3. Die Item Response Theorie

3.1. Einleitung

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

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

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

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

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

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

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

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

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

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

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

3.2. Die Klassische Test-Theorie (KTT)

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

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

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

3.2.1. Axiome der KTT

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

Die wichtigsten Axiome der KTT sind:

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

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

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

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

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

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

3.2.2. Grenzen der KTT

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

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

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

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

(a) Item- bzw. Teststatistiken und

(b) Testwerte von Personen.

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

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

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

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

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

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

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

3.3. Die Item Response Theorie (IRT)

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

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

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

3.3.1. Kernannahmen der IRT

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

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

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

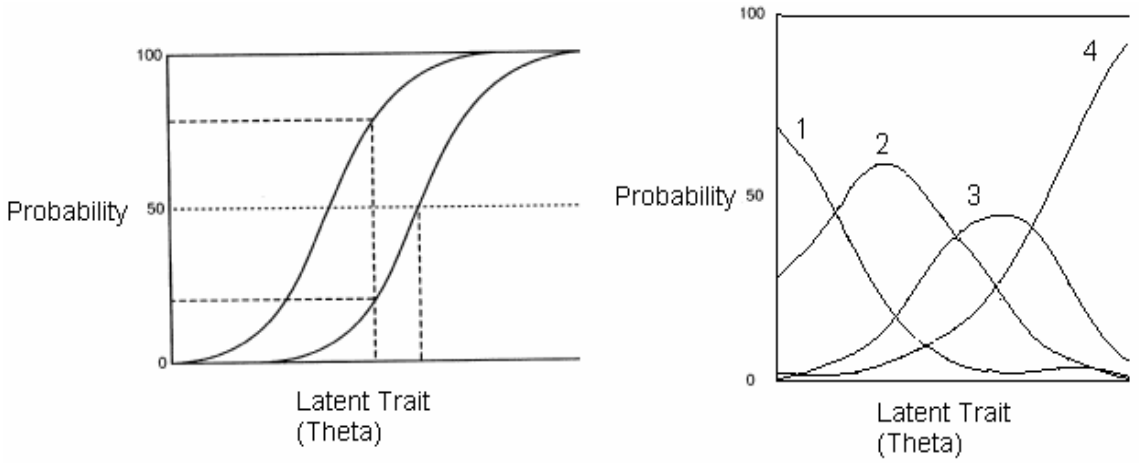

Abbildung 4: Item Response Curves (IRCs). Links: IRCs modelliert mit dem einparametrischen Rasch-Modell. Rechts: IRCs modelliert mit dem zweiparametrischen Generalized Partial Credit Modell (GPCM).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3.2. Voraussetzungen der IRT

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

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

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

3.3.3. Potentiale der IRT

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.3.4. Nachteile der IRT

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

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

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

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

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

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

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

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

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

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

3.4. IRT-Modelle

3.4.1. Ein Überblick

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

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

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

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

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

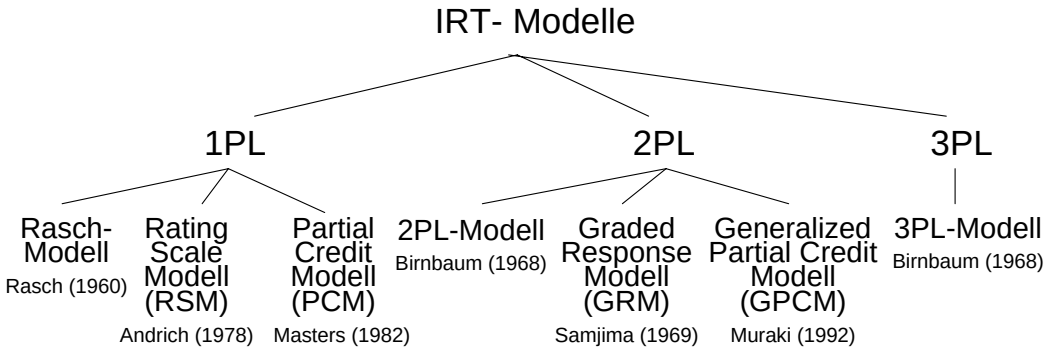

Abbildung 5: Überblick über die wichtigsten IRT-Modelle.

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

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

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

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

3.4.2. Das Rasch-Modell

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

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

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

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

3.4.3. Das Generalized Partial Credit Modell (GPCM)

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

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

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

Gleichung G.3.:

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

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

In der ICRF werden folgende Itemparameter[28] berücksichtigt:

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

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

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

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

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

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

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

3.4.4. IRT-Modelle im Vergleich

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

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

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

[29] Siehe Kapitel 5.3.1.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3.5. Aktueller Forschungsstand zur IRT

3.5.1. IRT Anwendungen in der Leistungsdiagnostik

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

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

3.5.2. IRT Anwendungen in der klinischen und Persönlichkeitsdiagnostik

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

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

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

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

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

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

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

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

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

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

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