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.