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, D. 1978). Die Anwendung dieses Modells impliziert, dass Items mit unterschiedlichen Antwortformaten in isolierten Gruppen analysiert werden müssen.

Als allgemeines Ein-Parameter-Modell steht das Partial Credit Modell (PCM; Masters, G. N. (1982) zur Verfügung. Sowohl das RSM wie auch das PCM können als Rasch-Modelle für polytome Daten charakterisiert werden. Von den Zwei-Parameter-Modellen kommen das Graded Response Modell (GRM; Samejima, F. (1996) und die Modifikation dieses Modells durch Muraki (1992; M-GRM) sowie das Generalized Partial Credit Modell (GPCM; Muraki, E. (1997) in Frage.[1]

### 3.3 Simulationsexperimente

Die Güte der erstellten Itembank und dessen Itemabfolgealgorithmus sollte im Rahmen von Simulationsexperimenten überprüft werden (Ben-Porath, Y. S., Slutske, W. S., & Butcher, J. N. 1989; Cook, K. F. 2004; Hornke, L. F. 1999; Ware, J. E., Jr. et al. 2000).

Die Simulation des computeradaptiven Antwortverhaltens gründet sich auf Itemantworten, die real (unter computerassistierten Erhebungsbedingungen) von der untersuchten Personenstichprobe abgegeben werden. Somit kann der adaptive Itemabfolgealgorithmus im Nachhinein an dieser Personenstichprobe simuliert werden. Für dieses Vorgehen muss angenommen werden, wie bei derartigen Simulationen üblich (Gardner, W., Kelleher, K. J., & Pajer, K. A. 2002), dass die damit verbundene andere Abfolge der Präsentation der Items unter realen Bedingungen und der kürzere Testumfang, nur einen nach geordneten Effekt für die Schätzung des Latent-Trait haben (Cook, K. F. 2004).

## 4 Methodik der Entwicklung des Stress-CAT

### 4.1 Stichprobe

Der Itemanalyse und -selektion liegen Daten zugrunde, die an insgesamt 1092 Patienten erhoben wurden, die sich in der Medizinischen Klinik mit Schwerpunkt Psychosomatik zur Diagnostik oder Therapie in dem Zeitraum von 06/2002 bis

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

04/2003 vorgestellt hatten. Sie wiesen unterschiedliche ICD-10-Diagnosen auf (30% depressive Störungen (F32-34/F43), 24% somatoforme Störungen (F45), 18% Essstörungen (F10/F50/F55), 13% Angststörungen (F40-41), 10% primär somatische Erkrankungen) und wurden in unterschiedlichen Settings untersucht (55,3% stationär; 33,4% ambulant; 11,3% konsiliarisch). Wie für psychosomatische Stichproben typisch, waren Frauen mit einem Verhältnis von 2:1 in der Stichprobe überrepräsentiert (w : 67,0%; m : 33,0%). Das Durchschnittsalter lag bei 42,1 Jahren (Standardabweichung (SD): 14,7). 38,7 % der Patienten waren verheiratet, 14,3 % lebten unverheiratet mit einem Partner zusammen, 23,7 % waren ledig (ohne Partner), 5,3 % der Patienten verheiratet, jedoch getrennt vom Partner lebend und 16 % waren geschieden oder verwitwet (2 % missing data).

## 4.2 Entwicklung der Itembank

Das Vorgehen bei der Entwicklung des IRT basierten Computer Adaptiven Tests zur Messung von *Stresserleben* (Stress-CAT) gliedert sich in drei prinzipielle Schritte (siehe Abbildung 7): (a) die inhaltliche Auswahl relevanter Items, (b) die sequentielle statistische Itemanalyse und –selektion dieser Items mit dem Ziel, die Items mit der besten psychometrischen Qualität zur Konstruktion einer Itembank zu nutzen, und (c) die Implementierung der Itembank in einen computergestützten adaptiven Itemabfolge-Algorithmus, der die Präsentation der Items und die Schätzung der individuellen Ausprägung von *Stresserleben* (Theta-Schätzung) von Testpersonen ermöglicht.

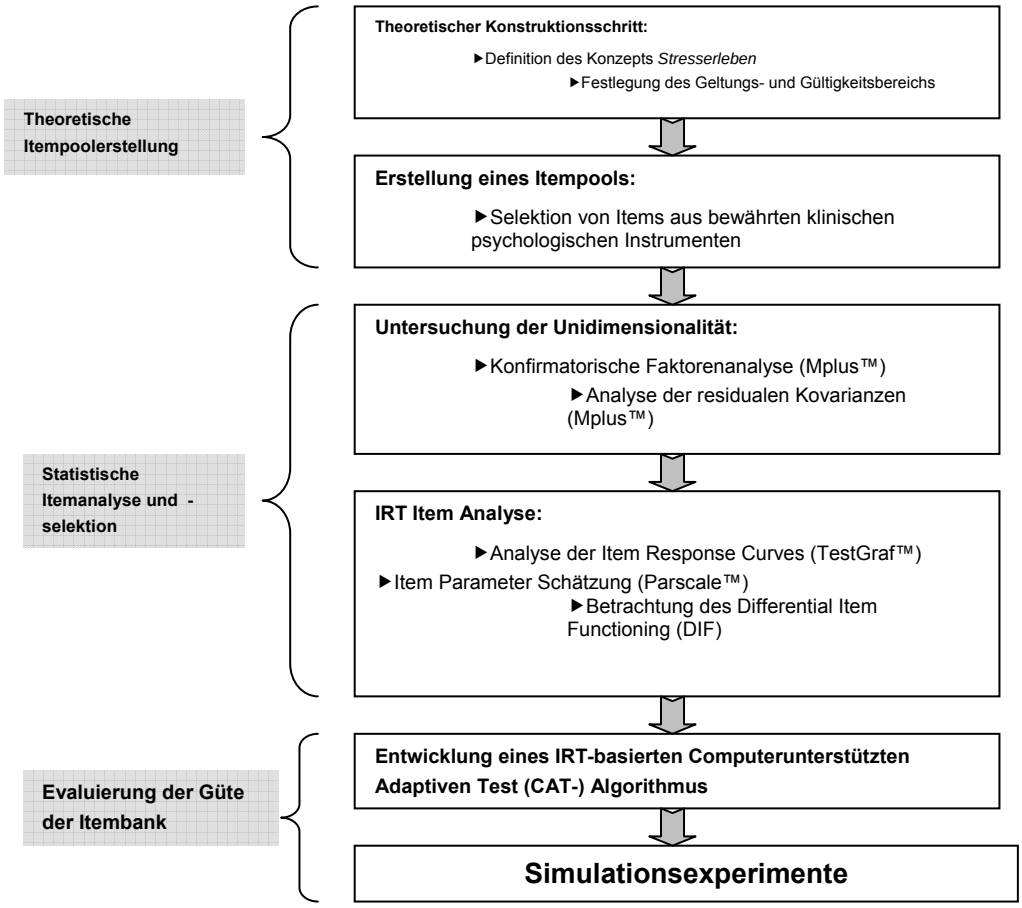


**Abbildung 7: Ablaufschema der Entwicklung des IRT-basierten Stress-CAT**

### 4.2.1 Theoriegeleitete Itemauswahl

In einem theoriegeleiteten Teil wird das Konstrukt *Stresserleben* theoretisch reflektiert und konzeptionell definiert (siehe auch Abschnitt 2.2).

Die vorliegende Arbeit folgt einem relationalen Verständnis von Stress. Operationalisiert werden soll das *Stresserleben* eines Individuums anhand der Dimensionen ‚erlebte Belastung' und ‚Belastungsreaktion' (siehe Abbildung 8).

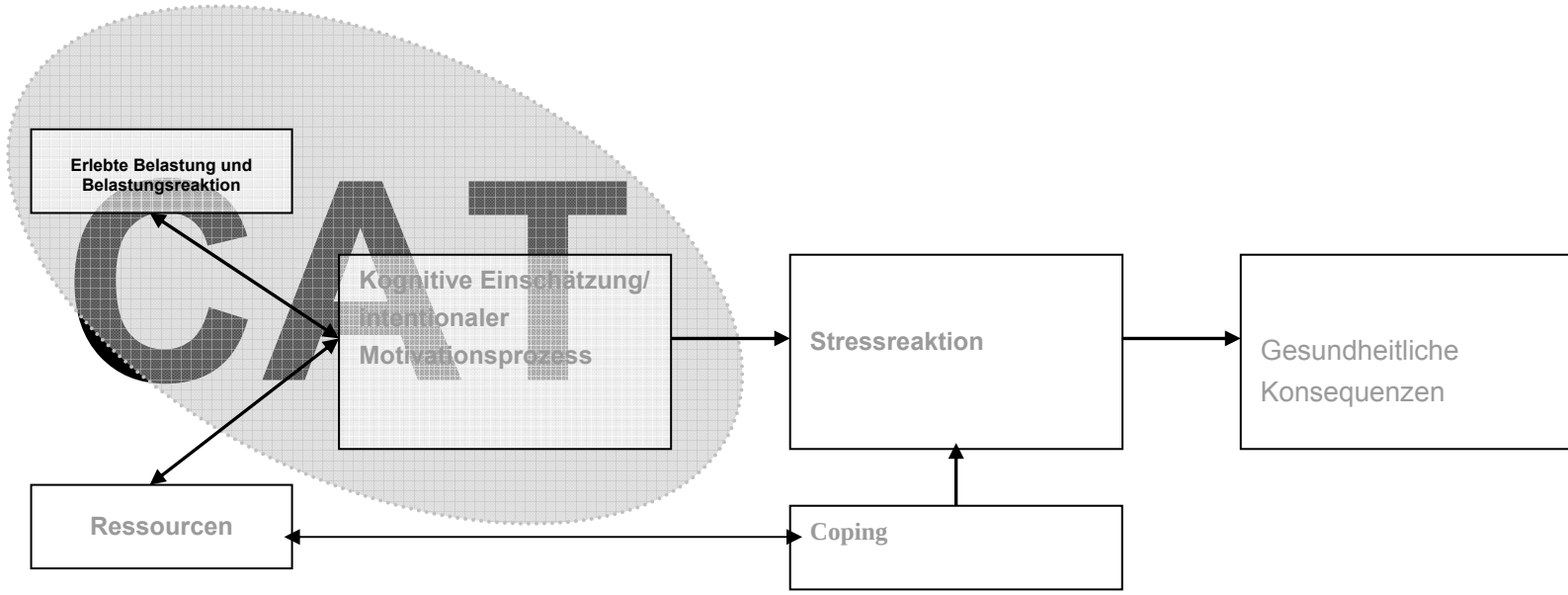


**Abbildung 8: Intendierter Geltungs- und Gültigkeitsbereich des Stress-CAT im Rahmen eines relationalen Verständnisses von Stress.**

Um die beiden Dimensionen unidimensional darstellen zu können, werden im Rahmen der Itembankkonstruktion zwei getrennte Itembanken für ‚erlebte Belastung' und ‚Belastungsreaktion' entwickelt, um den Bedingungen nach Homogenität und lokaler stochastischer Unabhängigkeit der Items bei der Konstruktion eines Computer Adaptiven Tests nachzukommen.

Übersetzte man die beiden Dimensionen ins Englische, wäre es am Treffendsten von ‚stress exposure' und ‚stress experience' zu sprechen. Dabei soll herausgestellt werden, dass nicht nur dieselben Belastungsbedingungen von verschiedenen Individuen unterschiedlich erlebt werden, sondern auch die wahrgenommenen Belastungsreaktionen variieren.

Im Zuge der theoretischen Itempoolerstellung des Stress-CAT erfolgt in einem ersten Schritt die Itemrekrutierung auf der Basis etablierter Fragebogen der Klassischen Testtheorie (Traue, H. et al. 2000) (Fliege, H. et al. 2001a) (Schulz, P. et al. 1999) (Hodapp, V. et al. 1980). Es werden keine neuen Items generiert. Die Fragebogen wurden prospektiv anhand der für den Stress-CAT geltenden Kriterien ausgewählt (vgl. Abschnitt 2.2.3): Wichtig war hierbei, dass die Dimensionen neben einer globalen Beurteilung subjektiv wahrgenommener Belastung auch spezifische Anforderungssituationen abbilden sollten.

(1) *Alltagsbelastungsfragebogen* (ABF) nach Traue, H. et al. (2000):

Skalen zur Erfassung von Alltagsbelastungen, wie das von Brantley, Waggoner, Jones und Rappaport (1987) veröffentlichte *Daily Stress Inventory*, sind nützliche Instrumente für Studien über den Zusammenhang zwischen Alltagsstress und der Symptomatik von psychosomatischen und chronischen Erkrankungen (Brantley, P., Waggoner, C. D., Jones, G. N., & Rappaport, N. B. 1987). Um das *Daily Stress Inventory* auch in Studien mit deutschsprachigen Probanden anwenden zu können, wurde von Traue et al. (2000) eine Übersetzung vorgelegt, die in zwei Studien mit einer Stichprobe von 115 gesunden Versuchspersonen und 451 psychosomatisch erkrankten Patienten validiert wurde.

Bereits Lazarus und Mitarbeiter nahmen an, dass solche Alltagsbelastungen in der Auslösung oder dem Verlauf einer Erkrankung möglicherweise eine größere Rolle spielen als kritische Lebensereignisse (DeLongis, A. et al. 1982). Es konnte gezeigt werden, dass Alltagsbelastungen einen direkteren Einfluss auf psychische Anpassungsleistung haben als belastende Lebensereignisse. Sie korrelieren mit psychischen Symptomen unabhängig von kritischen Lebensereignissen und können solche auch nach Bereinigung durch den Effekt von kritischen Lebensereignissen besser vorhersagen. Der Alltagsbelastungsfragebogen nach Traue et al. enthält 58 Items mit belastenden Alltagsereignissen, die potentiell für die Itemrekrutierung des Stress-CAT herangezogen wurden.

Die Alltagsbelastungen werden, falls aufgetreten, auf einer 5-stufigen Skala von 1 (das Ereignis ist nie aufgetreten) bis 5 (das Ereignis ist sehr häufig aufgetreten) bewertet (siehe auch Anhang).

(2) *Trierer Inventar zur Erfassung von chronischem Stress* (TICS) nach Schulz, P. et al. (1999):

Angeregt durch Befunde der Stressforschung, wonach chronischer Stress und Belastungen mit klinisch relevanten Beeinträchtigungen der Gesundheit in Zusammenhang steht (siehe auch Abschnitt 2.1), wurde das *Trierer Inventar zur Erfassung von chronischem Stress* (TICS) entwickelt. Der Fragebogen erfasst sechs Aspekte von chronischem Stress: Arbeitsüberlastung, Arbeitsunzufriedenheit, soziale Belastung, Fehlen sozialer Anerkennung,

Sorgen/Besorgnis und belastende Erinnerungen. Die Skala ‚Arbeitsüberlastung' vermag beispielsweise gut zu differenzieren zwischen Tinnitus-Patienten und Kontrollpersonen. Die Chronizität der Belastung wird durch die Häufigkeit retrospektiv erfragter Stresserfahrungen erhoben. Zur Beantwortung der Items stehen fünfstufige Ratingskalen zur Verfügung (1=das habe ich nie erlebt; 5=das habe ich sehr häufig erlebt).

(3) *Fragebogenskalen zur Erfassung der subjektiven Belastung* (FESB) nach Weyer, G. & Hodapp, V. (1975):
Auf der Grundlage des psychologischen Stress-Modells von Lazarus wurden 14 Skalen zur Erfassung der subjektiven Belastung konstruiert. Die Inhalte der Skalen beziehen sich auf mögliche alltägliche Belastungen im Bereich des Berufslebens (60 Items) sowie auf Belastungen des Hausfrauen- (41 Items) und familiären Bereichs (45 Items) und 23 zusätzlichen Items, die keinem Bereich zugeordnet werden konnten. In der vorliegenden Arbeit wird auf eine der 14 Skalen rekuriert: Skala B VIII: Überforderung durch Beschäftigung, die nach unserer Sicht, den höchsten Grad der Verallgemeinerungsfähigkeit hat.
Die Items haben dichotome Antwortformate (0=stimmt nicht; 1=stimmt).

(4) *Perceived Stress Questionnaire* (PSQ) nach Fliege, H. et al. (2001a):
Der *Perceived Stress Questionnaire* wurde ursprünglich von (Levenstein, S., Prantera, C., Varvo, V., Scribano, M. L., Berto, E., Luzi, C. et al. 1993a) entwickelt. Es handelt sich um ein Instrument zur Erfassung der aktuellen subjektiv erlebten Belastung. Der Fragebogen wurde von Fliege et al. (2001) in einer deutschen Fassung an 650 Probanden testatisch überprüft (n=249 stationär psychosomatischen Patienten, n=81 Frauen nach Fehlgeburt, n=74 Frauen nach komplikationsloser Entbindung, n=246 Medizinstudierende). Faktorenanalytisch fanden sich – abweichend vom Original – 4 Faktoren (Sorgen, Anspannung, Freude, Anforderungen). Der ursprüngliche Umfang wurde von 30 auf 20 Items reduziert (Fliege, H. et al. 2001a). Interkorrelationsmuster und Gruppendifferenzierungen legen nahe, dass die ersten drei Skalen die interne Stressreaktion des Individuums abbilden, während die Skala ‚Anforderungen' die Wahrnehmung äußerer Stressoren fokussiert. Zur Beantwortung der Items stehen vierstufige Ratingskalen zur

Verfügung (1=fast nie erlebt; 4=meistens erlebt).

Die Auswahl der relevanten Items geschieht anhand eines Delphi-Entscheidungsprozesses (Hasson, F., Keeney, S., & McKenna, H. 2000). Jedes Mitglied der Forschungsgruppe (zwei Diplom-Psychologinnen, 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) schätzen unabhängig voneinander ein, welche Items aus den ebenda genannten prospektiv ausgewählten KTT-basierten psychometrischen Verfahren theoretisch für die Erfassung von *Stresserleben* geeignet sind. Ein Item wird beibehalten, wenn vier der fünf Rater zustimmen.

### 4.2.2 Methoden der statistischen Itemanalyse und –selektion

Die statistische Itemanalyse und –selektion erfolgt an der oben beschriebenen Stichproben (siehe Abschnitt 4.1). Im Folgenden werden die ausgewählten Methoden und Programme der einzelnen Analyseschritte vorgestellt. Die Auswahl der Software 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, J. E., Jr. et al. 2000) sowie an der Entwicklung des Angst-CAT (Walter, O. B. et al. 2005).

#### 4.2.2.1 Unidimensionalitätsüberprüfung

Eine der Voraussetzungen für die Entwicklung Computer Adaptiver Tests auf der Grundlage der Item Response Theory (IRT) ist die Unidimensionalität (alle Items repräsentieren nur eine Dimension). Die ausgewählten Items werden als erstes auf das Vorliegen dieser Voraussetzung überprüft. Die Unidimensionalität kann auf unterschiedliche Weise getestet werden (siehe auch Kapitel 3.2.2.1). Zur Prüfung der Unidimensionalität werden in der vorliegenden Arbeit zunächst faktorenanalytisch (mittels konfirmatorischer Faktorenanalyse) das latente Merkmal und anschließend die residualen Kovariationen bestimmt (Mplus, (Muthén, L. K. & Muthén, B. O. 1998). Items mit

einer Ladung <.40, bzw. residualen Korrelationen > 0,25 werden nicht in die Itembank aufgenommen. Das Kriterium für die Faktorladung wurde in Anlehnung an Nunnally (1994) festgelegt, der eine Einteilung vorsieht, Items, die <.30 laden auszuschließen, bzw. Faktorladungen erst ab >.40 entsprechend zu interpretieren (Nunnally, J. C. et al. 1994). Das Kriterium für die residualen Kovariationen ist von der Autorin gesetzt. Allerdings experimentieren die beiden anderen Arbeitsgruppen, die im Feld der klinisch-psychometrischen Diagnostik arbeiten, auch mit ähnlichen Werten (Ware/Tufts & Harvard-University Boston: <.30/>.20; Cella/North Western University Chicago: <.40/>.30).

Mplus ist ein umfangreiches Programm zur Analyse kategorialer Daten. Die Faktoranalyse setzt eine latente (multivariate) Normalverteilung voraus, sie kann alle Implikationen der Unidimensionalitätsannahme überprüfen, aber der Fokus liegt auf den residualen Kovariationen (zur Bedeutung Residualer Kovariationen siehe auch (Bjorner, J. 2004; Ware, J. E., Jr. et al. 2000). Diese sollen möglichst niedrig sein. Items, deren residuale Korrelationen zu hoch sind, werden schrittweise und vorsichtig aus dem Modell entfernt, dabei wird darauf geachtet, welchen Inhalt die Items beschreiben. Dieser Schritt wird so oft durchlaufen, bis die residualen Kovariationen zufrieden stellend sind.

#### 4.2.2.2 Kategorienfunktion

Kategorienfunktionen einzelner Antwortkategorien können im Rahmen der IRT durch Item-Response-Curves (IRC) beschrieben werden. Diese zeigen die Antwortwahrscheinlichkeit der einzelnen Antwortkategorien (Ordinate) in Abhängigkeit vom latenten Traitkontinuum der Angst (Abszisse) (siehe auch Abschnitt 3.2.1.1.2). 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.

Das Programm TestGraf (Ramsay, J. O. 1995) stellt mittels einer nonparametrischen Glättungsfunktion namens *Kernel-Smoothing-Technique* IRCs grafisch dar und erlaubt die Berechnung oben genannter Statistiken. Jede Antwortkategorie soll an genau einer Ausprägung des geschätzten Scores die

maximale Antwortwahrscheinlichkeit haben. Die Maxima der Antwortkategorien müssen dabei auch in der ‚richtigen' Reihenfolge liegen, also erst Maximum der niedrigsten Kategorie (z.B. ‚trifft gar nicht zu'), dann der nächst höheren (z.B. ‚trifft manchmal zu') usw. Falls das nicht der Fall ist, müssen gegebenenfalls Kategorien zusammengefasst werden (*collapsed*) oder gegebenenfalls das ganze Item entfernt werden. Als ‚ungenügend' werden IRCs beurteilt, wenn sie nicht zwischen unterschiedlichen Ausprägungen des interessierenden Merkmals 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.

#### 4.2.2.3 Differential-Item-Functioning

Bei der Erstellung eines Fragebogens ist man bemüht, dass die Items möglichst für alle Untersuchten in gleicher Weise anwendbar sind. Manche Items weisen jedoch einen Bias auf, in der Form dass z.B. Männer und Frauen bestimmte Fragen unterschiedlich auffassen. Um die Items zu identifizieren, die in der IRT Terminologie eine ‚Differential Item Function (DIF)' aufweisen, existieren verschiedene Methoden (siehe auch Abschnitt 3.2.1.1.4).

Die DIF-Analysen in der vorliegenden Arbeit werden getrennt für die Dimensionen ‚erlebte Belastung' und ‚Belastungsreaktion' berechnet mittels logistischer Regressionsanalysen, einer Variante der Regressionsrechnung mit nominalen Kriteriumsvariablen. Auf eine Darstellung der logistischen Regression wir hier verzichtet. Ausführliche Hinweise hierzu, eine Anleitung zum Rechnen einer logistischen Regression mit dem Programmpaket SPSS sowie weitere Literatur findet sich bei (Rese, M. 2000).

#### 4.2.2.4 Itemparameterschätzung

Fragebögen zur Erfassung psychologischer Merkmale weisen typischerweise polytome, ordinal geordnete Antwortoptionen auf. Da im Gegensatz zu Leistungstests keine ‚richtigen' Antworten geraten werden können, kommen Ein- und Zwei-Parameter-Modelle für die Parameterschätzung in Frage. Bei den Ein-Parameter-Modellen unterscheiden sich die Items lediglich im Schwierigkeitsgrad (Item-Response-Thresholds bzw. Location-Parameter), aber

nicht in ihrer Diskriminationsfähigkeit, d.h. in ihrer Steilheit der Kurven (Steigungs- bzw. Slope Parameter). Zwei-Parameter-Modelle können sich besser den Daten anpassen, da bei ihnen der Slope-Parameter variieren kann (vgl. Abschnitt 3.2.1.1.4).

Das in der vorliegenden Arbeit verwendete Generalized Partial Credit Models (GPCM; Muraki, E. (1997) ist eine Verallgemeinerung des Partial Credit Models (PCM; Masters, G. N. (1982). Die Itemantwortfunktionen (Item Response Curves, IRC) $P_{jk}$ ($\theta$ ) des Generalized Partial Credit Models in Formel G2 beschreiben die Wahrscheinlichkeit, mit der die Antwortkategorie k eines Items j mit $m_j$+1 Antwortmöglichkeiten in Abhängigkeit von der Merkmalsausprägung $\theta$ gewählt wird. Geschätzt werden für jedes Item j die Steigungsparameter (slope) $a_j$ und die Schwellenparameter (category intersection parameters) $b_{jk}$. (Das Identitätszeichen ≡ in der unteren Zeile von G2 verweist auf die Konvention, den Summanden einer Summe, bei der Start- und Endwert des Summationsindex 0 sind, als 0 anzunehmen.)

G2:
$$P_{jk}(\theta) = \frac{\exp \sum_{c=0}^{k} [a_j(\theta - b_{jc})]}{\sum_{c=0}^{m_j} \exp \sum_{i=0}^{c} [a_j(\theta - b_{ji})]}$$

$$\left( 0 \le k \le m_j , \sum_{c=0}^{0} [a_j(\theta - b_{jc})] \equiv 0 \right)$$

In der Itembank verblieben nur die Items, die einen Steigungsparameter (slope) von > 0,70 aufwiesen. Der Schätzung der Itemparameter nach dem GPCM mit dem Programm Parscale (Muraki & Bock 1999) liegt implizit eine Transformation der Verteilung des Traitkontinuums in eine Standardnormalverteilung zugrunde, wie sie weiter oben für TestGraf beschrieben wurde.

Die Güte der erstellten Itembank des Stress-CAT und dessen Itemabfolgealgorithmus werden in der Folge in zwei Simulationsexperimenten nach der Methode von Wang (Wang, S. 1999) überprüft (vgl. Abschnitt 4.2.3 Simulationsexperimente).

### 4.3 Schätzung der Theta-Werte

Die EAP-Schätzung verwendet das Theorem von Bayes für bedingte Wahrscheinlichkeiten. Mit diesem Theorem kann eine Beziehung aufgestellt werden zwischen (a) der angenommenen Verteilung des Traitkontinuums $k(\theta)$ ohne Wissen über eine Antwort $u$ (prior distribution), (b) der Wahrscheinlichkeit $P(u)$, mit der eine Antwort $u$ auftritt, (c) der Wahrscheinlichkeit $P(\theta \mid u)$, mit der eine Merkmalsausprägung $\theta$ auftritt mit dem Wissen über die Antwort $u$ (posterior distribution), und (d) der Wahrscheinlichkeit $P(u \mid \theta)$, mit der eine Antwort $u$ auftritt unter der Annahme einer Merkmalsausprägung $\theta$ (G3):

G3: $$P(\theta \mid u) = \frac{P(u \mid \theta)k(\theta)}{P(u)}$$

Für einen beliebigen aber festen Antwortvektor $u_0$ kann $P(u_o \mid \theta)$ als eine Funktion von $\theta$ aufgefasst werden, die angibt, mit welcher Wahrscheinlichkeit der Antwortvektor $u_0$ in Abhängigkeit von der zugrunde liegenden Merkmalsausprägung $\theta$ auftritt (Likelihood-Funktion $L(u_o \mid \theta)$). Hieraus ergibt sich für den Antwortvektor $u_0$ die Proportionalität $P(\theta \mid u_0) \propto L(u_0 \mid \theta)k(\theta)$ (posterior distribution $\propto$ likelihood x prior distribution). Aufbauend auf diesen Überlegungen kann die EAP Schätzung folgendermaßen formuliert werden (G4) (Bock & Aitkin, 1981).

G4: $$EAP(\theta) = E(\theta \mid u) = \int_{-\infty}^{\infty} P(\theta \mid u)\theta d\theta = \frac{\int_{-\infty}^{\infty} L(u \mid \theta)k(\theta)\theta d\theta}{\int_{-\infty}^{\infty} L(u \mid \theta)k(\theta)d\theta}$$

In die Berechnung des Erwartungswertes $E(\theta \mid u)$ für $\theta$ unter der Annahme von $u$ geht das Integral der Likelihood-Funktion $L(u \mid \theta)$ ein, die mit der Verteilungsannahme $k(\theta)$ des Traitkontinuums gewichtet wird. Praktisch heißt dies, dass nachdem der Patient eine Antwortoption gewählt hat, über die Berechnung des Erwartungswertes für $\theta$ unter der Annahme der Antwort $u$ eine Schätzung für die zugrunde liegende Merkmalsausprägung getroffen werden kann. Wurde

bislang nur ein Item beantwort, liegt die Schätzung meist in der Nähe des Maximums der mit der gewählten Antwortoption in Beziehung stehenden Kategorienfunktion, wobei die Schätzung jedoch noch durch Verteilungsinformationen über das Merkmal $\theta$ modifiziert wird, also durch die Einbeziehung von Wissen, welche Theta-Werte häufig bzw. weniger häufig vorkommen. Aus diesem Grund weisen EAP-Schätzungen eine Tendenz zur Mitte der Population auf.

Stehen im weiteren Verlauf des CAT die Informationen von mehreren Itemantworten zur Verfügung, ändert sich die Likelihood-Funktion $L(u \mid \theta)$ und die Berechnungen erfolgen analog für jeden Antwortvektor neu. Wie präzise die Schätzung ist, ergibt sich aus der Formel G5.

G5: $$Var(\theta \mid u) = \int_{-\infty}^{\infty} \theta^2 P(\theta \mid u) d\theta - [E(\theta \mid u)]^2$$

Für die Berechnung der in den Gleichungen G3 und G4 auftretenden Integrale werden meist numerische Verfahren herangezogen. Bock & Mislevy (1982) geben eine solche numerische Integration mit Hilfe des Gauss-Hermite'schen Quadraturverfahrens an (Bock, R. D. & Mislevy, R. J. 1982). Hierbei werden die Flächen unter den Kurven durch die Summation von einfach zu berechnenden Teilflächen beliebig genau approximiert.

Die Messung wird beendet, wenn für die aktuelle Schätzung der Merkmalsausprägung $\theta$ der Standardfehler (SE) einen Wert von 0,32 unterschreitet. Ähnlich wie in der klassischen Testtheorie kann dieses auf dem SE beruhende Stoppkriterium mit der Reliabilität des Tests in Beziehung gesetzt werden. Nach der klassischen Testtheorie berechnet sich der Standardfehler einer Messung aus der Quadratwurzel von 1 minus der Reliabilität r, multipliziert mit der Standardabweichung des Tests:

G6: $$SE = \sigma\sqrt{1-r}$$

Dieser Ansatz wurde auch auf IRT-basierte Messungen übertragen, um einen über die Population gemittelten Standardfehler angeben zu können (Embretson, S. E. et al. 2000b). Hierbei geht in die Reliabilität einer IRT-basierten Messung das Verhältnis zwischen dem für die Testpopulation durchschnittlichen quadrierten Standardfehler $\overline{SE}^2$ und der Varianz $\sigma^2$ der zu messenden Merkmalsausprägung ein:

G7: $r' = 1 - \frac{\overline{SE}^2}{\sigma^2}$

Formal geht G7 aus G6 durch elementare Umformungen hervor, wenn man den Standardfehler aus der klassischen Testtheorie mit dem für die Testpopulation gemittelten Standardfehler im IRT Ansatz identifiziert. Im Gegensatz zur klassischen Testtheorie ist der Standardfehler eines Tests auf der Grundlage der IRT von der Merkmalsausprägung $\theta$ und dem verwendeten Schätzverfahren abhängig. Während des computeradaptiven Testprozesses wird für die nach Testkonstruktion N(0,1) verteilte Merkmalsausprägung $\theta$ (d.h. $\sigma = 1$) der für die aktuelle $\theta$-Schätzung gültige Standardfehler nach G5 aus dem aktuellen Antwortmuster des Probanden bestimmt. Der Testprozess bricht ab, wenn der Standardfehler den Wert von 0,32 unterschreitet oder wenn keine weiteren Items in der Itembank zur Verfügung stehen. Dieses Stoppkriterium entspricht nach G7 einer Reliabilität von 0,9.

### 4.4 Simulationsexperimente

Die Güte der erstellten Itembank des Stress-CAT und dessen Itemabfolgealgorithmus werden in zwei Simulationsexperimenten überprüft.

Zunächst wird nach der Methode von Wang (1999) eine simulierte Personenstichprobe mit verschiedenen Traitausprägungen generiert (Wang, S. 1999). Ein Simulee beschreibt das (teilweise zufällige) Antwortverhalten eines Patienten mit einer beliebigen aber festen Merkmalsausprägung $\theta_0$. Um die Antwort eines Simulees zu einem Item mit K Antwortmöglichkeiten zu erzeugen, wird eine Zufallszahl zwischen 0 und 1 erzeugt. Für jede der $1 \le k \le K$ Antwortoptionen wird die Wahrscheinlichkeit berechnet, mit der ein Proband diese oder eine höhere Antwortmöglichkeit auswählt. Fällt die Zufallszahl zwischen die kumulierten Wahrscheinlichkeiten für die k-1 und k Antwortmöglichkeit, wird Option k als Antwort des Simulees ausgegeben. D.h. mit diesem Verfahren wird bei hoch diskriminanten Items häufiger die Antwortkategorie mit der höchsten Wahrscheinlichkeit gewählt. Für 100 Simulees wird im Theta–Bereich zwischen -3.5 bis +3.5, in Theta-Abständen von 0,25, eine computeradaptive Testung durchgeführt und die geschätzten Merkmalsausprägungen mit den tatsächlichen Theta-Werten der Simulees

verglichen.

Die zweite Stichprobe, an der die Itembank sowie der computeradaptive Algorithmus getestet wird, ist die Stichprobe von N=1092 Patienten, die bereits der Entwicklung des Stress-CAT dient. Die Simulation des computeradaptiven Antwortverhaltens beruht also auf realen Itemantworten, die computerassistiert von der untersuchten Personenstichprobe abgegeben werden. Es wird nachträglich der adaptive Itemabfolgealgorithmus an dieser Personenstichprobe simuliert. Wie bei derartigen Simulationen üblich, muss für dieses Vorgehen angenommen werden, dass die damit verbundene andere Abfolge der Präsentation der Items unter realen Bedingungen und der kürzere Testumfang, nur einen nach geordneten Effekt für die Schätzung der Personenparameter haben (Gardner, W. et al. 2002).

# 5 Ergebnisse der Entwicklung der Itembank

## 5.1 Theoriegeleitete Itemauswahl

Aus einem Pool von 126 rekrutierten Items (s. Abschnitt 4.2.1) wurden insgesamt 104 Items von der Forschungsgruppe ausgewählt. Davon wurden 62 Items der Dimension ‚erlebte Belastung', 42 Items der Dimension ‚Belastungsreaktion' zugeordnet. Die Beurteilerübereinstimmung lag bei $\kappa$=.80 (T=7,43, p$\leq$.001)[2]. 22 Items konnten keiner der beiden Dimensionen eindeutig zugeordnet werden.

Tabelle 2 gibt einen Überblick über den Prozess der theoriegeleiteten Itemselektion.

[2] Allgemein wird in der Literatur für Kappa ($\kappa$) folgende Einteilung angegeben (Fleiss, J. L. & Cohen, J. 1973) (Frick, T. & Semmel, M. 1978):
- $\kappa$> 0.75 sehr gute Übereinstimmung
- $\kappa$ zwischen 0.6 und 0.75 gute Übereinstimmung
- $\kappa$ zwischen 0.4 und 0.6 akzeptable Übereinstimmung