3 Item Response Theory – Entwicklung und Anwendung von Fragebögen zur Erfassung von Patient-Reported-Outcomes (PRO)

Im Folgenden werden die Herausforderungen der Entwicklung und Anwendung von computeradaptiven Tests auf der Grundlage der IRT in der Gesundheitsforschung besprochen.

Die empirische Erfassung gesundheitsbezogener Merkmale im Allgemeinen und psychischer Merkmale im Besonderen erfolgt in der Regel mit Instrumenten, die auf der Grundlage der klassischen Testtheorie entwickelt wurden.

Seit den 60-er Jahren bietet sich mit der Item Response Theory (IRT) hierzu eine Alternative an, die verschiedene Vorteile verspricht (Rasch, G. 1960b; Birnbaum, A. 1968).

Die IRT erfuhr seit den 80er Jahren mit der Verfügbarkeit von Software zur computergestützten Anwendung von IRT-basierten Methoden zunächst in der Leistungs- und Eignungsdiagnostik eine weite Verbreitung (Hambleton, R. K. & Slater, S. C. 1997; Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L. et al. 1990). Vor allem größere Testorganisationen, welche umfangreiche Routinetestungen durchführen, wie der Educational Testing Service (ETS), das American College Test (ACT) Board, das College Board, die Psychological Corporation und der Law School Admissions Council (LSAC) nutzen die Potentiale der IRT zur Entwicklung und Evaluation von psychometrischen Tests (Embretson, S. E. et al. 2000b). Für eine umfassende Darstellung computergestützter Anwendungen in der Leistungs- und Eignungsdiagnostik siehe (Becker, J. 2004b).

Die Wurzeln der IRT liegen bei Rasch (1960) und Birnbaum (1968), welche erstmals mathematische, stochastische Modelle in die psychologische Forschung einführten. In einem wegbereitenden Textbuch von Lord und Novick (1968) wurde die IRT (Lord, F. N. & Novick, M. R. 1968), welche seither auch den Namen ‚probabilistische' Testtheorie trägt, einem breiten Fachpublikum zugänglich gemacht (Rost, J. & Spada, H. 1982).

Genau genommen ist die IRT nicht eine einzelne Theorie, sondern umfasst eine Familie von formalen, mathematischen, probabilistischen Messmodellen,

welche postulieren, dass dem beobachtbaren Testverhalten (manifeste *V*ariable) eine Fähigkeit / Eigenschaft bzw. Disposition (latente Variable) zugrunde liegt, die das Testverhalten ‚steuert' (Rost, J. et al. (1982), S. 60). Während die Messung in der KTT als eine *direkte* Messung zu verstehen ist, konzipiert die IRT die Messung als *indirekt*. Das beobachtbare Verhalten stellt also lediglich einen Indikator für ein - in IRT Begrifflichkeiten ausgedrückt - *latentes Trait* dar, auf dessen Ausprägung es zu schließen gilt (Müller, H. 1999). Eine Grundannahme der IRT stellt die Modellierung des Itemantwortverhaltens durch eine mathematische non-lineare Funktion, welche *Item Response Function* (IRF) genannt wird, dar. Die IRF kann als Antwortkategorienfunktion (*Item Response Curve* (IRC)) grafisch visualisiert werden.

Im Unterschied zu Fragebögen der klassischen Testtheorie mit einer zufälligen, aber starren Abfolge von Fragen, werden bei dem so genannten ‚Computeradaptiven Testen' – auf der Grundlage der IRT - die einzelnen Items in einer logischen, flexiblen Abfolge vorgelegt, die durch einen Rechenalgorithmus definiert ist (siehe auch Abschnitt 3.1). Eingangs wird ein Item präsentiert, das eine grobe Orientierung im gesamten Ausprägungsbereich des zu messenden Merkmals erlaubt. Diese Forderung würde das in Abbildung 5 wiedergegebene Item mit einer gleichmäßigen polynominalen Wahrscheinlichkeits-Verteilung von fünf Antwortenalternativen erfüllen.

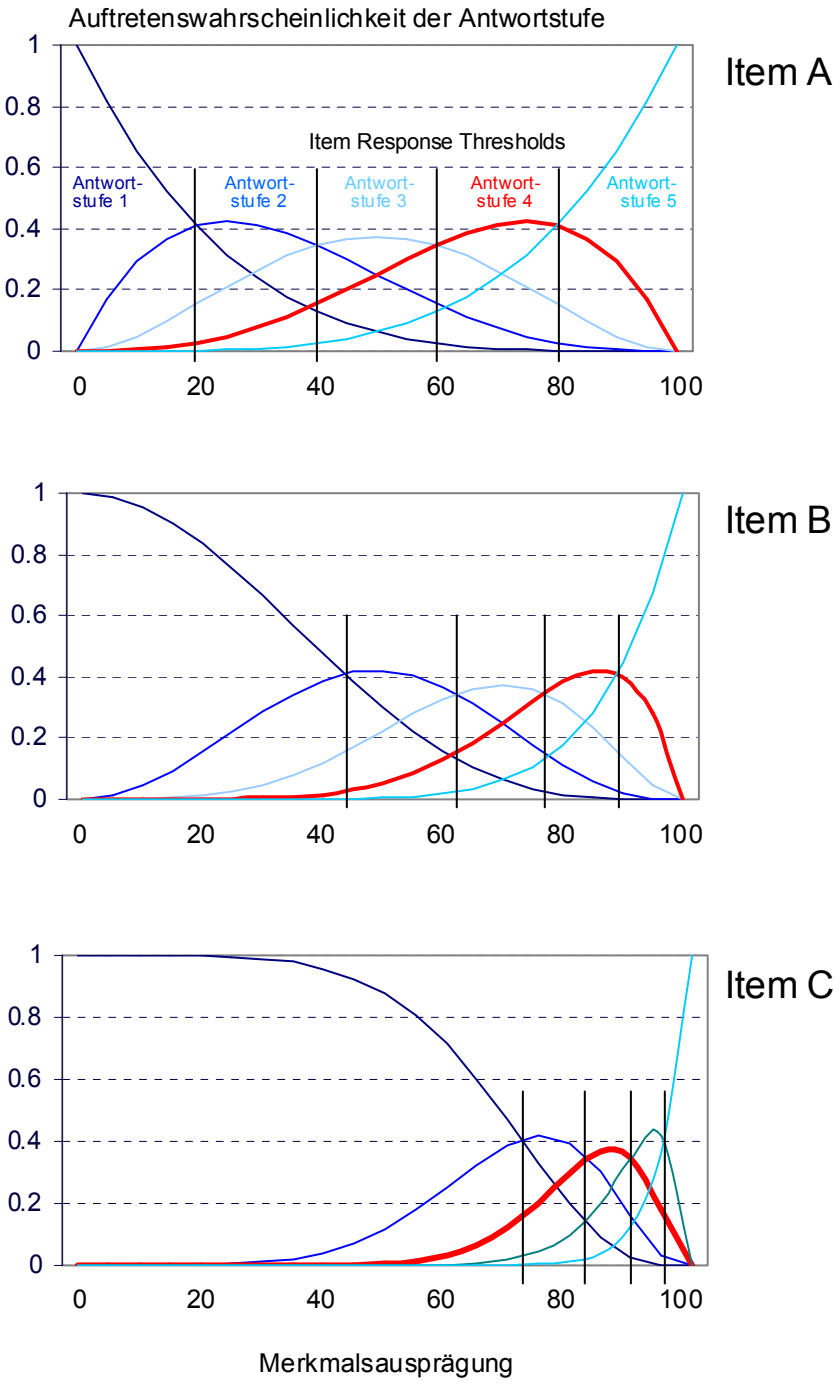

Abbildung 4: Zusammenhang zwischen der Ausprägung des zu messenden Merkmals und der Auftretenswahrscheinlichkeit verschiedener Antwortalternativen bei drei exemplarischen Items. Die x-Achse entspricht der Ausprägung des zu beobachtenden Konstruktes in der Population von 0 ‚gar nicht' bis 100 ‚vollständig', die y-Werte der Wahrscheinlichkeit, dass bei einer bestimmten Merkmalsausprägung (‚latent trait') des Probanden eine bestimmte Antwortstufe gewählt wird.

Wenn ein Proband bei der Frage A z.B. die Antwortstufe 4 wählt, ist es unwahrscheinlich, dass die Ausprägung des Merkmals insgesamt im unteren Bereich der Population liegen wird. Nach der einen Antwort ist am ehesten zu vermuten, dass die Ausprägung des Merkmals zwischen den Prozenträngen 60-80% der Population liegt (60 - 80 Item Response Thresholds). Deshalb würde die zweite Frage so ausgewählt werden, dass sie vor allem in diesem Bereich weiter zur Präzisierung und Absicherung der ersten Antwort beiträgt. In diesem Fall würde sich ein Item mit einem Antwortbereich um 70 und einer Antwortverteilung anbieten, wie es in Abb.1 bei Item B dargestellt ist.
Antwortet der Proband auch auf diese Frage mit der Antwortmöglichkeit 4 verschiebt sich der Bereich der Ausprägung des zu messenden Merkmals weiter nach oben und engt sich weiter ein ($\approx$ 78 - 90). Analog zum bisherigen Vorgehen, würde jetzt eine Frage mit Differenzierungspotential in diesem Bereich gestellt werden, wie dies Item C in Abb.1 darstellt. Antwortet jetzt der Proband mit der Antwortstufe 3, können die Wahrscheinlichkeiten aller drei Antworten miteinander in Beziehung gesetzt werden, so dass für diesen Beispiel-Probanden mit relativ hoher Sicherheit von einer Ausprägung des zu untersuchenden Merkmals zwischen 82 und 85 auszugehen ist.
D.h. nach Kenntnis der Antworten der ersten beiden Fragen trägt das Item C mit seinem sehr kleinen, aber sehr differenzierten Messbereich weiter zur Präzision der Messung bei. Würde hingegen diese Frage allen Patienten gestellt, würde sie in der Mehrzahl der Fälle keinen Differenzierungsbeitrag leisten, da die meisten Patienten der Population ohnehin die Kategorie 1 ankreuzen würden. Sie wäre also meist überflüssig.

Ein weiterer Vorteil der IRT liegt darin, dass sich Item- und Personenparameter auf einer gemeinsamen Skala liegend konzipiert. (Hambleton, R. K. et al. 1997). Dies hat vorteilhafte Implikationen für die Interpretation der Personen- und Itemparameter (siehe Kapitel 3.3.3.). Der Personenparameter wird in der IRT mit dem griechischen Buchstaben „θ“ (= Theta) gekennzeichnet und entspricht dem in der KTT üblichen Summenscore eines Tests. Die Theta-Skala hat per se keinen natürlichen Referenzpunkt (Suen, H. K. 1990), sondern wird üblicherweise in z-Werten dargestellt (M = 0; SD = 1). Die Theta-Werte sind wie folgt zu interpretieren: je größer die Theta-Werte, desto stärker ist das zu

messende Merkmal ausgeprägt bzw. desto ‚schwieriger' ist ein Item und umgekehrt: je geringer der Theta-Wert, desto weniger ist das zu messende Merkmal ausgeprägt bzw. desto ‚leichter' ist ein Item.

Obgleich beide Parameter auf einer gemeinsamen Skala positioniert werden, können sie unabhängig voneinander geschätzt werden (Separierbarkeit von Item- und Personenparametern; (Rasch, G. 1960a).

Diese dritte zentrale Charakteristik der IRT wird auch *Invarianz*-Eigenschaft genannt: Itemparameter und Personenparameter sind *stichprobenunabhängig* (Hambleton, R. K., Swaminathan, H., & Rogers, H. J. 1991).

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

Umgekehrt hängt die Schätzung der individuellen Merkmalsausprägung Theta nicht von dem spezifischen Set dargebotener Items ab. Dies erlaubt die Vergleichbarkeit von Theta-Werten von Personen, denen z. B. im Rahmen eines individuellen Itemselektionsprozesses beim adaptiven Testen unterschiedliche Items zur Beantwortung vorgelegt werden (siehe Kapitel 3.2.2). Nicht nur Theta-Werte von Personen, die unterschiedliche Itemsets beantwortet haben, können verglichen werden, sondern auch ein Vergleich von individuellen Standardmessfehlern, welche bei der Erhebung von Personen mit unterschiedlichen Merkmalsausprägungen eingegangen sind, ist im Rahmen der IRT möglich, da ein weiteres zentrales Messprinzip wie folgt lautet: Der Standardmessfehler variiert in Abhängigkeit von der Ausprägung auf dem latenten Trait θ (Embretson, S. E. et al. 2000b).

Während bei der praktischen Anwendung der KTT unterstellt wird, dass der Standardmessfehler für einen Gesamttest über alle Merkmalsausprägungen konstant ist, ermöglicht die IRT eine individuelle Erfassung desselben. Dies erlaubt beim adaptiven Testen die Kontrolle des Standardmessfehlers einer Messung und ermöglicht eine konstant hohe Messgenauigkeit über das gesamte Kontinuum der Merkmalsausprägung (zum Stoppkriterium, siehe Kapitel 3.1).

Eng verschwistert mit dem Konzept des Standardmessfehlers ist die Reliabilität. Die IRT eröffnet Möglichkeiten der Reliabilitätsbestimmung, welche sich von der in der KTT üblichen unterscheiden. Es gilt folgendes: Die Berechnung der Reliabilität macht keine parallelen Messungen nötig und die Reliabilität hängt nicht von der Testlänge ab.

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

Eine Reihe von Arbeiten haben daher bereits versucht die Vorteile der IRT auf traditionelle KTT-basierte Instrumente zu übertragen (siehe Tabelle 1).

Tabelle 1: Überblick über IRT-Anwendungen im Bereich der Persönlichkeits- und klinischen Diagnostik nach (Becker, J. 2004b)

Autoren	Jahr	Inventar	IRT-Modell
Gibbons, Clark, Cavanaugh & Davis	1985	Beck Depression Inventory (BDI)	Rasch-Modell
Bouman & Kok	1987	BDI	Rasch-Modell
Waller & Reise	1989	Absorption Scale	2 PL-Modell
Reise & Waller	1990	Multidimensional Personality Questionnaire (MPQ)	2 PL-Modell (Birnbaum, 1968)
King, King, Fairbank & Schlenger	1993	Mississippi Scale for Combat-Related Posttraumatic Stress Disorder	unklar
Ellis, Becker & Kimmel	1993	Trier Personality Inventory (TPI)	3 PL-Modell
Santor, Ramsay & Zuroff	1994	BDI	Nonparametrisches Modell (Ramsay, 1995)
Harvey, Murry & Markham	1994	Meyer-Briggs Type Indicator	unklar
Steinberg	1994	State Trait Anxiety Inventory (STAI-Trait)	Nonparametrisches Modell (Ramsay, 1995)
Santor, Zuroff, Ramsay, Cervantes & Palacious	1995	BDI, Center of Epidemiological Studies-Depression Scale (CES-D), NEO-PI (N)	Nonparametrisches Modell (Ramsay, 1995)
Waller, Tellegen, McDonald & Lykken	1996	Negative Emotionality Scale	2 PL-Modell
Gray-Little, Williams & Hancock	1997	Rosenberg Self-Esteem Scale	GRM (Samejima,1969)
Cooke & Michie	1997	Hare Psychopathy Checklist – Revised	GRM (Samejima,1969)
Schmit & Ryan	1997	NEO-PI Conscientiousness Scale	GRM (Samejima,1969)
Rost, Carstensen & Davier	1999	NEO-FFI	Eindim. Rasch-Modell & Mixed Rasch-Modell
Cooke, Michie, Hart & Hare	1999	Screening Version of the Hare Psychopathy Checklist (PCL:SV)	GRM (Samejima,1969)
Rouse, Finger & Butcher	1999	MMPI-Psy-5 Scale	2 PL-Modell
Reise & Henson	2000	NEO PI-R	GRM (Samejima,1969)

Autoren	Jahr	Inventar	IRT-Modell
Orlando, Sherbourne & Thissen	2000	CES-D	GRM (Samejima,1969)
Santor & Coyne	2000	Hamilton Rating Scale for Depression	Nonparametrisches Modell (Ramsay, 1995)
Childs, Dahlstrom, Kemp & Panter	2000	MMPI-Depression Scale	2 PL-Modell
Chernyshenko, Stark, Chan, Drasgow & Williams	2001	16 Personality Factor Questionnaire (16 PF), Big Five Personality Measure	2/3 PL-Modell: GRM (Samejima,1969), Maximum likelihood formula scoring (MFS, Levine, 1974)
Ferrando	2001	Neuroticism Scales of Maudsley Medical Questionnaire (MMQ), Maudsley Personality Inventory (MPI), Eysenck Personality Inventory (EPI), Eysenck Personality Questionnaire (EPQ)	2 PL-Modell
Cooke, Kosson & Michie	2001	Psychopathy Checklist-Revised (PCL-R)	GRM (Samejima,1969)
Marshall, Orlando, Jaycox, Foy & Belzberg	2002	Modified Version of the Peritraumatic Dissociative Experience Questionnaire (PDEQ)	GRM (Samejima,1969)
Orlando & Marshall	2002	Post Traumatic Stress Disorder Checklist (PTSD-C)	GRM (Samejima,1969)

An dieser Stelle konnten nur die wichtigsten Grundzüge der IRT vorgestellt werden. Für einen systematischen Überblick der Unterschiede zwischen Messprinzipien der KTT versus der IRT seien (Embretson, S. E. 1996; Embretson 1997) und (Embretson, S. E. et al. 2000b) empfohlen.

Zusammengenommen lässt sich sagen, dass die IRT deutlich erweiterte Möglichkeiten bietet, Skalen zu analysieren und zu bewerten. Itemdiskrimationsparameter und Testinformationskurven erlauben Aussagen über die Güte, der mit einem Item ermöglichten Differenzierung zwischen Merkmalsausprägungen von Testpersonen sowie über die Genauigkeit der Messung in Abhängigkeit von der Merkmalsausprägung der Testperson. Eine geringere Messgenauigkeit in den extremen Bereichen kann genau bestimmt und ggf. kontrolliert werden. Dabei erlauben IRT-basierte Verfahren zudem eine Interpretation der Skalenwerte bezogen auf die Iteminhalte, was das

Verständnis gerade von weniger erfahrenen Anwendern erheblich erleichtern kann. Daneben eröffnet die IRT auch die Möglichkeit, verschiedene etablierte Skalen miteinander anhand ihrer Testinformationskurven zu vergleichen und damit gezielte Aussagen über den Anwendungsbereich eines Tests zu treffen. Auch Items aus unterschiedlichen Erhebungen können miteinander auf einer gemeinsamen Skala vereint werden (mittels eines *Item-Link-Designs*;Hornke, L. F. et al. (2000)); dies ist besonders bei kulturübergreifenden Studien eine interessante Option. Des Weiteren kann der so genannte Personen-Fit bestimmt werden, z.B. zur Identifikation von Personen mit inkonsistentem Antwortmuster als Indiz intendierter Falschantworten (Embretson, S. E. 1996).
Und eben eine besonders interessante Option der IRT liegt darin, auf ihrer Grundlage computeradaptive Tests (CAT)' entwickeln zu können, welche die vorgelegten Items dem Antwortverhalten der Patienten anpassen (*tailored testing*; (Kubinger, K. D. 1996).

3.1 Praktische Implikationen und Nutzen der Item Response Theory und dem Computer Adaptiven Testen (CAT)

Idealer Weise sollte bei der Erfassung gesundheitsbezogener Merkmale im Allgemeinen und psychologischer Merkmal im Besonderen eine klinische (Fragebogen-)Routinediagnostik implementiert werden, (a) die nur wenige Items braucht, (b) die leicht auszufüllen ist vom Patienten, (c) bei der die Datenerhebung und -auswertung nur leichte oder keine zusätzliche Belastung vom Klinikpersonal erfordert und (d) wo unmittelbar nach der laufenden Untersuchung auf die Ergebnisse zurück gegriffen werden kann.
Der Schlüssel zu einer ökonomischen und effizienten Erfassung gesundheitsbezogener Merkmale im Allgemeinen und psychologischer Merkmale im Einzelnen, wird derzeit in der einschlägigen Literatur in computeradaptivem Testen (CAT) gesehen und diskutiert (McHorney, C. & Cohen, A. 2000; Hays, R., Morales, L., & Reise, S. 2000; Ware, J. E., Jr., Bjorner, J. B., & Kosinski, M. 2000; Walter, O. B., Becker, J., Fliege, H., Bjorner, J., Kosinski, M., Walter, M. et al. 2004).
Eine Grundannahme Computer Adaptiven Testens (CAT) besteht darin, dass der Computer die Items auswählt, die am geeigneten für einen bestimmten Probanden scheinen (siehe Abbildung 6). Welches Item jeweils während der

CAT-Bearbeitung als ‚optimal' gilt, hängt dabei sowohl von der individuellen Beantwortung vorangegangener Items, als auch von der vorher an einer Kalibrierungsstichprobe errechneten Iteminformation der einzelnen Items ab. Dadurch, dass einer Testperson nur die jeweils ‚passenden' Items vorgelegt werden, kann eine deutliche Itemreduktion bei einem gleichzeitig konstant hohen Messpräzisionsniveau erreicht werden (Cella, D. et al. 2000; Embretson, S. E. 1996).

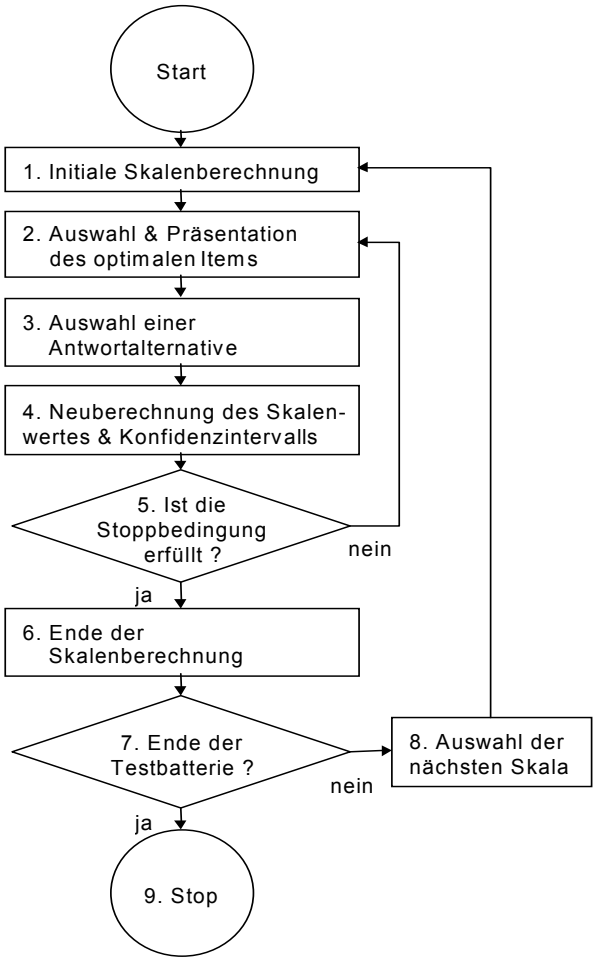

Abbildung 5: Logik eines CATs nach Wainer, H. et al. (1990)

Zusammengefasst sind folgende Aspekte charakteristisch für IRT-basierte CATs (nach Becker, J. (2004b):

1. die *sofortige* Registrierung jeder einzelnen Itemantwort,
2. die *iterative* Neuschätzung des Personenparameters mit Hilfe der Itemantwort(en) und der Itemcharakteristiken,
3. die iterative Auswahl des informativsten Items der erzielten

Schätzung,

4. die iterative Bestimmung des Konfidenzintervalls der erzielten Schätzung,
5. die regelgeleitete Entscheidung über Fortsetzung oder Abbruch der Testung,
6. die finale modellbasierte Personenparameterschätzung stellt das Testergebnis dar.

Auf der Grundlage der IRT werden in jüngster Zeit auch in der Gesundheitsforschung computeradaptive Tests (CAT)' entwickelt, welche die Auswahl der vorgelegten Items dem Antwortverhalten der Patienten ‚anpassen' und damit eine bessere Messgenauigkeit bei reduzierter Itemzahl versprechen (Ware, J. E. et al. 2003; Walter, O. B. et al. 2004).
Eine erhöhte Testökonomie kommt nicht nur dem Diagnostiker, sondern auch der Testperson zugute, da durch die alleinige Darbietung derjenigen Items, die für die individuelle Testperson am informativsten sind, die Testperson durch die Psychodiagnostik zeitlich wie emotional weniger belastet wird. D. h. Ärger und Langeweile bei der Präsentation inadäquater Items (sowie potentiell resultierende Verminderungen der Datenqualität z. B. durch Flüchtigkeitsfehler oder Motivationseffekte) können durch ein adaptives Testvorgehen vermieden werden (Wainer, H. et al. 1990). Im Idealfall fühle sich – so Hornke (1993) - die Testperson optimal angesprochen und schreibe der CAT-Testung bedingt durch eine hohe Standardisierung und Augenscheinvalidität eine hohe Testfairness zu (Hornke, L. F. 1993).

Im folgenden Abschnitt werden die einzelnen Entwicklungsschritte eines computeradaptiven Tests dargestellt.

3.2 Entwicklung einer Itembank und computeradaptives Testen

Das Vorgehen bei der Entwicklung einer Itembank lässt sich in drei prinzipielle Schritte gliedern. Im ersten Schritt wird ein Itempool zur Messung des interessierenden Konstruktes theoriegeleitet erstellt. Der zweite Schritt besteht aus der statistischen Itemanalyse und –selektion. Im dritten Schritt werden die Items, welche sich in den vorangegangenen Schritten bewährt haben, als

Itembank einem Itemabfolgealgorithmus zugrunde gelegt, welcher die Schätzung des so genannten Theta-Wertes (θ) ermöglicht, was in der KTT der Skalenberechnung entspricht. Zur Überprüfung der Güte der entwickelten Itembank und des computeradaptiven Itemabfolgealgorithmus werden hiermit Simulationsexperimente durchgeführt.

3.2.1 Erstellung der Itembank

Der Güte der Itembank kommt bei der Entwicklung eines CATs eine zentrale Rolle zu. So kann nach Embretson und Reise (2000) ein CAT nur so gut sein wie seine Itembank.

Leider existieren in der Psychologie wenig einheitliche Regeln, nach denen bei der Testkonstruktion eines CATs vorgegangen werden sollte.

Embretson und Reise (2000) machen einen allgemeinen Unterschied zwischen drei Testkonstruktionsansätze: a) den „empirical keying approach“, welcher sich auf die Vorhersage von Verhalten von Probanden fokussiert, jedoch ohne einen unidimensionalen Messanspruch zu stellen; b) den „construct approach“, darunter wird der traditionelle Testkonstruktionsansatz - wie er im Rahmen der Klassischen Test-Theorie (KTT) favorisiert wird - verstanden (bestehend aus der Berechnung von Faktorenanalysen, Inter- und Item-Test-Korrelationen etc.), und c) eine IRT-basierte Skalenkonstruktion, welche eine Kalibrierung von IRT-Parametern an einer zuvor erhobenen Kalibrierungsstichprobe umfasst.

Während die Anforderungen an die Kalibrierungsstichprobe relativ gering erscheinen, existiert eine Reihe von strengen psychometrischen Anforderungen an eine ‚gute’ Itemstichprobe (*Itembank)*, welche nach folgenden Aspekten zusammengefasst werden (Hambleton, R. K. & Zaal, J. N. 1990; Wainer, H. et al. 1990; Weiss, D. J. 1985; Embretson 1997; Embretson, S. E. et al. 2000b; Ware, J. E., Jr. et al. 2000):

1. Größe der Itembank,
2. Homogenität der Itembank,
3. Erfassung eines weiten Bereichs des Merkmalsausprägungskontinuums,
4. Hohe Diskriminationsfähigkeit der Items,
5. Ausschluss ‚ungenügender’ Items,
6. Evaluierung der Güte der Itembank.

Zu (1): Eine der Anforderungen an eine Itembank ist ihre Größe. Es ist das ideale Vorgehen, speziell für den CAT neue Items zu entwickeln. Dies ist jedoch oft aufgrund des damit verknüpften großen Erhebungsaufwandes nicht realisierbar.

In der Praxis folgt man der Annahme, dass in der Regel schon ein potentiell guter Itempool für die Erfassung bestimmter Konstrukte (d. h. gute Indikatoren für das latente Trait) geschrieben wurde, z. B. Items aus KTT-basierten Fragebögen, der - falls er bereits an einer ausreichend großen Kalibrierungsstichprobe erhoben wurde - zur Berechnung IRT-basierter Parameter genutzt werden kann (Weiss, D. J. 1985; Embretson, S. E. et al. 2000b). Die Kalibrierungsstichprobe (von Personen) muss nicht repräsentativ sein (aufgrund der in der IRT formulierten Unabhängigkeit der Item- und Personenparameterschätzung) und darf bzw. sollte möglichst heterogen in Bezug auf das zu messende Merkmal sein (Reise, S. P. 2000; Bjorner, J. 2004). Für die erwünschte Größe der Itembank liegen bisher nur Erfahrungswerte aus der Leistungsdiagnostik vor. Hier rät Weiss (1985) zu Itemmengen von N_{items} = 100-200, Hornke (1993) zu Itemmengen von N_{Items} = 70-200, während Embretson und Reise (2000) N_{Items} = 100 empfehlen, jedoch darauf hinweisen, dass für den Bereich der Persönlichkeitsdiagnostik weniger Items nötig seien, da diese in der Regel ein polytomes Antwortformat haben (Hornke, L. F. 1993; Dodd, B. D., De Ayala, R. J., & Koch, W. R. 1995; Masters, G. N. & Evans, J. 1986).

Zu (2): Weiterhin ist die Homogenität einer Itembank speziell bei der Entwicklung eines unidimensionalen CATs bedeutsam. Diese kann durch die Selektion anhand von inhaltlichen Itemtext-Kriterien (durch Expertenurteile), sowie mittels Unidimensionalitätsüberprüfungen (Faktorenanalysen, Analysen residualer Kovarianzen) gewährleistet werden.

Zu (3): Schließlich ist die Erfassung eines weiten Bereichs des Merkmalsausprägungsspektrums vor allem dann erwünscht, wenn es sich um die Konstruktion eines so genannten *equal precise* Tests handelt, also ein Test entwickelt werden soll, der anstrebt, die Merkmalsausprägung von Personen unterschiedlicher Ausprägungsniveaus gleich gut zu messen.

Zu (4): Diese Anforderung muss nicht erfüllt werden im Falle so genannter *peaked* Tests (kriteriumsbasierter Tests), welche das Ziel verfolgen, Personen

anhand eines bestimmten computergestützten Testscores (Kriteriumswertes) in zwei Gruppen zu klassifizieren. In diesem Fall wären nur Items mit einer hohen Information um den Kriteriumstestwert nötig (Embretson, S. E. & Reise, S. P. 2000a).

Zu (5): Schwieriger gestaltet sich schon der Ausschluss ‚ungenügender' Items. Denn es gibt in der IRT-Entwicklung von Itembanken bisher noch keine einheitlichen Bewertungsstandards der Qualität von Items (Reeve, B. 2004) (Orlando, M. 2004) (Chang, C.-H. 2004).

So können sich Selektionskriterien einerseits auf die Überprüfung der Unidimensionalität, die Kontrolle der Diskriminationsfähigkeit, die Passung an das ausgewählte IRT-Modell (Modell-Fit) oder ähnliches beziehen.

Zu (6): Mit Hilfe so genannter Simulationsexperimente werden die Güte der erstellten Itembank sowie der Itemabfolgealgorithmus überprüft.

Zusammenfassend ist hervorzuheben, dass speziell bei CATs hohe Anforderungen an die Items gestellt werden, da durch die adaptive Reduktion der Testlänge ‚ungenügende' Items vor allem zu Beginn der Testung den Testverlauf stärker negativ beeinflussen können als bei konventionellen Tests. Allerdings bieten IRT-basierte CATs die Möglichkeit, ihre bestehenden Itembanken kontinuierlich über das Hinzufügen eigens generierter Items und den Ausschluss ‚ungenügender' Items zu verbessern (Bjorner, J. 2004).

3.2.1.1 Statistische Itemanalyse und –selektion

Nachdem in einem ersten Schritt ein Itempool zur Messung des zu erfassenden Konstruktes theoriegeleitet erstellt wurde, besteht der zweite Schritt aus der Itemanalyse und –selektion sowie den Simulationsexperimenten.

Die statistische Itemanalyse und –selektion verläuft in drei sequentiellen Schritten: (a) die Untersuchung der Unidimensionalität, (b) die grafische Analyse der Item Response Curves (IRCs), (c) die Untersuchung von Differential-Item-Functioning (DIF) (d) die IRT-Modellierung, d.h. die Schätzung der Itemparameter unter Anwendung eines geeigneten IRT-Models. Anschließend erfolgen die Simulationsexperimente, die die Implementierung der Itembank in einen computergestützten adaptiven Itemabfolge-Algorithmus auf ihre Güte hin überprüfen.

3.2.1.1.1 Unidimensionalitätsüberprüfung

Ziel der statistischen Itemanalyse und –selektion ist die Items im Itempool zu erhalten, die die meiste Information liefern und die dahinter liegende Dimension am besten abbilden. Eine der Forderungen für einen IRT basierten Test ist die Sicherstellung der Unidimensionalität und lokalen Unabhängigkeit der Items. Seit einiger Zeit scheint sich in methodisch orientierten Forscherkreisen (Stout, W. 1987) zunehmend die Meinung durchzusetzen, dass für eine erfolgreiche unidimensionale IRT-Modellierung keine *perfekte* Unidimensionalität, sondern lediglich eine *approximative* (McDonald, R. P. 1989) oder *essentielle Unidimensionalität* erforderlich sei (Ferrando, P. J. 2001). Das bedeutet, dass für eine IRT-Modellierung die Anforderungen an die Unidimensionalität nicht so streng sein müssen wie ursprünglich angenommen, sondern dass eine IRT-Modellierung bereits dann erlaubt sei, wenn eine *major dimension* im Sinne eines dominanten Faktors existiere (unabhängig von der Existenz von mehreren *minor dimensions*; Ferrando, 2001), der den größten Teil der gemeinsamen Varianz aufkläre (Reise, S. P., Widaman, K. F., & Pugh, R. H. 1993; Embretson, S. E. et al. 2000b). Nach Stout (1990) ist es psychometrisch begründet und angemessen, die *strenge* Forderung nach lokaler Unabhängigkeit der Daten durch die Forderung nach *essentieller* Unidimensionalität abzuschwächen (Stout, W. 1990).

Es ist umstritten, welches Verfahren für die Bestimmung der Dimensionalität einer Datenmatrix am geeignetsten erscheint. So hat Hattie bereits 1984 ein Dutzend der derzeit angewandten Verfahren zur Testung der Unidimensionalität überprüft (Hattie, J. 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.

Laut Embretson und Reise (2000) könne man bei der Gesamtsicht der Arbeiten in diesem Bereich den Schluss ziehen (Stout, W. 1987; Nandakumar, R. 1994), dass eine Analyse der residualen Kovarianzen derzeit die sinnvollste Aussage über die Dimensionalität der Daten erlaube. Nachdem die gemeinsame Varianz der Items einem Hauptfaktor zugeordnet wurde, der das zu messende Merkmal

repräsentiert, wobei es offenbar eine nach geordnete Rolle spielt, 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 (Waller, N. G., Tellegen, A., McDonald, R. P., & Lykken, D. T. 1996). Und Hambleton, Swaminathan und Rogers (1991) verweisen insbesondere auf den hohen Stellenwert der Analyse von Residuen im Rahmen der Untersuchung der Unidimensionalität (Hambleton, R. K. et al. 1991).

3.2.1.1.2 Antwortkategorienfunktion (Item-Response-Curves, (IRC))

Während in KTT-basierten Verfahren ein Messergebnis in der Regel in Bezug auf eine Normstichprobe interpretiert wird (so genannte komparative Messung), kann in der IRT – aufgrund der Positionierung der Item- und Personenparameter auf einer gemeinsamen Skala – zusätzlich zur normbezogenen Interpretation - eine Interpretation der Theta-Schätzung auf Iteminhalte bezogen erfolgen. Während in der KTT also eine Aussage getroffen wird, die beispielsweise wie folgt lautet: „Person j hat ein Messergebnis auf der Skala *Stresserleben*, welches größer ist als bei 85% aller Personen einer Normstichprobe“, kann in einem IRT-basierten Test die geschätzte Merkmalsausprägung mit Hilfe des Inhalts der Items beschrieben werden, die durch ihre Itemparameter in der Nähe der geschätzten Merkmalsausprägung lokalisiert sind. Ein Beispiel für eine solche inhaltsbezogene direkte Interpretation wäre: „Die Merkmalsausprägung der erlebten Belastung von Person j kann behaftet mit einem Vorhersagefehler v durch die Items „Aufgabe nicht gut bewältigt“ (Item i1), „hetzen müssen“ (Item i2) und „missverstanden werden“ (Item i3) am besten beschrieben werden“. Eine solche Beschreibung der Merkmalsausprägung kann eine informationsreiche Ergänzung zur üblichen normbezogenen Interpretation von Testwerten sein.

Kategorienfunktionen einzelner Antwortkategorien sowie Item- und Testinformationskurven können mit Hilfe grafischer Betrachtung von Item Response Curves (IRCs) untersucht werden. Außerdem können Standardmessfehler und Reliabilität einer Skala in Abhängigkeit vom geschätzten Merkmalsausprägungsniveau berechnet werden.

Item Response Curves (IRCs) sind grafische Darstellungen der (Antwort-)

Kategorienfunktionen von Items (siehe auch Abschnitt 3). Sie veranschaulichen die Antwortwahrscheinlichkeit der einzelnen Antwortkategorien (Ordinate) in Abhängigkeit von der latenten Merkmalsausprägung (Theta/Abszisse).

Das latente Merkmalsausprägungskontinuum wird in Einheiten der Standardnormalverteilung ($\overline{X}$ = 0; SD = 1) dargestellt.

Die Kategorienfunktionen können nicht nur grafisch dargestellt werden, sondern auch in Form einer mathematischen Gleichung beschrieben werden, welche der darauf folgenden Schätzung der Itemparameter dient.

Diese statistische Kenngröße ist vorteilhaft für das adaptive Testen und ist die mit dem Standardmessfehler und der Reliabilität eng verwandt ist. Es ist die *Iteminformationsfunktion* $I(\theta, i)$. Sie beschreibt die Information, welche ein Item i zur Diskrimination zwischen verschiedenen Merkmalsausprägungen bei der Theta-Schätzung beiträgt, in Abhängigkeit von Theta (Suen, H. K. 1990). Obgleich sie mathematisch auf unterschiedliche Weise abgeleitet werden kann, stellt sie konzeptuell das Verhältnis der Steigung der IRC zum erwarteten Standardmessfehler auf der jeweiligen Ausprägung des Theta-Kontinuums dar. Sie berechnet sich durch folgende Formel:

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

$P'_i(\theta)^2$ = 1. Ableitung der IRC; $P_i(\theta)$ = Wahrscheinlichkeit einer richtigen Antwort; $Q_i(\theta)$ = Wahrscheinlichkeit einer falschen Antwort ($Q_i(\theta) = 1 - P_i(\theta)$).

Die Iteminformation ist der Kennwert, welcher zur Itemselektion, d. h. zur Auswahl des ‚passendsten' Items für ein Individuum, im Rahmen des IRT-basierten adaptiven Testens genutzt werden kann. Ferner ist sie bei der Itembankentwicklung von Tests interessant, da sie erlaubt, Items mit einem geringen Informationsgehalt bei der Testkonstruktion auszuschließen. Auch zur Bewertung der Indikation verschiedener Tests kann sie aufschlussreich sein. Durch die pure Summierung der *Item*informationen aller Items kann nämlich die *Testinformation* berechnet werden, welche genutzt werden kann, um zu bewerten, welcher Test in welchen Bereichen der Merkmalsausprägung den höchsten Informationswert bietet (Embretson, S. E. et al. 2000b) siehe auch Abschnitt 3.2.2.3).

Die zu schätzenden Itemparameter finden sich in der grafischen Darstellung der

Kategorienfunktionen (Item Response Curves, (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 bei 2-Parameter Modellen (s.u.) durch den *Slope Parameter* (Steigungsparameter) ausgedrückt.

Die grafische Darstellung der Kategorienfunktionen kann zur differenzierten Beurteilung der psychometrischen Qualität der Items genutzt werden. Items mit modellkonformen Kategorienfunktionen zeichnen sich durch IRCs aus, welche pro Antwortkategorie eingipflige, glockenförmige, jedoch nicht unbedingt symmetrische Kurvenverläufe aufweisen (Santor, D. A. & Coyne, J. C. 2001). Zudem sollte die Anordnung der einzelnen IRCs auf dem geschätzten latenten Kontinuum der Merkmalsausprägung 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 6, IRC Nr. 1 und 5).

Als ‚ungenügend' werden Kategorienfunktionen beurteilt, wenn sie nicht zwischen unterschiedlichen Ausprägungen des interessierenden Merkmals auf dem latenten Kontinuum zu diskriminieren vermögen. Ungenügend sind Kategorienfunktionen also dann, wenn die Kurvenverläufe pro Antwortkategorie mehrgipflig sind und sich die Kurvenverläufe verschiedener Antwortkategorien mehrfach überschneiden.

3.2.1.1.3 Testinformationsfunktion

Die Beurteilung der Iteminformationsfunktion gibt an, wie viel 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 (Muraki, E. 1993). Eine Auswahl der Items mit modellkonformen Kategorienfunktionen, 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.

3.2.1.1.4 Differential-Item-Functioning (DIF)

Für einen gegebenen IRT-Skalenwert, sollten Item-Antworten unabhängig von der Gruppenzugehörigkeit sein, d.h. Itemparameter und Personenparameter sollen stichprobenunabhängig sein (Hambleton, R. K. et al. 1991). Es bedeutet, dass die in der IRT geschätzten Itemstatistiken von der untersuchten Personenstichprobe unabhängig sind, d. h. im Falle, dass die Daten den vom IRT-Modell spezifizierten Annahmen entsprechen, die berechneten Itemstatistiken wie z. B. die Schwierigkeit oder Diskriminationsfähigkeit von einzelnen Items über verschiedene Stichproben von Personen generalisierbar sind (siehe Kapitel 3.1).

Umgekehrt hängt die Schätzung der individuellen Merkmalsausprägung Theta nicht von dem spezifischen Set dargebotener Items ab (Embretson, S. E. et al. 2000a). Dies erlaubt die Vergleichbarkeit von Theta-Werten von Personen, denen z. B. im Rahmen eines individuellen Itemselektionsprozesses beim adaptiven Testen unterschiedliche Items zur Beantwortung vorgelegt werden (Ware, J. E., Jr. et al. 2000; Bjorner, J. B., Kosinski, M., & Ware, J. E. 2002).

Nicht nur Theta-Werte von Personen, welche unterschiedliche Itemsets beantwortet haben, können verglichen werden, da sie auf einer gemeinsamen Skala abgebildet werden, sondern auch ein Vergleich von individuellen Standardmessfehlern, welche bei der Erhebung von Personen mit unterschiedlichen Merkmalsausprägungen eingegangen werden, ist im Rahmen der IRT möglich.

Die Evaluation von Differential-Item-Functioning (DIF) muss folglich ein Teil der Itembank Entwicklung sein (Bjorner, J. 2004). Items mit DIF, die die Eigenschaft der Stichprobenunabhängigkeit gefährden, können dann von der Itembank entfernt werden, wobei die Item Response Theory auch genutzt werden kann, um DIF zu korrigieren. In den USA beispielsweise ist die Überprüfung von DIF zentral und ist den Förderrichtlinien des National Institute of Health (NIH) vorgeschrieben, wenn es gilt ethnische Zusammenhänge in Verbindung mit dem Antwortverhalten aufzudecken. In Deutschland spielt vordem die Aufdeckung von Unterschieden zwischen Mann und Frau, bzw. in Anhängigkeit vom Alter eine Rolle.

Die Definition(en) von DIF variieren je nach methodischem Zugang, in

Abhängigkeit von der Antwortkategorienzahl, also, ob es sich um dichotome oder polytome Items handelt, die in der Regel ordinal skaliert sind.

Einigkeit besteht darüber, dass DIF definiert werden kann als bedingte Wahrscheinlichkeit oder genauer bedingter erwarteter Item Score, der variiert in Hinblick auf die Gruppenzugehörigkeit (Teresi, J. 2004). Konzeptionell könnte man sagen, dass DIF das Antwortverhalten vorhersagt in Bezug auf die Gruppenzugehörigkeit – wobei die Fähigkeit des Probanden kontrolliert wird.

Wenn DIF kontrolliert wird für das Niveau des Gesundheitszustandes, stellt sich also die Frage, ob die Antwort auf ein einzelnes Item mit der Gruppenzugehörigkeit des Probanden zusammenhängt (z.B. Alter, Geschlecht). Die Überprüfung von DIF entspricht in der KTT Terminologie der Prüfung eines möglichen Item-Bias.

Bis dato herrscht kein Konsens, welche Methode die geeignete ist, um DIF zu überprüfen (Camilli, G. & Shepard, L. 1994; Holland , P. & Wainer, H. 1993; Millsap, R. & Everson, H. 1993; Potenza, M. & Dorans, N. 1995; Thissen, D., Steinberg, L., & Wainer, H. 1993). Ein Überblick über DIF Methoden, die in der Gesundheitsforschung angewendet wurden, findet sich bei Teresi, J. (2001).

Die Interpretation von DIF fällt bisweilen widersprüchlich aus. Gierl und Khaliq (2001) konnten zeigen, dass erfahrene Gutachter wenig erfolgreich waren bei der Vorhersage von Items mit DIF (Gierl, M. & Khaliq, S. N. 2001). Aber nicht nur die Vorhersage erschien schwierig, sondern auch die Begründung *nach* statistischer Analyse war nicht immer eindeutig.

3.2.1.1.5 IRT-Modellierung

Die Entwicklung von IRT-Modellen begann in den 50er und 60er Jahren mit Vertretern wie Lord (1952), der als Vater des *Normal Ogive Modells* (NOM) angesehen werden kann, und Rasch (1960) und Birnbaum (1968), welche alternativ zum mathematisch komplexen NOM die logistische Funktion einführten (Lord, F. M. 1952; Rasch, G. 1960b; Birnbaum, A. 1968).

Die meisten Modelle, die in dieser Anfangsphase der IRT-Geschichte entstanden, sind eindimensional konzipierte Modelle, welche für die Modellierung des Antwortverhaltens von Items mit dichotomem Antwortformat entwickelt wurden. Erst in den 80er Jahren gelang es einer Reihe von Forschern (Samejima, F. 1969; Samejima, F. 1996; Andrich, D. 1978; Masters,

G. N. 1982) IRT-Modelle zu entwickeln, die auch auf Items mit polytomem Antwortformat anwendbar sind, und seither vielfach erprobt wurden. Etwas später entstanden IRT-Modelle, welche für die Modellierung multidimensionaler Daten entwickelt wurden (Bock, R. D. & Mislevy, R. J. 1988; Carstensen, C. H. 2000; Kelderman, H. 1997; McKinley, R. L. & Way, W. D. 1992; Reckase, M. D. 1997; Rost, J. & Carstensen, C. H. 2002; Segall, D. O. 1996;Segall, D. O. 2001).

Mittlerweile existieren eine Fülle von unterschiedlichen IRT-Modellen, welche sich nach verschiedenen Aspekten taxonomisch ordnen lassen, wie z. B. der der Art der Variablen (Rost, J. 1996), der Anzahl der Itemparameter (Weiss, D. J. & Davison, M. L. 1981) und der Separierbarkeit von Itemparametern (Müller, H. 1999). Für einen detaillierten Überblick siehe Becker, J. (2004a).

Die Klassifikation der verschiedenen Modelle erfolgt am häufigsten nach der Zahl der in der IRF spezifizierten Itemparameter (siehe Abbildung 6).

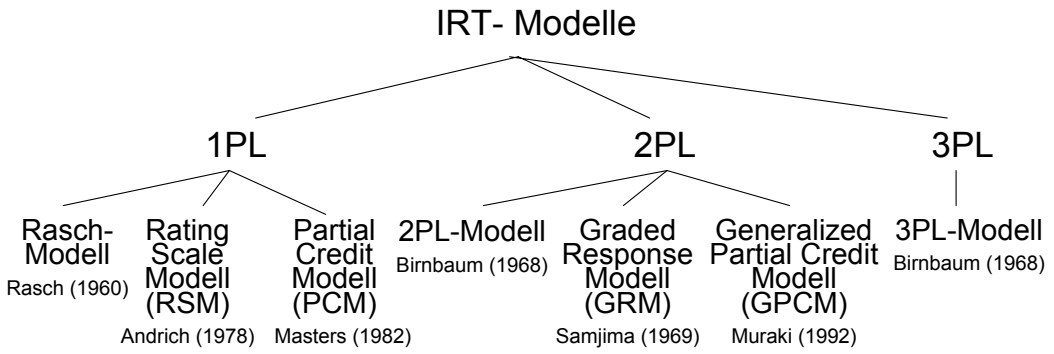

Abbildung 6: Überblick über die wichtigsten 1-Parametrischen-, 2-Parametrischen- und 3-Parametrischen Modelle IRT-Modelle (nach Becker, 2004)

IRT-Modelle unterscheiden sich in ihren jeweils postulierten mathematischen Annahmen. Eine zentrale Voraussetzung, welche von allen IRT-Modellen gleichermaßen postuliert wird, ist die lokale stochastische Unabhängigkeit. Sie wird definiert als die Unabhängigkeit der Antwortwahrscheinlichkeit eines Items von der Antwortwahrscheinlichkeit eines vorangegangenen Items bei konstanter Merkmalsausprägung. Das heißt, die Wahrscheinlichkeit, ein Item in einer bestimmten Weise zu beantworten, hängt nicht davon ab, ob das

vorangegangene Item in ähnlicher Weise beantwortet wurde, wenn die Merkmalsausprägung von Personen gleich ist (Rost, J. et al. 1982). Oder anders ausgedrückt, es wird vorausgesetzt, dass das *latent trait* der einzige Faktor ist, welcher das Antwortverhalten beeinflusst (Hambleton, R. K. et al. 1991). Methodisch kann dies überprüft werden, indem beispielsweise in einer Faktorenanalyse nach der Herauspartialisierung des dominanten Faktors keine Restkorrelationen zwischen den Items verbleiben. Aus dieser Eigenschaft kann auf die Homogenität von Items geschlossen werden (Nunnally, J. C. & Bernstein, I. H. 1994). Wobei die Homogenität als die Eigenschaft von Items definiert wird, dieselbe Fähigkeit bzw. dasselbe Merkmal zu erfassen (Rost, J. et al. 1982). Die Unidimensionalität, ist eng mit diesen beiden Konzepten verwandt. Sie ist gegeben, wenn dem Antwortverhalten nur ein einziger *latent trait* zugrunde liegt. Methodisch untersucht wird sie meist durch die Suche nach einem dominanten Faktor (siehe auch Kapitel 3.2.2.1). Ist die Forderung der meisten IRT-Modelle nach Unidimensionalität erfüllt, so ist auch die lokale stochastische Unabhängigkeit gegeben. Jedoch kann die lokale stochastische Unabhängigkeit auch erreicht werden, wenn die Daten nicht eindimensional sind (Hambleton, R. K. et al. 1991). Unidimensionalität wird nicht von allen IRT-Modellen verlangt, sondern nur von eindimensional konzipierten Modellen gefordert. Die lokale stochastische Unabhängigkeit und die Homogenität sind notwendige Bedingungen bei der Anwendung jeglicher IRT-Modelle, da sie eine zentrale Voraussetzung für die Stichprobenunabhängigkeitsannahme (siehe Kapitel 3.1) darstellen.

3.2.1.1.5.1 Itemparameterschätzung

Welches IRT-Modell das geeignete zur Darstellung der Daten ist, hängt im Wesentlichen von der Art der Daten ab. So weisen Fragebögen zur Erfassung psychologischer Konstrukte, wie Stimmungen, Beschwerden etc. typischerweise polytome Antwortformate auf. Da hier keine ‚richtigen' Antworten geraten werden können, wie dies z. B. bei Leistungstests der Fall ist, kommen prinzipiell so genannte Ein- und Zwei-Parameter-Modelle in Frage. 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, 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).