4. Computerdiagnostik

4.1. Einleitung

Unter Computerdiagnostik im psychologischen Bereich versteht Jäger (1990):

> „eine strategische Variante innerhalb der Diagnostik [...], um psychologisch relevante Variablen zu erfassen, deren Auswahl zu steuern, die erhaltenen Informationen zu einem Urteil zu verdichten und gegebenenfalls schriftlich und / oder bildlich darzustellen.“ (S. 91)

Nach ihm ist kein Abschnitt des psychologischen diagnostischen Prozesses ungeeignet, um ihn innerhalb der Computerdiagnostik zu realisieren (Jäger, 1990, S. 93). Die Geschichte der computergestützten psychologischen Diagnostik begann in den 20er Jahren, als erstmals automatisierte Testscorerechenmaschinen zur Berufseignungsdiagnostik in den U.S.A. eingesetzt wurden (SVIB: Strong Vocational Interest Blanks; Moreland, 1992). Seither trägt die zunehmende weltweite Verbreitung von Computern aufgrund stetiger technischer Fortschritte in der Hard- und Software-Entwicklung begleitet von einer allgemeinen Kostenreduktion dazu bei, dass in vielen psychologischen Feldern Computer als technische Hilfsmittel zur Diagnostik eingesetzt werden. Der Höhepunkt in der Computerdiagnostik ist aufgrund der fortschreitenden Soft- und Hardware-Entwicklung noch nicht abzusehen (Kubinger, 1993). Dies trifft vor allem auf den klinisch-psychologischen Bereich zu, in dem Computerdiagnostik bislang eher vernachlässigt wurde (Jäger & Krieger, 1994; Hänsgen & Bernascomi, 2000).

Die erste computerdiagnostische Anwendung im klinisch-psychologischen Bereich lässt sich in die 60er Jahre zurückdatieren, als in der Mayo-Klinik in Minnesota (U.S.A.) das international weit verbreitete Minnesota Multiphasic Personality Inventory (MMPI), ein umfangreicher klinischer Persönlichkeits-fragebogen, erstmals computergestützt erhoben wurde (Swenson, Rome, Pearson & Brannick, 1965). Inzwischen existieren weltweit Hunderte von psychodiagnostischen Computeranwendungen, welche grob in die folgenden Einsatzbereiche eingeteilt werden können:[31]

[31] Der dokumentarische und organisatorische Einsatz von Computern in der psychologischen Praxis und Forschung wurde hier nicht extra aufgeführt, da dieser mittlerweile selbstverständlich erscheint (Farrell konstatierte z.B. bereits 1989, dass jeder vierte klinische Psychologe regelmäßig zu dokumentarischen Zwecken einen Computer nutzt). Und klassische

1. Computergestütztes Testen:
   a) Testentwicklung,
   b) Testdurchführung,
   c) Testauswertung,
   d) Testevaluation,
      - Computergestütztes Adaptives Testen (CAT)
2. Computergestützte Interviews,
3. Computer Basierte Test Interpretationsprogramme (CBTI),
4. Computergestützte Expertensysteme.

Um einen Überblick über die genannten Computeranwendungen zu erleichtern, entspricht die formale Aufzählungsreihenfolge (1.-4) ihrem Verbreitungsgrad. Der internationale „Markt“ computergestützter Tests, die von Psychologen / Medizinern / Informatikern und auch fachfremden (!) Anbietern entwickelt werden, ist mittlerweile so groß, dass er kaum noch überschaubar erscheint. In einem über 10 Jahre alten Kompendium wurden bereits mehr als 1.000 computergestützte Tests weltweit aufgelistet (Sweetland & Keyser, 1991), dennoch ist deren Einsatz im Rahmen klinisch-psychologischer Diagnostik im europäischen Raum noch relativ selten (Hänsgen & Bernascomi, 2000).

Als ein Spezialfall computergestützter Tests können Computergestützte Adaptive Testverfahren (CAT) in den Kanon der Computerdiagnostik eingegliedert werden. Deren Verbreitungsgrad ist bislang noch sehr begrenzt (siehe Kapitel 4.6.). Spezifisch für CATs ist, dass sie sich die enorme Rechen- und Speicherkapazitäten von Computern zunutze machen, um Testungen möglichst individuell an die jeweilige Testperson „anzupassen“ (adaptiv). Die „Anpassung“ der Testung erfolgt, indem einzelnen Testpersonen jeweils nur die Fragen gestellt werden, welche für die Messung am informativsten sind (siehe Kapitel 3.3.). An zweiter Stelle der Popularität computergestützter Diagnostik stehen meines Wissens computerdiagnostische klinische Interviews (z. B.

Testentwicklungen – wie die des NEO-FFIs sind heutzutage ohne computergestützt berechnete Faktorenanalysen nur noch schwer vorstellbar). Der Einsatz von Computern im klinisch-therapeutischen Bereich stößt dagegen schnell an seine Grenzen. Für einen (leider veralteten) Überblick wird Bloom (1992) empfohlen. Hier werden Software Programme aus den 80er Jahren zur Unterstützung der Beck’schen kognitiven Therapie (Selmi, Klein, Greist, Johnson & Harris, 1982), der systematischen Desensibilisierung zur Behandlung von Phobien (Ghosh, Marks & Carr, 1984) und ein PC-Therapieprogramm mit zirkulären Fragen (Colby, Watt & Gilbert, 1966) erwähnt.

Psyndex[32] Recherche zwischen 1977 und 2003: 151 Artikel zur Nutzung des computergestützten Interviews CIDI; Wittchen & Pfister, 1996), welche in der Regel hoch strukturiert sind, und entweder vom Diagnostiker während des Gesprächs genutzt oder vom Patienten alleine interaktiv mit dem Computer bearbeitet werden. Der klinische Nutzen und die Validität solcher Interviews ist derzeit jedoch noch umstritten (Wetzler & Marlowe, 1994). Ebenso umstritten, aber noch weniger etabliert sind Computer Basierte Test Interpretationsprogramme (CBTI), die aufgrund häufig fehlender Validierungsuntersuchungen in die Kritik gerieten (Wetzler & Marlowe, 1994; Hornke, 1993; Garb, 2000).

Am wenigsten verbreitet, obgleich erwiesen wurde, dass allgemein die statistische Modellierung des diagnostischen Prozesses einer rein intuitiven klinischen Diagnostik überlegen ist (Wiggins, 1981), sind computergestützte diagnostische Expertensysteme. Sie wurden im deutsprachigen Raum bislang vor allem für den schulpsychologischen Bereich entwickelt, wo sie in der Einzelfalldiagnostik einerseits als wissensbasierte, interaktive Systeme den Diagnostiker in seinen Entscheidungen (bzgl. Hypothesenauswahl, Testindikationsentscheidungen und Testbewertungen) während des gesamten diagnostischen Prozesses regelgeleitet unterstützen (z. B. DIASYS; Hageböck, 1994, Westmeyer & Hageböck, 1992) oder auch „nur“ der statistischen Analyse und Interpretation von einzelnen psychometrischen Testbefunden dienen (z. B. PSYMEDIA, Hageböck, 1990).

[32] Psyndex: Datenbank der Zentralstelle für Psychologische Information und Dokumentation der Universität Trier. Sie enthält Nachweise und Abstracts zu deutschsprachigen Publikationen aus der Psychologie und ihren Randgebieten. Hier werden Artikel aus 250 Zeitschriften, Monographien, Beiträge aus Sammelwerken sowie Dissertationen und Reportliteratur aus Deutschland, Österreich und der Schweiz sowie Beschreibungen von in deutschsprachigen Ländern seit 1945 gebräuchlichen psychologischen Testverfahren dokumentiert.

4.2. Computergestütztes Testen

4.2.1. Vorteile

Viele Wissenschaftler sind sich einig, dass computergestützte Tests die folgenden Vorteile bieten:

1. Verbesserung der Datenqualität durch eine Erhöhung der Gütekriterien:
    - Objektivität,
    - Reliabilität,
    - Validität;
2. Ökonomische Vorteile:
    - Zeitersparnis,
    - Arbeitserleichterung,
    - Kostenreduktion,
    - Nützlichkeit;
3. Nutzung von Potentialen durch:
    - Multimedia,
    - Interaktive und Adaptive Strategien (z. B. durch CAT).

Einer der drei aus meiner Sicht wesentlichsten Vorteile computergestützter Tests ist die Verbesserung der klassischen Gütekriterien (Lienert & Raatz, 1994). Indem der Testleiter, welcher konventionell Papier-und-Bleistift-Testdarbietungen leitete, durch einen Computer ersetzt wird, entfallen mögliche Testleitereffekte (Schötzau-Fürwentsches & Grubitzsch, 1991; Kubinger, 1993). Dies bedeutet, dass mögliche Faktoren, welche die soziale Interaktion beeinflussen können, als „Störvariablen“ wegfallen, da z. B. ein Computer niemals müde, gelangweilt oder frustriert ist, sich jeder (moralischen) Wertung enthält und darüber hinaus über ein „konsistentes, perfektes Gedächtnis“ verfügt (Wetzler & Marlowe, 1994, S. 56ff). So wird die Testerhebung maximal standardisiert und die *Objektivität* steigt.

Indirekt wird dadurch auch die *Reliabilität* günstig beeinflusst (Retest-/ Interrater-Reliabilität). Einen direkten Einfluss auf die Reliabilität hat die Reduktion von (menschlichen) routinebedingten Auswertungs- bzw. Messfehlern (Butcher, 1987, S.17, schätzt, dass Auswertungsfehler aufgrund menschlichen Versagens in durchschnittlich 10% der Fälle vorkommen), d. h. der Computer bietet eine hohe Verrechnungs- bzw. Auswertungssicherheit (Kubinger, 1993; Gregory, 1996; Garb, 2000). Direkte Validitätsverbesserungen

haben sich einige Wissenschaftler (Johnson & Johnson, 1981; Lucas, Mullin, Luna & McInroy, 1977) zeitweise dadurch erhofft, dass „anonyme“ Computerbearbeitungen die Bereitschaft erhöhen könnten, offener intime / persönliche Fragen zu beantworten. Dies konnten Menghin und Kubinger (1996) jedoch empirisch nicht bestätigen. Weiterhin wird vermutet, dass die „hohe face validity“ (Kubinger, 1993) von computergestützten Tests sowie deren ansprechende mobile Darbietung (z. B. per Taschencomputer, siehe Rose et al., 1999, 2003) aufgrund des impliziten spielerischen Moments motivationsfördernd sein kann, und sich somit die Datenqualität und indirekt auch die *Validität* verbessert. Aufgrund eines diesbezüglichen Forschungs-defizits lassen sich darüber jedoch noch keine empirischen Aussagen treffen.

Zu den möglichen erheblichen *ökonomischen* Vorteilen zählt die *Zeitersparnis* bei der Testdurchführung und –auswertung für den Diagnostiker (Rose et al., 1999: Zeiteinsparungen von 2/3) und die Testpersonen (Butcher, 1987, S. 19: Zeiteinsparungen von 15-50%). Desweiteren können computergestützte Tests insofern zu einer massiven *Arbeitserleichterung* des Diagnostikers führen, als sie von gleichförmigen (organisatorischen und administrativen) Routine-tätigkeiten befreien (Schötzau-Fürwentsches & Grubitzsch, 1991; Jäger & Krieger, 1994) und durch die Standortflexibilität des Computers bzw. mobilen Taschencomputers die Arbeitskapazität des Diagnostikers von der Fragebogenbearbeitungszeit der Testperson(en) entkoppeln (Kleinmuntz & McLean, 1968). Eine Arbeitserleichterung stellt auch die schnelle Berechnung komplizierter Auswertungsalgorithmen, die einfache Doku-mentation (Speicherung), Verwaltung (Organisation in Datenbanken) und Verknüpfung großer Testdatenmengen (z. B. zur „online“-Aktualisierungen von Testnormen) sowie deren schnelle Abrufbarkeit dar. In diesem Zusammenhang ist die Vermeidung von „missing data“ durch computergestütztes Testen interessant. Rose und Mitarbeiter (1999) berichten beispielsweise über eine Zunahme der Vollständigkeit von Testdatensätzen von 15% (Papier-und-Bleistift-Tests: 80%; computergestützte Tests: 95%). Sie kann evoziert werden, indem der Computer so eingestellt wird, dass die nächste Frage nur erscheint, wenn die vorherige beantwortet wurde (Itemdarbietungskontrolle).

Verglichen mit umfangreichen Papier-und-Bleistift-Testheften weist Butcher (1987) auch darauf hin, dass bei der computergestützten Testvorgabe einzelner Items ein „Verrutschen“ auf dem herkömmlichen Antwortbogen vermieden wird. Schließlich führen Einsparungen von Testmaterial und Personalkosten zu Test*kosten*reduktionen von bis zu 50% (Gregory, 1996; Hornke, 1993, 1999; Rose et al., 1999; Weiss & Vale, 1987; zu den Nachteilen computergestützter Tests, siehe Kapitel 4.2.2.). Dies kann sich nach Hornke (1993, S. 115) bei 200.000 Testuntersuchungen jährlich in Kosteneinsparungen von 1,1 Mio. DM (pro Jahr) niederschlagen.[33]

Hieraus mag man leicht auf die *Nützlichkeit* von computergestützten Tests allgemein schließen. Kubinger (1993) merkt dazu jedoch an, dass die bloße Computerisierung von Papier-und-Bleistift-Tests einen Test als solchen nicht „nützlicher“ mache (S. 133). Ebenso wenig ist es nützlich, denselben Test mehrfach zu computerisieren (z. B. von verschiedenen Anbietern). Ein Test wird computergestützt dann nützlich, wenn anfangs erläuterte *Vorteile* genutzt werden können oder Potentiale genutzt werden, welche sich aus den Möglichkeiten des Computers ergeben. Dazu zählt z. B. die Nutzung von Multimedia (Gregory, 1996) durch die Ausschöpfung visueller (Tabellen, Grafiken, Video, Animationen), akustischer (Geräusche, Töne, Sprache, Musik), taktiler (z. B. Messung des Tastendrucks, z. B. mit „touchpads“), zeitlicher (Messung von Antwortlatenz bzw. Festlegung verschiedener Bearbeitungs-geschwindigkeiten z. B. bei der Leistungsdiagnostik), interaktiver und adaptiver Potentiale (zu den Vorteilen von CAT siehe Kapitel 4.4.). Dadurch kann Diagnostik realitätsgerechter - z. B. durch (Arbeitsalltags-) Simulationen im Rahmen der Berufseignungsdiagnostik - und individueller - z. B. durch adaptives Messen - werden.

[33] Rechenbeispiel zu Einspareffekten nach Hornke (1993): Eine Einsparung von 5 Items bei 200.000 Probanden macht einen Gewinn von 200.000[Pbn]*5[eingesparte Items]* 20sek.[Testzeit pro Item] = 5555 eingesparte Teststunden (z. B. beim Graduate Record of Examination pro Jahr mühelos erreicht). Wird ein Organisationsstundensatz von 200 DM zugrunde gelegt, so sind das Einsparungen von 1,1 Mio. DM pro Jahr.

4.2.2. Nachteile

Neben den genannten Vorteilen computergestützter Tests wird in der Literatur auch auf eine Reihe von möglichen Nachteilen hingewiesen.

Diese können in Kategorien negativer Auswirkungen in Bezug auf a) den Diagnostiker, b) die Testpersonen und c) die Datenqualität gegliedert werden.

Computerdiagnostik setzt eine gewisse technische Kompetenz im Umgang mit Computern voraus. Ist der *Diagnostiker* wenig vertraut mit Computern, so kann allein der Umstand, dass ein Computer eingesetzt wird, zu (technokratischer) Angst, Zurückhaltung, Skepsis, Vorbehalten und schließlich Ablehnung führen (Butcher, 1987; Hornke, 1993; Jäger & Krieger, 1994). Wird der Einsatz von spezifischer Software als „undurchschaubar“ erlebt, so entsteht Angst vor Kontrollverlust (Butcher, 1987). Da zunehmend auch „Fachfremde“ (Mediziner, Informatiker, Mathematiker, Laien aus der Privatwirtschaft etc.) computergestützte Tests entwickeln, ist die Gefahr einer Entprofessionalisierung (Schötzau-Fürwentsches & Grubitzsch, 1991) nicht von der Hand zu weisen. Auch eine gewisse Selbstwertbedrohung (Garb, 2000) scheint verständlich, wenn die Sorge entsteht, durch einen Computer ersetzt zu werden (Butcher, 1987; Gregory, 1996) und in der jeweiligen Institution nicht darauf fokussiert wird, die durch den Computereinsatz frei gewordenen Personalressourcen für wichtigere, interessantere und kreativere (z. B. therapeutische) als rein administrative Aufgaben zu nutzen (siehe Kapitel 4.2.3.).

Neben diesen potentiellen negativen Auswirkungen auf (a) den Diagnostiker müssen auch mögliche Nachteile für (b) die *Testpersonen* diskutiert werden. Kubinger (1993) weist beispielsweise auf die Möglichkeit einer ungewollten psychischen Stressinduktion hin, räumt aber ein, dass bislang empirisch nicht belegt werden konnte, dass Testpersonen sich subjektiv durch den Computereinsatz überfordert fühlen. Weiterhin beklagen einige Autoren (Butcher, Keller & Bacon, 1985; Kubinger, 1999), dass Variablen der sozialen Interaktion (z. B. durch Verhaltensbeobachtungen) bei der Anwendung von computergestützten Tests nicht erfasst werden. Dem ist entgegen zu halten, dass bei den klassischen Papier-und-Bleistift-Tests (ausgenommen projektiven Verfahren) Verhaltensbeobachtungen der Testpersonen ebenfalls nicht standardisiert gesammelt werden, sondern höchstens ein subjektiver Eindruck der Testbearbeitung beim Diagnostiker entsteht.

Ein wichtiger Faktor, den es in diesem Zusammenhang zu berücksichtigen gilt und der häufig befürchtet wird, ist eine mögliche Abhängigkeit zwischen Testergebnis und Computer*erfahrung*. Erste Untersuchungen weisen darauf hin, dass nach der vorangegangenen Applikation eines entsprechenden Lernprogramms zum Gebrauch der Software keine signifikanten Testniveauunterschiede zwischen Personen mit und ohne Computererfahrung resultieren (Hergovich, 1992). Hier ist jedoch besonders im Leistungsbereich weitere Forschung nötig.

Potentielle Gefahren im Hinblick auf die Testfairness sollten stets reflektiert werden. So gibt Kubinger (1993) zu bedenken, dass ethische, kulturelle, geschlechtsspezifische und sensorische Faktoren ein Testergebnis verzerren können. Interessant ist die These, dass durch die rein visuelle Darbietung der Testinstruktion beim computergestützten Testen möglicherweise „auditive" Wahrnehmungstypen diskriminiert werden könnten, da die Instruktion computergestützter Tests nur visuell, Papier-und-Bleistift-Testinstruktionen jedoch in der Regel auditiv *und* visuell erfolgen.

Schließlich mag der Computereinsatz, wie Kubinger (1999) vermutet, dazu führen, dass Items weniger sorgfältig bearbeitet werden als in Papier-und-Bleistift-Testversionen, d. h. der Computereinsatz per se zu vorschnellen Antworten und Überlesen verleiten kann. Dies führt zur dritten groben Klasse der Nachteile: die Gefahr der Verringerung der Datenqualität (c).

Diese droht, wenn...

1. entwickelte computergestützte Tests nicht ausreichend validiert werden (Gregory, 1996),
2. unkritisch Normen von Papier-und-Bleistift-Tests auf die vermeintlich äquivalente Computerversion übertragen werden (zur Äquivalenz-forschung siehe u. a. Mead & Drasgow, 1993; Kubinger, 1993, Jäger & Krieger, 1994; Rose et al., 1999, 2003; Schwenkmezger & Hank, 1993),
3. sich durch den Einsatz eines fehlerhaften Computer-Programms wieder-holt Fehler reproduzieren (Schötzau-Fürwentsches & Grubitzsch, 1991),
4. ein Computerausdruck gerade bei Kenntnismangel und unter Zeitdruck dazu verleitet, „blind" der Technik zu vertrauen, da er autorisiert (auch

ohne Unterschrift > Diffusion der Verantwortlichkeit; Butcher, 1987; Gregory, 1996; Schötzau-Fürwentsches & Grubitzsch, 1991) erscheint.

Insbesondere der letzte Punkt ist eng mit der Gefahr eines Testmissbrauchs verknüpft, der im medizinischen Bereich dadurch provoziert werden kann, dass Mediziner Psychodiagnostik als einen Gebührenposten kassenärztlich „abrechnen“ können (Computerausdrucke werden hier also im doppelten Sinne als „bare Münze“ genommen; Schötzau-Fürwentsches & Grubitzsch, 1991, S. 309).

Da keine strikten berufspolitischen juristischen Grenzen zum Gebrauch von computergestützten Tests existieren, ist auch die Gefahr des Missbrauchs gegeben. Diese ist jedoch nicht nur auf computergestützte Tests beschränkt, sondern gilt gleichermaßen auch für Papier-und-Bleistift-Tests.

Ein Aspekt, der jüngst im Zeitalter der Computer-Hacker und Wireless Local Area Networks (LAN) psychometrischer Daten bei computergestützten Tests in den Vordergrund gerückt wird, ist der der Datensicherheit (Gregory, 1996). Allgemein muss speziell bei der Benutzung von institutionseigenen Netzwerken diese weitestgehend durch Datenverschlüsselungen und Zugriffsbegrenzungen (Passwords) gewährleistet sein.

4.2.3. Zum Umgang mit computergestützten Tests

Da für die Entwicklung von computergestützten Tests oftmals nicht nur psychologisches Fachwissen, sondern auch Fachwissen aus der Medizin, Mathematik und Informatik benötigt wird, implizieren Gedanken über computergestützte Tests auch berufspolitische Überlegungen. Schötzau-Fürwentsches und Grubitzsch (1991) betonen in Übereinstimmung mit einem Großteil von Psychodiagnostikern, dass unabdingbare Voraussetzung für die Anwendung psychodiagnostischer Verfahren (hier speziell computergestützter Tests) eine qualifizierte psychologische Ausbildung sei. Auf eine wissenschaftlich abgesicherte und fundierte computergestützte Psycho-diagnostik wurde schon vor mehr als 30 Jahren großer Wert gelegt. So formulierten 1986 das Testkuratorium und die American Psychological Association (APA) zeitgleich Richtlinien zur computergestützten Diagnostik (APA, 1986; Testkuratorium, 1986). In ihnen wird auf die Bedeutung eines wohlüberlegten, verantwortungsbewussten, nachvollziehbaren, transparenten und reflektierten Umgangs mit computergestützten Tests hingewiesen und

Empfehlungen in Bezug auf die Kontrolle und Bewertung von Ergebnissen ausgesprochen.

Mehrere Autoren (Jäger & Krieger, 1994; Wetzler & Marlowe, 1994) betonen in diesem Zusammenhang, dass der Computer lediglich ein technisches Hilfsmittel im Rahmen des diagnostischen Prozesses darstelle, welches bei begründeter Indikation als Ausgangspunkt der diagnostischen Hypothesenbildung fungieren könne. Der Computereinsatz solle einseitig abhängig vom Urteil des Psychodiagnostikers sein und keinen Selbstzweck erfüllen, sondern im Interesse der Testperson(en) stattfinden. Ergebnisse sind persönlich, gruppiert nach Konstrukten, verständlich auf Item- und Skalenniveau mit der Angabe von Vergleichsgruppen/-werten ökonomisch und für den Laien verständlich rückzumelden. Die unreflektierte Anwendung undurchschaubarer von Laien entwickelter computergestützter Tests, die einer „black box“ ähneln, verbiete sich, und die Verwendung automatisierter nicht valider Interpretationsprogramme sei zu vermeiden (Jäger & Krieger, 1994). Letztendlich ist jeder Testentwickler von computergestützten Tests (bzw. CATs) herausgefordert, qualitativ hochwertige Tests nach wissenschaftlichen Kriterien in transparenter Weise zu konstruieren und zu validieren, sowie die Soft- und Hardware leicht verständlich und benutzerfreundlich zu gestalten. Der wissenschaftlichen Fundierung computergestützter Psychodiagnostik kommt in jedem Fall das Primat über technische Überlegungen zu.

4.2.4. Computergestützte Tests zur Angstmessung

Im deutschen Sprachraum existieren bereits eine Reihe von computergestützten Testverfahren zur Angstmessung, welche auf den Prinzipien der KTT entwickelt wurden. Im Rahmen des Computerbasierten Ratingsystems zur Psychopathologie (CORA, Hänsgen & Merten, 1994) liegen computergestützte Versionen der folgenden fünf Fragebögen vor:

- Hamilton-Angst-Skala (HAMA; Hamilton, 1959, 1977),
- Selbstbeurteilungs-Angst-Skala (SAS; Collegium-Internationale-Psychiatriae-Scalarum (CIPS), 1996),
- Interaktions-Angst-Fragebogen (IAF; Becker, 1997),
- State-Trait-Angst-Inventar (STAI-State; Laux et al., 1981),
- Fragebogen zur Angst vor körperlichen Symptomen.

4.3. Computergestütztes Adaptives Testen (CAT)

4.3.1. Einleitung

Das allgemeine Prinzip einer Adaptivität / Adaptation (= Anpassung) findet sich in der psychologischen Diagnostik auf zwei verschiedenen Ebenen realisiert. So kommen nach Kisser (1995) adaptive Strategien auf einer „Makroebene“ zum Einsatz, wenn die Auswahl der Untersuchungsbereiche (z. B. Fähigkeiten, Einstellungen) und die Art und Reihenfolge einzusetzender Erhebungsinstrumente (Fragebogen, Verhaltensbeobachtung, Interview,...) von spezifischen diagnostischen Fragestellungen abhängig gemacht wird. Ein Diagnostiker sollte demnach im Idealfall sein diagnostisches (und damit treatmententscheidendes) Vorgehen dem individuellen Fall „anpassen“.

Auf der „Mikroebene“ ist Adaptivität gegeben, wenn die Darbietung einzelner Fragen, Experimente und Testaufgaben an den Einzelfall angepasst wird. Die Grundidee des adaptiven Testens besteht in der Annahme, dass ein Test am besten misst, wenn der Testperson im Laufe eines Tests genau diejenigen Fragen (= Items) dargeboten werden, welche über die Testleistung der Testperson das meiste aussagen, welche also am „informativsten“ für die Diagnostik sind.

Daraus ergibt sich die Frage, welche Items am „informativsten“ (und übrigens auch am subjektiv interessantesten / motivierendsten) für eine Person sind.

Nach Birnbaum (1968) sind es diejenigen Fragen / Aufgaben, welche einen mittleren Schwierigkeitsgrad für eine spezifische Person aufweisen. Da die Einschätzung der mittleren Schwierigkeit einer Testaufgabe von der individuellen Fähigkeit abhängt, wird die mittlere Schwierigkeit allgemein in Abhängigkeit von der Lösungswahrscheinlichkeit einer Testaufgabe definiert. So besitzt ein Item i für eine bestimmte Person j eine mittlere Schwierigkeit, wenn die Wahrscheinlichkeit einer Person j dieses Item i zu lösen p_{ij} (richtig) = 0,5 entspricht, d. h. wenn es gleich wahrscheinlich ist, dass die Testperson das Item löst bzw. nicht löst (p_{ij} (richtig) = p_{ij} (falsch) = 0,5; Birnbaum, 1968). Hier zeigt sich bereits, dass die Wahrscheinlichkeitstheorie eine wesentliche Grundlage des adaptiven Testens darstellt, weshalb IRT-basierte Tests von manchen Autoren auch als Realisierungen eines „stochastischen Testdesigns“ (Wainer, 1990, S. 130) bezeichnet werden.

Es lässt sich zusammenfassen, dass beim adaptiven Testen eine Anpassung der Itemdarbietung an das Fähigkeitsniveau einer Testperson wie folgt geschieht:

> „Adaptives Testen ist interaktiv, indem Testpersonen diejenigen Items dargeboten werden, von denen man auf der Grundlage des Wissens um die Beantwortung bereits beantworteter Items annimmt, dass sie für die zu testende Person am informativsten sind." (Freie Übersetzung nach Embretson, 1992, S. 129)

Konkret folgt daraus folgendes strategisches Vorgehen:

> Wenn die Testperson ein Item „falsch" beantwortet, wird ihr als nächstes ein „einfacheres" Item gestellt, antwortet die Testperson auf das Item hingegen „richtig" wird ein „schwierigeres" Item dargeboten.

Die Anfänge des adaptiven Testens finden sich zu Beginn des letzten Jahrhunderts in Frankreich, wo Binet 1909 einen adaptiven Papier-und-Bleistift-Test zur Messung von Intelligenz im Rahmen der Schuleignungsdiagnostik (Pädagogik) entwickelte. Er realisierte eine sogenannte „upward / downward"-Strategie (Gregory, 1996, S. 589), bei der für jede Testperson eine „obere" und „untere" Fähigkeitsgrenze erhoben wurde, indem jeder Testperson einerseits so lange immer schwierigere Items gestellt wurden, bis sie eine bestimmte Anzahl von Testaufgaben mit gleicher Schwierigkeit immer falsch beantwortete („upward"), und andererseits einer Testperson so lange immer leichtere Items gestellt wurden, bis sie eine bestimmte Anzahl von Testaufgaben mit gleicher Schwierigkeit immer richtig beantwortete („downward"; zu unterschiedlichen Formen adaptiven Testens siehe Kapitel 4.3.2.). Dieser Intelligenztest blieb lange Zeit der einzige adaptive Test seiner Art, bis in den 60er Jahren durch das Aufkommen der Item Response Theorie (IRT, siehe Kapitel 3) und der rapiden technischen Entwicklung von Computern ein idealer Nährboden für die weitere Erforschung von Computergestützten Adaptiven Tests (CATs) entstand. Im Rahmen eines umfangreichen Forschungsprogramms verfolgte als erster Forscher Lord (1980) in den 60er Jahren die Entwicklung von IRT-basierten CATs in der Schuleignungsdiagnostik in den U.S.A. (Educational Testing Service). Dies initiierte unterstützt von dem U.S. Armed Services und der U.S.

Civil Service Commission (Hambleton & Zaal, 1990) die Entwicklung einer Reihe weiterer IRT-basierter computergestützter adaptiver Leistungs- und Eignungstests (Scholastic Aptitude Test, SAT; California Achievement Tests, CAT; Stanford Achievement Tests and the Woodcock-Johnson-Psycho-Educational-Battery).

Dabei impliziert adaptives Testen per se nicht den Einsatz eines Computers. So wurde der erste adaptive Test in der Leistungsdiagnostik wie eingangs erwähnt als Papier-und-Bleistift-Verfahren entwickelt (IQ-Test von Binet, 1909).

Computer erleichtern jedoch aufgrund ihrer hohen Rechen- und Speicherkapazität (besonders bei der Anwendung von IRT-basierten Tests ist diese aufgrund der hohen Rechenanforderungen beinahe unabdingbar) das adaptive Testen ungemein.

Dabei dient der Computer folgenden Aufgaben (Weiss & Vale, 1987):

- Selektion der Items,
- Präsentation der Items,
- Registrierung der Itemantwort,
- Berechnung eines Fähigkeitsscores (während der Testdarbietung),
- Beenden des Tests.

4.3.2. Varianten des Adaptiven Testens

Seit den 70er Jahren entwickelten sich eine Reihe von verschiedenen Formen adaptiver Tests, denen gemein ist, dass sie den „Spagat“ zwischen Individual- und Gruppendiagnostik zu lösen versuchen, indem sie über eine große Itemzahl verfügen (Itembank), welche alle Schwierigkeitsgrade abdecken sollten und aus deren Menge jeweils die Items ausgewählt und dargeboten werden, welche dem Fähigkeitsniveau einer Person optimal entsprechen („tailored testing“: maßgeschneidertes Testen; Weiss, 1985).

Die bislang entwickelten adaptiven Tests, welche teilweise in Papier-und-Bleistift-Format und teilweise in Form von CATs vorliegen, können in verschiedene Gruppen klassifiziert werden, welche sich in ihrer Art der Realisierung der Adaptivität unterscheiden. Die folgende Abbildung 6 gibt einen groben Überblick über die verschiedenen Formen adaptiver Tests.

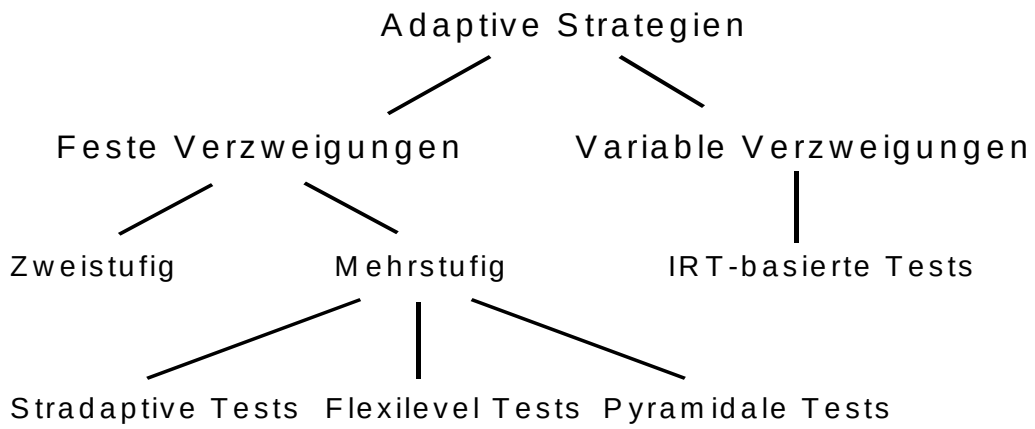

Abbildung 6: Überblick über verschiedene Formen von adaptiven Testsstrategien.

Allgemein lassen sich zwei grundlegende adaptive Teststrategien unterscheiden: Tests beruhend auf festen (vorher fixierten) Verzweigungsstrukturen, welche die Itemauswahl bestimmen, und Tests mit variablen Verzweigungswegen, die auf der Grundlage der Item Response Theorie (IRT) berechnet werden.

Im Folgenden wird zunächst das Grundprinzip von Tests mit *festen* Verzweigungsstrukturen vorgestellt, bevor der Schwerpunkt auf die Testform mit *variablen* Verzweigungswegen gelegt wird, welche in vorliegender Dissertation realisiert wurde: ein IRT-basierter CAT (zur IRT siehe Kapitel 3).

Adaptive Tests, welche sich feste Verzweigungsstrategien zunutze machen („branching tests“; Thissen & Mislevy, 1990), beruhen auf einer durch die Schwierigkeit von Items festgelegten Struktur und Hierarchisierung des Itempools, d. h. diesen Tests liegt ein statisches Verzweigungsschema, zugrunde, welches während der Testkonstruktion entwickelt wurde. Adaptive Tests mit festen Verzweigungen können in Zweistufige und Mehrstufige unterschieden werden. *Zweistufige* fest verzweigte adaptive Tests sind minimal adaptiv („two stage procedure“; Lord, 1980; Hambleton & Zaal, 1990).

Sie bestehen meist aus einem anfänglichen Set von Screening-Aufgaben, welche alle Schwierigkeitsgrade grob abdecken („routing test“), und einem in Abhängigkeit von den Antworten auf diese Anfangsaufgaben nachgeschalteten für die Testperson optimalen Subset von Fragen, das am besten dem (vor-) ermitteltem Fähigkeitsniveau entspricht, und damit eine differenziertere (End-) Testung erlaubt.

Unter *mehrstufigen* adaptiven Tests mit festen Verzweigungsregeln versteht man klassischerweise Tests, welche sich durch Verzweigungen auf der Itemebene auszeichnen (denkbar sind aber auch Verzweigungen auf der Skalenebene). Hier kann entweder anhand inhaltlicher Gesichtspunkte die Itemmenge so strukturiert sein, dass eine Gruppe von Items einem spezifischen Inhaltsbereich („testlet“) angehört, so dass der Itempool in verschiedene Subsets von Items geordnet werden kann („stratified / stradaptive Tests“; Lord, 1980), welche je nach „Anpassung“ bearbeitet werden, oder die Strukturierung der Items erfolgt in Abhängigkeit von der Schwierigkeit. Letzteres ist das grundlegende Prinzip der „flexilevel Tests“ und der „pyramidalen Tests“ (Lord, 1980). *Flexilevel* Tests verfügen über jeweils *ein* Item auf jeder Schwierigkeitsstufe. Die Itempräsentation beginnt mit einem mittelschwierigen Item und vollzieht sich entweder in Richtung schwierigere („downward“) oder leichtere („upward“) Items (Binet, 1909). Durch dieses Vorgehen kann ein Test in seiner Testlänge halbiert werden. *Pyramidalen* Tests liegt eine pyramidenartige Strukturierung des Itempools zugrunde, da sie über *mehrere* Items pro Schwierigkeitsstufe verfügen, und damit die Itemauswahl in Form eines „Entscheidungsbaumes“ mit multiplen Verzweigungen die rein binäre Itemauswahlstrategie der Flexilevel Tests übertreffen (z. B. Adaptives-Intelligenz-Diagnostikum, AID; Kubinger & Wurst, 2000).

Natürlich wurden in der Vergangenheit noch eine Reihe weiterer Formen adaptiver Tests („Robbins-Monro branching method“; „Implied orders tailored testing“ etc.) erprobt. In jüngster Vergangenheit seien hier interessante Ansätze, bei denen die Itembankstrukturierung theoriegeleitet nach Prinzipien der strukturellen Informationstheorie erfolgte (Guthke, Räder, Caruso & Schmidt, 1991), sowie ein Ansatz erwähnt, der sich das methodische Prinzip des „Cluster-Branchings“ als Grundlage der Itembankstrukturierung zunutze machte (Laatsch & Choca, 1994). Abgesehen von diesen Publikationen finden sich jedoch in diesem Forschungsfeld vor allem eher veraltete adaptive Ansätze, welche zum Teil verworfen wurden bzw. heute nur noch von historischem Wert sind. Daher wird hier auf eine ausführliche Darstellung dieser verzichtet (für einen historischen Überblick wird Lord, 1980, empfohlen).[34]

[34] Desweiteren finden bei Butcher und Mitarbeiter (1985) allgemeine adaptive Teststrategien Erwähnung, welche vor allem das Ziel verfolgen, Testpersonen zu klassifizieren, so z. B. die „Countdown Strategie“, welche eine Testung von Personen impliziert bis ein „Cut Score“

Zusammenfassend ist die grundlegende Gemeinsamkeit adaptiver Tests mit festen Verzweigungen ein nach der Item*schwierigkeit* (andere Itemparameter wie z. B. die Iteminformation bei einer IRT-basierten CAT-Anwendung, siehe Kapitel 4.3.3.3., werden nicht genutzt) vorstrukturierter Itempool, der die Grundlage der Itemauswahl bildet. Meist ist die Testlänge auf eine bestimmte dargebotene Itemanzahl fixiert und nicht durch eine logische Stoppfunktion (wie z. B. durch ein bestimmtes Messgenauigkeitskriterium wie bei IRT-basierten CATs siehe Kapitel 3.3.3.) begründet. Weiterhin nachteilig erscheint, dass dem adaptiven Testprozess keine gemeinsame Metrik (wie bei IRT-basierten CATs) zugrunde liegt, was die Vergleichbarkeit der Testergebnisse im strengen Sinn unmöglich macht.

Die IRT vermag diese drei „Mängel“ der fixierten adaptiven Tests zu beheben, da sie folgende Möglichkeiten eröffnet:

1. die Berechnung mehrerer Itemparameter:
   - Implikation: Nutzung derselben zur gezielten Itemauswahl;
2. die Berechnung von Messgenauigkeiten (bzw. Reliabilitäten) in Abhängigkeit zur Merkmalsausprägung:
   - Implikation: Nutzung dieser als Stoppfunktion;
3. die Positionierung von Items und Personen auf einer gemeinsamen Metrik:
   - Implikation: Vergleichbarkeit von Testergebnissen.

Obgleich im folgenden Kapitel zunächst auf die methodischen Grundzüge IRT-basierter CATs fokussiert wird, sei schon anhand der drei beschriebenen Potentiale der IRT hervorgehoben, dass diese „neue“ Testtheorie seit ihrer Entstehung als die eleganteste (und aufwendigste) Methodologie bei der Realisierung von CATs gilt (zur IRT siehe Kapitel 3).

---

erreicht wird, und der „adaptiv typologische Ansatz“, bei dem aufgrund von vorausgehenden Item-Diskriminationsanalysen Items in Form von Stufentests vorgeschlagen werden. Beide adaptive Formen finden kaum Anwendung.

4.3.3. Grundzüge IRT-basierter CATs

Die Wurzeln IRT-basierter CATs finden sich bei Lord und Novick (1968), welche durch ein bahnbrechendes Textbuch, mit einem Kapitel von Rasch und vier Kapiteln[35] von Birnbaum (1968), die statistischen Grundlagen der stochastischen Testtheorie in die psychologische Forschung einführten und damit den Grundstein der IRT legten (Wainer, 1990). Die IRT bietet als eine „Familie“ mathematischer Modelle eine kohärente Methodologie, welche das Testverhalten einer Person zu beschreiben versucht, und die Berechnung von Itemcharakteristiken ermöglicht, die über die konventionellen Statistiken bei der Testkonstruktion auf der Basis der Klassischen Test-Theorie (KTT) hinausgehen (siehe Kapitel 3.).

Durch die Anwendung der IRT zur Testkonstruktion können - verglichen mit den in Kapitel 4.3.2. erörterten verschiedenen Formen adaptiver Strategien - mögliche Gewinne von computergestützten adaptiven Tests maximiert werden.

Charakteristisch für CATs ist, dass eine spezifische Interaktionsregel zwischen Computer und Testperson eingehalten wird, die lautet: „Präsentiere dem Pbn nur solche Items, die geeignet für ihn sind!“ (Hornke, 1994, S. 321). Um die Itemeignung bei *IRT*-basierten CATs zu bestimmen, sind in der Regel umfangreiche (Vor-) Kalibrierungsuntersuchungen an den später zu präsentierenden Items nötig (Kubinger, 1996). Sie dienen der Berechnung von Itemcharakteristiken, welche folgendermaßen genutzt werden können:

1. zur Selektion der „besten“ Items für die Itembank,
2. zur Programmierung des Itemselektionsalgorithmus und
3. zur Berechnung des Skalenwertes einer Person (Personenparameterschätzung).

Der Veranschaulichung eines IRT-basierten computergestützten adaptiven Testablaufs dient Abbildung 7, welche im Folgenden erläutert wird.

Die Nummern im Text beziehen sich auf die Nummern in der Abbildung:

(1.) Die initiale Skalenberechnung geht z.B. von dem Mittelwert der klinischen Population aus ($\theta_0 = 0$; zur Startfunktion siehe Kapitel 4.3.3.2.). (2.) Die Wahl des ersten Items in der Regel auf ein Item, welches mit seinen Antwortalternativen in diesem Bereich die höchste Information verspricht

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

(z. B Fisher-Information, zur Itemselektionsstrategie siehe Kapitel 4.3.3.3.). Nach (3.) der Auswahl einer Antwortalternative auf das erste Items durch die Testperson, wird (4.) der aktuelle Skalenwert anhand eines Personenparameter-Schätzalgorithmus (siehe Kapitel 4.3.3.4.) und das Messgenauigkeitsniveau der jeweiligen Schätzung berechnet. Eine dementsprechende Itemdarbietung und Neuschätzung des Skalenwertes geschieht iterativ und sukzessiv bis (5.) eine bestimmte Stoppfunktion, wie z.B. die maximale Anzahl von Items dargeboten wurde und / oder die Messpräzision hinreichend erfüllt ist. Dann wird (6.) der CAT-Prozess beendet (siehe Kapitel 4.3.3.6.). (7.) Ist die Skala Teil einer Testbatterie so wird (8.) die nächste Skala zur Messung eines weiteren Konstruktes ausgewählt. Wird nur eine Skala in einem CAT-Prozess angewandt, so wird (9.) der CAT-Prozess nach Erfüllung des Stoppkriteriums beendet.

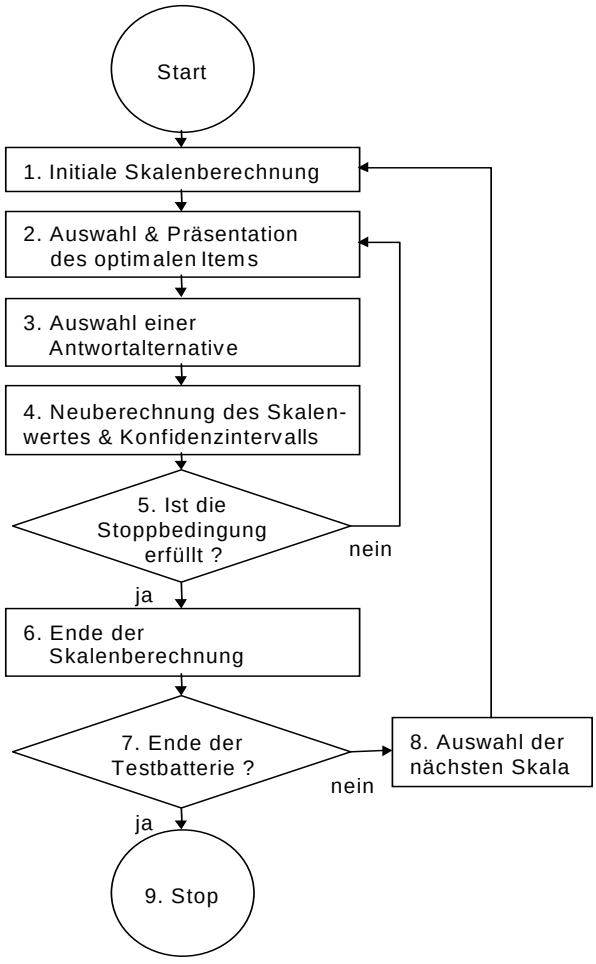

Abbildung 7: Flussdiagramm eines IRT-basierten computergestützten adaptiven Testprozesses (Wainer, 1990, S. 108).

Zusammenfassend lässt sich sagen, dass für IRT-basierte CATs folgende Aspekte charakteristisch sind:

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.

Im Folgenden werden einige der bereits eingeführten Themen IRT-basierter CATs näher fokussiert.

4.3.3.1. 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, d. h. die Güte der Itembank entscheidet letztendlich über die Effektivität des CATs.

Leider existieren in der Psychologie wenig einheitliche Regeln, nach denen bei der Testkonstruktion vorgegangen werden sollte. Embretson und Reise (2000) unterscheiden 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 Testkonstruktions-ansatz - wie er im Rahmen der Klassischen Test-Theorie (KTT) favorisiert wird - verstanden (bestehend aus der Berechnung von Faktorenanalysen, Inter-Item- und Item-Test-Korrelationen etc.), und c) eine IRT-basierte Skalenkonstruktion, welche eine umfangreiche Kalibrierung von IRT-Parametern an einer zuvor erhobenen *Kalibrierungsstichprobe* umfasst.

Ein Vorteil IRT-basierter Itembanken gegenüber KTT-basierten Itempools liegt in dem Potential, Items mit unterschiedlichen Antwortformaten auf einer Skala zu integrieren. Ein Nachteil ist mit dem Umstand der Kalibrierung verknüpft.

Da eine der Anforderungen an eine „gute“ Itembank ihre Größe ist, ist das eigentlich ideale Vorgehen, speziell für den CAT neue Items zu entwickeln, 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; Weiss, 1985; Embretson & Reise, 2000), der - falls er bereits an einer ausreichend großen Kalibrierungsstichprobe erhoben wurde - zur Berechnung IRT-basierter Parameter genutzt werden kann. Dabei sind die Anforderungen, welche an eine *Kalibrierungsstichprobe* gestellt werden, nach Embretson und Reise (2000) nicht sehr hoch. 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.

Während die Anforderungen an die Kalibrierungsstichprobe gering erscheinen, existieren eine Reihe von strengen psychometrischen Anforderungen an eine „gute“ Itemstichprobe (*Itembank)*, welche nach folgenden Aspekten zusammengefasst werden (Hambleton & Zaal, 1990; Wainer, 1990; Weiss, 1985; Embretson & Reise, 2000):

1. Größe der Itembank,
2. Homogenität der Itembank,
3. Erfassung eines *weiten* Bereichs des Merkmalsausprägungskontinuums,
4. Hohe Diskriminationsfähigkeit der Items,
5. Ausschluss „schlechter“ Items,
6. Validität der Itembank.

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 (Dodd, De Ayala & Koch, 1995; Master & Evans, 1986).

Weiterhin ist die *Homogenität* einer Itembank speziell bei der Entwicklung eines unidimensionalen CATs zentral. Diese kann durch die Selektion anhand von

inhaltlichen Itemtext-Kriterien (durch Expertenurteile), sowie mittels Unidimensionalitätsüberprüfungen (Faktorenanalysen, Analysen residualer Kovarianzen) gewährleistet werden. Schließlich ist die Erfassung eines *weiten* Bereichs des *Merkmalsausprägungsspektrums* vor allem dann erwünscht, wenn es sich um die Konstruktion eines sogenannten „equal precise“ Tests handelt, also ein Test entwickelt werden soll, der anstrebt, die Merkmalsausprägung von Personen unterschiedlicher Ausprägungsniveaus gleich gut zu messen. Diese Anforderung muss nicht erfüllt werden im Falle sogenannter „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 & Reise, 2000).

Die Anforderung einer hohen *Diskriminationsfähigkeit* versteht sich vor diesem Hintergrund von selbst. Schwieriger gestaltet sich schon der Ausschluss „*schlechter*“ Items. Denn es gibt in der IRT-Entwicklung von Itembanken bisher noch keine einheitlichen Bewertungsstandards der Qualität von Items. 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. Weitere Forschung ist in diesem Feld dringend erforderlich.

Einig sind sich die meisten Forscher, dass die Itembank eines CATs einer umfangreichen *Validierung* unterzogen werden sollte, um sicher zu stellen, dass das CAT wirklich das misst, was es zu messen vorgibt (siehe Kapitel 6.).

Zusammenfassend ist hervorzuheben, dass speziell bei CATs hohe Anforderungen an die Items gestellt werden, da durch die adaptive Reduktion der Testlänge „schlechte“ Items vor allem zu Beginn der Testung den Testverlauf stärker negativ beeinflussen können als bei konventionellen Tests (Wainer, 1990). Allerdings bieten IRT-basierte CATs die Möglichkeit, ihre bestehenden Itembanken kontinuierlich über das Hinzufügen speziell „gezüchteter guter“ Items (durch sogenannte Item-Link-Designs; siehe Kapitel 3.3.3. und 5.3.2.3.3.) und den Ausschluss „schlechter“ Items zu verbessern.

4.3.3.2. Startfunktion

Je kürzer ein adaptiver Test ist, desto mehr Einfluss hat das erste dargebotene Item auf das Messergebnis (Lord, 1980). Aus diesem Grund wird der Startfunktion an dieser Stelle ein eigenes Unterkapitel gewidmet. Nach Embretson und Reise (2000) existieren drei Möglichkeiten, wie ein CAT begründeterweise beginnen kann:

a) mit der Darbietung eines leichten Items,
b) mit der Darbietung eines Items in Abhängigkeit vom Vorwissen,
c) mit der Darbietung eines Items mit mittlerer Schwierigkeit.

Die Darbietung eines leichten Items als „Start-Item“ bei Leistungstests wird von Wainer und Kiely (1987) empfohlen. Indem Frustrationen durch die anfängliche Vermeidung der Darbietung schwerer Items vermieden werden, solle sich die initiale Testangst reduzieren. Zudem sollte bei Leistungstests darauf geachtet werden, dass das erste Item keinem Lerneffekt unterliegen kann, so dass es bei Retests nicht in seiner Aussagekraft reduziert ist.

Eine Präsentation des ersten Items in Abhängigkeit vom Vorwissen aus einer vorangegangenen Testung erscheint sinnvoll, um Redundanz in Mehrfachmessungen zu vermeiden. Da jedoch in den meisten Fällen kein Vorwissen um die Merkmalsausprägung einer Testperson besteht, werden CATs in den meisten Fällen mit der Darbietung eines Items mittlerer Schwierigkeit begonnen. Dies ist vor dem Hintergrund der Annahme einer Normalverteilung der Merkmalsausprägung insofern sinnvoll, da ein Item mittlerer Schwierigkeit initial die beste Schätzung der Merkmalsausprägung erlaubt (Thissen & Mislevy, 1990).

4.3.3.3. Itemselektion

Der Itemselektion liegt in der Regel einer von mehreren möglichen Algorithmen zugrunde, welche speziell für IRT-basierte CATs als Software entwickelt (programmiert) werden müssen. Nach Thissen und Mislevy (1990, S. 103) werden derartige Algorithmen als Regelwerk definiert, welches festlegt, welche Fragen in welcher Reihenfolge von Probanden beantworten werden sollen.

Es lassen sich bei IRT-basierten CATs zwei grundlegende Algorithmen / Verfahren der Itemselektion unterscheiden:[36] [37]

[36] Für einen Überblick über verschiedene Itemselektionsverfahren siehe Thissen und Mislevy (1990) sowie Schnipke und Green (1995).

1. das Maximum-Information-Verfahren (MI) und
2. das Bayes'sche Sequentialverfahren (BE).

Die Idee des Maximum-Information-Verfahrens (MI) stammt wahrscheinlich ursprünglich von Urry (1977)[38], der vorschlug, immer diejenigen Items zu präsentieren, welche für die jeweilige Schätzung der Merkmalsausprägung die höchste Iteminformation aufweisen (d. h. p_{ij} (richtig) = 0,5; entspricht einer mittleren Itemschwierigkeit). Die Iteminformation (meist: Fisher-Information, möglich ist aber auch die Kullback-Leibler Information o. ä.) entnimmt der Computer entweder einer vorher an einer Kalibrierungsstichprobe berechneten Iteminformationstabelle oder er errechnet die Iteminformation simultan während des computergestützten adaptiven Prozesses. Die erste Realisierung des MI-Verfahrens erfolgte im Jahre 1977 durch Brown und Weiss, welche diese Itemselektionsstrategie (mittels eines Rückgriffs auf eine Iteminformations-tabelle durch den Testadministrator) in Papier-und-Bleistift-Format umsetzten. Um zu vermeiden, dass ein Item mehrfach dargeboten wird, da es in mehreren Bereichen die höchste Iteminformation besitzt, kann dieses Verfahren so abgewandelt werden, dass eine Zufallsauswahl des „besten" Items pro Schwierigkeitsbereich realisiert wird. Dies setzt jedoch voraus, dass mehrere Items mit einem ähnlich hohen Informationsgehalt pro Schwierigkeitsbereich in der Itembank vorliegen. Veerkamp und Berger (1997) schlagen eine Abwandlung des Selektionsalgorithmus vor, in dem die Items mit jeweils der höchsten mittleren Information innerhalb eines bestimmten Konfidenzintervalls des Merkmalsausprägungskontinuums ausgewählt werden.

Das Bayes'sche Sequentialverfahren (Bayesian Estimation, BE) wurde erstmals 1969 von Owen publiziert. Es basiert auf der Annahme einer bestimmten Form und Verortung der Merkmalsausprägungsverteilung („a priori"-Verteilung; Weiss & Vale, 1987) - in der Regel einer Normalverteilung (Thissen & Milevy, 1990) - und kombiniert diese in einem komplizierten Rechenalgorithmus mit den bekannten Itemcharakteristiken und dem Antwortverhalten einer Person. Die Itemauswahl verfolgt hierbei das Ziel, die „a posteriori belief distribution"

[37] Neben diesen beiden am häufigsten zur Anwendung kommenden Itemselektionsverfahren (1. & 2.) sei der Vollständigkeit halber darauf verwiesen, dass es auch die Möglichkeit gibt, die Itemselektion gänzlich in Abhängigkeit von Inhalts- und Zeitkriterien zu gestalten (Eggen, van der Linden, Scrams & Schnipke, 1999 zitiert nach Meijer und Nering, 1999).

[38] Urry (1977) selbst nutzte jedoch auch das Bayes'sche Sequentialverfahren und nicht die MI-Itemselektionsstrategie.

(Hambleton & Zaal, 1990, S. 350) so weit wie möglich einzuengen. Dazu wird jeweils das Item mit der kleinsten erwarteten „a posteriori"-Varianz gewählt, so dass der Standardmessfehler minimiert (Thissen & Mislevy, 1990) und eine möglichst genaue Schätzung ermöglicht wird. Diese Art der Itemselektion hängt logischerweise stark von der Adäquatheit der Vorannahme über die „a priori"-Verteilung ab. Van der Linden und Hambleton (1997) schlagen in diesem Zusammenhang vor, Wissen um bereits bekannte „a priori"-Verteilungen zu nutzen. Vergleicht man MI und BE miteinander, so heben Meijer und Nering (1999) hervor, dass beide Itemselektionsverfahren als stabil gelten und sich insbesondere, wenn sich die „Start-Items" gleichen, bei längeren Tests (N = 20 Items; Thissen & Mislevy, 1990) kaum unterscheiden. In kürzeren CATs sei jedoch eine Anwendung des BEs dem MI vorzuziehen (Hambleton et al., 1991). Abschließend sei eingeräumt, dass die Güte der beiden Itemselektionsstrategien in starkem Maße davon abhängt, inwiefern das Antwortverhalten den IRT-Modellannahmen entspricht.

Im Umgang mit diesen ausgefeilten mathematischen Itemselektionsverfahren weisen Thissen und Mislevy (1990) darauf hin, dass die Itemselektion sich nie gänzlich unreflektiert auf mathematische Berechnungen gründen sollte, sondern Forscher den Itemselektionsalgorithmus inhaltlich reflektieren und gegebenenfalls durch eine Iteminhaltsbalancierung[39] die Itemdarbietung kontrollieren sollten.

4.3.3.4. Personenparameterschätzung

Zur Schätzung der Merkmalsausprägung einer Person, in der IRT auch „Personenparameterschätzung" oder „θ (=Theta)"-Schätzung genannt, kommen in der adaptiven Forschung zur Zeit die folgenden vier verschiedenen Verfahren zum Einsatz:

1. **die Maximum-Likelihood-Schätzung (MLE),**
2. **die Weighted-Maximum-Likelihood-Schätzung (WLE),**
3. **die Expected-A-Posteriori-Schätzung (EAP) und**
4. **die Maximum-A-Posteriori-Schätzung (MAP).**

[39] Iteminhaltsbalancierung ist eine freie Übersetzung des Begriffs „Content Balancing" (Wainer, 1990, S. 122). Bei adaptiven Tests mit heterogenem Iteminhalt besteht die Gefahr, dass der Itemselektionsalgorithmus allein aufgrund statistischer Kennwerte die Itemselektion gestaltet und damit unter Umständen der gesamte Inhaltsbereichs des zu messenden Konstrukts nicht hinreichend erfasst wird. Um dem vorzubeugen, können Strategien zur Iteminhaltsbalancierung - wie z. B. die Strukturierung des Itempools in homogene Testlets, aus denen adaptiv Items gewählt werden - angewandt werden.

Die ersten beiden Ansätze (MLE und WLE) basieren auf dem Likelihood-Schätzverfahren und gehen auf ein von Lord (1980) formuliertes Grundprinzip zurück, der vorschlug, die Wahrscheinlichkeit einer bestimmten Merkmalsausprägung aus einer mathematischen „Kombination“ („joined likelihood function“) der Wahrscheinlichkeit des individuellen Antwortmusters einer Person und des Wissens um die Itemcharakteristiken der dargebotenen Items zu schätzen. Es wird jeweils der Merkmalsausprägungswert auf dem Theta-Kontinuum als beste Schätzung angenommen, an dem die Likelihood Funktion ihr Maximum aufweist.

Der dritte und vierte Ansatz (EAP und MAP) hat seine Wurzeln bei Owen (1969). Ihm liegt das Bayes’sche Schätzverfahren der Merkmalsausprägung auf der Grundlage einer „a priori“-Verteilung zugrunde. Beide Ansätze greifen bei der Theta-Schätzung auf Maße der zentralen Tendenz (EAP: Arithmetischer Mittelwert; MAP: Modalwert) der angenommenen „a priori“-Verteilung (Normalverteilung) zurück. Was wiederum kritisch ist, wenn die vermutete „a priori“-Verteilung nicht der tatsächlichen empirischen Merkmalsausprägungs-verteilung entspricht. Allerdings nimmt mit steigender Testlänge der potentiell verzerrende Einfluss der „a priori“-Verteilungsannahme ab und die „Likelihood“-Verteilung gewinnt an Einfluss.

Alle Ansätze gelten als konsistent und effektiv in ihrer Anwendung (Chen, 1997), ihre Robustheit ist jedoch sowohl von der (IRT-) Modellkonformität des Antwortverhaltens als auch der dargebotenen Itemanzahl abhängig. So weisen eine Reihe von Autoren (Thissen & Mislevy, 1990; Wang, 1995, 1999) darauf hin, dass mit zunehmender Itemdarbietungszahl die Robustheit der Schätzung steigt und die Unterschiede zwischen den einzelnen Algorithmen abnehmen.

Vergleicht man die verschiedenen Ansätze, so tendiert der MLE-Ansatz allgemein zu einer Schätztendenz zu den Extremen (Lord, 1983). Desweiteren funktioniert seine Anwendung in folgenden drei Spezialfällen nicht: a) wenn nur ein Item dargeboten wird (also als Anfangsschätzalgorithmus; Voraussetzung für das Funktionieren des MLE-Algorithmus ist mindestens eine richtige und eine falsche Antwort auf jeweils ein Item), b) wenn alle Items richtig, und c) wenn alle Items falsch beantwortet werden (da in diesen Fällen die Schätzung gegen unendlich läuft).

Die Weighted-Likelihood-Schätzung (WLE; Warm, 1989) gilt als eine Weiterentwicklung des MLE-Ansatzes, der die Wurzel der Testinformationsfunktion als Gewichtung in die Schätzung (bei ein- bzw. zweiparametrischen Modellanwendungen) einfließen lässt, so dass seine Anwendung auch in den oben genannten drei „Spezialfällen“ möglich ist. Nach Meijer und Nering (1999) produziert dieser Ansatz weniger „bias“ (Testergebnisverzerrung).

Auch die EAP- (Bock & Mislevy, 1982) und MAP-Algorithmen können bereits nach der ersten Antwort auf ein Start-Item genutzt werden, da sie auf die vermutete „a priori“-Verteilung zurückgreifen. Dies kann zu einer Verbesserung der Theta-Schätzung führen (Meijer & Nering, 1999). Zudem kommt es zu keinen „Unendlichkeitsschätzungen“. Der Nachteil dieser Verfahren liegt jedoch, im Falle der Darbietung nur weniger Items und einer starken Abweichung des Mittelwerts der „a priori“-Verteilung von der geschätzten Likelihood, in einer „Schätztendenz zur Mitte“. Vergleicht man EAP- und MAP-Algorithmus, so ist der MAP- dem EAP- Algorithmus durch eine geringere Verzerrungstendenz überlegen, während umgekehrt der MAP- den EAP-Algorithmus durch einen etwas geringeren Standardmessfehler übertrifft (Meijer & Nering, 1999).

Möchte man Vorteile beider Ansätze (EAP / MAP und MLE / WLE) nutzen, so kann unter Umständen eine „Step-size-procedure“ (Embretson & Reise, 2000, S. 266f) empfehlenswert erscheinen, bei der die Anfangsschätzung auf der Basis von EAP bzw. MAP erfolgt, bis eine Schätzung auf Basis des MLE- bzw. WLE-Algorithmus möglich wird.

4.3.3.5. Itemdarbietung

Bislang findet sich wenig Forschung zur Itemdarbietung und deren Kontrolle (Thissen & Mislevy, 1990). Meijer und Nering (1999) sowie Embretson und Reise (2000) regen bei der Erforschung dieses Feldes folgende Fragestellungen an:

1. Dürfen bekannte Items mehrmals in einem CAT-Prozess dargeboten werden?
2. Welchen Einfluss haben Vorwissen bzw. Lerneffekte auf das CAT?
3. Sollen alle Items im Laufe eines bestimmten Zeitintervalls dargeboten werden, z. B. durch ein Itembankrotationssystem?
4. Welchen Einfluss hat die Darbietungszeit auf die Itemantwort?

5. Kann während der Itemdarbietung inkonsistentes Antwortverhalten identifiziert und eventuell beeinflusst werden?
6. Welchen Einfluss haben Itempositions-/-reihenfolgeeffekte?

Es bleibt zu hoffen, dass dieses spannende Forschungsfeld in naher Zukunft weitere Forschungsarbeiten motiviert.

4.3.3.6. Stoppfunktion

Um einen computergestützten adaptiven Algorithmus zu beenden, bieten sich prinzipiell drei Stoppkriterien an (Hambleton & Zaal, 1990):

1. ein festgelegtes Messfehlerkriterium (> Reliabilitätskriterium),
2. eine bestimmte Testlänge (minimale bzw. maximale Itemanzahl),
3. ein bestimmtes Klassifikationskriterium („Cut-Off-Wert“).

IRT-basierte CATs bieten gegenüber KTT-basierten Verfahren den großen Vorteil der Berechnung des *individuellen Messfehlers* einer Personenparameterschätzung. Dies ermöglicht die Realisierung eines „equal precise“ Tests, d. h. eines Tests, der empirisch gesichert auf allen Merkmalsausprägungsstufen gleich gut misst. Um dieses zu gewährleisten, kann die eigentliche Testlänge eines IRT-basierten CATs variabel gehalten werden. Konventionelle „fixed-length“ Testverfahren bieten diese Möglichkeit nicht.

Ein IRT-basierter CAT kann aber natürlich genauso in seiner *Testlänge* auf eine bestimmte maximale und / oder minimale Itemdarbietungszahl festgelegt werden, was vor allem bei großen Forschungserhebungen aus ökonomischen Gründen wünschenswert sein kann. Aufgrund der relativen Kürze eines CATs ist eine maximale Begrenzung jedoch häufig nicht nötig. Im Gegenteil merken Hambleton und Zaal (1990) an, dass für Laien die extreme Kürze von CATs mitunter unglaubwürdig oder sogar suspekt wirken könne, so dass eventuell eher (auch im Sinne des Vorbeugens eines „bias“) eine Limitierung im Hinblick auf die minimale Anzahl dargebotener Items angezeigt sei, um die „face validity“ (Augenscheinvalidität) zu erhöhen.

Als Stoppkriterium kann ebenfalls eine Kombination aus minimaler Testlänge und einem bestimmten Messfehlerkriterium gewählt werden.

Und schließlich können im Rahmen kriteriumsorientierter Tests auch sogenannte *„Cut-Off-Werte“* (bezüglich eines Testwertes oder Konfidenzintervalls) als Abbruchkriterien fungieren, welche der reinen

Klassifikation von Personen in zwei (oder mehr) Gruppen dienen. Bei der Nutzung solcher „Cut-Off-Werte“ als Stoppfunktion, erhöht sich meist die Testlänge / -zeit, je näher die Schätzung der Merkmalsausprägung einer Person dem vorher festgelegten „Cut-Off-Wert“ kommt (Weiss & Vale, 1987).

4.3.3.7. Wahl der Soft- und Hardware

Hornke (1996) hebt hervor, dass die Anforderungen, welche CATs an die Hardware stellen, weniger problematisch sind als diejenigen, die CATs an die Software-Programmierung stellen. Er fasst zusammen, dass die Hardware langlebig sein, und sich ihre Benutzeroberflächen für Laien (Testpersonen) handhabbar gestalten sollte (ergonomische Erwägungen, Benutzer-freundlichkeit, gute Lesbarkeit von Itemtexten, einfache Tastenbedienung etc.; Wainer, 1990).

Die einzigen universellen Software-Pakete, welche der Umsetzung des CAT-Prozesses nach bereits stattgefundener Itemkalibrierung dienen können, sind meines Wissens das „Micro-CAT“ (Hambleton & Zaal, 1990), welches 1988 von der Assessment Systems Corporation entwickelt wurde, und der „ADTEST“, der 1994 von Ponsoda, Olea und Revuelta vorgestellt wurde. Mitunter wird von den einzelnen Forschergruppen computergestützte adaptive Testsoftware auch selbst entwickelt (Ware et al., 2000, 2003).

Allgemein gilt die Empfehlung, Software zu entwickeln, welche nicht als „Inselprodukt“ oder „Exot“ auf dem Markt wahrgenommen wird, sondern Software (sowie auch Hardware) so zu standardisieren, dass sie über Schnittstellen zu anderen Komponenten und zu unterschiedlichen Zeitversionen (z. B. von Computersystemen: Windows; Linux etc.) kompatibel ist. In diesem Sinne sollten CATs wie „Haushaltsgeräte mit Bedienungsanleitung“ für ausgebildete Psychodiagnostiker leicht zu handhaben sein, jedoch stets einer professionellen Pflege und einer ernsthaften, verantwortungsbewussten Administration unterliegen.

4.4. Vorteile IRT-basierter CATs

Die zwei Hauptvorteile, welche von vielen Autoren (Weiss & Vale, 1987; Kubinger, 1993; Kisser, 1995; Hornke, 1999; Gregory, 1996; Amelang und Zielinski, 1996; Embretson & Reise, 2000) für IRT-basierte CATs ins Feld geführt werden, sind die Verbesserung a) der *Testökonomie* bzw. –effizienz und b) der *Messgenauigkeit*.

Wie im Kapitel zu computergestützten Tests bereits angedeutet, können diese zu *Zeit*- und *Kosteneinsparungen* von bis zu 50%, IRT-basierte CATs sogar zu Einsparungen von 50-80% führen (Weiss & Vale, 1987; Hornke, 1993, 1996; Gregory, 1996), da durch adaptives Testen die Zeit der Testadministration sowie der Testauswertung und –dokumentation erheblich verringert werden kann und die laufenden Materialkosten (Papier, Bleistifte etc.) gegen eine einmalige Anschaffungsgebühr der Software und Hardware entfallen. Dies ist in großen Forschungsprogrammen von Belang, aber auch für den unmittelbaren klinischen Alltag relevant. Denn durch adaptives Testen (und die adaptive Auswahl von Testverfahren) wird Testen auf Nachfrage möglich, und dies kann zu einer Erleichterung der klinischen Fokusbildung führen. Embretson und Reise (2000) beschreiben exemplarisch einen solchen Nutzen an dem Beispiel eines kognitiven Screening-Instruments, dem bei diagnostischen Hinweisen auf kognitive Defizite ein Gedächtnistest adaptiv nachgeschaltet werden kann.

Desweiteren wird ähnlich wie in Kapitel 4.2.1. darauf verwiesen, dass der eingesparte Aufwand an Routineadministration Zeit für weitere Diagnostik oder Therapie bietet. Eine erhöhte Testökonomie kann nicht nur dem Diagnostiker, sondern auch der Testperson zugute kommen, 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. Über- und Unterforderung und damit einhergehende Frustration und Verwirrung bei der Darbietung zu schwieriger Items, sowie Ärger und Langeweile bei der Präsentation zu leichter Items (sowie potentiell resultierende Verminderungen der Datenqualität z. B. durch Flüchtigkeitsfehler oder Motivationseffekte) können durch ein adaptives Testvorgehen vermieden werden (Wainer, 1990). Im Idealfall fühle sich - so Hornke (1993) - die Testperson optimal gefordert und schreibe der CAT-

Testung bedingt durch eine hohe Standardisierung und Augenscheinvalidität eine hohe Testfairness zu.

Die Bestimmung und Kontrolle der *Messgenauigkeit* (Reliabilität) resultiert aus den Möglichkeiten der IRT (siehe Kapitel 3.3.3.). Sie wird durch eine Reihe von Autoren (Weiss, 1985; Weiss & Vale, 1987; Kisser, 1995; Amelang & Zielinski, 1996; Gregory, 1996; Meijer & Nering, 1999; Embretson & Reise, 2000) als der zweite Hauptvorteil adaptiven Testens genannt.

Während des adaptiven Testens ist eine Erhöhung der Messgenauigkeit durch einzelne Items kumulativ abschätzbar, so dass sowohl Aussagen darüber getroffen werden können, wie stark einzelne Items die Messgenauigkeit beeinflussen, als auch mit welcher Messgenauigkeit der gesamte individuelle CAT-Prozess einhergeht. Ersteres erlaubt die Auswahl der Items, welche für ein bestimmtes Merkmalsausprägungsniveau die höchste Messgenauigkeit aufweisen, woraus die eingangs erwähnte Testökonomie resultiert (Amelang & Zielinski, 1996). Eine solche Erhöhung der Messgenauigkeit kann sich auch positiv auf die Validität auswirken (Weiss & Vale, 1987). Desweiteren erlaubt die Kontrolle einer konstanten Messgenauigkeit über verschiedene Merkmalsausprägungsniveaus hinweg den interindividuellen Vergleich von einzelnen Testpersonen sowie den Vergleich von Gruppenkollektiven (trotz unterschiedlicher Art und Anzahl dargebotener Items, d. h. trotz variabler Testlänge; Kisser, 1995). Durch die dadurch bedingte Vermeidung von Decken- oder Bodeneffekten können z. B. Gruppenvergleiche wie in der Lebensspannenforschung, der (Therapie-) Evaluationsforschung und bei Wachstums- /Veränderungsmessungen verbessert werden (Embretson, 1992; Embretson & Reise, 2000).

Neben der Messgenauigkeitsberechnung eröffnet ein IRT-basiertes Vorgehen auch die Möglichkeit der Berechnung weiterer Parameter, wie z. B. der Iteminformationsfunktion, die zur Itemselektion genutzt wird, sowie der Testinformationsfunktion (siehe Kapitel 3.3.3.), welche die Vergleichbarkeit der Messgenauigkeit unterschiedlicher Tests in Bezug auf unterschiedliche Merkmalsausprägungsbereiche oder Personenkollektive und somit gezielte Test-Indikationsentscheidungen ermöglicht (dies übersteigt die Möglichkeiten der KTT). Desweiteren wird darauf hingewiesen, dass IRT-basierte Testscores die empirische Wirklichkeit adäquater als KTT-basierte Testscores abzubilden

vermögen (Kubinger, 1993), was u. a. aus dem Einbezug einer größeren Anzahl von Parametern (z. B. des Rateparameters bei dreiparametrischen IRT-Modellen; Wainer, 1990) resultiere.

Neben den zwei genannten Hauptvorteilen (a) der Testökonomie und (b) der Messgenauigkeit und deren positiven Implikationen (Entlastung des Diagnostikers und der Testperson, Vergleichbarkeit von Messwerten) sowie (c) den zuletzt genannten Vorteilen, welche mit der Berechnung zusätzlicher *IRT-spezifischer Parameter* verbunden sind, werden IRT-basierten CATs in der Literatur eine Reihe von weiteren Vorteilen zugeschrieben.

Diese können grob in Vorteile unterteilt werden, welche sich auf: d) Unterschiede in der *Testform* (Testlänge, Antwortformate und Testinstruktionen) und e) Unterschiede in der *Durchführung* und *Auswertung* von IRT-basierten CATs beziehen.

Vergleicht man konventionelle Verfahren (KTT-basierte Papier-und-Bleistift-Versionen) mit IRT-basierten CATs so stechen die Vorteile einer variablen, adaptiven und damit kürzeren Testlänge,[40] eines variablen Antwortformats (Hambleton & Zaal, 1990; Hambleton et al., 1991) und einer möglichen „maßgeschneiderten“ adaptiven Instruktion (Kisser, 1995) ins Auge.

IRT-basierte CATs unterscheiden sich weiterhin von konventionellen Verfahren, indem umfangreiche Antwortbögen, welche die Gefahr des „Verrutschens“ in der Itemtext- / Antwortzeile mit sich bringen können (Embretson & Reise, 2000) durch eine „Item-by-Item“ Präsentation ersetzt werden. Simultan zur Darbietung einzelner Items vollzieht sich die Schätzung der Merkmalsausprägung der Testperson, so dass eine schnelle / sofortige Testergebnisberechnung, -dokumentation und –rückmeldung (Feedback) ermöglicht wird (Hambleton & Zaal, 1990; Hambleton et al., 1991; Embretson & Reise, 2000).

Aus der kontinuierlichen Verrechnung der Itemantworten einer Testperson ergeben sich zwei weitere Vorteile: zum einen lässt sich dadurch inkonsistentes Antwortverhalten einer Testperson bereits während des CAT-Prozesses identifizieren und eventuell korrigieren (Meijer & Nering, 1999) und zum anderen resultieren daraus (potentielle) Vorteile bezüglich der Itembankentwicklung. So machten bereits 1985 Butcher, Keller und Bacon darauf aufmerksam, dass im Rahmen IRT-basierter CATs eine kontinuierliche

[40] Hornke (1999) zeigt an der Entwicklung von drei CATs eine adaptive Itemreduktion um 2/3 der vorherigen Testlänge (durchschnittliche Anzahl dargebotener Items: 7).

Aktualisierung der Itembank möglich sei, z. B. durch die Einspeisung von „neuen“ Testitems und deren simultaner Kalibrierung im Rahmen des CAT-Prozesses. Durch eine „Züchtung“ „guter“ Items und die Identifikation und den Ausschluss „schlechter“ Items (z. B. durch die Berechnung von Item Response Curves, IRCs; siehe Kapitel 3.3.1.) kann die einem CAT zugrunde liegende Itembank ständig verbessert werden (Thissen & Mislevy, 1990). Meines Wissens wurde das Potential einer simultan zum CAT-Prozess möglichen Aktualisierung der Itembank jedoch in der Praxis noch nicht erprobt.

4.5. Nachteile IRT-basierter CATs

Der größte Nachteil IRT-basierter CATs liegt in den hohen Anfangskosten, welche die Entwicklung und Implementierung solcher Verfahren begleiten (Meijer & Nering, 1999). Diese sind sowohl finanzieller (Kosten von Soft- und Hardware) wie auch personeller (psychodiagnostische, statistische, technische Qualifikationen) Art. Am aufwendigsten ist wohl die umfangreiche Itembankkalibrierung, welche die Erhebung einer Vielzahl von Items an einem großen Personenkollektiv voraussetzt. So wird im individuellen Fall mit detaillierten Kosten-Nutzen-Analysen (Thissen & Mislevy, 1990) abzuwägen sein, ob sich die Entwicklung und Implementierung von IRT-basierten CATs in der jeweiligen Institution bzw. Organisation lohnt.

Während vor einigen Jahrzehnten die technischen Möglichkeiten (begrenzte Rechnerkapazitäten) noch die Grenzen IRT-basierter CAT-Entwicklungen steckten, stellen Hardware-Begrenzungen heutzutage aufgrund des raschen technischen Fortschritts und der ubiquitären Verbreitung von Computern kein ernsthaftes Hindernis mehr dar.

Problematisch ist in diesem Zusammenhang wohl eher die relative Benutzer*un*freundlichkeit der Software, mit der IRT-basiert Itembanken kalibriert werden, sowie der relative Unbekanntheitsgrad der - verglichen mit der KTT - eher komplizierten IRT. Diese beiden Umstände führten bislang zumindest im klinischen Bereich nur zu einer geringen Verbreitung der Methodik (Rost, 1999; siehe Kapitel 3.3.4. und 3.5.). Das damit verbundene Forschungsdefizit lässt viele Fragen offen.

So zweifelt beispielsweise Kisser (1995), ob die erhoffte Zeitersparnis bei IRT-basierten CATs sich bei deren Anwendung in der Realität tatsächlich zeigt. Es existieren zwar einige Belege für eine kürzere Bearbeitungszeit von CATs

(Hornke, 1993, 1996, 1999), allerdings vermutet Kisser (1995), dass eine geringe Anzahl von Items (wie beim CAT) nicht unweigerlich zu einer Testzeitverkürzung führe, wenn die Bearbeitung von Items mit *unterschiedlichen* Antwortformaten mehr Zeit als die Beantwortung von Items mit dem *gleichen* Antwortformat (wie bei konventionellen Verfahren) in Anspruch nähme. Bezüglich der Untersuchung der Zeitersparnis fand Hornke (1996) in einer seiner Studien, dass im Laufe des CAT-Prozesses die Itembearbeitungszeit abnähme, gleichzeitig verringerte sich jedoch auch die Konstanz der Messergebnisse, was Hornke auf einen Sorgfalts-, Aufmerksamkeits- und / oder Motivationsverlust (> Flüchtigkeitsfehler) der Testpersonen in der Interaktion mit dem Computer zurückführte. Auch dies ist kritisch zu bewerten.

Neben der Überprüfung der tatsächlichen Zeitersparnis, sind bislang weitere grundlegende Aspekte von IRT-basierten CATs unerforscht. So bestehen beispielsweise große Forschungsdefizite im Hinblick auf...

1. die methodischen Standards der IRT-basierten Itembankentwicklung (Selektionskriterien),
2. die Itembanksicherheit (v. a. bei „wireless LAN-Applikationen“[41]),
3. die Robustheit von Item- und Personenparameterschätzungen…
    a) über verschiedene Zeiten, Kontexte und Stichproben (Gefahr des „Parameterdrifts“; Bock & Mislevy, 1988),
    b) bei unterschiedlichen Itembankgrößen,
    c) bei unterschiedlichen Antwortformaten (Kisser, 1995),
    d) auf der Grundlage unterschiedlicher Itemselektionsalgorithmen,
    e) bei unterschiedlichen Itemreihenfolgen und –positionen,
    f) bei unterschiedlichen Itemdarbietungskontrollen (Zeitrestriktionen, Unmöglichkeit des Zurückblätterns / Korrigierens, Iteminhalts-balancierung),
    g) im Falle von Vorwissen um Items (Lerneffekte),
    h) im Falle von (Computer-) Testangst,
    i) bzgl. der Verletzung von IRT-Modellannahmen wie z. B.:
        - Unidimensionalitätsverletzungen,
        - Item-Misfits,
        - Personen-Misfits (Wainer, 1990; Kubinger, 1999).

[41] Wireless LAN (Local Area Network) = kabellose Datenübertragung in lokalen Computernetzwerken.

4. die Anwendung von IRT-Modellen auf polytome Items (Dodd et al., 1995),
5. die beste Kommunizierbarkeit IRT-basierter Testscores (Theta), welche in Einheiten der Standardnormalverteilung (z-Werte) ausgegeben werden und für den Laien (Testpersonen) nicht intuitiv verständlich erscheinen (Embretson & Reise, 2000),
6. die Äquivalenzprüfung (Kubinger, 1999) von Papier-und-Bleistift-Verfahren, computergestützten Tests und CATs,
7. die prospektive Validität von IRT-basierten CATs und
8. die allgemeine Qualitätssicherung IRT-basierter CATs.

Wie die aufgeführten Forschungsdefizite (siehe auch Kapitel 3.3.4.) verdeutlichen, stecken IRT-basierte CATs (v. a. im klinisch-psychologischen Bereich) noch weitgehend in den Kinderschuhen (siehe Kapitel 4.6.). Das junge Forschungsfeld IRT-basierter CATs ist durch ein Mosaik technischer Artikel (Embretson & Hershberger, 1999) gekennzeichnet. Empirische Befunde zur Anwendung IRT-basierter CATs im psychologischen Bereich beschränken sich größtenteils auf Simulationsstudien (Kisser, 1996; Hornke, 1999; Gardner et al., 2002). Daher kommt der Entwicklung und Erprobung „echter“ CATs in der Praxis bei der Beantwortung oben genannten Forschungsfragen ein großer Stellenwert zu.

4.6. Aktueller Forschungsstand zu IRT-basierten CATs

Die Sichtung der Literatur zum Thema IRT-basierter CATs gestaltet sich etwas verwirrend, da CATs entwickelt wurden, welche sich andere adaptive Strategien zunutze machen als die IRT (siehe Kapitel 4.3.2.). Zum Beispiel wandten Ben-Porath, Slutske und Butcher (1989) die „Countdown“-Methode an, um ein CAT des Minnesota Multiphasic Personality Inventory (MMPI) zu entwickeln (Roper, Ben-Porath & Butcher, 1991; Handel, Ben-Porath & Watt, 1999).

Zudem existieren eine Reihe von Forschungsarbeiten zur Anwendung der IRT bei der Itembankentwicklung, die in der Entwicklung von CATs mündeten, bei denen jedoch der Itemselektionsalgorithmus und die Personenparameter-schätzung nicht IRT-basiert erfolgen, sondern „konventionell“ programmiert sind. Tabelle 6 gibt einen Überblick über solche CATs im deutschsprachigen Raum. Sie begrenzt sich auf einen Überblick der im internationalen Raum

aktuell eingesetzten CATs, welche *gänzlich* IRT-basiert sind. Das heißt, es werden nur CATs aufgeführt, bei denen die IRT sowohl bei der Itemparameterkalibrierung, als auch bei der Itemselektion und Personenparameterschätzung im Rahmen des CAT-Prozesses angewandt wurde.

Tabelle 6: Überblick über CATs im deutschen Sprachraum, bei denen die Itembankentwicklung IRT-basiert erfolgte (die Itemselektion und Testergebnisberechnung jedoch nicht IRT-basiert sind).

Inventar	Bereich	Autoren	Jahr	Ort
Verbal Memory Test	Gedächtnistest	Hornke & Etzel	1999a	Institut für Psychologie, Rheinisch-Westfälische TH Aachen, Deutschland.
Visueller Gedächtnis Test	Gedächtnistest	Hornke & Etzel	1999b	Schuhfried-Testverlag, Mödling, Österreich.
Adaptive Three-Dimensional Cube Comparison Test (A3DW)	Eindimensionaler Intelligenztest	Gittler	1999	Institut für Psychologie der Universität Wien, Österreich.
Adaptiver Matrizentest (AMT)	Wehrpsychologische Eignungsdiagnostik	Hornke & Habon	1984	Institut für Psychologie, Rheinisch-Westfälische Technische Hochschule, TH Aachen, Deutschland.
Adaptiver Zahlenfolgen-Lerntest (AZAFO)	Lernfähigkeitstest	Vahle & Rittner	1995	Schuhfried-Testverlag, Mödling, Österreich.
Computergest. Intelligenz-Lerntest-Batterie (ACIL)	Intelligenztest	Beckmann & Guthke	1999	Schuhfried-Testverlag, Mödling, Österreich.
Adaptiver Analogien-Lerntest (ADANA)	Lernfähigkeitstest	Stein	1995	Schuhfried-Testverlag, Mödling, Österreich.
Syllogismen	Eindimensionaler Intelligenztest	Srp & Hörndler	1994	Swets Test Services, Frankfurt am Main, Deutschland.
Begriffs-Bildungs-Test (BBT).	Informations-verarbeitungtest	Kubinger, Fischer & Schuhfried	1993	Schuhfried-Testverlag, Mödling, Österreich.

4.6.1. IRT-basierte CATs in der Leistungs- und Eignungsdiagnostik

IRT-basierte CATs sind vor allem im Bereich der Fähigkeitseinschätzung zur *Eignungsdiagnostik* auf internationaler Ebene mittlerweile gut etabliert. Die zwei größten Anwendungsgebiete liegen im Bereich der *Schul*diagnostik und der *militärischen* Eignungsdiagnostik.

In der *Schuleignungsdiagnostik* sind eine Reihe IRT-basierter CATs in der Anwendung wie z. B. in den U.S.A. der „Scholastic Aptitude Test“ (SAT), die „Graduate Record Examination“ (GRE, 1996; Educational Testing Service, ETS, 2001), der „Computerized Placement Test“ (College Board, 1993) und verschiedene Mathematik, Lese- und Schreibtests innerhalb des „COMPASS“-Programms (American College Testing, 1993; Dodd et al., 1995), in Südafrika der „Learning Potential Computerized Adaptive Test“ (LPCAT; de Beer, 2000) sowie in den Niederlanden zwei Mathematikleistungstests (National Institute for Educational Measurement; Verschoor & Straetmans, 1999).

Im Rahmen von *wehrpsychologischen Untersuchungen* werden sowohl in Deutschland als auch in den U.S.A. IRT-basierte CATs eingesetzt. In Deutschland zählen Hornke und seine Mitarbeiter zu den Hauptvertretern dieser Richtung, welche IRT-basierte CATs zur Diagnostik „Verbaler Analogien“ (Hornke, 1989), zur Messung der Gedächtnis- und Orientierungsleistung (Hornke, 1999) und zur Intelligenz („Matrizentest“; Hornke, 1999) entwickelten. In der U.S. Armee wird zur Eingangsdiagnostik ein IRT-basierter CAT namens „Armed Services Vocational Aptitude Battery“ (ASVAB; Curran & Wise, 1994; Sands, Waters & McBride, 1997) angewandt. Da der initiale Entwicklungsaufwand IRT-basierter CATs recht hoch ist, finden sich die meisten Anwendungen IRT-basierter CATs in *größeren Organisationen* und Institutionen, welche regelmäßig umfangreiche psychodiagnostische Testungen durchführen. So machen sich neben amerikanischen Schulbehörden (z. B. Portland Public School District, Kingsbury & Houser, 1993) und Militäreinrichtungen (z. B. U.S. Department of Defense) auch Prüfungsbüros von medizinischen Ausbildungseinrichtungen die IRT-basierte CAT-Methodik bei der Durchführung von Examina zunutze (z. B. American Society of Clinical Pathologists; Lunz, Bergstrom & Wright, 1992; National Council of State Boards of Nursing; Zara, 1988; American Board of Internal Medicine; Reshetar, Norcini & Shea, 1993).

4.6.2. IRT-basierte CATs in der klinischen und Persönlichkeitsdiagnostik

Während die Anwendungsbeispiele von IRT-basierten CATs im Bereich der Leistungs- und Eignungsdiagnostik zeigen, dass diese Methodik in diesem Bereich bereits verbreitet ist, gilt dies nicht für den Bereich der klinischen Diagnostik sowie der Messung von Einstellung und Persönlichkeitseigenschaften.

Im Bereich der *klinisch-medizinischen Diagnostik* existieren meines Wissens (neben der Forschungsgruppe an der Charité Berlin, in dessen Rahmen vorliegende Arbeit geschrieben wurde), nur drei Forschungsgruppen, welche folgende IRT-basierte CATs entwickelt haben:

- Ware, Bjorner und Kosinski (2000):
  Dynamic Health Assessment (DynHA):
  - Headache Impact Test (HIT)[42],
  - Dynamic SF-36 Health Survey,
  - Depression Impact Test etc;
- Simms und Clark (submitted):
  Schedule for Nonadaptive and Adaptive Personality;
- Gardner, Kelleher und Pajer (2002):
  Pediatric Symptom Checklist (PSC).

Im Bereich der *Einstellungsdiagnostik* (sowie Leistungsmessung) findet sich neben Reise und Waller (1990), welche die Absorption Scale des MPQ (Tellegen, 1982) IRT-basiert computer-adaptiv erprobten, und Andrich (1978) eine rege Forschungstätigkeit nur in der Forschungsgruppe um Dodd, Ayala und Koch (1995). Diese sticht jedoch dafür durch eine hohe Publikationsfreudigkeit hervor (De Ayala, 1989, 1992; Dodd, 1990; Dodd et al., 1988, 1989, 1993; Koch & Dodd, 1985, 1989; Koch et al., 1990), indem sie gezielt die Anwendung verschiedener IRT-Modelle auf polytome Items fokussiert (bislang werden in der IRT-basierten CAT-Eignungsdiagnostik fast ausschließlich dichotome Items genutzt).

Polytome Items werden auch vielfach in der *Persönlichkeitsdiagnostik* verwandt. Jedoch hinkt die IRT-basierte CAT-Forschung in diesem Bereich dem Forschungsstand, wie er z. B. bereits in der Eignungsdiagnostik gediehen ist,

[42] Ware, Kosinski, Bjorner, Bayliss, Batenhorst, Dahlöt, Tepper & Dowson (2003).

stark hinterher. Ursächlich hierfür könnte u. a. die Diskussion um die (Uni-) Dimensionalität von Persönlichkeitskonstrukten (die meisten IRT-basierten CATs sind bislang unidimensional konstruiert; komplexe multidimensionale IRT-Ansätze finden sich meines Wissens nur bei Gardner et al., 2002) sowie das allgemein geringe wirtschaftliche Interesse an der Persönlichkeitsdiagnostik sein (Persönlichkeitsdiagnostik ist im Rahmen von Eignungsdiagnostik umstritten; die psychologische Diagnostik wird im chronisch unterfinanzierten öffentlichen Gesundheitswesen eher vernachlässigt).

Genuine IRT-basiert entwickelte CATs zur Messung von Persönlichkeitsvariablen existieren meines Wissens weder im deutschen noch im internationalen Sprachraum.

Jedoch publizierten kürzlich Reise und Henson (2000) in einer Simulationsstudie eine computergestützte adaptive Version des bereits etablierten NEO-PIs, dessen Itemselektion und Personenparameterschätzung IRT-basiert anhand des Graded Response Modells erfolgt, dessen Itembank jedoch nicht mit IRT-Methoden entwickelt wurde. Zudem bereiten Simms und Clark eine Publikation vor, in der ein Persönlichkeitsfragebogen („Schedule for Nonadaptive and Adaptive Personality“, SNAP; Clark, 1993) als IRT-basierte CAT-Version anhand des 2PL-Modells von Birnbaum entwickelt und an N=413 Studenten erfolgreich validiert wurde. Die meisten Forschungsarbeiten in diesem Gebiet sind noch nicht so weit fortgeschritten und beschränken sich größtenteils auf die Erprobung von IRT-Methoden im Rahmen der (Re-)Analyse bzw. Bewertung bereits etablierter KTT-basierter Verfahren (zum aktuellen Forschungsstand bzgl. IRT-Anwendungen und zu möglichen Gründen dieses Forschungsdefizits siehe Kapitel 3.5.2.).

Zusammenfassend lässt sich resümieren, dass sich das in vorliegender Arbeit entwickelte IRT-basierte CAT zur Angstmessung (Angst-CAT) als eine *klinisch-psychologische* Pionierarbeit in die oben genannten US-amerikanischen Forschungsarbeiten zur klinisch-medizinischen Diagnostik, Einstellungs- und Persönlichkeitsdiagnostik einreihen lässt.