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.

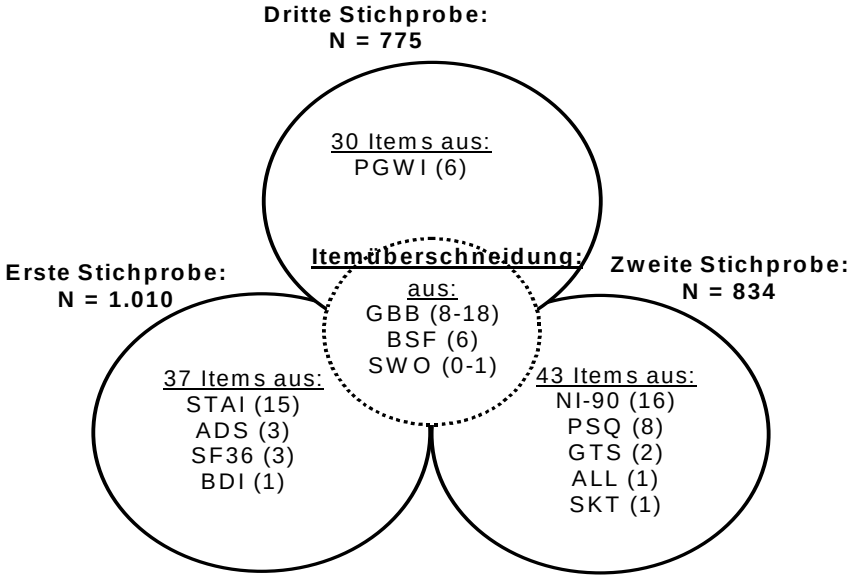


Abbildung 8: Überblick über die drei Teilstichproben, an denen die statistische Itemanalyse und -selektion erfolgte (Testabkürzungen siehe Fußnoten S. 106).

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.

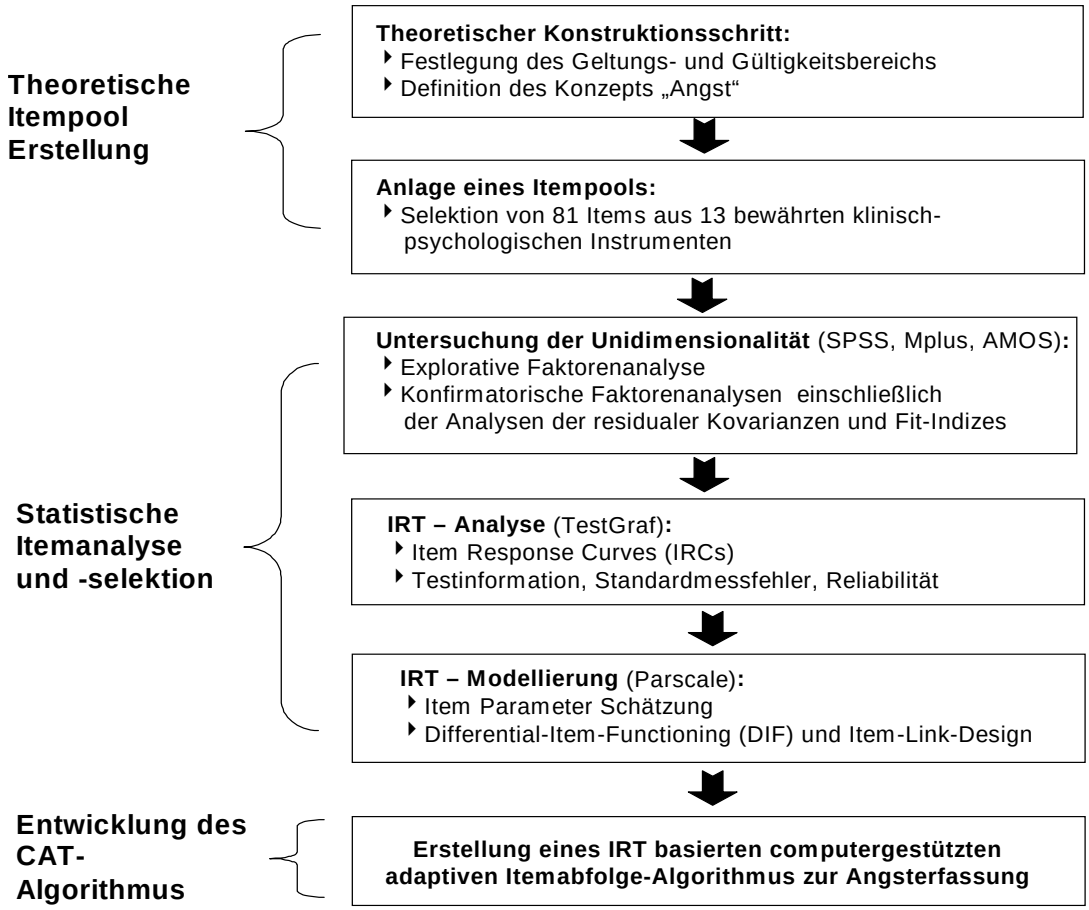


Abbildung 9: Ablaufschema der Entwicklung des IRT-basierten Angst-CATs.

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).

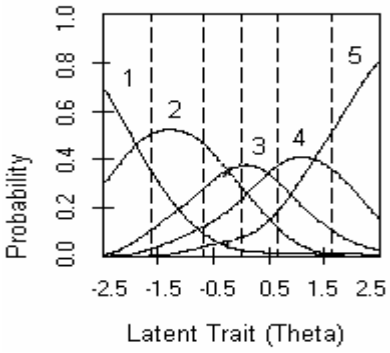


Abbildung 10: Exemplarische Darstellung eines polytomen Items mit modellkonformen Item Response Curves (IRCs).[72]

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).

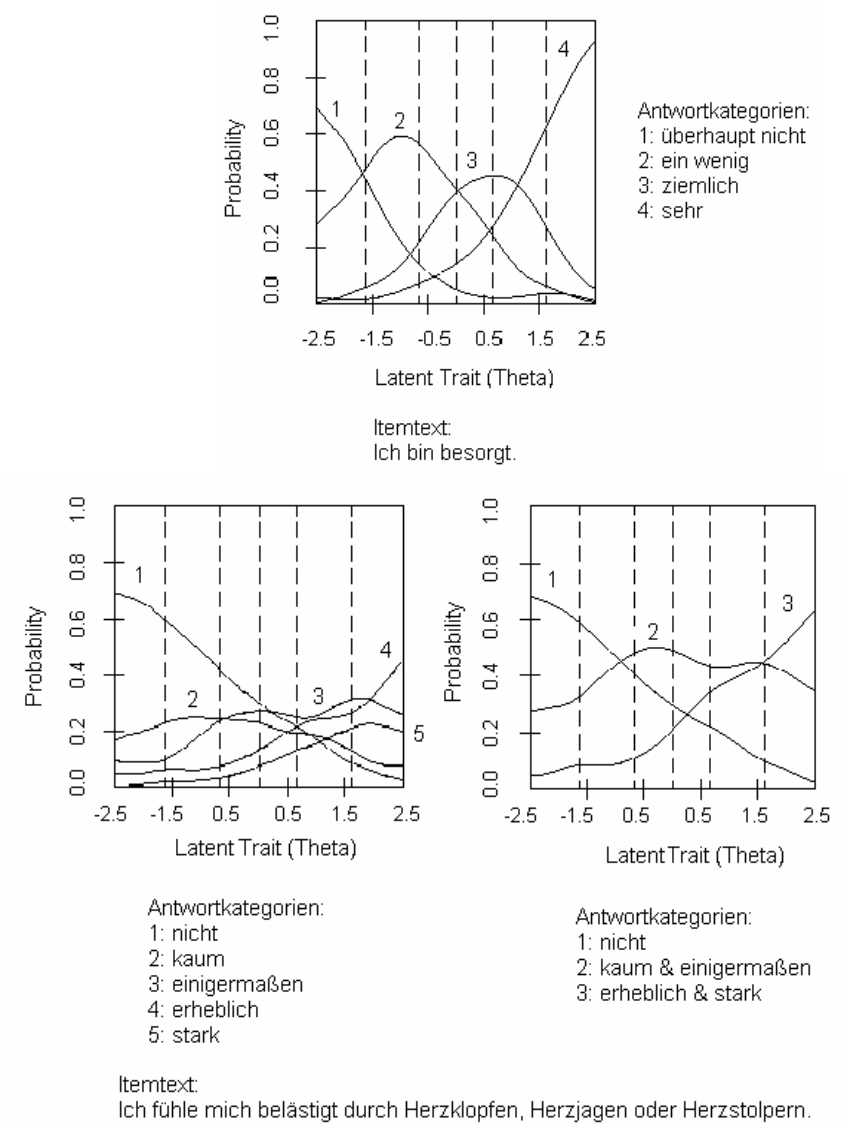


Abbildung 11: IRCs eines Items mit modellkonformer Itemcharakteristik (oben) und eines Items mit nicht modellkonformer Itemcharakteristik[76] (unten links), die ggf. durch das Zusammenlegen von Antwortkategorien verbessert werden kann (unten rechts).

[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.

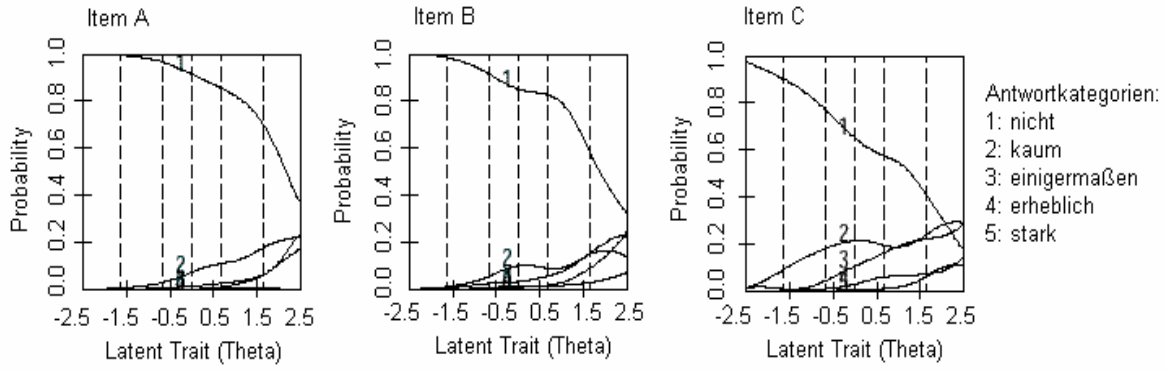


Abbildung 12: Ungenügende IRCs der Items „Ohnmachtsanfälle“ (A), „Anfälle“ (B) und „Leichtes Erröten“ (C).

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.

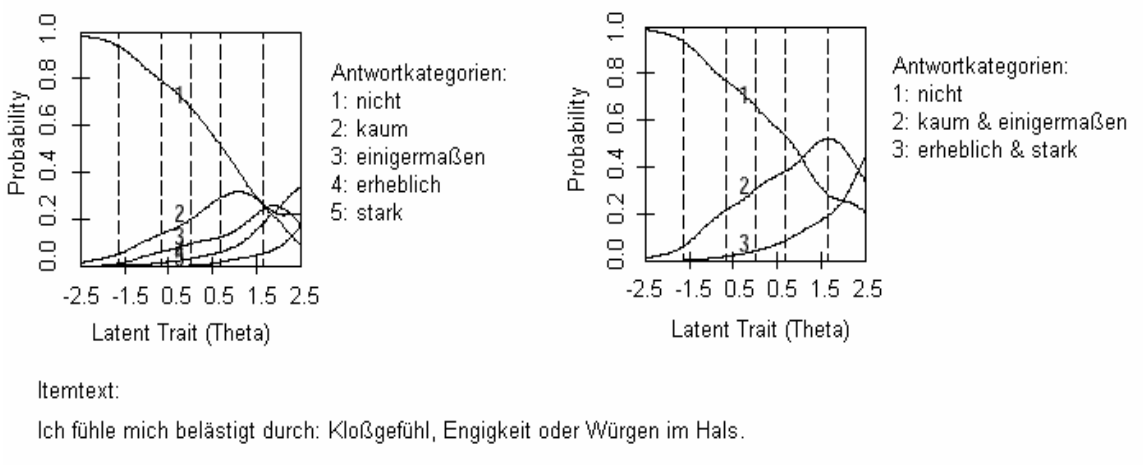


Abbildung 13: Beispiel für eine mögliche Modifikation der IRCs des Items „Kloßgefühl im Hals“.

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.).

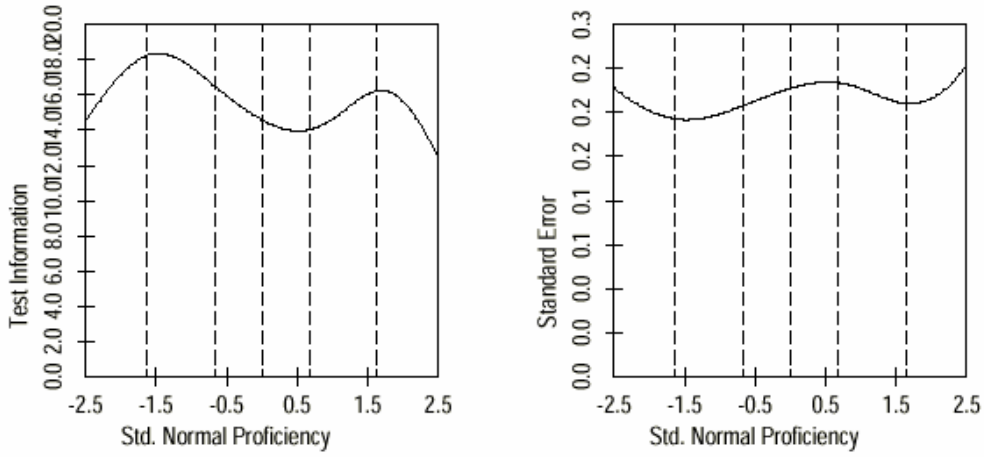


Abbildung 14: Testinformationsniveau (links) und Standardmessfehler (rechts) der selektierten Items der ersten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

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.

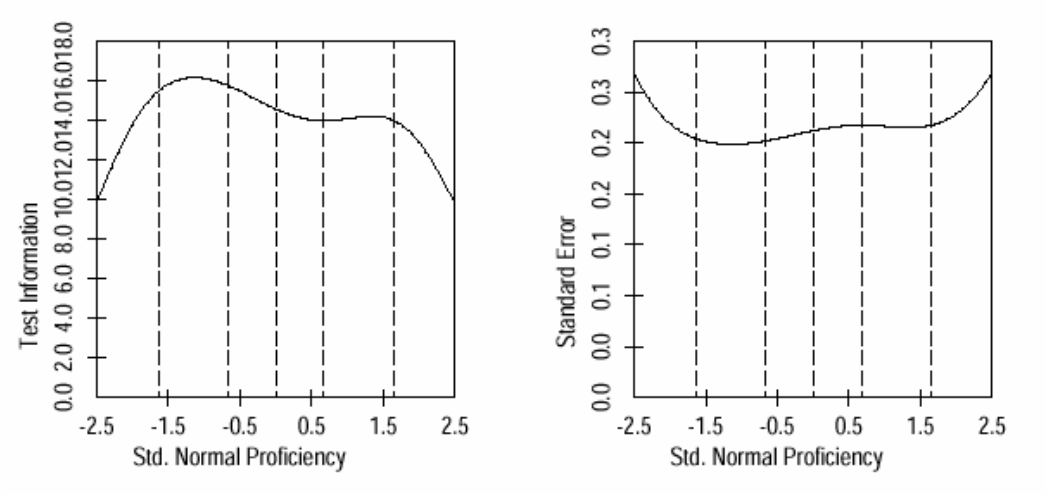


Abbildung 15: Testinformationsniveau (links) und Standardmessfehler (rechts) der selektierten Items der zweiten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung in Einheiten der Standardnormalverteilung).

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.

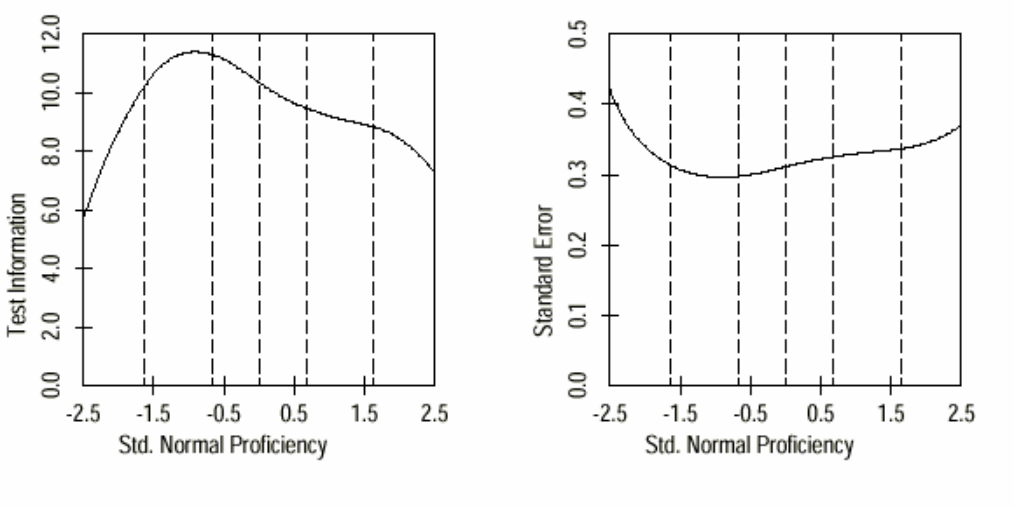
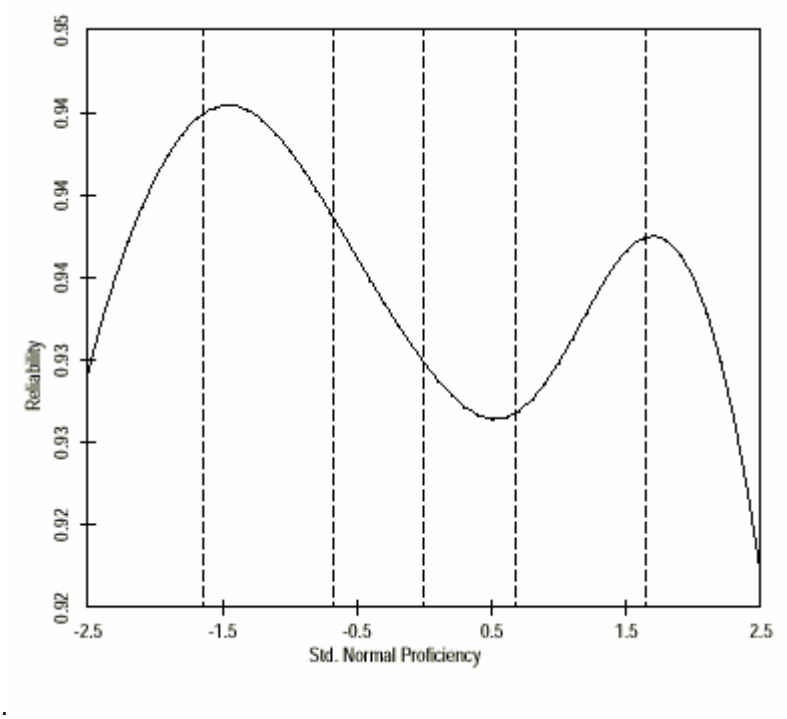


Abbildung 16: Testinformationsniveau (links) und Standardmessfehler (rechts) der selektierten Items der dritten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

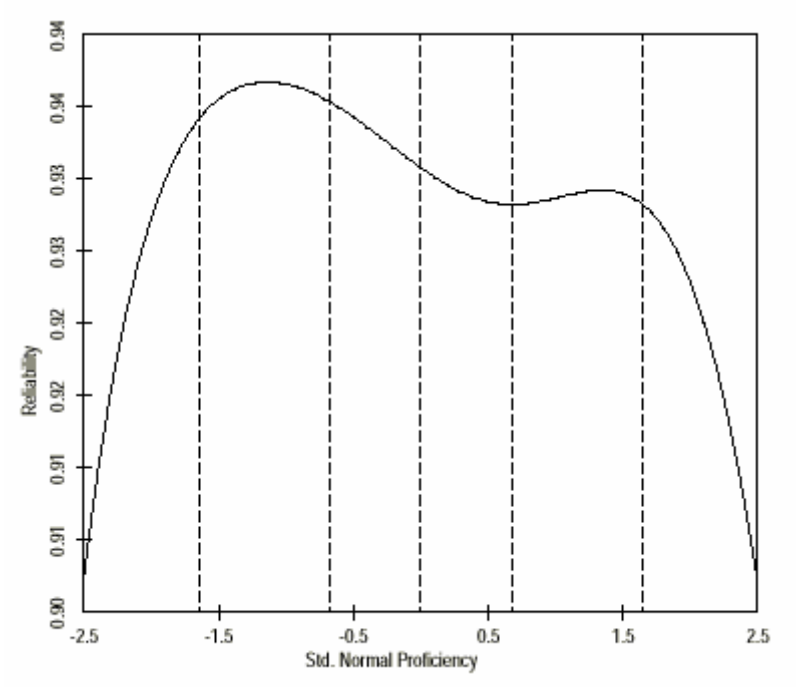
5.4.2.3. Reliabilität

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).

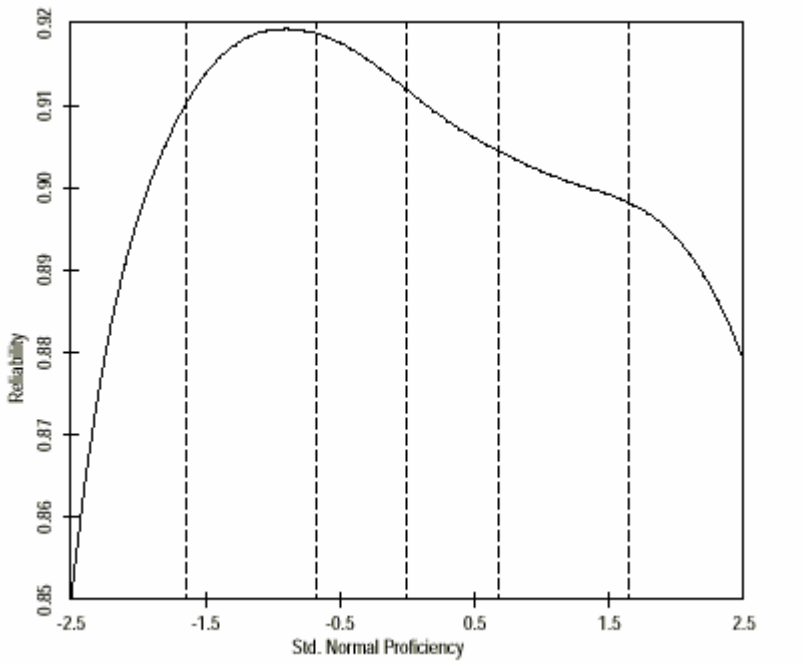


Abbildung 19: Reliabilitäten der selektierten Items aus der dritten Teilstichprobe in Abhängigkeit zur Angstausprägung (Theta-Schätzung; in Einheiten der Standardnormalverteilung).

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.

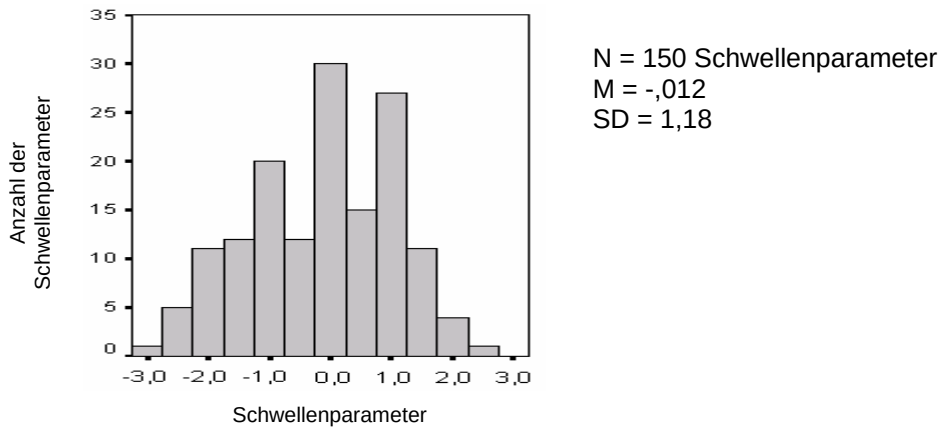


Abbildung 20: Verteilung der Schwellenparameter der Itembank des Angst-CATs.

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.