & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. Journal of Applied Psychology, 68, 363-373.

Educational Testing Service (ETS; 1996). Graduate Record Examinations (GRE) 1996-1997: Information and registration Bulletin. Princeton, N.J.: Author.

Eggert, D. (1983). Eysenck-Persönlichkeitsinventar: EPI. Göttingen: Hogrefe.

Ellis, B. B., Becker, P. & Kimmel, H. D. (1989). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). International Journal of Psychology, 24, 665-684.

Embretson, S. E. (1992). Computerized adaptive testing: Its potential substantive contributions to psychological research and assessment. Current Directions in Psychological Science, 4, 129-131.

Embretson, S. E. (1996). The new rules of measurement. Psychological Assessment, 8, 341-349.

Embretson, S. E. & Hershberger, S. L. (1997). The new rules of measurement. What every psychologist and educator should know. Mahwah, N.J. : Lawrence Erlbaum Associates.

Embretson, S. E. & Reise, S. P. (2000). Item response theory for psychologists. London: Lawrence Erlbaum Associates.

Endler, N. S., Edwards, J. M. & Vitelli, R. (1991). Endler Multidimensional Anxiety Scales: EMAS. Los Angeles, C.A.: Western Psychological Services.

Endler, N. S., Hunt, J. M. & Rosenstein, A. J. (1962). An S-R inventory of anxiousness. Psychological Monographs: General and Applied, 76.

Endler, N. S., Magnusson, D., Ekehammar, B. O. & Okada, M. (1976). The multidimensionality of state and trait anxiety. Scandinavian Journal of Psychology, 17, 81-96.

Ettrich, K.-U., Krauss, H. & Sandau, T. (1992). Analysen zur Geburts-Angst-Skala (GAS-R) des Projektes Kinderwege (Forschungsbericht 2/92). Leipzig: Universität Leipzig, Fachbereich Psychologie.

Everett, J. E. (1983). Factor comparability as a means of determining the number of factors and their rotation. Multivariate Behavioral Research, 18, 197-218.

Eysenck, H. J. (1947). Dimensions of personality. London: Routledge.

Eysenck, H. J. & Eysenck, M. W. (1985). Personality and individual differences. New York: Plenum Press.

Fahrenberg, J. (1967). Physiologische Persönlichkeitsforschung. Göttingen: Hogrefe.

Fahrenberg, J. (1994). Ambulantes Assessment. Computerunterstützte Datenerfassung unter Alltagsbedingungen. Diagnostica, 40, 195-216.

Fahrenberg, J., Hampel, R. & Selg, H. (1989). Das Freiburger Persönlichkeitsinventar Revidierte Fassung: FPI-R (5. Aufl.). Göttingen: Hogrefe.

Faller, H. (1997). Subjektive Krankheitstheorien bei Patienten einer psychotherapeutischen Ambulanz. Zeitschrift für klinische Psychologie, Psychiatrie und Psychotherapie, 45, 264-278.

Farrell, A. D. (1989). Impact of standards for computer-based tests on practice: Consequences of the information gap. Computers in Human Behavior, 5, 1-11.

Feldman, J. M. (1992). Constructive processes as a source of context effects in survey research: Explorations in self-gernerated validity. In N. Schwarz & S. Sudman (Eds.), Context effects in social and psychological research (pp. 49-62). New York; N.Y.: Springer.

Feldman, J. M. & Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention and behavior. Journal of Applied Psychology, 73, 421-435.

Fenz, W. D. & Epstein, S. (1965). Manifest anxiety: Unifactorial or multifactorial composition? Perceptual and Motor Skills, 20, 773-780.

Ferrando, P. J. (1994). Fitting item response models to the EPI-A impulsivity subscale. Educational and Psychological Measurement, 54, 118-127.

Ferrando, P. J. (2001). The measurement of neuroticism using MMQ, MPI, EPI and EPQ items: A psychometric analysis based on item response theory. Personality and Individual Differences, 30, 641-656.

Ferrando, P. J., Lorenzo, U. & Molina, G. (2001). An item response theory analysis of response stability in personality measurement. Applied Psychological Measurement, 25, 3-17.

Finch, J. F. & West, S. G. (1997). The investigation of personality structure: Statistical models. Journal of Research in Personality, 31, 439-485.

Finney, J. C. (1962). Prolegomena to epidemiology in mental health. Journal of Nervous and Mental Disease, 135, 99-104.

Finney, J. C. (1985). Anxiety: Its measurement by objective personality tests and self-report. In A.H.Tuma & J.Maser (Eds.), Anxiety and anxiety disorders (pp. 645-679). London: Lawrence Erlbaum.

Finzen, A. (1988). Angst als gesellschaftliches Phänomen. In W.Pöldinger (Hrsg.), Angst und Angstbewätligung (S. 73-88). Bern: Huber.

Fischer, G. H. (1983). Neuere Testtheorie. In H. Feger & J. Bredenkamp (Hrsg.), Enzyklopädie der Psychologie, Serie: Forschungsmethoden der Psychologie, Bd. 3: Messen und Testen (S. 604-692). Göttingen: Hogrefe.

Fliege, H., Rose, M., Bronner, E. & Klapp, B. F. (2002). Prädiktoren des Behandlungs-ergebnisses stationärer psychosomatischer Therapie. Psychotherapie, Psychosomatik und medizinische Psychologie, 52, 47-55.

Forsyth, R., Saisangjan, U. & Gillmer, J. (1981). Some empirical results related to the robustness of the Rasch model. Aplied Psychological Measurement, 5, 175-186.

Franke, G. H. (1995). SCL-90-R. Die Symptom-Checkliste von Derogatis (Deutsche Version). Weinheim: Beltz-Verlag.

Fraser, C. & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. Multivariate Behavioral Research, 23, 267-269.

Freud, A. (1936). Das Ich und die Abwehr. München: Kindler.

Freud, S. (1940). Hemmung, Symptom und Angst. In Freud, S. (Hrsg), Gesammelte Werke, XIV (S. 111-205). London: Imago.

Freyberger, H. J. & Stieglitz, R.-D. (1996). Kompendium der Psychiatrie und Psychotherapie (10. Aufl.). Basel: Karger.

Garb, H. N. (2000). Computers will become increasingly important for psychological assessment: Not that there's anything wrong with that! Psychological Assessment, 12, 31-39.

Garber, J., Miller, S. M. & Abramson, L. Y. (1980). On the distinction between anxiety and depression: Perceived control, certainty, and probability of goal attainment. In J.Garber & E. P. Seligman (Eds.), Human helplessness theory and applications (pp. 131-169). New York: Academic Press.

Gardner, W., Kelleher, K. J. & Pajer, K. A. (2002). Multidimensional adaptive testing for mental health problems in primary care. Medical Care, 40, 812-823.

Ghosh, A., Marks, U. M. & Carr, A. C. (1984). Controlled study of self-exposure treatment for phobics: Preliminary communication. Journal of Royal Society of Medicine, 77, 483-487.

Gibbons, R. D., Clark, D. C., Cavanaugh, S. V. & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. Journal of Psychiatric Research, 19, 43-55.

Gittler, G. (1999). Adaptiver 3-dimensionaler Würfeltest. A3DW. Mödling: Schuhfried-Verlag.

Gitzinger, I. (1990). Akzeptanz der Darbietung eines Tests auf dem Personalcomputer von stationären Patient/-innen. Psychotherapie, Psychosomatik und medizinische Psychologie, 40, 143-145.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinominal distribution. Psychometrika, 53, 525-546.

Gray-Little, B., Wililams, V. S. L. & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg Self-Esteem Scale. Personality and Social Psychology Bulletin, 23, 443-451.

Gray, J. A. (1981). The psychophysiology of anxiety. In R. Lynn (Ed.), Dimensions of personality – Papers in honor of H.J. Eysenck (pp. 233-252). Oxford: Pergamon.

Gregory, R. J. (1996). Special topics and issues in testing: Computer-aided psychological assessment. In R. J. Gregory (Ed.), Psychological testing. History, principles and applications (2nd ed., pp. 572-591). London: Allyn & Bacon.

Guilford, J. S., Zimmermann, P. S. & Guilford, J. P. (1976). The Guilford Zimmermann temperament survey handbook. San Diego: Cal. Edits Publishers.

Gulliksen, H. (1950). Theory of mental tests. New York, N.Y.: Wiley.

Gulliksen, H. & Tukey, J.W. (1958). Reliability for the law of comparative judgement. Psychometrika, 23, 95-110.

Guthke, J., Räder, E., Caruso, M. & Schmidt, K.-D. (1991). Entwicklung eines adaptiven computergestützten Lerntests auf der Basis der strukturellen Informationstheorie. Diagnostica, 37, 1-28.

Guttman, L. (1954). Some necessary conditions for common factor analysis. Psychometrika, 19, 149-161.

Hageböck, J. (1990). PSYMEDIA: Ein Computer-Programmsystem für die psychometrische Einzelfalldiagnostik. Diagnostica, 36, 220-227.

Hageböck, J. (1994). Computerunterstützte Diagnostik in der Psychologie. Göttingen: Hogrefe.

Hambleton, R. K. & Slater, S. C. (1997). Item response theory models and testing practices: Current international status and future directions. European Journal of Psychological Assessment, 13, 21-28.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Hingham, M.A.: Kluwer.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, C.A.: Sage Publications.

Hambleton, R. K. & Zaal, J. N. (1990). Computerized adaptive testing: Theory, applications and standards. In R. K. Hambleton & J. N. Zaal (Eds.), Advances in educational and psychological testing (pp. 341-366). London: Kluwer Academic Press.

Hamilton, M. (1959). Hamilton-Angst-Skala: HAMA. Fremdbeurteilungsskala. Berlin: Autor.

Hamilton, M. (1977). Hamilton-Angst-Skala. Fremdbeurteilungs-Skala (F). In Collegium Internationale Psychiatriae Scalarum (CIPS) (Hrsg.), Internationale Skalen für Psychiatrie. Berlin: Autor.

Handel, R. W., Ben Porath, Y. S. & Watt, M. (1999). Computerized adaptive assessment with the MMPI-2 in a clinical setting. Psychological Assessment, 11, 369-380.

Harvey, R. J., Murry, W. D. & Markham, S. E. (1994). Evaluation of three short-form versions of the Meyer-Briggs Type Indicator. Journal of Personality Assessment, 63, 181-184.

Hasson, F., Keeney, S. & McKenna, H. (2000). Research guidelines for the delphi survey technique. Journal of Advances in Nursing, 32, 1008-1015.

Hathaway, S. R. & McKinley, J. C. (1983). The Minnesota Multiphasic Personality Inventory Manual. New York: Psychological Corporation.

Hathaway, S. R., McKinley, J. C. & Engel, R. R. (2001). Minnesota-Multiphasic Personality Inventory 2: MMPI-2. Minneapolis: National Computer Systems Inc., Professional Assessment Services Division.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hautzinger, M. & Bailer, M. (1993). Allgemeine-Depressionsskala: ADS. Weinheim: Beltz-Verlag.

Hautzinger, M., Bailer, M., Worall, H. & Keller, F. (1994). Beck-Depressions-Inventar: BDI. Bern: Huber.

Häcker, H., Schmidt, L. R., Schwenkmezger, P. & Lutz, H. E. (1975). Objektive Testbatterie: OATB 75. Weinheim: Beltz-Verlag.

Häcker, H. & Stapf, K. H. (1998). Dorsch Psychologisches Wörterbuch (13. Aufl.). Bern: Huber.

Hänsgen, K. D. & Merten, T. (1994). Computerbasiertes Ratingsystem zur Psychopathologie: CORA (2. Aufl.). Göttingen: Apparatezentrum.

Hänsgen, K.D. & Bernascomi, M. (2000). Befragung zur Situation der Psychodiagnostik in der Schweiz. Freiburg, Schweiz: Zentrum für Testentwicklung und Diagnostik am Departement für Psychologie, Universität Freiburg.

Heidegger, M. (1979). Sein und Zeit (15. Aufl.). Tübingen: Niemeyer.

Heinerth, K. (1972). Prüfungsangst von Studenten. Psychologische Rundschau, 23, 79-90.

Helmchen, H. & Linden, M. (1986). Die Differenzierung von Angst und Depression. Heidelberg: Springer-Verlag.

Hergovich, H. (1992). Computer-Häuschentest Dissertation, Universität Wien.

Hermann, Ch., Buss, U. & Snaith, R. P. (1995). Hospital Anxiety and Depression Scale: HADS. Bern: Huber.

Hermann, Ch., Scholz, K.-H. & Kreuzer, H. (1991). Screening von Patienten einer kardiologischen Akutklinik mit einer deutschen Fassung der „Hospital Anxiety and Depression“ (HAD)-Skala. Psychotherapie, Psychosomatik und medizinische Psychologie, 41, 83-92.

Hetter, R. D., Segall, D. O. & Bloxom, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. Applied Psychological Measurement, 18, 197-204.

Hinz, A. & Schwarz, R. (2001). Angst und Depression in der Allgemeinbevölkerung. Eine Normierungsstudie zur Hospital Anxiety and Depression Scale. Psychotherapie, Psychosomatik, Medizinische Psychologie, 51, 193-200.

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweitere und modifizierte Version mit vier Komponenten. Zeitschrift für Pädagogische Psychologie, 5, 121-130.

Hogen, H. (2001). Der Brockhaus Psychologie. Leipzig: Brockhaus.

Holland, P. & Wainer, H. (1993). Differential item functioning. Hillsdale, N.J.: Erlbaum.

Holtzman, W. H., Thorper, J. S. & Swartz, J. D. (1961). Holtzman-Inkblot-Technique: HIT. Austin, TX: University of Texas Press.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Hornke, L. F. (1989). Konstruktion eines Tests mit verbalen Analogien. CAT-A2: Weitere Untersuchungen. Untersuchungen des psychologischen Dienstes der Bundeswehr, 24, 49-137.

Hornke, L. F. (1996). Stand der Technik zum Computergestützten Adaptiven Testen (CAT). 28./30. Jahrgang 1993 / 1995. In K. Puzicha (Hrsg.), Bundesministerium der Verteidigung. Untersuchungen des Psychologischen Dienstes der Bundeswehr (2. Aufl.). München: Verlag für Wehrwissenschaften.

Hornke, L. F. (1999). Benefits from computerized adaptive testing as seen in simulation studies. European Journal of Psychological Assessment, 15, 91-98.

Hornke, L. F. & Etzel, S. (1999a). Verbaler Gedächtnis Test: VERGED. Mödling: Schuhfried-Verlag.

Hornke, L. F. & Etzel, S. (1999b). Visueller Gedächtnis Test: VISGED. Mödling: Schuhfried-Verlag.

Hornke, L. F. & Habon, M. W. (1984). Regelgeleitete Konstruktion und Evaluation von nicht-verbalen Denkaufgaben. Wehrpsychologische Untersuchungen, 19, 1-153.

Hornke, L. F., Küppers, A. & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. Diagnostica, 46, 182-188.

Hornke, L. F. (1981). Computer Unterstütztes Testen (CUT) von Prüfungsangst. Zeitschrift für Differentielle und Diagnostische Psychologie, 2, 325-335.

Hornke, L. F. (1993). Mögliche Einspareffekte beim computergestützten Testen. Diagnostica, 39, 109-119.

Hornke, L. F. (1994). Erfahrungen mit der computergestützten adaptiven Diagnostik im Leistungsbereich. In D. Bartussek & M. Amelang (Hrsg.), Fortschritte der Differentiellen Psychologie und psychologischen Diagnostik (S. 321-332). Göttingen: Hogrefe.

Hörhold, M. & Klapp, B. F. (1993). Testungen der Invarianz und der Hierarchie eines mehrdimensionalen Stimmungsmodells auf der Basis von Zweipunkterhebungen an Patienten- und Studentenstichproben. Zeitschrift für Medizinische Psychologie, 2, 27-35.

Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

Hull, S. L. (1943). Principles of behavior. New York: Appleton-Century-Crofts.

Humphreys, L. G. & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behavioral Research, 10, 193-205.

Janke, W. & Debus, G. (1978). Die Eigenschafts-Wörter-Liste: EWL. Göttingen: Hogrefe.

Jaspers, K. (1973). Philosophie (4. Aufl.). Berlin: Springer-Verlag.

Jäger, R. S. (1990). Computerdiagnostik – Eine Einführung. Diagnostica, 36, 91-95.

Jäger, R. S. & Krieger, W. (1994). Zukunftsperspektiven der computerunterstützten Diagnostik, dargestellt am Beispiel der treatmentorientierten Diagnostik. Diagnostica, 40, 217-243.

Johnson, J. H. & Johnson, J. N. (1981). Psychological considerations related to the development of computerized testing stations. Behavior Research Methods & Instrumentation, 13, 421-424.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-202.

Jöreskog, K. & Sörbom, D. (2002). Prelis 2: User's Reference Guide. Lincolnwood: Scientific Software International.

Jöreskog, K., Sörbom, D., du Toit, S. & du Toit, M. (2000). Lisrel 8: New Statistical Features. Lincolnwood: Scientific Software International.

Kaplan, D. (2000). Structural equation modeling: Foundation and extensions. Thousand Oaks, CA.: Sage Publications.

Kaskowitz, G. S. & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. Applied Psychological Measurement, 25, 39-52.

Kazdin, A. E. (2000). Encyclopedia of psychology. Washington, D.C.: American Psychological Association and Oxford University Press.

Kelderman, H. (1984). Loglinear Rasch model tests. Psychometrika, 49, 223-245.

Kelderman, H. (1997). Log-linear multidimensional model for polytomous scored items. In W. J.Van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory. New York, N.Y.: Springer.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshelman, S., Wittchen, H. & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-II-R-psychiatric disorders in the United States. Archives of Generic Psychiatry, 51, 8-19.

Kierkegaard, S. (1844). Begriff der Angst (Gesammelte Werke, Abt. 11/12). Gütersloh: Gütersloher Taschenbücher Siebenstern.

King, D. W., King, L. A., Fairbank, J. A. & Schlenger, W. E. (1993). Enhancing the precision of the Mississippi Scale for combat-related posttraumatic stress disorder: An application of item response theory. Psychological Assessment, 5, 457-471.

Kingsbury, G. G. & Houser, R. L. (1993). Assessing the utility of item response models. Educational Measurement: Issues and Practice, 12, 21-27.

Kisser, R. (1995). Adaptive Strategien. In J. Petermann (Hrsg.), Psychologische Diagnostik (S. 161-170). Weinheim: Psychologie-Verlags-Union.

Klages, L. (1926). Grundlagen der Charakterkunde. Bonn: Bouvier.

Klapp, B.F. & Danzer, G. (1999). Psychosomatische Grundlagen. In M. v. Classen, V. Diehl & K. Kochsiek (Hrsg.), Innere Medizin. München: Urban-Schwarzenberg Verlag.

Kleinmuntz, B. & McLean, R. S. (1968). Computers in behavioral science: Diagnostic interviewing by digital computer. Behavioral Science, 13, 75-80.

Knapp, G. (2001). Angst und Depression. Grundformen und Pathologie. Sternenfels: Verlag Wissenschaft & Praxis.

Knowles, E. S. (1988). Item context effects in personality scales: Measuring changes the measure. Journal of Personality and Social Psychology, 55, 312-320.

Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., Neville, J. W. & Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman (Eds.), Context effects in social and psychological research (pp. 465-479). New York: Springer.

Knowles, E. S. & Condon, C. A. (1999). Why people say “yes”: A dual-process theory of acquiescence. Journal of Personality and Social Psychology, 77, 379-386.

Knowles, E. S. & Condon, C. A. (2000). Does the rose still smell as sweet? Item variablitiy across test forms and revisions. Psychological Assessment, 12, 245-252.

Koch, W. R. & Dodd, B. D. (1985). Computerized adaptive attitude measurement. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Koch, W. R. & Dodd, B. D. (1989). An investigation of procedures for computerized adaptive testings using partial credit scoring. Educational and Psychological Measurement, 2, 335-357.

Koch, W. R., Dodd, B. D. & Fitzpatrick, S. J. (1990). Computerized adaptive testing using the successive intervals Rasch model. Measurement and Evaluation in Counselling and Development, 23, 20-30.

Kolen, M. J. (1986). Traditional equating methodology. Educational Measurement: Issues and Practice, 7, 29-36.

Kraepelin, E. (1918). Hundert Jahre Psychiatrie. Berlin: Springer-Verlag.

Kranz, H. T. (1979). Einführung in die klassische Testtheorie. Frankfurt a.M.: Fachbuch-handlung für Psychologie.

Kristof, W. (1983). Klassische Testtheorie und Testkonstruktion. In H. Feger & J. Bredenkamp (Hrsg.), Enzyklopädie der Psychologie, Serie: Forschungsmethoden der Psychologie (Bd. 3: Messen und Testen, S. 544-603). Göttingen: Hogrefe.

Krohne, H. W. (1993). Vigilance and cognitive avoidance concepts in coping research. In H. W. Krohne (Ed.), Attention and avoidance strategies in coping with aversiveness. Seattle: Hogrefe & Huber.

Krohne, H. W. & Hindel, C. (1990). Die Erfassung störender Kognitionen bei Leistungssportlern im Tischtennis. Sportwissenschaft, 20, 56-63.

Krohne, H. W. (1996). Angst und Angstbewältigung. Stuttgart: Kohlhammer Verlag.

Krueger, R. F. & Finger, M. S. (2001). Using item response theory to understand comorbidity among anxiety and unipolar mood disorders. Psychological Assessment, 13, 140-151.

Kubinger, K. D. (1993). Testtheoretische Probleme der Computerdiagnostik. Zeitschrift für Arbeits- und Organisationspsychologie, 37, 130-137.

Kubinger, K. D. (1996). Methoden der Psychologischen Diagnostik. In E. Erdfelder, R. Mausfeld, T. Meister & G. Rudinger (Hrsg.), Handbuch Quantitative Methoden (S. 567-576). Weinheim: Psychologie-Verlags-Union.

Kubinger, K. D. (1999). Forschung in der psychologischen Diagnostik. Psychologische Rundschau, 50, 131-139.

Kubinger, K. D., Fischer, D. & Schuhfried-Verlag, G. (1993). Begriffs-Bildungs-Test: BBT. Mödling: Schuhfried-Verlag.

Kubinger, K. D. & Wurst, E. (2000). Adaptives Intelligenz Diagnostikum 2: AID2. Göttingen: Beltz-Verlag.

Kubinger, K. D. (1986). Adaptive Intelligenzdiagnostik. Diagnostica, 32, 330-344.

Laatsch, L. & Choca, J. (1994). Cluster-branching methodology for adaptive testing and the development of the adaptive category test. Psychological Assessment, 345-351.

Lautenschlager, G. J. (1989). A comparison of alternatives to conductiong Monte Carlo analysis for determining parallel analysis criteria. Multivariate Behavioral Research, 24, 365-395.

Laux, L. & Glanzmann, P. (1996). Angst und Ängstlichkeit. In M. Amelang (Hrsg.), Enzyklopädie der Psychologie. Themenbereich C. Serie VIII, Bd. 3 (S. 107-146). Göttingen: Hubert.

Laux, L., Glanzmann, P., Schaffner, P. & Spielberger, C. D. (1981). State-Trait-Angstinventar: STAI. Weinheim: Beltz-Verlag.

Lehmann, G. (1983). Testtheorie: Eine systematische Übersicht. In H. Feger & J. Bredenkamp (Hrsg.), Enzyklopädie der Psychologie. Themenbereich B: Methodologie und Methoden. Serie I: Forschungsmethoden der Psychologie. (Bd. 3: Messen und Testen, S. 427-543). Göttingen: Verlag für Psychologie.

Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C. & Andreoli, A. (1993). Development of the Perceived Stress Questionnaire (PSQ): A new tool for psychosomatic research. Journal of Psychosomatic Research, 1, 19-32.

Levine, M. V., Drasgow, F., Williams, B., McCusker, C. & Thomasson, G. L. (1992). Distinguishing between item response theory models. Applied Psychological Measurement, 16, 261-278.

Lieb, R. & Wittchen, H.-U. (1998). Angststörungen. Klassifikation und Diagnostik. In U. Baumann & M. Perrez (Hrsg.), Klinische Psychologie – Psychotherapie (S. 882-892). Bern: Huber.

Liebert, R. M. & Morris, L. W. (1967). Cognitive and emotional components of anxiety tests. A distinction and some initial data. Psychological Reports, 20, 975-978.

Lienert, G. A. & Raatz, U. (1994). Testaufbau und Testanalyse. Weinheim: Psychologie-Verlags-Union.

Linacre, J. M. (1994). Sample size and item calibration stability. Rasch Measurement Transactions, 7, 4, p. 328.

Longman, R. S., Cota, A. A., Holden, R. R. & Fekken, G. C. (1989). A regression for the parallel analysis criterion in principal components analysis: Mean and 95th percentile eigenvalues. Multivariate Behavioral Research, 24, 59-79.

Lord, F. M. (1952). A theory of test scores (Psychometric Monograph No.7). Iowa City, IA.: Psychometric Society.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance and their parallel forms reliability. Psychometrika, 48, 233-245.

Lord, F. N. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading. MA: Addison-Wesley.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 1, 135-143.

Lucas, R. W., Mullin, P. J., Luna, C. B. X. & McInroy, D. C. (1977). Psychiatrists and a computer as interrogators of patients with alcohol-related illnesses: A comparison. British Journal of Psychiatry, 131, 160-171.

Ludwig, M., Geier, S. & Bullinger, M. (1990). Skalen zur Erfassung des Wohlbefindens: Psychometrisches Analysen zum „Profile of Mood States“ (POMS) und zum „Psychological General Well-Being Index“ (PGWI). Zeitschrift für Differentielle und Diagnostische Psychologie, 11, 53-61.

Lumsden, J. (1976). Test theory. Annual Review of Psychology, 27, 251-280.

Lunz, M. E., Bergstrom, B. A. & Wright, B. D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. Applied Psychological Measurement, 16, 33-40.

Lushene, R. E. (1970). The effects of physical and psychological threat on the autonomic, motoric and ideational components of state anxiety. Unpublished dissertation, Florida State University.

Lück, H. E. & Timaeus, E. (1969). Skalen zur Messung Manifester Angst (MAS) und sozialer Wünschbarkeit (SDE-E und SDS-SM). Diagnostica, 17, 53-59.

Mandler, G. & Sarason, S. B. (1952). A study of anxiety and learning. Journal of Abnormal and Social Psychology, 47, 166-173.

Margraf, J. (2000). Lehrbuch der Verhaltenstherapie (Bd. 1 & 2). Berlin: Springer-Verlag.

Margraf, J. & Bandelow, B. (1997). Empfehlungen für die Verwendung von Messinstrumenten in der klinischen Angstforschung. Zeitschrift für klinische Psychologie, 26, 150-156.

Margraf, J. & Ehlers, A. (1995). Beck Angst Inventar: BAI. Frankfurt: Swets & Zeitlinger.

Margraf, J. & Ehlers, A. (in Druck). Beck Angst Inventar: BAI (2.Aufl.). Frankfurt: Swets & Zeitlinger.

Margraf, J., Ehlers, A. & Schneider, S. (1994). Diagnostisches Interview bei psychischen Störungen (DIPS) (2. Aufl.). Berlin: Springer-Verlag.

Margraf, J. & Schneider (1990). Panik. Angstanfälle und ihre Behandlung (2. Aufl.). Berlin: Springer-Verlag.

Marks, I. M. (1970). The classification of phobic disorders. British Journal of Psychiatry, 116, 377-386.

Marshall, G. N., Orlando, M., Jaycox, L. H., Foy, D. W. & Belzberg, H. (2002). Development and validation of a modified version of the peritraumatic dissociative experiences questionnaire. Psychological Assessment, 14, 123-134.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G. N. & Evans, J. (1986). Banking non-dichotomously scored items. Applied Psychological Measurement, 10, 355-367.

May, R. (1950). The meaning of anxiety. New York, N.Y.: Ronald Press.

Maydeu-Olivares, A., Drasgow, F. & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. Applied Psychological Measurement, 18, 245-256.

McDonald, R. P. (1989). Future directions for item response theory. International Journal of Educational Research, 13, 205-220.

McDonald, R.P. (1994). Testing for approximate unidimensionality. In D. Laveault, B. Zumbo, M. E. Gessaroli & M. W. Boss (Eds.). Modern theories of measurement: Problems and issues (pp. 63-86). Ottawa, Edumetrics.

McKinley, R. L. & Way, W. D. (1992). The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models (TOEFL technical resport TR-5). Princeton, N.J.: Educational Testing Service.

McNemar, Q. (1946). Opinion-attitude methodology. Psychological Bulletin, 43, 289-374.

MacCallum, R.C., Browne, M.W. & Sugawara, H.M. (1996). Power analysis and determination of sample size for covariance structure modeling. Psychological Bulletin, 100, 107-120.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114, 449-458.

Meijer, R. R. (1996). Person-fit research: An introduction. Applied Measurement in Education, 9, 3-8.

Meijer, R. R. & Nering, M. L. (1999). Computerized adaptive testing. Overview and introduction. Applied Psychological Measurement, 23, 187-194.

Melfsen, S., Florin, I. & Warnke, A. (2001). Sozialphobie und –angstinventar für Kinder. SPAIK. Göttingen: Hogrefe.

Menghin, S. & Kubinger, K. D. (1996). Zur Legende: „Testpersonen beantworten dem Computer persönliche und intime Fragen offener als einem Testleiter“ – Ergebnisse eines Experiments. Zeitschrift für Differentielle und Diagnostische Psychologie, 17, 163-169.

Mineka, S., Watson, D. & Clark, L. A. (1998). Comorbidity of anxiety and unipolar mood disorders. Annual Review of Psychology, 49, 377-412.

Mislevy, R. J. & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Chicago, IL.: Scientific Software Incorporation.

Molenaar, W. (1974). De logistische en de normale kromme. [The logistic and the normal curve]. Nederlands Tijdschrift voor de Psychologie, 29, 415-420.

Moosbrugger, H. (1984). Konzeptuelle Probleme und praktische Brauchbarkeit von Modellen zur Erfassung von Persönlichkeitsmerkmalen. In M. Amelang & H. J. Ahrens (Hrsg.), Brennpunkte der Persönlichkeitsforschung. Göttingen: Hogrefe.

Moreland, K. L. (1992). Computer-assisted psychological assessment. In M. Zeidner & R. Most (Eds.), Psychological testing: An inside view. Palo Alto, CA.: Consulting Psychologists Press.

Morris, L. W., Davis, M. A. & Hutchings, C. H. (1981). Cognitive and emotional components of anxiety: Literature review and a revised worry-emotionality scale. Journal of Educational Psychology, 73, 541-555.

Morris, L. W., Franklin, M. S. & Ponath, P. (1983). The relationship between trait and state indices of worry and emotionality. In H.M.van der Plög, R. Schwarzer & C. D. Spielberger (Eds.), Advances in test anxiety research (pp. 3-13). Lisse, NL.: Swets & Zeitlinger.

Morris, L. W. & Liebert, R. M. (1970). Effects of anxiety on timed and untimed intelligence tests: Another look. Journal of Consulting and Clinical Psychology, 35, 332-337.

Möller, H. J., Laux, G. & Deister, A. (1996). Psychiatrie. Stuttgart: Hippokrates.

Mrazek, J. (1985). AF-HI. Die subjektive Wahrnehmung des Herzinfarkts und die Angst des Infarktkranken. In W. Langosch (Hrsg.), Psychische Bewältigung der chronischen Herzerkrankung (S. 159-169). Heidelberg: Springer-Verlag.

Muraki, E. (1990). Fitting a polytomous item response model into Likert-type data. Applied Psychological Measurement, 16, 59-71.

Muraki, E. (1992). A Generalized Partial Credit Model (GPCM): Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. (1993). Information functions of the Generalized Partial Credit Model (GPCM). Applied Psychological Measurement, 17, 351-363.

Muraki, E. (1997). A Generalized Partial Credit Model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 153-164). Berlin: Springer.

Muraki, E. & Bock, R. D. (1999). Parscale: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks [Manual and Software]. Chicago: Scientific Software Int.

Murray, H. A. (1991). Thematic Apperception Test: TAT. Cambridge: Harvard University Press.

Muthén, L. K. & Muthén, B. O. (1998). Mplus. The comprehensive modeling program for applied researchers. User's guide [Manual and Software]. Los Angeles: Authors.

Muthny, F. A. (1991). Lebenszufriedenheit bei koronarer Herzkrankheit: Ein Vergleich mit anderen lebensbedrohlichen Erkrankungen. In M. Bullinger, M. Ludwig & N. v. Steinbüchel (Hrsg.), Lebensqualität bei kardiovaskulären Erkrankungen. Grundlagen, Messverfahren und Ergebnisse (S. 196-210). Göttingen: Hogrefe.

Müller, H. (1999). Probabilistische Testmodelle für diskrete und kontinuierliche Ratingskalen. Bern: Huber.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. Applied Psychological Measurement, 17, 29-38.

Nandakumar, R. (1994). Assessing dimensionality of a set of items – Comparison of different approaches. Journal of Educational Measurement, 31, 17-35.

Nandakumar, R. & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics, 18, 41-68.

Neumer, S. P. (2000). Beiträge zur Gemischten Angst-Depression als DSM-IV-Forschungsdiagnose. Probleme und Perspektiven. Berlin: Wissenschaftsverlag.

Newmark, C.S., Faschingbauer, T.R., Finch, A.J. & Kendall, P.C. (1979). Factor analysis of the MMPI-STAI. Journal of Clinical Psychology, 31, 3, 449-452.

Nicewander, W. A. & Thomasson, G. L. (1999). Some reliability estimates for computer adaptive tests. Applied Psychological Measurement, 23, 239-247.

Novick, M. R. (1966). The axioms and principal results of classical test theory. Journal of Mathematical Psychology, 3, 1-18.

Orlando, M. & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD Checklist: Detection and evaluation of impact. Psychological Assessment, 14, 50-59.

Orlando, M., Sherbourne, C. D. & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. Psychological Assessment, 12, 354-359.

Orlando, M. & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. Applied Psychological Measurement, 24, 50-64.

Osman, A., Hoffman, J., Barrios, F. X., Kopper, B. A., Breitenstein, J. L. & Hahn, S. K. (2002). Factor structure, reliability and validity of the Beck Anxiety Inventory in adolescent psychiatric inpatients. Journal of Clinical Psychology, 58, 443-456.

Owen, R. J. (1969). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association 351-356.

Ozer, D. J. & Reise, S. P. (1994). Personality assessment. Annual Review of Psychology, 45, 357-388.

Peters, U. H. (2000). Peters Lexikon. Psychiatrie, Psychotherapie, Medizinische Psychologie (5. Aufl.). München: Urban & Fischer Verlag.

Ponsoda, V., Olea, J. & Revuelta, J. (1994). ADTEST: A computer-adaptive test based on the maximum information principle. Educational and Psychological Measurement, 54, 680-686.

Ramsay, J. O. (1995). TestGraf. A program for the graphical analysis of multiple choice test and questionnaire data [Manual and Software]. Montreal: Author.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rauchfleisch, U. (1992). Handwörterbuch der Psychiatrie. Stuttgart: Enke Verlag.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. Applied Psychological Measurement, 21, 25-36.

Reise, S. P. (1999). Personality measurement issues viewed through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement. Hillsdale: LEA.

Reise, S. P. (2000). Application of IRT in personality and attitude assessment. In S. E. Embretson & S. P. Reise (Eds.), Psychometric methods: Item response theory for psychologists. Mahway, N.J.: Lawrence Erlbaum.

Reise, S. P. & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. Assessment, 7, 347-364.

Reise, S. P. & Waller, N. G. (1990). Fitting the two-parameter model to personality data: The parameterization of the Multidimensional Personality Questionnaire (MPQ). Applied Psychological Measurement, 14, 45-58.

Reise, S. P., Widaman, K. F. & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114, 352-356.

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Rentz, R. R. & Barshaw, W. L. (1977). The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 14, 161-180.

Reshetar, R. A., Norcini, J. J. & Shea, J. A. (1993). A simulated comparison of two content balancing and maximum information item selection procedures for an adaptive certification examination. Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta.

Revicki, D. A. & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking and computer adaptive testing. Quality of Life Research, 6, 595-600.

Reynolds, C. R. & Richmond, B. O. (1978). What I think and feel: A revised measure of children's manifest anxiety. Journal of Abnormal Child Psychology, 6, 271-280.

Roper, B. L., Ben-Porath, Y. S. & Butcher, J. N. (1991). Comparability of computerized adaptive and conventional testing with the MMPI-2. Journal of Personality Assessment, 57, 278-290.

Rorschach, H. (1954). Psychodiagnostik. Der Rorschach-Test. Bern: Huber.

Rose, M., Fliege, H., Walter, O. B., Becker, J., Bjorner, J., Ravens-Sieberer, U. & Klapp, B. F. (2002). Using the item response theory to develop a computer adaptive test for depression. Quality of Life Research, 11, 626.

Rose, M., Fliege, H., Walter, O. B., Hörhold, M. & Klapp, B. F. (in Druck). Erfassung verschiedener Stimmungsdimensionen mit dem Berliner Stimmungsfragebogen (BSF).

Rose, M., Hess, V., Hörhold, M., Brähler, E. & Klapp, B. F. (1999). Mobile computergestützte psychometrische Diagnostik. Ökonomische Vorteile und Ergebnisse zur Teststabilität. Psychotherapie, Psychosomatik, Medizinische Psychologie, 49, 202-207.

Rose, M., Walter, O. B., Fliege, H., Becker, J., Hess, V. & Klapp, B. F. (2003). 7 years of experience using Personal Digital Assistants (PDA) for psychometric diagnostics in 6000 inpatients and polyclinic patients. In H.-B. Bludau & A. Koop (Eds.), Mobile Computing in Medicine. Second conference on mobile computing in medicine, Heidelberg, Germany. Gesellschaft für Informatik (pp. 35-44). Bonn: Köllen Verlag.

Roskam, E. E. (1985). Current issues in item response theory. In E. E. Roskam (Ed.), Measurement and personality assessment. Amsterdam: North-Holland.

Rost, D. H. & Schermer, F. J. (1987). Auf dem Wege zu einer differentiellen Diagnostik der Leistungsangst. Psychologische Rundschau, 38, 14-36.

Rost, D. H. & Spada, H. (1978). Probabilistische Testtheorie. In K. J. Klauer (Hrsg.), Handbuch der pädagogischen Diagnostik (Bd. 1, S. 59-97). Düsseldorf: Schwann.

Rost, J. (1996). Lehrbuch Testtheorie und Testkonstruktion. Bern: Huber.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? Psychologische Rundschau, 50, 140-156.

Rost, J. & Carstensen, C. H. (2002). Multidimensional Rasch measurement via item component models and faceted designs. Applied Psychological Measurement, 26, 42-56.

Rost, J. & Luo, G. (1997). An application of a Rasch-based unfolding model to a questionnaire on adolescent centrism. In J. Rost & R. Langeheine (Eds.), Applications of latent trait and latent class models in the social sciences (pp. 278-286). Münster: Waxmann.

Rost, J. & Spada, H. (1982). Probabilistische Testtheorie. In K. J.Klauer (Hrsg.), Handbuch der Pädagogischen Diagnostik (1. Aufl.). Düsseldorf: Schwann.

Rost, J., Carstensen, C. H. & von Davier, M. (1999). Sind die Big Five Rasch-skalierbar? Eine Reanalyse der NEO-FFI-Normierungsdaten. Diagnostica, 45, 119-127.

Rouse, S. V., Finger, M. S. & Butcher, J. N. (1999). Advances in clinical personality measurement: An item response theory analysis of the MMPI-2 PSY-5 scales. Journal of Personality Assessment, 72, 307.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, 17.

Samejima, F. (1996). Graded Response Model. In W. J.van der Linden & R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory (pp. 85-100). New York: Springer.

Sands, W. A., Waters, B. K. & McBride, J. R. (1997). Computerized adaptive testing – From theory to operation. Washington, D.C.: American Psychological Association.

Santor, D. A. & Coyne, J. C. (2000). Examining symptom expression as a function of symptom severity: Item performance on the Hamilton Rating Scale for Depression. Psychological Assessment, 13, 127-139.

Santor, D. A. & Ramsay, J. O. (1998). Progress in the technology of measurement: Applications of item response models. Psychological Assessment, 10, 345-359.

Santor, D. A., Ramsay, J. O. & Zuroff, D. C. (1994). Nonparametric item analyses of the Beck Depression Inventory: Evaluating gender item bias and response option weights. Psychological Assessment, 6, 255-270.

Santor, D. A., Zuroff, D. C., Ramsay, J. O., Cervantes, P. & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. Psychological Assessment, 7, 131-139.

Sarason, I. G. (1978). Test Anxiety Scale (TAS): Concept and research. In C. D.Spielberger & I. G. Sarason (Eds.), Stress and anxiety (5th ed., pp. 193-216). Washington, D.C.: Hemisphere.

Sarason, I. G. (1984). Stress, anxiety and cognitive interference: Reactions to tests. Journal of Personality an Social Psychology, 46, 929-938.

Sartre, J. P. (1962). Das Sein und das Nichts. Hamburg: Rowohlt.

Saß, H., Wittchen, H. U. & Zaudig, M. (1996). DSM-IV. Diagnostisches und Statistisches Manual psychischer Störungen IV. Göttingen: Hogrefe.

Schermelleh-Engel, K., Moosbrugger, H. & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods of Psychological Research Online 2003, 8, 23-74.

Schmit, M.J. & Ryan, A.M. (1997). Specifity of item content in personality tests: An IRT analysis. Paper presendted at the 12th Annual SIOP Conference, St. Louis, M.O..

Schneewind, K. A. & Graf, J. (1998). Der 16-Persönlichkeits-Faktoren-Test (Revidierte Fassung). Bern: Hübner.

Schnipke, D. L. & Green, B. F. (1995). A comparison of item selection routines in linear and adaptive tests. Journal of Educational Measurement, 32, 227-242.

Scholler, G., Fliege, H. & Klapp, B. F. (1999). Fragebogen zu Selbstwirksamkeit, Optimismus und Pessimismus. Restrukturierung, Itemselektion und Validierung eines Instrumentes an Untersuchungen klinischer Stichproben. Psychotherapie, Psychosomatik, Medizinische Psychologie, 49, 275-283.

Schöneich, F., Rose, M., Danzer, G., Thier, P., Weber, C. & Klapp, B. F. (2000). Narzissmusinventar-90. NI-90. Empiriegeleitete Itemreduktion und Identifikation veränderungssensitiver Items des Narzissmusinventars zur Messung selbstregulativer Parameter. Psychotherapie, Psychosomatik, Medizinische Psychologie, 50, 396-405.

Schötzau-Fürwentsches, P. & Grubitzsch, S. (1991). Der Einsatz des Computers in der psychologischen Diagnostik. In S. Grubitzsch (Hrsg.), Testtheorie und Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick (S. 297-313). Hamburg: Reinbeck.

Schwenkmezger, P. & Hank, P. (1993). Papier-Bleistift- versus computerunterstützte Darbietung von State-Trait-Fragebogen: Eine Äquivalenzprüfung. Diagnostica, 39, 189-210.

Schwenkmezger, P. & Hodapp, V. & Spielberger, C.D. (1992). Das State-Trait-Ärgerausdrucks-Inventar (STAXI). Bern: Huber-Verlag.

Sedlmayer, E. (1980). The development of scales for measuring motor, cognitive and physiological anxiety states. Behavioral Analysis and Modification, 4, 141-151.

Segall, D. O. (1996). Multidimensional adaptive testing. Psychometrika, 61, 331-354.

Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. Psychometrika, 66, 79-97.

Seligman, M. E. P. (1975). Helplessness. On depression, development and death. San Francisco, CA.: Freeman.

Selmi, P. M., Klein, M. H., Greist, J. H., Johnson, J. H. & Harris, W. G. (1982). An investigation of computer-assisted cognitive-behavior therapy in the treatment of depression. Behavior Research Methods & Instrumentation, 14, 181-185.

Selye, H. (1957). The stress of life. London: Longmans, Green & Co.

Simms, L. J. & Clark, L. A. (submitted). Validation of a Computerized Adaptive Version of the Schedule for Nonadaptive and Adaptive Personality.

Sims, A. & Snaith, P. (1993). Angsttherapie in der klinischen Praxis. München: Quintessenz Verlag.

Sinar, E.F. & Zickar, M.J. (2002). Evaluating the robustness of graded response model and classical test theory parameter estimates to deviant items. Applied Psychological Measurement, 26, 2, 181-191.

Singh, J. (1993). Some initial experiments with adaptive survey designs for structured questionnaires. Paper presented at the New Methods and Applications in Consumer Research Conference, Cambridge, M.A..

Slangen, K., Kleemann, P. P. & Krohne, H. W. (1993). Coping with surgical stress. In H. W. Krohne (Ed.), Attention and avoidance. Strategies in coping with aversiveness (pp. 321-348). Seattle, W.A.: Hogrefe & Huber.

Slinde, J. A. & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 15, 23-35.

Spearman, C. (1904). General intelligence, objectively determined and measured. American Journal of Psychology, 15, 201-293.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. American Journal of Psychology, 18, 161-169.

Spielberger, C. D. (1972). Anxiety. Current trends in theory and research (Vols. 1 & 2). London: Academic Press.

Spielberger, C. D. (1980). Furcht und Angst. In C. D. Spielberger (Hrsg.), Stress und Angst. Risiken unserer Zeit (S. 63-78). Weinheim: Beltz Psychologie-Verlags-Union.

Spielberger, C. D., Gorsuch, R. L. & Lushene, R. E. (1970). STAI manual for the State-Trait Anxiety Inventory. Pao Alto, CA.: Consulting Psychology Press.

SPSS Inc. (1999). SPSS Advanced Statistics (Version 10.0). Chicago, ILL.: SPSS Inc..

Srp, G. & Hörndler, H. (1994). Syllogismen. Frankfurt: Swets Test Services.

Steer.R.A., Beck, A. T., Riskind, J. H. & Brown, G. (1986). Differentiation of depressive disorders from generalized anxiety by the Beck Depression Inventory. Journal of Clinical Psychology, 42, 475-478.

Steiger, J.H. & Lind, J.C. (1980). Statistically-based test for the number of common factors. Paper presented at the Annual meeting of Psychometric Society, Iowa City, I.A.

Stein, H. (1995). Adaptiver Analogien-Lerntest: ADANA. Mödling: Schuhfried-Verlag.

Steinberg, L. (1994). Context and serial effects in personality measurement: Limits on the generality of “measuring changes the measure”. Journal of Personality and Social Psychology, 66, 341-349.

Steinberg, L. & Thissen, D. (1995). Item response theory in personality research. In P. E. Shrout & S.T. Fiske (Eds.), Personality, research, methods and theory. A festschrift honoring Donald W. Fiske (pp. 161-181). Hillsdale, N.J.: Lawrence Erlbaum.

Steyer, R. & Eid, M. (1993). Messen und Testen. Berlin: Springer-Verlag.

Steyer, R., Schmidt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. European Journal of Personality, 13, 389-408.

Stocking, M. L. (1997). Revising item responses in computerized adaptive tests: A comparison of three models. Applied Psychological Measurement, 21, 129-142.

Stotland, E. (1969). The psychology of hope. San Francisco, CA.: Jossey-Bass. Inc..

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589-617.

Stout, W., Douglas, J., Junker, B. & Roussos, L. (1993). DIMTEST. Urbana: University of Illinois.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Stöber, J. & Schwarzer, R. (2000). Ausgewählte Emotionen: Angst. In J. H.Otto, H. A. Euler & H. Mandl (Hrsg.), Emotionspsychologie. Ein Handbuch (S. 189-198). Beltz Psychologie-Verlags-Union.

Ströbe, W., Hewstone, M. & Stephenson, G. M. (1996). Sozialpsychologie. Berlin: Springer-Verlag.

Stumm, G. & Pritz, A. (2000). Wörterbuch der Psychotherapie. Wien: Springer-Verlag.

Stumpf, H. (1996). Klassische Testtheorie. In E. Erdfelder, R. Mausfeld, T. Meister & G. Rudinger (Hrsg.), Handbuch Quantitative Methoden (S. 411-430). München: Psychologie-Verlags-Union.

Suen, H. K. (1990). Principles of test theories. Hillsdale: LEA.

Swaminathan, H. & Rogers, J. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Sweetland, R. C. & Keyser, D. J. (1991). Tests: A comprehensive reference for assessments in psychology, education and business (3rd ed.). Austin, TX.: Pro-Ed..

Swenson, W. M., Rome, H., Pearson, J. & Brannick, T. (1965). A totally automated psychological test: Experience in a medical center. Journal of the American Medical Association, 191, 925-927.

Swinson, R. P., Cox, B. J. & Fergus, K. D. (1993). Diagnostic criteria in generalized anxiety disorder treatment studies. Journal of Clinical Psychopharmacology, 13 (6), 455.

Taylor, C. W. (1953). Variables related to creativity and productivity among men in two research laboratories. The second University of Utah Research Conference on the identification of creative scientific talent, Salt Lake City: University of Utah.

Tellegen, A. (1982). Brief manual for the Multidimensional Personality Questionnaire. Unpublished manuscript, University of Minnesota, Minneapolis.

Tellegen, A. & Waller, N. G. (2001). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In S. R. Briggs & J. M. Cheek (Eds.), Personality measures: Development and evaluation. Greenwich, C.T.: JAI Press.

Testkuratorium der Föderation Deutscher Psychologenvereinigungen (1996). Richtlinien für den Einsatz elektronischer Datenverarbeitung in der psychologischen Diagnostik. Psychologische Rundschau 163-165.

Tewes, U. & Wildgrube, K. (1999). Psychologie Lexikon (2. Aufl.). München: Oldenburg Wissenschaftsverlag.

Thissen, D. (1991). MULTILOG: Multiple, categorical item analysis and test scoring using item response theory. Chicago: Scientific Software International.

Thissen, D. & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), Computerized adaptive testing: A primer (pp. 103-134). Hillsdale, N.J.: Erlbaum.

Thissen, D. & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.

Thissen, D., Steinberg, L. & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. Psychological Bulletin, 99, 118-128.

Thissen, D., Steinberg, L., Pyzczynski, T. & Greenberg, J. (1983). An item response theory in the study of group differences in trace lines. Applied Psychological Measurement, 7, 211-226.

Thurner, F. & Tewes, U. (2000). Der Kinder-Angst-Test-II: K-A-T-II. Göttingen: Hogrefe.

Tönnies, S. (1995). Vom gesunden und kranken Denken. Die Bedeutung der Kognitionen und ihre Selbstkommunikation für die seelische Gesundheit. In R. Lutz & N. Mark (Hrsg.), Wie gesund sind Kranke? Zur seelischen Gesundheit psychisch Kranker (S. 123-137). Göttingen: Verlag für Angewandte Psychologie.

Tonidandel, S., Quinones, M. A. & Adams, A. A. (2002). Computer-Adaptive Testing: The impact of test characteristics on perceived performance and test takers' reactions. Journal of Applied Psychology, 87, 320-332.

Tourangeau, R. & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin, 103, 299-314.

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Tunner, W. (1978). Angst, Angstabwehr und ihre therapeutische Veränderung. In L. Pongratz (Hrsg.), Handbuch der Psychologie, Bd. VII, 2. Klinische Psychologie. Göttingen: Hogrefe.

Tupes, E. C. & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. Lackland Air Force Base, TX: Aeronautical Systems Devision, Personnel Laboratory.

Uhlenhuth, E. H. (1985). The measurement of anxiety: Reply to Finney. In A. H.Tuma & J. Maser (Eds.), Anxiety and anxiety disorders (pp. 675-679). London: Lawrence Erlbaum Associates.

Ulich, D. (1989). Angst. In D. Ulich (Hrsg.), Das Gefühl. Eine Einführung in die Emotionspsychologie (2. Aufl., S. 206-219). München: Psychologie-Verlags-Union.

Urry, V. W. (1977). Tailored testing: A successful application of item response theory. Journal of Educational Measurement, 14, 181-196.

Usala, P. L. & Hertzog, C. (1991). Evidence of differential stability of state and trait anxiety in adults. Journal of Personality and Social Psychology, 60, 471-479.

Vahle, H. & Rittner, S. (1995). Adaptiver Zahlenfolgen-Lerntest: AZAFO. Mödling: Schuhfried-Verlag.

Vale, C. D. (1986). Linking item parameters onto a common scale. Applied Psychological Measurement, 10, 333-344.

Van der Linden, W. J. & Glas, C. A. W. (2003). Computer adaptive testing: Theory and practice. Boston: Kluwer Academic Press.

Van der Linden, W. J. & Hambleton, R. K. (1997). Handbook of modern item response theory. Berlin: Springer.

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics, 22, 203-226.

Verschoor, A. & Straetmans, G. (1999). Math CAT: A flexible testing system for adult mathematics education. In W. J. van der Linden & C. A. W. Glas (Eds.), Computer adaptive testing: Theory and practice. Boston: Kluwer Academic.

Verstralen, H., Bechger, T. & Maris, G. (2001). The combined use of classical test theory and item response theory. Arnhem, N.L.: Authors.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. & Thissen, D. (1990). Computerized adaptive testing: A primer. Hillsdale: Lawrence Erlbaum Associates.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Waller, N. G. (1997). Searching for structure in the MMPI. In S. E. Embretson & S. L. Hershberger (Eds.), The new rules of measurement (pp. 185-218). Mahwah, N.J.: Lawrence Erlbaum Associates.

Waller, N. G. & Reise, S. P. (1989). Computerized adaptive personality assessment: An illustration with the absorption scale. Journal of Personality and Social Psychology, 57, 1051-1058.

Waller, N. G., Tellegen, A., McDonald, R. P. & Lykken, D. T. (1996). Exploring nonlinear models in personality assessment: Development and validation of a negative emotionality scale. Journal of Personality, 64, 545-576.

Walter, O. B., Becker, J., Fliege, H., Klapp, B. F., Bjorner, J. & Rose, M. (submitted). Evaluating a computer adaptive test for 'anxiety' in simulation studies. European Journal of Psychological Assessment.

Walter, O. B., Becker, J., Fliege, H., Klapp, B. F. & Rose, M. (eingereicht). Entwicklung eines Computer Adaptiven Tests zur Erfassung von „Angst": Angst-CAT. Diagnostica.

Walter, R., Leifert, J. & Linster, H. (1975). An S-R-Inventory of anxiousness. Psychological Monographs, 76.

Wang, S. (1999). The accuracy of ability estimation methods for computerized adaptive testing using the Generalized Partial Credit Model. University of Pittsburgh.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. Applied Psychological Measurement, 25, 317-331.

Wang, T.-Y. (1995). The precision of ability estimation methods in computerized adaptive testing (Dissertation). Iowa: The University of Iowa.

Wang, T., Hanson, B. A. & Che-Ming, A. L. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. Applied Psychological Measurement, 23, 263-278.

Ware, J. E., Jr., Bjorner, J. B. & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of the widely used headache impact scales. Medical Care, 38, 1173-1182.

Ware, J.E., Kosinski, M., Bjorner, J.B., Bayliss, M.S., Batenhorst, A., Dahlöt, C. G. H., Teppers, S. & Dowson, S. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. Quality of Life Research, 12, 935-952.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54, 427-450.

Watson, D., Clark, D. C., Weber, K., Assenheimer, J. S., Strauss, M. E. & McComick, R. A. (1995). Testing a tripartite model: Exploring the symptom structure of anxiety and depression in student, adult and patient samples. Journal of Abnormal Psychology, 104, 14.

Watson, D. & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. Psychological Bulletin, 96, 465-490.

Weiss, D. J. (1985). Adaptive testing by computer. Journal of Consulting and Clinical Psychology, 53, 774-789.

Weiss, D. J. & Davison, M. L. (1981). Test theory and methods. Annual Review of Psychology, 32, 629-658.

Weiss, D. J. & Vale, D. (1987). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N. Butcher (Ed.), Computerized psychological assessment (pp. 325-343). New York: Basic Books.

Welsh, G. S. (1952). An anxiety index and an internalization ration for the MMPI. Journal of Consulting Psychology, 16, 72.

Westhoff, G. (1993). Handbuch psychosozialer Messinstrumente. Göttingen: Hogrefe.

Westmeyer, H. & Hageböck, J. (1992). Computer-assisted assessment: A normative perspective. European Journal of Psychological Assessment, 8, 1-16.

Wetzler, S. & Marlowe, D. B. (1994). Clinical psychology by computer? The state of the “art”. European Journal of Psychological Assessment, 10, 55-61.

Wiggins, J. S. (1981). Clinical and statistical prediction: Where are we and where do we go from here? Clinical Psychology Review, 1, 3-18.

Wilson, D. T., Wood, R. & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics and item factor analysis. Chicago: Scientific Software International.

Windle, C. (1954). Test-retest effect on personality questionnaires. Educational and Psychological Measurement, 14, 617-633.

Wittchen, H. U. & Pfister, H. (1996). M-CIDI. PC-Version des Diagnostisches Expertensystem für Psychische Störungen DIA-X. Frankfurt: Swets & Zeitlinger.

Wittchen, H.U., Schuster, P. & Vossen, A. (1997). Generalisierte Angst – Ihr Therapieratgeber. Bristol-Myers Squibb, ZNS-Service. München: Mosaik.

Wittchen, H. U., Wunderlich, U., Gruschwitz, S. & Zaudig, M. (1997). Strukturiertes Klinisches Interview für DSM-IV. SKID. Göttingen: Hogrefe.

Woodcock, R. W. & Johnson, M. B. (1989). Woodcock-Johnson-Psycho-Educational-Battery. Revised. Allen, TX: DLM Teaching Resources.

Wright, B. D. (1996). Sample size again. Rasch Measurement Transactions, 9, 4, p. 468.

Zara, A. R. (1988). Introduction to item response theory and computerized adaptive testing as applied in licensure and certification testing. National Clearing-house of Examination Information Newsletters, 6, 11-17.

Zigmond, A. S. & Snaith, R. P. (1983). The hospital anxiety and depression scale. Acta Psychiatrica Scandinavica, 67, 361-370.

Zimmerman, D. W. (1975). Test theory with minimal assumptions. Educational and Psychological Measurement, 36, 85-96.

Zinbarg, R. E. & Barlow, D. H. (1996). Structure of anxiety and the anxiety disorders: A hierarchical model. Journal of Abnormal Psychology, 105, 181-193.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Dissertation, Ottawa, Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zung, W. W. K. (1965). A self-rating depression scale. Archives of General Psychiatry, 12, 63-70.

Zwick, W. R. & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99, 432-442.

9. Anhang

9.1. Initialer Itempool des Angst-CATs

Tabelle 31: Initialer Itempool, aus dem in einem konsensuellen (Delphi-) Entscheidungsprozess „angstrelevante“ Items selektiert wurden (N = 125).

Fragebögen	Items
ADS	**Während der letzten Woche:**
ADS_1	Haben mich Dinge beunruhigt, die mir sonst nichts ausmachen.
ADS_3	Hatte ich Mühe, mich zu konzentrieren.
ADS_5	War alles anstrengend für mich.
ADS_7	Hatte ich Angst.
ADS_12	Habe ich das Leben genossen.
ALL	**Konnten Sie in der letzten Woche:**
ALL_2	Sich länger auf eine Aufgabe konzentrieren?
ALL_16	Ihre Aufgaben im Beruf und Haushalt verrichten?
ALL_18	Sich am Leben freuen?
ALL_21	Es sich bequem machen und sich entspannen?
ALL_24	Einkäufe und Besorgungen außer Haus erledigen?
ALL_25	Zuversichtlich in die Zukunft sehen?
ALL_27	Etwas Schönes tun und es genießen?
ALL_28	Beim Planen und Problemlösen klar denken?
ALL_36	Ihren Hobbys und Lieblingsbeschäftigungen nachgehen?
ALL_39	Sich selbstsicher fühlen?
ALL_41	Ihre Verpflichtungen zu Ihrer Zufriedenheit erfüllen?
BDI	**Wie haben Sie sich in dieser Woche einschließlich heute gefühlt?**
BDI_15	Ich bin unfähig zu arbeiten.
BDI_20	Ich mache mir so große Sorgen über gesundheitliche Probleme, dass ich an nichts anderes mehr denken kann.
BSF	**Ich fühle mich:**
BSF_1	Matt.
BSF_2	Konzentriert.
BSF_3	Gelöst.
BSF_5	Besorgt.
BSF_7	Schlaff.
BSF_11	Müde.
BSF_12	Beunruhigt.
BSF_14	Kribbelig.
BSF_17	Abgespannt.
BSF_20	Ausgeglichen.
BSF_23	Unsicher.
BSF_27	Aufmerksam.
BSF_29	Erschöpft.
GBB	**Ich fühle mich durch folgende Beschwerden belästigt:**
GBB_1	Schwächegefühl.
GBB_2	Herzklopfen, Herzjagen oder Herzstolpern.
GBB_3	Druck- oder Völlegefühl im Leib.
GBB_6	Ohnmachtsanfälle.
GBB_10	Schwindelgefühl.
GBB_12	Starkes Schwitzen.
GBB_17	Anfälle.
GBB_18	Übelkeit.
GBB_20	Kloßgefühl, Engigkeit oder Würgen im Hals.
GBB_21	Drang zum Wasserlassen.
GBB_28	Überempfindlichkeit gegen Wärme.
GBB_30	Schlafstörungen.
GBB_34	Schluckbeschwerden.
GBB_36	Gefühl der Benommenheit.
GBB_37	Taubheitsgefühl (Einschlafen, Absterben, Brennen oder Kribbeln in Händen und Füßen).
GBB_38	Verstopfung.

Tabelle 31 (Fortsetzung 1): Initialer Itempool, aus dem in einem konsensuellen (Delphi-) Entscheidungsprozess „angstrelevante“ Items selektiert wurden (N = 125).

Fragebögen	Items
GBB	**Ich fühle mich durch folgende Beschwerden belästigt:**
GBB_39	Appetitlosigkeit.
GBB_40	Aufsteigende Hitze, Hitzewallungen.
GBB_43	Durchfälle.
GBB_45	Stiche, Schmerzen oder Ziehen in der Brust.
GBB_46	Zittern.
GBB_48	Leichtes Erröten.
GBB_51	Magenschmerzen.
GBB_52	Anfallsweise Atemnot.
GBB_53	Unterleibsschmerzen.
GBB_56	Anfallsweise Herzbeschwerden.
GT	**Die Aussage stimmt...**
GT_5	Ich habe den Eindruck, dass ich mir eher selten über meine Probleme Gedanken mache.
GT_8	Ich halte mich für wenig ängstlich.
GT_23	Ich glaube, ich bin eher darauf eingestellt, dass man mich für minderwertig hält.
GT_32	Ich glaube, ich mache mir verhältnismäßig selten Sorgen um andere Menschen.
GT_38	Ich glaube, ich habe es im Vergleich mit anderen eher leicht, bei einer Sache zu bleiben.
LZI_5[121]	Ich bin augenblicklich zufrieden mit: meiner Stimmung.
NI	**Die Aussage stimmt...**
NI-90_1	Ich habe manchmal plötzlich furchtbare Angst, schwer krank werden zu können.
NI-90_4	Es könnte mir schon gefallen, einmal so richtig im Mittelpunkt zu stehen.
NI-90_6	Man kann sich furchtbar schämen, wenn man glaubt, versagt zu haben.
NI-90_8	Manchmal quält mich das unbestimmte Gefühl, irgendetwas sei mit meinem Körper nicht in Ordnung.
NI-90_11	In manchen Zeiten sehe ich alles so schwarz, dass mich eine furchtbare Panik ergreift.
NI-90_13	Es gibt Stunden, in denen ich das Gefühl habe, gar nicht wirklich da zu sein.
NI-90_14	Menschenansammlungen schrecken mich eher ab.
NI-90_21	Ich beobachte meinen Körper ziemlich genau, um verdächtige Krankheiten früh zu entdecken.
NI-90_22	Ich erlebe mich manchmal wie eine fremde Person.
NI-90_31	Wenn ich mich im Spiegel sehe, habe ich manchmal das Gefühl, als wäre ich das gar nicht richtig selbst.
NI-90_32	Die Vorstellung selbst mal im Rampenlicht zu stehen, ist schon verführerisch.
NI-90_43	Ich schäme mich, wenn andere merken, dass ich etwas nicht kann.
NI-90_45	Es ist mir meistens unheimlich peinlich, wenn ich vor einer Gruppe etwas Dummes gesagt habe.
NI-90_48	Mitunter bin ich so von Angst und Unruhe getrieben, dass ich weder ein noch aus weiß.
NI-90_49	Ich würde mich auf sehr viel mehr Herausforderungen einlassen, wenn ich nicht Angst hätte, meine Gesundheit würde das nicht durchstehen.
NI-90_62	Es macht mich völlig unsicher, wenn sich in einer Gruppe die Aufmerksamkeit aller plötzlich auf mich richtet.
NI-90_63	Menschen, die attraktiv sind, machen mich unsicher.
NI-90_70	Manchmal erscheint mir mein Körper plötzlich fremd und nicht zu mir dazugehörig.
NI-90_71	Ich bin sehr sprunghaft in meinen Gedanken und Gefühlen.
NI-90_87	Es beunruhigt mich, dass heutzutage von so vielen neuen Krankheiten berichtet wird.
PGWI_4[122]	Haben Sie im vergangenen Monat Ihr Verhalten, Ihre Gedanken und Ihre Gefühle fest im Griff gehabt?
PGWI_5	Haben Sie im vergangenen Monat unter Nervosität oder Ihren „Nerven“ gelitten?
PGWI_8	Waren Sie im allgemeinen angespannt oder haben Sie im vergangenen Monat irgendwelche Spannungen verspürt?
PGWI_13	Haben Sie im vergangenen Monat wegen Ihrer Gesundheit Sorgen oder Befürchtungen gehabt?
PGWI_17	Waren Sie im vergangenen Monat ängstlich, besorgt oder aufgeregt?
PGWI_18	Im vergangenen Monat war ich ausgeglichen und mir meiner selbst sicher.
PGWI_19	Haben Sie sich im vergangenen Monat entspannt und gelassen oder angespannt und aufgeregt gefühlt?

[121] Instruktion des LZIs ist bereits in Itemtext enthalten.
[122] Instruktionen des PGWIs sind bereits im Itemtext enthalten.

Tabelle 31 (Fortsetzung 2): Initialer Itempool, aus dem in einem konsensuellen (Delphi-) Entscheidungsprozess „angstrelevante“ Items selektiert wurden (N = 125).

Fragebögen	Items
PSQ	**Wie häufig trifft diese Feststellung im allgemeinen auf Sie zu?**
PSQ_9	Sie fürchten Ihre Ziele nicht erreichen zu können.
PSQ_10	Sie fühlen sich ruhig.
PSQ_14	Sie fühlen sich angespannt.
PSQ_17	Sie fühlen sich sicher und geschützt.
PSQ_18	Sie haben viele Sorgen
PSQ_22	Sie haben Angst vor der Zukunft.
PSQ_25	Sie sind leichten Herzens.
PSQ_27	Sie haben Probleme sich zu entspannen.
SF36_9B[123]	Wie oft waren Sie in den vergangenen 4 Wochen sehr nervös?
SF36_9D	Wie oft waren Sie in den vergangenen 4 Wochen ruhig und gelassen?
SF36_11C	Ich erwarte, dass meine Gesundheit nachläßt.
SKT_6[124]	Könnten Ihre Beschwerden daher kommen, dass Sie an inneren Ängsten leiden?
SKT_8	Könnten Ihre Beschwerden daher kommen, dass es Ihnen an Selbstvertrauen fehlt?
SKT_9	Könnten Ihre Beschwerden daher kommen, dass Sie durch Sorgen und Probleme in Partnerschaft und Familie belastet sind?
STAI	**Wie fühlen Sie sich jetzt, d. h. in diesem Moment?**
STAI_1	Ich bin ruhig.
STAI_2	Ich fühle mich geborgen.
STAI_3	Ich fühle mich angespannt.
STAI_4	Ich bin bekümmert.
STAI_5	Ich bin gelöst.
STAI_6	Ich bin aufgeregt.
STAI_7	Ich bin besorgt, dass etwas schiefgehen könnte.
STAI_8	Ich fühle mich ausgeruht.
STAI_9	Ich bin beunruhigt.
STAI_10	Ich fühle mich wohl.
STAI_11	Ich fühle mich selbstsicher.
STAI_12	Ich bin nervös.
STAI_13	Ich bin zappelig.
STAI_14	Ich bin verkrampft.
STAI_15	Ich bin entspannt.
STAI_16	Ich bin zufrieden.
STAI_17	Ich bin besorgt.
STAI_18	Ich bin überreizt.
STAI_19	Ich bin froh.
STAI_20	Ich bin vergnügt.
SWO_8[125]	Schwierigkeiten sehe ich gelassen entgegen, weil ich mich immer auf meine Fähigkeiten verlassen kann.

ADS: Allgemeine-Depressions-Skala (Hautzinger & Bailer, 1993).
ALL: Fragebogen zum Alltagsleben (Bullinger, Kirchberger & Steinbüchel, 1993).
BDI: Beck-Depressions-Inventar (Hautzinger, Bailer, Worall & Keller, 1994).
BSF: Berliner-Stimmungs-Fragebogen (Hörhold & Klapp, 1993; Rose et al., 2003).
GBB: Gießener-Beschwerde-Bogen (Brähler & Scheer, 1995).
GT: Gießen-Test Selbst & Idealselbst (Beckmann, Brähler & Richter, 1991).
LZI: Lebens-Zufriedentheits-Inventar (Muthny, 1991).
NI: Narzissmus-Inventar (NI: Deneke & Hilgenstock, 1989; NI-90: Schöneich et al., 2000).
PGWI: Psychological General Wellbeing Index (Ludwig, Geier & Bullinger, 1990).
PSQ: Perceived Stress Questionnaire (Levenstein et al., 1993).
SF36: Fragebogen zum Gesundheitszustand (Bullinger & Kirchberger, 1998).
SKT: Subjektive-Krankheitstheorien-Ursachenvorstellung (Faller, 1997).
STAI: State Trait Anxiety Inventory (Laux, Glanzmann, Schaffner & Spielberger, 1981).
SWO: Fragebogen zu Selbstwirksamkeit, Optimismus und Pessimismus (Scholler et al., 1999).

[123] SF-36: Die Instruktion ist bereits im Itemtext enthalten.
[124] SKT: Die Instruktion ist bereits im Itemtext enthalten.
[125] Instruktion des SWOs lautet: „Bei den folgenden Fragen bitten wir um Ihre Einschätzung von Einstellungen und Gefühlen. Hierzu können Sie jeweils einen Wert von 0 bis 3 auf der folgenden Skala angeben“.

9.2. Ergebnisse der Analyse residualer Kovarianzen

9.2.1. Erste Teilstichprobe

	STAI01	STAI02	STAI03	STAI05	STAI06
STAI01					
STAI02	.046				
STAI03	.066	-.038			
STAI05	.023	.116	-.003		
STAI06	.048	-.124	.039	-.123	
STAI07	-.052	-.072	-.023	-.116	.096
STAI09	-.060	-.084	.030	-.119	.045
STAI10	.023	.139	-.016	.108	-.142
STAI11	-.005	.098	-.103	.072	-.046
STAI12	.064	-.064	.061	-.119	.140
STAI13	.070	-.094	.032	-.145	.081
STAI14	.016	-.030	.067	-.008	.028
STAI15	.059	.068	.022	.096	-.070
STAI17	-.079	-.088	.012	-.103	-.015
STAI18	.030	-.049	.115	-.073	.032
GBB02	-.006	-.041	-.011	-.043	.028
GBB36	-.068	.007	-.013	-.046	-.088
BSF03	-.007	.096	-.044	.138	-.167
BSF05	-.145	-.094	-.052	-.124	-.079
BSF12	-.085	-.140	-.040	-.135	-.032
BSF14	.012	-.118	.028	-.097	.118
BSF20	-.003	.106	-.084	.087	-.109
BSF23	-.057	.003	-.061	-.068	.020
SF09B	-.006	-.028	-.043	-.103	-.013
SF09D	-.009	.015	-.053	.005	-.154
ADS01	-.027	-.078	-.055	-.056	-.054
ADS03	-.034	.008	-.004	-.046	-.102
ADS07	-.056	-.030	-.082	-.055	-.058
BDI20	-.053	-.059	-.032	-.056	-.039
SWO08	-.059	.079	-.092	.008	-.061

	STAI07	STAI09	STAI10	STAI11	STAI12
STAI09	.147				
STAI10	-.110	-.064			
STAI11	-.037	-.111	.094		
STAI12	.019	.001	-.134	-.082	
STAI13	-.031	-.042	-.121	-.108	.176
STAI14	-.019	-.027	-.046	-.009	.033
STAI15	-.077	-.105	.086	.077	-.080
STAI17	.131	.147	-.050	-.117	-.048
STAI18	.027	-.001	-.026	-.111	.074
GBB02	-.106	-.036	-.026	-.107	.025
GBB36	-.088	-.076	.058	-.053	-.057
BSF03	-.142	-.097	.060	.036	-.146
BSF05	-.013	.008	-.097	-.150	-.109
BSF12	-.006	.046	-.094	-.165	-.068
BSF14	-.046	-.065	-.127	-.146	.050
BSF20	-.140	-.136	.076	.082	-.115
BSF23	.025	-.055	-.058	.146	-.038
SF09B	-.074	-.099	-.103	-.091	.056
SF09D	-.100	-.065	.012	-.012	-.082
ADS01	.004	.002	-.080	-.045	-.054
ADS03	-.122	-.127	-.017	.000	-.073
ADS07	-.004	.008	-.066	.029	-.072
BDI20	.038	.045	.025	-.036	-.090
SWO08	-.043	-.115	-.020	.263	-.087

	STAI13	STAI14	STAI15	STAI17	STAI18
STAI14	.089				
STAI15	-.073	.062			
STAI17	-.108	-.061	-.129		
STAI18	.107	.053	-.040	.050	
GBB02	.018	-.007	-.081	-.091	-.038
GBB36	.002	.045	-.014	-.090	.007
BSF03	-.207	-.048	.103	-.117	-.096
BSF05	-.148	-.086	-.163	.135	-.052
BSF12	-.113	-.073	-.158	.075	-.040
BSF14	.181	.012	-.066	-.100	.017
BSF20	-.121	-.054	.042	-.090	-.051
BSF23	-.084	-.005	-.022	-.058	-.088
SF09B	.068	-.007	-.089	-.118	.046
SF09D	-.042	-.018	.024	-.091	-.033
ADS01	-.051	-.038	-.060	.009	-.024
ADS03	-.007	-.032	.011	-.118	-.022
ADS07	-.118	-.038	-.073	-.024	-.117
BDI20	-.035	.030	-.103	.074	.033
SWO08	-.099	-.063	-.011	-.138	-.114

	GBB02	GBB36	BSF03	BSF05	BSF12
GBB36	.152				
BSF03	-.034	.016			
BSF05	-.002	-.027	-.060		
BSF12	.050	.007	-.074	.241	
BSF14	.093	.016	-.110	.026	.086
BSF20	-.036	-.006	.157	-.056	-.077
BSF23	-.014	.042	.021	.015	.026
SF09B	.085	.011	-.059	-.009	-.012
SF09D	-.009	.002	.056	-.068	-.051
ADS01	.025	.041	.008	-.011	.027
ADS03	.023	.181	.067	-.050	-.046
ADS07	.095	.069	-.035	.037	.044
BDI20	.078	.054	-.086	.060	.101
SWO08	-.045	-.035	.074	-.093	-.131

	BSF14	BSF20	BSF23	SF09B	SF09D
BSF20	-.069				
BSF23	-.016	.017			
SF09B	.123	-.011	-.049		
SF09D	-.014	.094	-.020	.140	
ADS01	-.039	-.043	-.029	.056	.058
ADS03	-.018	.048	.005	.076	.079
ADS07	-.038	-.027	.073	.060	.045
BDI20	-.010	-.020	-.023	-.035	-.012
SWO08	-.136	.064	.181	.007	.057

	ADS01	ADS03	ADS07	BDI20	SWO08
ADS03	.135				
ADS07	.175	.073			
BDI20	.049	-.028	.064		
SWO08	-.020	.062	.055	-.013	

9.2.2. Zweite Teilstichprobe

	ALLT21	BSF03	BSF05	BSF12	BSF14
ALLT21					
BSF03	.071				
BSF05	-.051	-.038			
BSF12	-.020	.012	.192		
BSF14	.015	-.007	.036	.094	
BSF20	.084	.190	-.077	-.014	.044
BSF23	-.038	-.004	.060	.062	.068
GBB20	-.022	-.019	-.006	.030	.098
GBB36	.065	.064	.019	.061	.094
NI1	-.041	-.108	.060	.001	-.039
NI11	-.043	-.114	-.106	-.093	-.050
NI13	-.075	-.116	-.134	-.135	-.099
NI14	-.060	-.091	-.093	-.101	-.010
NI22	-.101	-.123	-.163	-.166	-.086
NI48	-.039	-.119	-.088	-.099	.001
NI62	-.040	-.131	-.047	-.125	-.044
NI70	-.108	-.192	-.163	-.200	-.104
SKT06	-.083	-.078	-.104	-.098	-.048
PSQ09	-.044	-.095	-.038	-.047	-.052
PSQ10	.098	.079	-.116	-.041	.076
PSQ14	-.057	.006	.038	.053	-.054
PSQ17	.019	.027	-.111	-.088	-.070
PSQ18	-.045	-.126	.027	-.033	-.040
PSQ22	-.063	-.114	-.024	-.031	-.056
PSQ25	.038	.107	-.124	-.091	-.029
PSQ27	.132	.015	-.090	-.067	.024
SWO08	.001	-.054	-.093	-.078	-.071

	BSF20	BSF23	GBB20	GBB36	NI1
BSF23	-.002				
GBB20	-.015	.039			
GBB36	.024	.051	.253		
NI1	-.115	.010	.067	-.013	
NI11	-.113	-.029	-.032	-.082	.106
NI13	-.126	-.032	-.020	.022	-.071
NI14	-.053	-.004	.005	-.055	.037
NI22	-.122	-.064	-.011	-.054	.000
NI48	-.098	-.024	.007	-.034	.108
NI62	-.110	.075	-.067	-.124	.077
NI70	-.166	-.049	-.014	-.079	.007
SKT06	-.089	-.005	.009	-.053	.037
PSQ09	-.086	.014	-.079	-.052	-.027
PSQ10	.120	-.092	-.051	.034	-.060
PSQ14	.036	.029	-.018	-.042	.022
PSQ17	.049	-.034	-.052	.005	-.073
PSQ18	-.096	-.006	-.064	-.070	.004
PSQ22	-.106	.004	-.037	-.082	.033
PSQ25	.069	-.061	-.027	-.005	-.076
PSQ27	-.009	-.026	-.021	.018	-.039
SWO08	.002	.023	-.053	-.068	-.032

	NI11	NI13	NI14	NI22	NI48
NI13	.069				
NI14	.049	.082			
NI22	.018	.166	.019		
NI48	.104	.037	.058	-.002	
NI62	.033	.070	.240	.035	.029
NI70	.028	.171	.022	.274	.006
SKT06	.072	-.041	.051	-.048	.107
PSQ09	-.010	-.039	-.029	-.052	-.050
PSQ10	-.093	-.124	-.063	-.090	-.065
PSQ14	.083	.041	-.007	.079	.044
PSQ17	-.030	-.062	-.008	-.089	-.045
PSQ18	-.004	-.031	-.012	-.098	.006
PSQ22	.016	-.020	-.039	-.059	-.013
PSQ25	-.085	-.109	-.047	-.118	-.063
PSQ27	-.036	-.061	-.013	-.107	-.041
SWO08	.019	-.017	-.006	-.053	-.005

	NI62	NI70	SKT06	PSQ09	PSQ10
NI70	.079				
SKT06	.007	-.091			
PSQ09	.032	-.040	-.005		
PSQ10	-.109	-.121	-.046	-.076	
PSQ14	.048	.125	.014	-.048	-.055
PSQ17	-.079	-.133	.027	-.015	.113
PSQ18	.000	-.093	.037	.093	-.102
PSQ22	-.002	-.057	.042	.122	-.079
PSQ25	-.042	-.148	-.057	-.022	.153
PSQ27	-.023	-.141	.020	-.022	.100
SWO08	.064	-.052	.046	.072	-.011

	PSQ14	PSQ17	PSQ18	PSQ22	PSQ25
PSQ17	.036				
PSQ18	-.048	.074			
PSQ22	.024	.007	.107		
PSQ25	-.073	.165	-.043	-.047	
PSQ27	-.143	-.009	-.001	-.026	.049
SWO08	.034	.078	-.001	.036	.078

	PSQ27	SWO08
SWO08	-.022	

9.2.3. Dritte Teilstichprobe

	BSF03	BSF05	BSF12	BSF14	BSF20
BSF03					
BSF05	-.073				
BSF12	-.067	.194			
BSF14	-.032	.068	.085		
BSF20	.183	-.123	-.111	-.036	
BSF23	-.012	.098	.080	.088	-.015
GBB02	-.065	-.099	-.079	.020	-.053
GBB06	-.055	-.106	-.095	.015	-.101
GBB17	-.080	-.093	-.096	-.073	-.023
GBB18	-.073	-.105	-.096	.029	-.019
GBB20	-.013	-.126	-.099	-.024	-.027
GBB36	-.049	-.123	-.081	.017	-.025
GBB37	-.009	-.104	-.077	.037	-.071
GBB40	-.067	-.099	-.086	-.014	-.063
GBB46	-.029	-.059	-.090	.090	-.057
GBB48	-.063	-.071	-.034	.036	-.026
PGWI05	-.076	-.107	-.084	-.020	-.046
PGWI08	-.042	-.118	-.104	-.049	-.033
PGWI13	-.065	-.029	.023	-.052	-.097
PGWI17	-.076	-.052	-.038	-.055	-.069
PGWI18	.020	-.162	-.109	-.121	.053
PGWI19	-.008	-.158	-.125	-.037	-.011

	BSF23	GBB02	GBB06	GBB17	GBB18
GBB02	-.085				
GBB06	-.081	.112			
GBB17	-.102	.078	.219		
GBB18	.014	.035	.125	.119	
GBB20	-.044	.186	.053	.135	.174
GBB36	-.022	.111	.166	.125	.154
GBB37	-.048	.192	.115	.080	.001
GBB40	-.080	.235	.066	.219	-.011
GBB46	-.033	.168	.126	.188	.032
GBB48	.063	.095	.080	.067	.049
PGWI05	-.050	-.040	-.067	-.064	-.040
PGWI08	-.089	-.093	-.151	-.157	-.058
PGWI13	-.023	-.044	.007	-.024	-.017
PGWI17	.001	-.054	-.071	-.078	-.018
PGWI18	.022	-.157	-.146	-.105	-.030
PGWI19	-.074	-.096	-.068	-.091	-.030

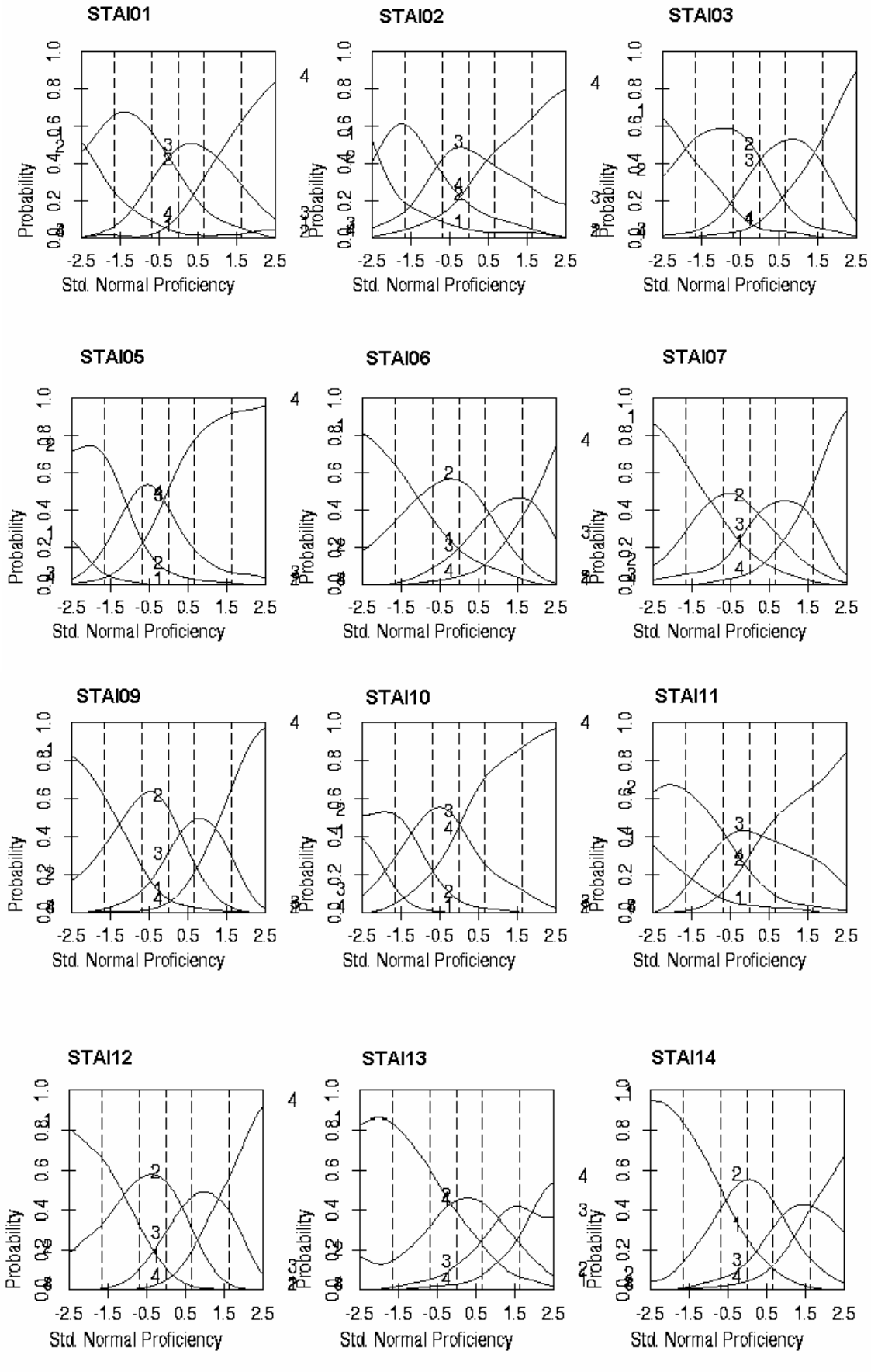
	GBB20	GBB36	GBB37	GBB40	GBB46
GBB36	.128				
GBB37	.150	.175			
GBB40	.163	.153	.174		
GBB46	.129	.116	.166	.173	
GBB48	.109	.061	.107	.151	.114
PGWI05	-.066	-.027	-.020	-.053	-.009
PGWI08	-.132	-.098	-.103	-.126	-.134
PGWI13	-.059	.025	.022	-.012	-.058
PGWI17	-.076	-.088	-.146	-.078	-.057
PGWI18	-.086	-.067	-.178	-.146	-.164
PGWI19	-.087	-.141	-.079	-.128	-.155

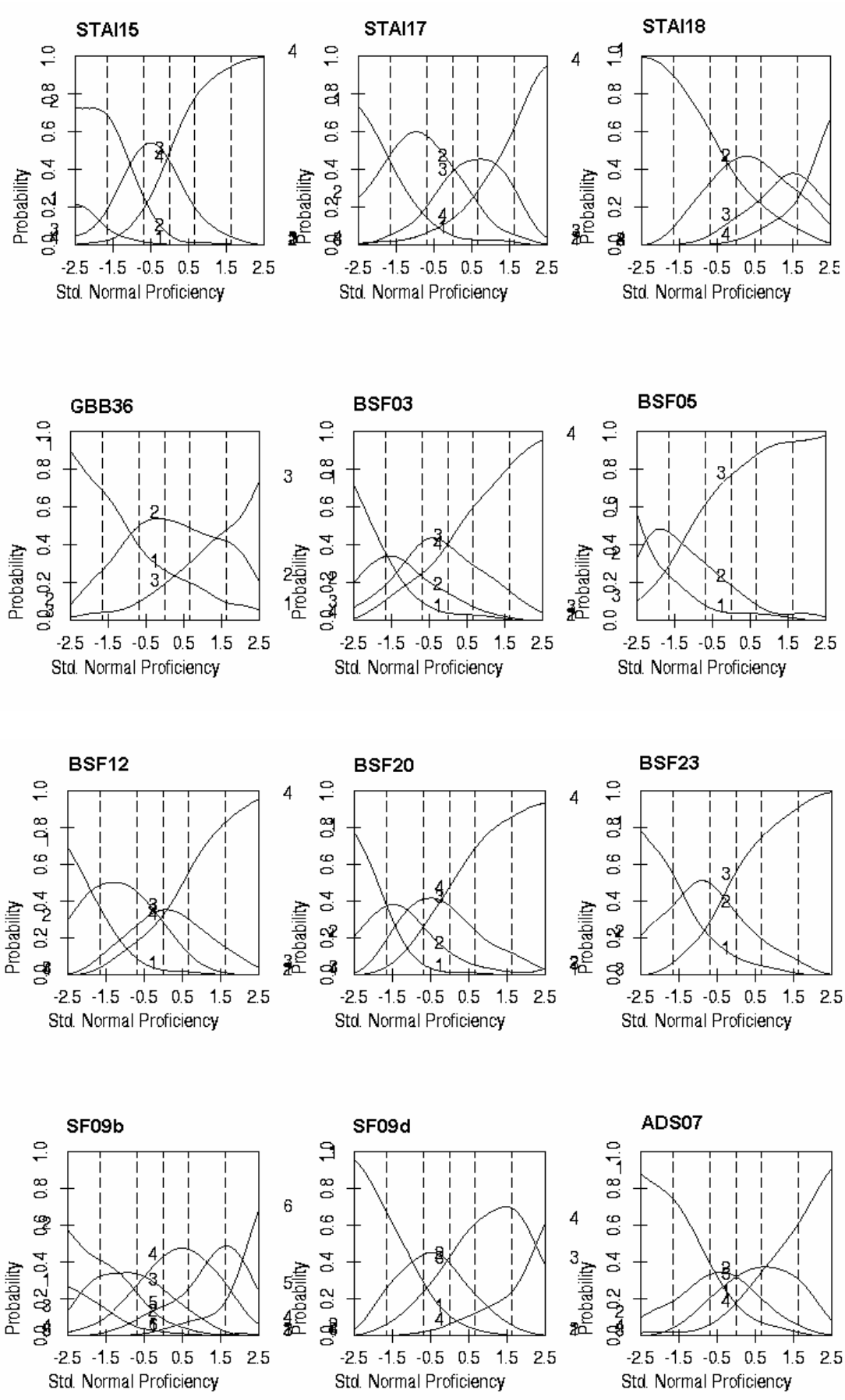
	GBB48	PGWI05	PGWI08	PGWI13	PGWI17
PGWI05	-.056				
PGWI08	-.113	.075			
PGWI13	-.025	.031	.026		
PGWI17	-.037	.063	.028	.075	
PGWI18	-.104	.032	.022	.030	.050
PGWI19	-.075	.020	.116	-.023	.021

	PGWI18	PGWI19
PGWI19	.063	

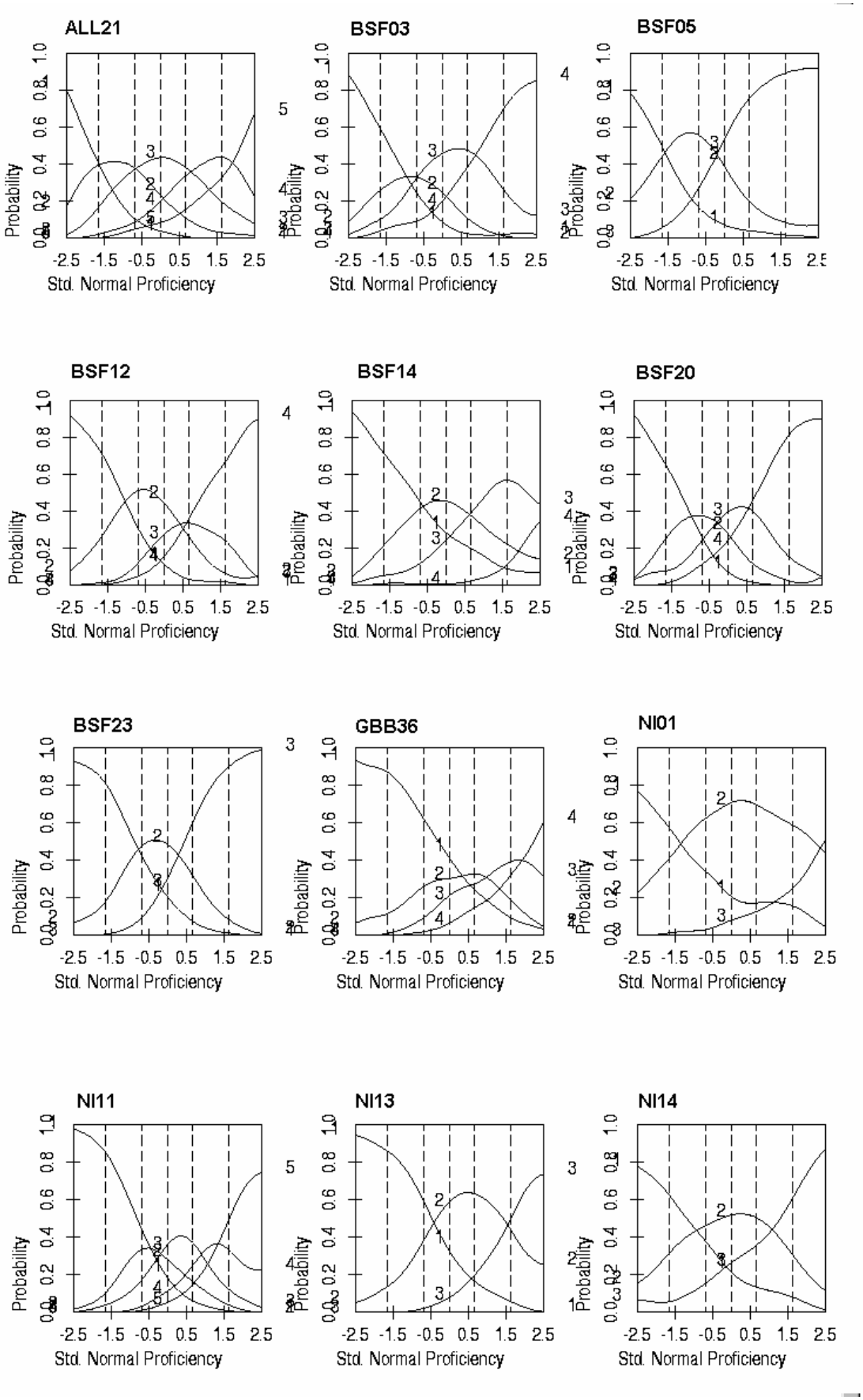
9.3. Ergebnisse der Item Response Curves (IRCs)

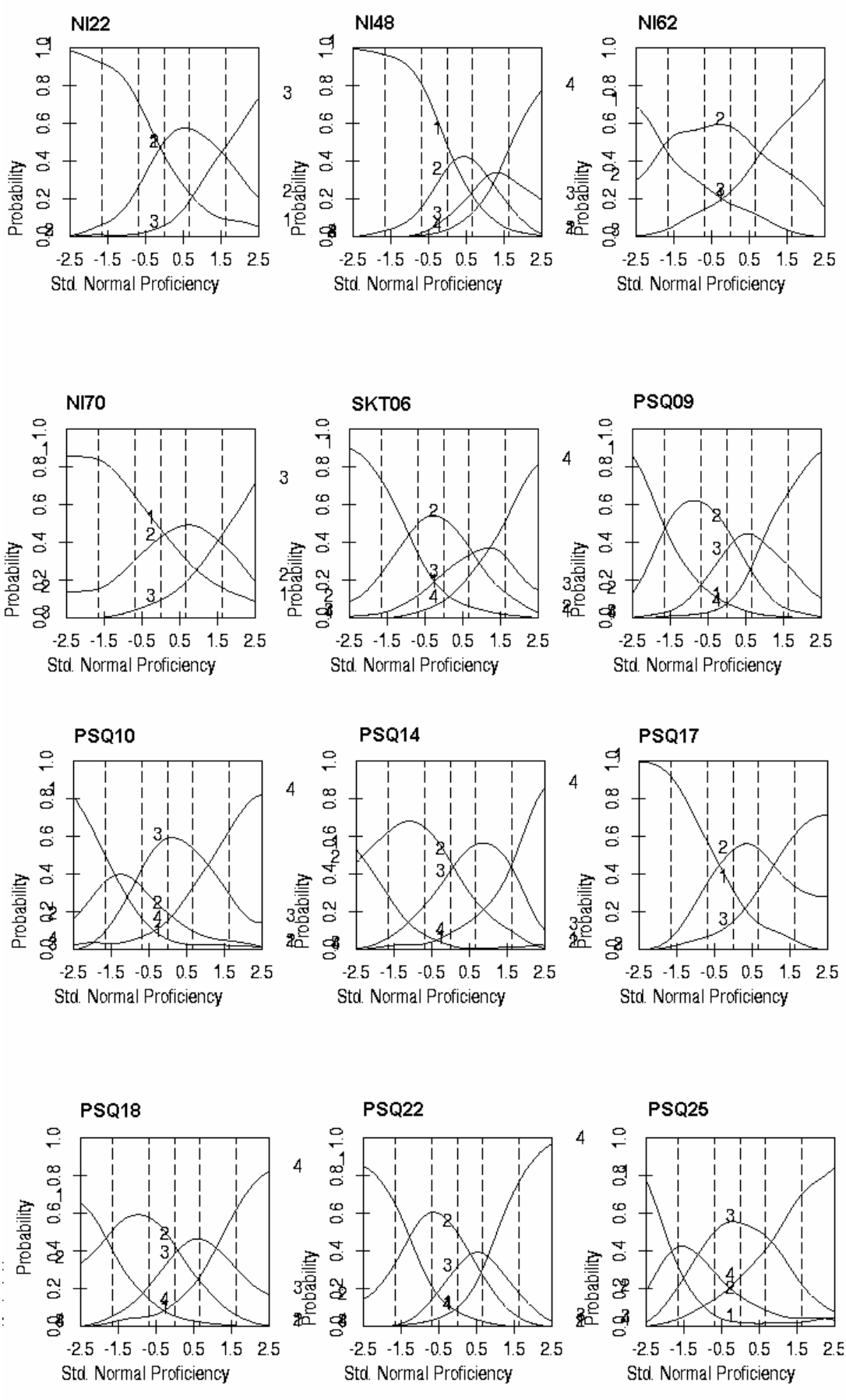
9.3.1. Erste Teilstichprobe

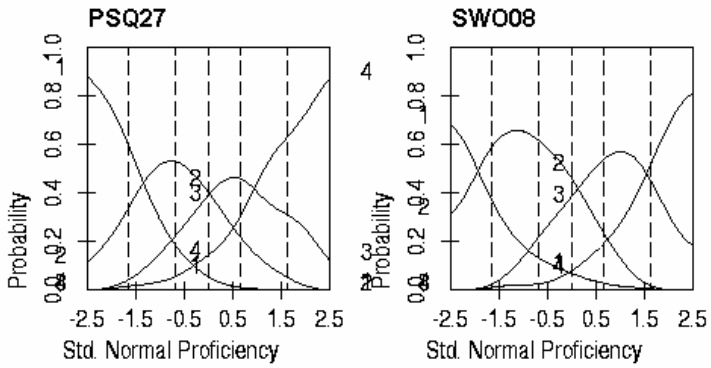



STAI15
STAI17
STAI18
GBB36
BSF03
BSF05
BSF12
BSF20
BSF23
SF09b
SF09d
ADS07
Probability
Std. Normal Proficiency
-2.5 -1.5 -0.5 0.5 1.5 2.5

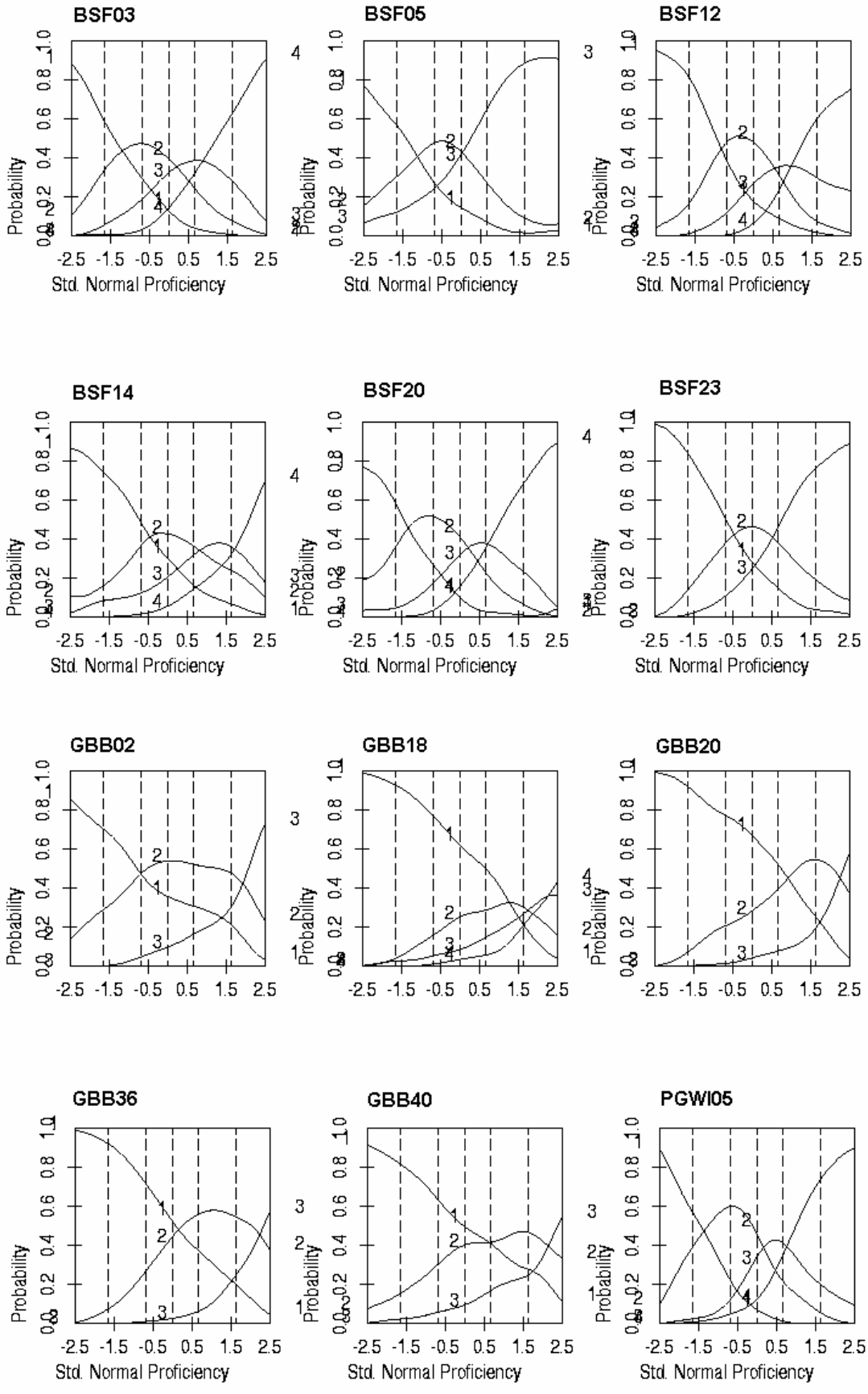
9.3.2. Zweite Teilstichprobe

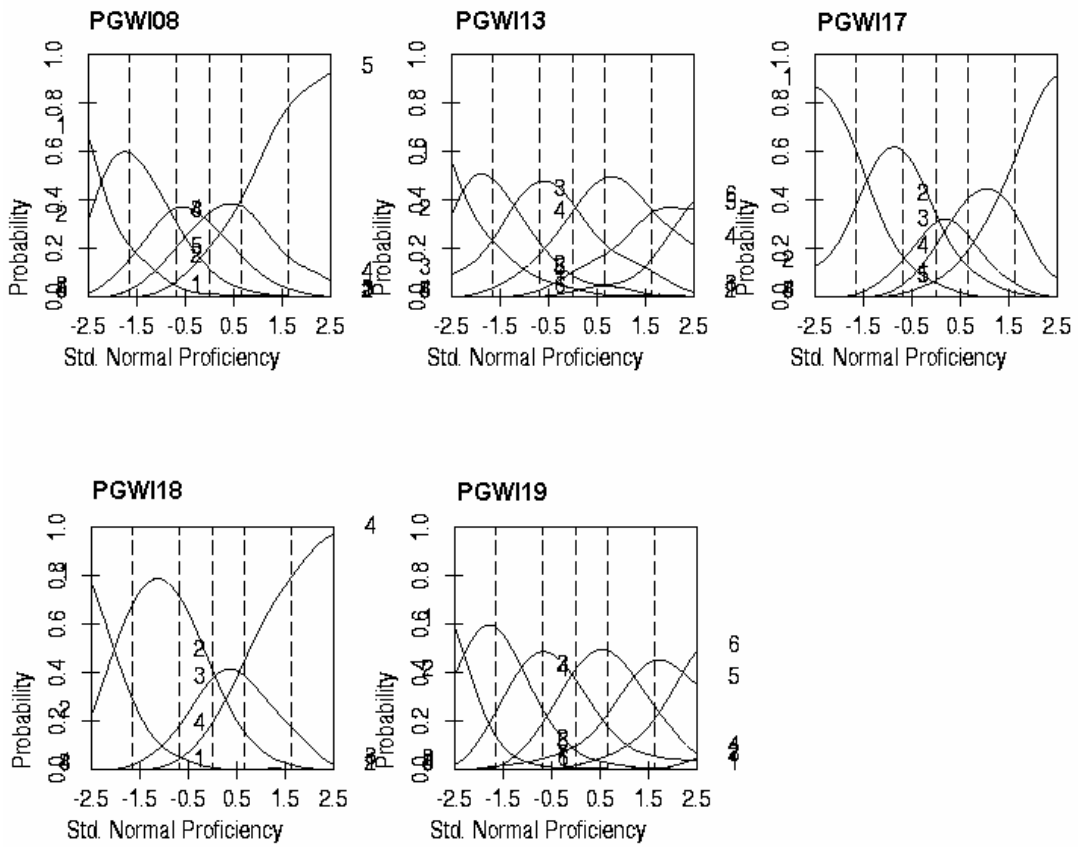



NI22
NI48
NI62
NI70
SKT06
PSQ09
PSQ10
PSQ14
PSQ17
PSQ18
PSQ22
PSQ25
Probability
Std. Normal Proficiency
-2.5 -1.5 -0.5 0.5 1.5 2.5
0.0 0.2 0.4 0.6 0.8 1.0


PSQ27
Probability
0.0 0.2 0.4 0.6 0.8 1.0
-2.5 -1.5 -0.5 0.5 1.5 2.5
Std. Normal Proficiency
SWO08
Probability
0.0 0.2 0.4 0.6 0.8 1.0
-2.5 -1.5 -0.5 0.5 1.5 2.5
Std. Normal Proficiency

9.3.3. Dritte Teilstichprobe





Zu den Abbildungen 26. des Anhangs 9.3.:

Item Response Curves (IRC) der analysierten Items der drei Teilstichproben. IRCs sind grafische Darstellungen der (Antwort-) Kategorienfunktionen von Items und veranschaulichen die Antwortwahrscheinlichkeit einzelner Antwortkategorien („Probability“, Ordinate) in Abhängigkeit von der latenten Merkmalsausprägung („Std. Normal Proficiency“, Abszisse) der Angst, welches in Einheiten der abweichungsnormierten Standardnormalverteilung dargestellt ist.

9.4. Abbildungsverzeichnis

9.5. Tabellenverzeichnis

Eidesstattliche Erklärung

Ich erkläre an Eides Statt, dass ich die beiliegende Dissertation selbständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen nicht benutzt, und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Dipl.-Psych. Janine Becker

Curriculum Vitae

06.05.1977	geboren in Duisburg.
1983-1987	Besuch der Primarstufe der *Städt. Gemeinschaftsgrundschule* in Mülheim a. d. Ruhr.
1987-1996	Besuch der Sekundarstufe I und II des *Otto-Pankok-Gymnasiums* in Mülheim a. d. Ruhr.
09/1993-12/1993	Auslandsaufenthalt in Großbritannien am *Kent College in Canterbury*.
05/1996	Abitur mit einem Notendurchschnitt von 1,9.
WS 1996	Studium der Psychologie, Anglistik und Geographie auf Lehramt (Sekundarstufe I und II) an der *Gerhard-Mercator-Universität Gesamthochschule Duisburg.*
1997-2001	Studiums der Psychologie an der *Freien Universität Berlin* mit klinischer und psychodiagnostischer Schwerpunktbildung.
08/1998	Vordiplom mit einem Notendurchschnitt von 1,1.
02/2000-05/2000	Praktikum an der *Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin.*
10/2000-04/2001	Diplomarbeit über das Thema „Selbstbild, Idealbild und Selbstwertregulation autodestruktiver Patienten im Vergleich zu einer psychosomatischen Stichprobe“ an der *Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin,* betreut von Dr. rer. nat. H. Fliege.
10/2001	Diplom mit einem Notendurchschnitt von 1,1.
seit 12/2001	Wissenschaftliche Mitarbeiterin an der *Medizinischen Klinik mit Schwerpunkt Psychosomatik der Charité Berlin* in der Forschungsgruppe „IRT-basierte Computergestützte Adaptive Tests“ (Projektleiter: Dr. med. habil. M. Rose).
04/2002-07/2002	Dreimonatiger Werkvertrag als wissenschaftliche Mitarbeiterin am *Robert-Koch-Institut Berlin* im Projekt „Indikatoren subjektiver Kinder- und Jugendgesundheit“ (Projektleiterin: Dr. phil. U. Ravens-Sieberer).
01/2004	Dissertation mit dem Titel „Computergestütztes Adaptives Testen (CAT) von Angst entwickelt auf der Grundlage der Item Response Theorie (IRT)“ (Betreuer: Prof. Dr. H. Westmeyer und Dr. med. habil. M. Rose).